UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

MÁVERICK ANDRÉ DIONÍSIO FERREIRA

NASC: Análise de rede para identificar o discurso sociocognitivo presente nos papéis desempenhados por estudantes em fóruns de discussão

Recife

2023

MÁVERICK ANDRÉ DIONÍSIO FERREIRA

NASC: Análise de rede para identificar o discurso sociocognitivo presente nos papéis desempenhados por estudantes em fóruns de discussão

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutor em Ciência da Computação.

Área de Concentração: Inteligência Computacional

Orientador (a): Prof. Dr. Rafael Dueire Lins

Coorientador (a): Prof. Dr. Rafael Ferreira Leite de Mello

Recife

2023

Catalogação na fonte
Bibliotecária Mônica Uchôa, CRB4-1010

F383n Ferreira, Máverick André Dionísio

NASC: análise de rede para identificar o discurso sociocognitivo presente nos papéis desempenhados por estudantes em fóruns de discussão / Máverick André Dionísio Ferreira. – 2023.

154 f.: il., fig., tab.

Orientador: Rafael Dueire Lins.

Tese (Doutorado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2023.

Inclui referências e apêndices.

1. Fóruns de discussão. 2. Comunidade de inquérito. 3. Processamento de linguagem natural. 4. Análises de redes epistêmicas. 5. Papéis roteirizados. 6. Papéis emergentes. I. Lins, Rafael Dueire (Orientador). II. Título.

004.68 CDD (23. ed.) UFPE - CCEN 2024 – 001

Máverick André Dionísio Ferreira

"NASC: Análise de rede para identificar o discurso sociocognitivo presente nos papéis desempenhados por estudantes em fóruns de discussão"

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutor em Ciência da Computação. Área de Concentração: Inteligência Computacional.

Aprovada em: 25/09/2023.

____

Orientador: Prof. Dr. Rafael Dueire Lins

BANCA EXAMINADORA

____

Profa. Dra. Patrícia Cabral de Azevedo Restelli Tedesco
Centro de Informática/ UFPE

____

Prof. Dr. Adenilton Jose da Silva
Centro de Informática/ UFPE

____

Prof. Dr. André Câmara Alves do Nascimento
Departamento de Computação / UFRPE

____

Prof. Dr. Filipe Dwan Pereira
Departamento de Ciência da Computação / UFRR

____

Prof. Dr. Rafael Dias Araújo
Faculdade de Computação / UFU

Dedico este trabalho a toda a minha família, em especial a minha mãe, por todo o apoio dado durante o percurso.

AGRADECIMENTOS

A Deus, pelo dom da vida e por estar ao meu lado em cada passo dado ao longo da minha caminhada até aqui. Aos meus familiares, em especial a minha mãe, por terem sido a minha base em meio a todas as dificuldades enfrentadas durante o doutorado. Aos orientadores, os professores Dr. Rafael Dueire Lins e Dr. Rafael Ferreira Leite de Mello por não terem desistido de seguirem comigo e por todos os ensinamentos compartilhados. As amizades construídas ao longo deste percurso, em especial aos amigos: Débora, Alisson, Anderson, Vitor e Elaine. Vocês contribuíram diretamente para a minha formação como pesquisador e como pessoa.

RESUMO

Os fóruns de discussões são ferramentas importantes para promover a interação social no contexto da Educação à Distância, pois permitem que estudantes e professores possam estabelecer um ambiente de colaboração *online*. Onde, para impulsionar debates com níveis elevados de aprendizagem, a literatura tem defendido o acompanhamento dos fóruns sob o olhar de uma teoria educacional denominada de Comunidade de Inquérito (CoI), que é composta por três dimensões: Presença Social, Presença Cognitiva e Presença de Ensino. Acredita-se também que a definição prévia dos papéis/funções que os estudantes devem assumir durante os debates, e o acompanhamento dos papéis que os discentes assumirão espontaneamente (emergentes), são relevantes para o processo de mediação nos fóruns. Apesar disso, estudos têm pontuado a dificuldade dos professores em acompanhar as discussões ao passo em que o número de postagens aumenta. Podendo essa dificuldade de acompanhamento resultar em diversos problemas, dentre eles: demora para os professores fornecerem *feedback* aos estudantes ocasionando, assim, desmotivação; e desconhecimento por parte dos professores dos desdobramentos das discussões, papéis assumidos pelos estudantes, impossibilitando a devida (re)adequação do design instrucional. Diante disso, nos últimos anos, diversos estudos foram realizados com o objetivo de construir soluções voltadas para apoiar acompanhamento das interações sociais e dos papéis desempenhados pelos estudantes durante os debates. Tendo em vista que grande parte dos conteúdos inseridos nos fóruns são textuais, a construção de tais soluções ocorrem por meio de uma área denominada de Processamento de linguagem natural (PLN) a qual possui predominância de ferramentas para o idioma Inglês. Portanto, esta tese propõe uma abordagem capaz de permitir o acompanhamento dos papéis roteirizados e emergentes desempenhados pelos estudantes em fóruns (Português e Inglês), enquanto as discussões evoluem, com base nas presenças Social e Cognitiva do modelo CoI. A referida abordagem combina classificadores multilingua (Português e Inglês) com uma estrutura de análise visual a qual é formada pela análise de clusters e pela a análise de redes espistêmicas.

Palavras-chave: fóruns de discussão; comunidade de inquérito; processamento de linguagem natural; análises de redes epistêmicas; papéis roteirizados; papéis emergentes.

ABSTRACT

Discussion forums are important tools for promoting social interaction in the context of Distance Education, as they allow students and teachers to establish a collaborative environment *online*. To promote debates with high levels of learning, the literature has defended the monitoring of forums from the perspective of an educational theory called Community of Inquiry (CoI), which is composed of three dimensions: Social Presence, Cognitive Presence, and Presence of Teaching. It is also believed that the prior definition of the roles/functions that students must assume during debates, and the monitoring of the roles that students will assume spontaneously (emerging), are relevant to the mediation process in the forums. Despite this, studies have highlighted the difficulty teachers have in following discussions as the number of posts increases. This difficulty in monitoring can result in several problems, including it takes time for teachers to provide *feedback* to students, thus causing demotivation; and lack of knowledge on the part of teachers of the developments of discussions, roles assumed by students, making it impossible to properly (re)adapt the instructional design. Therefore, in recent years, several studies have been carried out with the aim of building solutions aimed at supporting the monitoring of social interactions and the roles played by students during debates. Considering that much of the content inserted in the forums is textual, the construction of such solutions occurs through an area called Natural Language Processing (NLP), which has a predominance of tools for the English language. Therefore, this thesis proposes an approach capable of allowing the monitoring of scripted and emerging roles played by students in forums (Portuguese and English), while discussions evolve, based on the Social and Cognitive presences of the CoI model. This approach combines multilingual classifiers (Portuguese and English) with a visual analysis structure which is formed by cluster analysis and Epistemic Network Analysis (ENA).

Keywords: discussion forums; community of inquiry; natural language processing; epistemic network analysis; script roles; emerging roles.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

ACAC	Aprendizagem colaborativa apoiada por computador
AdaBoost	*Adaptive Boosting*
AM	Aprendizado de Máquina
API	*Application Programming Interface*
AVA	Ambientes Virtuais de Aprendizagem
BERT	*Bidirectional Encoder Representations from Transformers*
CoI	Comunidade de Inquérito
EAD	Educação à distância
ENA	*Epistemic Network Analysis*
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
LA	*Learning Analytics*
LIWC	*Linguistic inquiry and word count*
LSA	*Latent semantic analysis*
MDG	*Mean Decrease in Gini*
NASC	Análise de rede de discurso sociocognitivo
NER	Número de entidades nomeadas
NLTK	*Natural Language Toolkit*
PLN	Processamento de Linguagem Natural
RCD	Recursos de contexto de discussão
SNA	*Social Network Analysis*
SVM	*Support Vector Machine*
XGBoost	*eXtreme Gradient Boosting*

LISTA DE SÍMBOLOS

κ	Cohen's Kappa

SUMÁRIO

1 INTRODUÇÃO

A Educação à distância (EAD) é uma estrutura de ensino acessível, para diferentes públicos, uma vez que permite o acesso ao ecossistema de ensino das instituições de Educação a estudantes com diferentes disponibilidades de horário e localização geográfica (PANIGRAHI; SRIVASTAVA; SHARMA, 2018). Por isso, tem havido um crescimento no número de matrículas na EAD, considerando dados até o ano de 2014, em países como (GABA; LI, 2015; QAYYUM; ZAWACKI-RICHTER, 2019): Austrália, Brasil, Canadá, China, Alemanha, Índia, Rússia, África do Sul, Coreia do Sul, Turquia, Reino Unido e Estados Unidos. Sendo ainda citado um maior aumento em países emergentes, tais como o Brasil 63.8% e a Turquia 20.1% (QAYYUM; ZAWACKI-RICHTER, 2019). Olhando com mais detalhes para o cenário brasileiro, de acordo com o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), a EAD cresceu 474% em uma década (INEP, 2022).

Apesar dos benefícios presentes na EAD, estudos têm destacado os desafios que as instituições enfrentam para criar um ambiente de colaboração e pertencimento ao grupo por parte dos estudantes (GAYTAN, 2015). Sendo a sensação de pertencimento ao grupo um dos fatores considerados mais importantes no processo de retenção estudantil na EAD (STOYTCHEVA, 2021; PEACOCK et al., 2020; THOMAS; HERBERT; TERAS, 2014; KARA et al., 2019; ALENCAR; SANTOS; NETTO, 2015). Por exemplo, nas pesquisas de Kara *et al.* (2019) e Alencar *et al.* (2015), foram mencionados fatores com forte influência na decisão do estudante de evadir e dentre eles: baixa interação do tipo estudante-estudante e estudante-tutor, o sentimento de isolamento e a falta de apoio institucional. Pode-se citar também o estudo de Gaytan (2015) em que, após uma investigação sobre as percepções dos professores e dos estudantes acerca da retenção na EAD, elencou-se como um dos principais pontos o anseio dos discentes por mais instruções vindas por parte do corpo docente com viés corretivo (*feedback*).

Nesse contexto, os Ambientes Virtuais de Aprendizagem (AVA) surgiram como uma forma de apoiar interações entre estudantes e professores além de possibilitar o gerenciamento de todo o processo educacional *online* (JAIN, 2015). Podem-se citar como exemplos de AVA os ambientes Blackboard, WebCT, Moodle e o Sakai (BRI et al., 2009). Dentre as ferramentas existentes nesses ambientes, destaca-se o fórum de discussão por ser capaz de permitir interações assíncronas entre estudantes e professores (ANDERSON; KANUKA, 1997; XIA; FIELDER; SIRAGUSA, 2013), sendo assim uma ferramenta propícia para a aprendizagem colaborativa, envolvimento social entre os participantes, compartilhamento de informações sobre o curso e fornecimento de *feedback* por parte dos professores (HUGHES et al., 2002; PANITZ, 1999; GOKHALE, 1995).

Ao longo dos últimos anos pesquisas têm investigado formas de aumentar a promoção de conhecimento de alto nível e a interação social nos fóruns (NETO et al., 2018a; KOVANOVIĆ et al., 2016). Nesse sentido, no trabalho de Murphy (2004) foi proposto um arcabouço teórico para a avaliação da colaboração em fóruns de discussão. A proposta de Murphy (2004) foi experimentada posteriormente em diversas pesquisas (MURPHY, 2004; FERGUSON, 2009; MINHOTO; MEIRINHOS, 2012; GUIMARÃES; CAÇÃO; COUTINHO, 2013; ALRUSHIEDAT; OLFMAN, 2013). Na mesma linha, Garrison *et al.* (2001) propuseram um modelo teórico, denominado de CoI, para a avaliação de discussões em fóruns. Este modelo foi avaliado por pelo menos uma centena de pesquisas nas últimas 2 décadas (STENBOM, 2018; GARRISON; ANDERSON; ARCHER, 2001). Pesquisas as quais evidenciaram a validade das suas três dimensões, conhecidas por presenças Social, Cognitiva e de Ensino, como instrumento para mensurar a interação social, atividade docente e a produção de conhecimento em discussões assíncronas (ARBAUGH, 2013; ARBAUGH; BANGERT; CLEVELAND-INNES, 2010; BURGESS et al., 2010; CARLON et al., 2012; FERREIRA et al., 2022; FERREIRA et al., 2021; BARBOSA et al., 2021; FERREIRA et al., 2021a; FERREIRA et al., 2021b; FERREIRA et al., 2020a; TEIXEIRA et al., 2020).

Com o surgimento de tais metodologias e/ou estratégias para impulsionar o aprendizado e a colaboração nos fóruns, evidencia-se a necessidade dos professores acompanharem o impacto das suas respectivas aplicações na aprendizagem dos estudantes (REIS, 2008). Contudo, ao passo em que as discussões nos fóruns são aprofundadas, e com o aumento no número de contribuições textuais inseridas pelos estudantes, aumenta a dificuldade de análise e do acompanhamento docente dos fóruns (REIS, 2008; TENÓRIO; JUNIOR; TENÓRIO, 2015; ALMATRAFI; JOHRI; RANGWALA, 2017; CHEN et al., 2018; RAPANTA et al., 2020).

Dessa forma, com o objetivo de facilitar o acompanhamento dos fóruns sob a óptica da teoria educacional proposta por Garrison *et al.* (2001), pesquisadores propuseram abordagens para a classificação automática de postagens de acordo com a Presença Cognitiva do CoI, sendo identificados sete trabalhos para a língua inglesa e um para a língua portuguesa (NETO et al., 2018b; KOVANOVIĆ et al., 2016; GUTIÉRREZ-SANTIUSTE; RODRÍGUEZ-SABIOTE; GALLEGO-ARRUFAT, 2015; AKYOL; GARRISON, 2011; MCKLIN, 2004; LIU et al., 2022; WATERS et al., 2015; NETO et al., 2018b). No momento da concepção desta tese, também foi mapeamento um trabalho focado na identificação da Presença Social em discussões escritas em Português

(ZOU et al., 2021b). Por se tratarem de análises de contruições textuais, os referidos estudos utilizaram técnicas de uma área da Computação denominada de Processamento de Linguagem Natural (PLN) (FERREIRA-MELLO et al., 2019). Onde um dos fatores relacionados ao menor número de estudos para línguas que não sejam o inglês, como o Português, está na quantidade e na maior maturidade das ferramentas/técnicas de PLN existentes para o idioma Inglês (ALBUQUERQUE et al., 2023). Tendo em vista o foco desta tese nas interações do tipo estudante-estudante, as quais são contempladas pelas presenças Social e Cognitiva do CoI, bem como o crescimento da EAD no Brasil onde o idioma principal é o Português, emerge a primeira questão de pesqquisa:

Pergunta de Pesquisa 1 (PP1) Como analisar as contribuições textuais inseridas nos fóruns, escritas em Português e Inglês, considerando as presenças Social e Cognitiva do modelo CoI?

Outro ponto importante no contexto dos fóruns é que apesar da construção do design instrucional com base em teorias de aprendizagem (ex. CoI) e do acompanhamento intenso por parte do corpo docente, a não aquisição prévia de competências/habilidades por parte dos estudantes relacionadas autorregulação da aprendizagem pode comprometer o aprendizado de alto nível (SHEA; BIDJERANO, 2009; GAŠEVIĆ et al., 2019). Nesse sentido, pesquisas têm estudado a roteirização dos papéis que os estudantes devem adotar durante as discussões, bem como os pápeis que podem emergir em função das interações sociais (LIN, 2020; PETERSON; ROSETH, 2016; SCHELLENS et al., 2007). Desse modo, GAŠEVIĆ *et al.* (2019) estudaram a criação e acompanhamento de papéis (roteirizados e emergentes) desempenhados pelos estudantes durante as discussões sob o olhar da Presença Cognitiva do CoI (GAŠEVIĆ et al., 2019). Os estudos supracitados consistem de abordagens a serem aplicadas manualmente após a finalização das discussões, além de englobar apenas a dimensão cognitiva do CoI, o que possibilita apenas o (re)adequação dos designs instrucionais de discussões futuras. Sendo assim, emerge a segunda questão de pesquisa:

Pergunta de Pesquisa 2 (PP2) Como analisar de forma automática os papéis roteirizados e emergentes, em fóruns de dicussões, considerando as presenças Social e Cognitiva a medida em que as discussões evoluem?

1.1 OBJETIVOS

O principal objetivo desta pesquisa é desenvolver e validar uma abordagem que permita a classificação automática de postagens de fóruns, escritas em Português e Inglês, e a análise automatizada dos papéis roteirizados e emergentes a medida em que as discussões evoluem com base nas presenças Social e Cognitiva do CoI. Para atingir este objetivo duas questões de pesquisa foram elencadas, são elas:

PP1. Como analisar as contribuições textuais inseridas nos fóruns, escritas em Português e Inglês, considerando as presenças Social e Cognitiva do modelo CoI? Tendo como objetivos específicos associados:

- Construir um classificador para a Presença Social para mensagens escritas em Inglês (Capítulo 3 / Apêndice A e captítulo 4 / Apêndice B);
- Construir um classificador para a Presença Social para mensagens escritas em Português (Capítulos 5 / Apêndice C);
- Identificar as características linguísticas utilizadas nas pesquisas que propuseram classificadores para a Presença Cognitiva (Português e Inglês) (Capítulo 6 / Apêndice D);
- Construir uma abordagem multilingua capaz de classificar automaticamente postagens de fóruns de discussão de acordo com as categorias das presenças Social e Cognitiva do modelo CoI (Capítulo 6 / Apêndice D);

PP2. Como analisar de forma automática os papéis roteirizados e emergentes, em fóruns de dicussões, considerando as presenças Social e Cognitiva do CoI a medida em que as discussões evoluem? Tendo como objetivo específico associado:

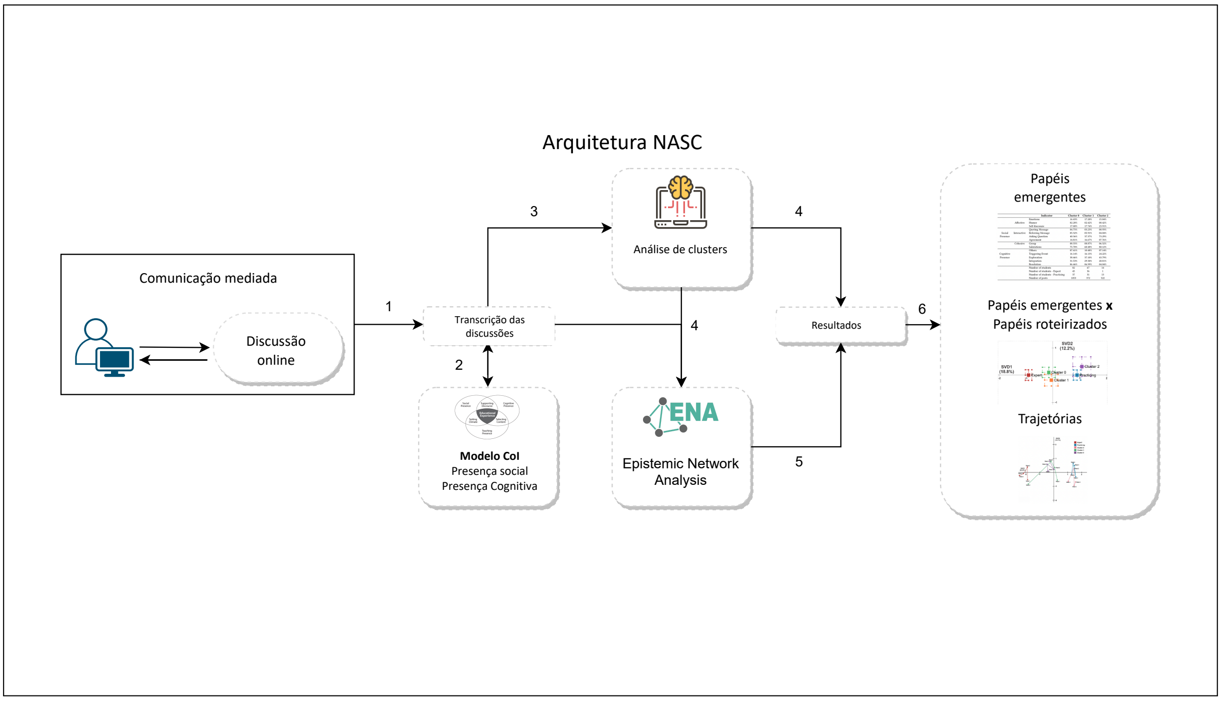
- Produzir e avaliar uma abordagem para a análise visual e automática dos papéis roteirizados e emergentes, ao passo em que as discussões evoluem, considerando as presenças Social e Cognitiva do CoI (Capítulo 7 / Apêndice G).

1.2 VISÃO GERAL

Tendo em vista os problemas de pesquisa elencados e a partir do atingimento dos objetivos traçados, esta tese consiste em uma abordagem denominada de NASC. Conforme mostra a

Figura 1, a NASC é composta pela combinação de classificadores multilingua (Português e Inglês) e uma estrutura de análise visual a qual é formada pela análise de clusters e pela a análise de redes espistêmicas com o objetivo de possibilitar o acompanhamento dos papéis roteirizados e emergentes, em fóruns educacionais, com base nos indicadores das presenças Social e Cognitiva do modelo CoI.

Figura 1 – Abordagem NASC



Fonte: o autor (2023)

De acordo com os objetivos traçados, a construção da NASC iniciou pela criação de classificadores individuais baseados na Presença Social do CoI para os idiomas Inglês e Português. Para tal, as fases descritas abaixo foram utilizadas:

- **Seleção de recursos linguísticos** - nesta fase, foram utilizados os algoritmos estado da arte na área de árvores de decisão *Random Forest*, *Adaptive Boosting* (AdaBoost) e *eXtreme Gradient Boosting* (XGBoost). Tais algoritmos foram adotados por serem considerados caixa-branca e, com isso, permitirem uma interpretação dos recursos linguísticos mais importantes no momento da modelagem. Ainda, realizou-se um comparativo dos algoritmos citados com uma rede neural de última geração denominada de *Bidirectional Encoder Representations from Transformers* (BERT). Para construir os modelos, um total de 13 recursos linguísticos foram amplamente combinados e avaliados como forma de identificar quais os mais relevantes para a construção de um classificador capaz de categorizar postagens de acordo com a Presença Social do modelo CoI. Dentre os recursos, podem-se citar o *Linguistic inquiry and word count* (LIWC) (TAUSCZIK; PENNEBAKER, 2010), *Coh-Metrix* (MCNAMARA et al., 2014), *Latent semantic analysis* (LSA) (LANDAUER; FOLTZ; LAHAM, 1998), Recursos de contexto de discussão (RCD), Número de entidades nomeadas (NER) (SHELAR et al., 2020), *Social Network Analysis* (SNA) (SCOTT, 1988), Análise de sentimentos (MEDHAT; HASSAN; KORASHY, 2014), Frequência das palavras (MANNING; SCHUTZE, 1999) e entre outros.

- **Avaliação dos modelos** - nesta fase, os modelos foram avaliados por meio das principais métricas da área de Aprendizado de Máquina (AM) (*Precision*, *Recall* e κ).

Ao considerar as limitações dos recursos linguísticos em línguas que não sejam o Inglês, a segunda fase consistiu em criar uma abordagem multilingua (Português e Inglês) para as presenças Social e Cognitiva do modelo CoI. Para tal, percorreu-se os passos expostos a seguir:

- **Tradução dos textos** - nesta fase, com o objetivo de mensurar a possibilidade de transitar postagens de fóruns de discussões de suas línguas nativas (Português ou Inglês) para uma língua alvo, postagens foram traduzidas via script através da *Application Programming Interface* (API) do *Google Translator*.
- **Extração de recursos linguísticos** - nesta fase, foram extraídas características das postagens considerando os recursos LIWC (TAUSCZIK; PENNEBAKER, 2010), Coh-Metrix (MCNAMARA et al., 2014), RCD e SNA (SCOTT, 1988).
- **Avaliação dos modelos** - nesta fase, os modelos foram avaliados por meio das métricas Acurácia e κ.

Por fim, na última etapa foi construída uma abordagem capaz de avaliar a evolução, bem como a relação dos papéis roteirizados e emergentes sob o olhar das presenças Social e Cognitiva do CoI. Para isso, foram seguidas as fases apresentadas logo após:

- **Análise de clusters** - nesta fase, os estudantes são agrupados considerando as suas contribuições textuais nos fóruns. Sendo utilizadas como características as codificações atribuídas, automaticamente, para cada postagem extraída de um fórum educacional acerca das presenças Social e Cognitiva do modelo CoI.

- *Epistemic Network Analysis* (ENA) - nesta fase, são gerados gráficos ENA com foco em evidenciar as relações entre os papéis roteirizados e mergentes com as presenças Social e Cognitiva.

1.3 CONTRIBUIÇÕES DESTA TESE PARA A LITERATURA

Com o atingimento dos objetivos traçados, podemos elencar as principais contribuições desta tese:

Contribuições dos capítulos 3, 4 e 5.

- **Contribuição técnica** - avaliação do impacto do uso de características linguísticas para a análise de sentimentos no contexto do CoI; obtenção de um classificador para a Presença Social em Inglês com resultados superiores aos relatados na literatura; obtenção de um classificador para a Presença Social em Português; avaliação de algoritmos baseados em árvores (XGBoost, AdaBoost e *Random Forest*) para lidar com dados esparsos no contexto educacional;

- **Contribuição educacional**: Avaliação da relação entre as características linguísticas mais importantes para classificar a Presença Social com a teoria do CoI. Com isso, contribuímos com a validação da consistência da teoria; Criação de um modelo de classificação que pode futuramente ser acoplado em AVA para auxiliar professores na análise da Presença Social em fóruns.

Contribuições do capítulo 6.

- **Contribuição técnica** - validação da possibilidade de utilizar o *Google Translate* para traduzir automaticamente textos, da língua portuguesa para a língua inglesa, como forma de aplicar os recursos de PLN desenvolvidos para o idioma Inglês na tarefa de classificar postagens de acordo com as presenças Social e Cognitiva do CoI; validação do impacto positivo da tradução das postagens da língua portuguesa para a língua inglesa na classificação das postagens de acordo com as presenças Social e Cognitiva do CoI.

- **Contribuição educacional** - criação de uma abordagem capaz de classificar postagens de acordo com as presenças Social e Cognitiva do CoI a qual pode ser acoplada em AVA para auxiliar a mediação em fóruns.

Contribuições do capítulo 7.

- **Contribuição técnica** - avaliação do desempenho da combinação entre algoritmos de AM e Análise de Redes para fornecer um entendimento das interações nos fóruns ao passo em que as discussões evoluem;
- **Contribuição educacional** - construção de uma abordagem capaz de possibilitar aos professores o acompanhamento dos papéis roteirizados e emergentes, sob a óptica do CoI, a medida em que as discussões evoluem.

1.4 PUBLICAÇÕES

Esta seção apresenta todas as publicações desenvolvidas durante o doutorado.

- **Publicação 1** Text mining in education (FERREIRA-MELLO et al., 2019). *Wires Data Mining and Knowledge Discovery* (Qualis A1).
- **Publicação 2** Towards Automatic Content Analysis of Social Presence in Transcripts of Online Discussions (FERREIRA et al., 2020a). *International Learning Analytics and Knowledge Conference* (Qualis A1).
- **Publicação 3** Toward Automatic Classification of Online Discussion Messages for Social Presence (FERREIRA et al., 2021). *IEEE Transactions on Learning Technologies* (Qualis A1).
- **Publicação 4** Classificação Automática da Presença Social em Discussões Online Escritas em Portuguese (TEIXEIRA et al., 2020). Simpósio Brasileiro de Informática na Educação (Qualis A3).
- **Publicação 5** The impact of automatic text translation on classification of online discussions for social and cognitive presences (BARBOSA et al., 2021). *International Learning Analytics and Knowledge Conference* (Qualis A1).
- **Publicação 6** NASC: Network analytics to uncover socio-cognitive discourse of student roles (FERREIRA et al., 2022). *International Learning Analytics and Knowledge Conference* (Qualis A1).

- **Publicação 7** Analytics of Emerging and Scripted Roles in Online Discussions: An Epistemic Network Analysis Approach (FERREIRA et al., 2021a). *International Conference on Artificial Intelligence in Education* (Qualis A3).

- **Publicação 8** Adopting *Learning Analytics* to Promote Collaboration in Online Discussions Written in Portuguese (FERREIRA et al., 2021b). *Computer-Based Learning in Context*.

1.5 ORGANIZAÇÃO DO TEXTO

Esta tese está organizada como uma coletânea composta pelos artigos publicados durante a pesquisa e mencionados na seção anterior. Dessa forma, o primeiro capítulo apresenta a contextualização, as questões de pesquisa e os objetivos traçados. O capítulo 2 apresenta a fundamentação teórica e os demais capítulos (3 até o 7) apresentam resumos dos artigos publicados os quais estão disponíveis na íntegra nos apêndices. O capítulo 8, por sua vez, apresenta as considerações finais.

Do capítulo 3 até o capítulo 5 são apresentados os resumos dos estudos que tiveram como objetivo principal identificar as principais características linguísticas para a construção de classificadores para as presenças Social e Cognitiva. De forma mais específica, o capítulo 3 apresenta uma abordagem para classificar postagens de discussões *online*, escritas em Inglês, de acordo com as categorias de Presença Social do modelo CoI (artigo completo no Apêndice A). O capítulo 4 também apresenta um classificador social para o Inglês. Tendo como diferencial a avaliação de diferentes algoritmos de classificação e a utilização de características linguísticas e estruturais (artigo completo no Apêndice B). O capítulo 5, diferentemente do anterior, compartilha uma classificador social focado apenas na língua portuguesa (artigo completo no Apêndice C). O capítulo 6 relata um estudo experimental, baseada em tradução automática de texto, cujo objetivo foi propor e validar um abordagem capaz de classificar postagens, escritas em Português e Inglês, conforme as categorias das presenças Social e Cognitiva do modelo CoI (artigo completo no Apêndice D). Por fim, no capítulo 7 é apresentada uma abordagem, alicerçada em análise de clusters e na análise de redes epistêmicas, focada em analisar os papéis roteirizados e emergentes em fóruns com base nas presenças Social e Cognitiva (artigo completo no Apêndice G).

2 FUNDAMENTAÇÃO TEÓRICA

Nesta seção são apresentados os principais conceitos que envolvem esta pesquisa: *Learning Analytics* (LA), CoI, Papéis roteirizados e emergentes, Análise de fóruns educacionais usando PLN e a teoria de papéis roteirizados e emergentes.

2.1 *LEARNING ANALYTICS*

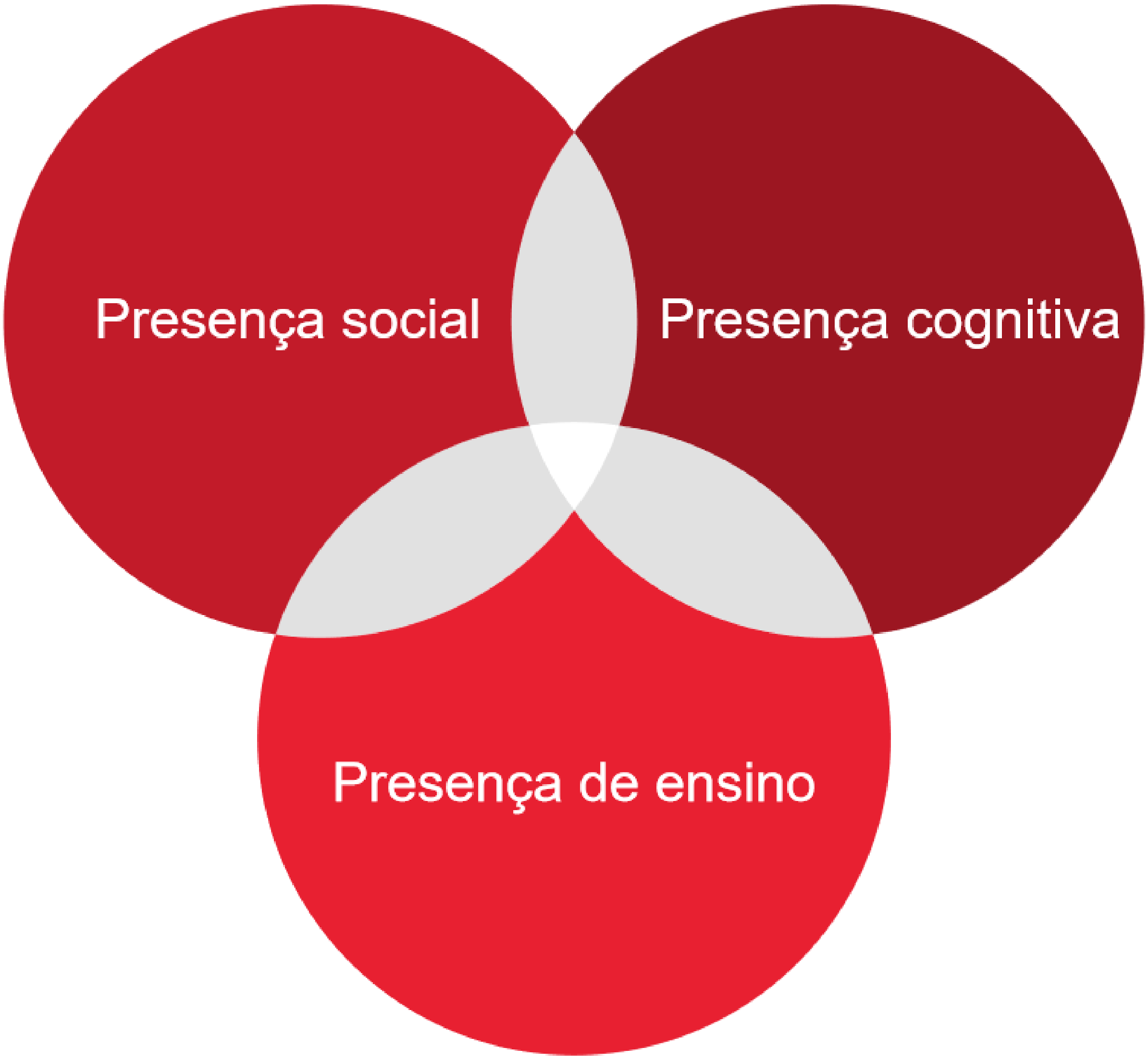
Ao longo dos últimos anos, com o crescimento das plataformas de Educação *online*, as instituições de ensino têm armazenado dados relacionados ao processo de ensino e aprendizagem (ROMERO; VENTURA, 2020), tais como: interações em fóruns, quiz, logs de interação, padrão de utilização dos ambientes e entre outros. A análise dos desses dados, por meio de técnicas de AM e Estatística, permite a identificação das preferências e das dificuldades demonstradas pelos estudantes em relação ao design instrucional, conteúdos didáticos e entre outros aspectos adotados no decorrer do proceso de ensino (ALBLAWI; ALHAMED, 2017). Com a identificação dos padrões estudantis se torna possível a personalização do ensino de modo a facilitar a construção de conhecimento (CLOW, 2013).

Sendo a área dedicada dedicada a análise de dados sobre a aprendizagem dos estudantes denominada de LA (LEITNER; KHALIL; EBNER, 2017; GAŠEVIĆ; DAWSON; SIEMENS, 2015; DYCKHOFF et al., 2012; FREITAS et al., 2017; GAŠEVIĆ; DAWSON; SIEMENS, 2015; ELIAS, 2011; SIEMENS et al., 2011). De acordo com (LANG et al., 2017), existem basicamente três abordagens para LA: **Análise de rede** - representa os atores do contexto educacional numa rede formada por nós ligados sob diferentes perspectivas: vínculos de afiliação, amizade, interação profissional, interação comportamental ou compartilhamento de informações. **Análise dos processos** - analisa as ações dos envolvidos no processo de ensino e aprendizagem por meio de *logs* do sistema e; **Análise do conteúdo** - extrai análises de conteúdos gerados pelos aprendizes, sendo comumente utilizadas técnicas de Mineração de Textos.

2.2 COMUNIDADE DE INQUÉRITO (COI)

O modelo de CoI fornece três dimensões (chamadas de presenças) conhecidas como Presença Social, Presença Cognitiva e Presença de Ensino. Como mostra a Figura 2, as presenças do CoI possuem interseções o que na prática indica uma influência mútua (GARRISON; ANDERSON; ARCHER, 2010).

Figura 2 – Dimensões do CoI



Fonte: (GARRISON; ANDERSON; ARCHER, 2010)

As referidas presenças têm como objetivo explicar as interações entre estudantes e instrutores em ambientes de AVA (GARRISON; ANDERSON; ARCHER, 1999a), conforme mostrado a seguir:

- **Presença Social** analisa a capacidade de humanizar as relações entre os participantes de uma discussão. Ele se concentra nas interações sociais e visa destacar a importância do clima social dentro de um grupo de estudantes (ROURKE et al., 1999).

- **Presença Cognitiva** visa investigar e apoiar as habilidades de resolução de problemas e pensamento crítico dos estudantes. Avalia o andamento das interações no processo Cognitivo dos estudantes, como a construção do conhecimento (GARRISON; ANDERSON; ARCHER, 2010).
- **Presença de Ensino** preocupa-se principalmente com os papéis dos instrutores antes (ou seja, design do curso) e por meio de interações sociais (ou seja, facilitação e instrução direta) dentro do curso (GARRISON; ANDERSON; ARCHER, 1999a).

O valor central da Presença Social nas discussões *online*, em contraste com as interações face a face, é possibilitar o estabelecimento de uma interação sócio-emocional por meio da comunicação baseada em texto (GARRISON; ARBAUGH, 2007). A Presença Social é analisada por meio de três categorias e doze indicadores, conforme exibido na Tabela 1. A categoria **Afetiva** examina como a tradução de emoções e sentimentos reais mapeia o texto das mensagens trocadas. Os indicadores desta categoria incluem a expressão de emoções, o senso de humor e a autorrevelação.

A categoria **Interativa** é caracterizada por indicadores como citar mensagens, fazer perguntas e expressar concordância. Destina-se a compreender e melhorar a comunicação aberta entre os participantes. Em geral, a categoria Interativa é a que mais ocorre nas discussões dentro de um CoI. Por fim, a categoria **Coesão de grupo** investiga o sentido de união e engajamento grupal. Os indicadores nesta categoria incluem vocativos e endereçamento ou referência ao grupo usando pronomes inclusivos.

A Presença Cognitiva tem um papel central no cenário CoI porque captura como "*participantes se movem deliberadamente desde a compreensão do problema ou questão até a exploração, integração e aplicação.*" (GARRISON; ARBAUGH, 2007, p. 162). Desenvolve-se por meio de um ciclo de quatro fases: 1) evento desencadeador, que inicia o ciclo apresentando problemas ou questões propostas pelo instrutor e alunos; 2) exploração, onde os participantes são estimulados a propor soluções para o problema, envolvendo *brainstorming* e troca de descobertas; 3) integração, nessa fase o estudante passa a construir novos conceitos a partir das informações coletadas; e 4) resolução, onde os estudantes propõem testes de hipóteses ou aplicações no mundo real com base no problema/dilema que desencadeou o ciclo de aprendizagem.

Apesar da importância da Presença Cognitiva, estudos anteriores destacaram que as aplicações de CoI devem estar comprometidas com o *framework* como um todo e não apenas com

Tabela 1 – Indicadores da Presença Social (ROURKE et al., 1999).

Categoria	Indicadores	Exemplos
Afetiva	1 - Expressão de emoções	Pontuações, emojis
	2 - Uso de humor	Piadas, ironias..
	3 - Autorrevelação	Falar sobre a vida além da discussão
Interativa	4 - Continuando um tópico	Responder a uma postagem ou tópico
	5 - Citando mensagens de outras pessoas	Citando mensagens de outros participantes
	6 - Referindo-se explicitamente à mensagens dos outros	Fazer referência ao conteúdo de outras mensagens
	7 - Fazendo perguntas	Perguntas direcionadas aos estudantes/professores
	8 - Elogiar, expressar apreciação	Elogiar outros participantes
	9 - Expressando acordo	Concordar com o conteúdo de outras mensagens
Coesão de grupo	10 - Vocativos	Referir-se aos demais participantes pelo nome
	11 - Referência ao grupo com pronomes inclusivos	Referir-se a todos os estudantes por meio de pronomes
	12 - Saudações	Saudações e despedidas

presenças individuais (GARRISON; ARBAUGH, 2007; GARRISON; ANDERSON; ARCHER, 2010; ROLIM et al., 2019a). Nesse contexto, tópicos como a previsão dos resultados dos estudantes (GARRISON; ANDERSON; ARCHER, 2010), o impacto da Presença Social no desenvolvimento da Presença Cognitiva (GARRISON; ARBAUGH, 2007; GARRISON; ANDERSON; ARCHER, 2010) e a análise do progresso dos estudantes nas presenças Social e Cognitiva (ROLIM et al., 2019a) têm sido amplamente estudadas.

2.3 PAPÉIS ROTEIRIZADOS E EMERGENTES

A discussão assíncrona é um elemento chave para apoiar a colaboração em ambientes *online*. No entanto, uma interação espontânea dos estudantes em discussões *online* não leva necessariamente a um alto nível de construção de conhecimento social, Presença Social e desenvolvimento cognitivo (GARRISON, 2015; FISCHER et al., 2013). Em vez disso, são necessárias abordagens eficazes para orientar os estudantes em discussões *online* produtivas. Nesse contexto, a teoria do scripts de orientação de Fischer *et al.* (2013) ganhou destaque significativo na literatura (FISCHER et al., 2013). Do ponto de vista da pesquisa em Aprendizagem colaborativa apoiada por computador (ACAC), a teoria do script de orientação postula que discussões produtivas estudante-estudante podem ser apoiadas e um alto nível de construção de conhecimento social por meio de scripts e atribuição de funções aos estudantes (STRIJBOS; WEINBERGER, 2010; FISCHER et al., 2007; FISCHER et al., 2013). De acordo com Strijbos e Weinberger (2010) os papéis são "funções ou responsabilidades mais ou menos declaradas que orientam o comportamento individual e regulam a interação do grupo", enquanto *papéis com script* são intencionalmente "projetados para melhorar os processos e resultados de aprendizagem"[p. 492]. A atribuição de papéis com script de colaboração tem sido extensivamente estudada na literatura ACAC como uma abordagem eficaz para aumentar o nível de construção de conhecimento, processamento cognitivo e argumentação (WEVER et al., 2007; ROLIM et al., 2019a; SCHELLENS et al., 2007). A atribuição de papéis e o script também são usados em pesquisas sobre CoI como uma forma de orientar os estudantes a alcançar altos níveis de Presença Social e cognitiva (GAŠEVIĆ et al., 2015c).

Em contraste com os papéis roteirizados, os papéis emergentes surgem espontaneamente de trocas interpessoais sem orientação prévia, capturam aspectos que não são predefinidos no início do trabalho em grupo (STRIJBOS; WEINBERGER, 2010). Papéis emergentes tambéms ão importantes para promover a autonomia do estudante, uma vez que papéis roteirizados podem ser vistos como promotores de conformidade (WISE; SCHWARZ, 2017) e até mesmo desmotivadores (RADKOWITSCH; VOGEL; FISCHER, 2020), apesar de serem benéficos para o aprendizado (RADKOWITSCH; VOGEL; FISCHER, 2020). Os trabalhos na área de LA permitiram a detecção automatizada de funções emergentes a partir da análise de rastros digitais registrados em fóruns de discussão *online* (SCHNEIDER; DOWELL; THOMPSON, 2021; MARTINEZ-MALDONADO et al., 2021). As abordagens existentes são baseadas em PLN, SNA, algoritmos de agrupamento e ENA (SAQR; VIBERG, 2020; ROLIM et al., 2019a; DOWELL; POQUET, 2021; MARCOS-GARCÍA; MARTÍNEZ-MONÉS; DIMITRIADIS, 2015; GAŠEVIĆ et al., 2019). Por exemplo, Dowell e Poquet (2021) propuseram uma abordagem que combina análise automatizada de conteúdo e análise de rede social para identificar papéis emergentes na discussão *online*. Gasevic *et al.* (2019) sugeriram uma abordagem que combina técnicas de análise de rede social e epistêmica, técnicas de análise de conteúdo automatizado e análise de cluster para estudar papéis emergentes sob o olhar das dimensões sociais e cognitivas da colaboração *online*.

2.4 ANÁLISE DE FÓRUNS EDUCACIONAIS USANDO PLN

Diversos estudos pontuam o potencial dos fóruns para contribuir para o engajamento e a aprendizagem *online* (REIS, 2008; TENÓRIO; JUNIOR; TENÓRIO, 2015; ALMATRAFI; JOHRI; RANGWALA, 2017; CHEN et al., 2018; RAPANTA et al., 2020), ao passo em que destacam a dificuldade para os professores em acompanhar o grande número de contribuições textuais inseridas pelos estudantes durante as discussões. Tal problema tem motivado a condução de pesquisas, no âmbito da comunidade de Informática na Educação, focadas em facilitar o acompanhamento das interações e o consequente exercício da atividade docente nos fóruns (FERREIRA-MELLO et al., 2019). Dentre as linhas de pesquisa adotadas nesse contexto, podem-se citar:

- 1) Detecção de urgência - tendo em vista a importância de fornecer *feedback* constantes para os estudantes, nesta linha de pesquisa o objetivo tem sido classificar as postagens de acordo com o seu grau de gravidade (ALMATRAFI; JOHRI; RANGWALA, 2018).
- 2) Extração de tópicos - devido ao grande volume de dados gerados durante as discussões, nesta linha de pesquisa as pesquisas estão buscando identificar quais os temas estão se sobressaindo nos debates de modo a possibilitar uma melhor segmentação (VYTASEK; WISE; WOLOSHEN, 2017; ZARRA et al., 2018; PENG; XU; GAN, 2021).
- 3) Identificação do gênero das postagens - em um fórum de discussão o tratamento dado a uma postagem deve considerar o seu teor. Por exemplo, uma postagem enquadrada como uma pergunta deve ser tratada diferentemente de uma caracterizada como afirmação. Desse modo, é possível identificar na literatura trabalhos voltados para classificar automaticamente o gênero das postagens (LIN; HSIEH; CHUANG, 2009; QU; LIU, 2012)
- 4) Fornecimento de *feedback* - com o andamento das discussões nos fóruns é importante o fornecimento de *feedback* dos professores para os estudantes. Contudo, a medida em que o número de postagens aumenta se torna inviável fornecer *feedback* em tempo hábil. Diante disso, existem trabalhos direcionados para produzir *feedback* de forma automatizada (NOURI; SAQR; FORS, 2019; WANG et al., 2022).

Sendo possível identificar também estudos cujo foco principal consistiu da identificação de aspectos colaborativos nos fóruns sob a óptica do CoI (ALMATRAFI; JOHRI, 2018; JELODAR et al., 2020; ZOU et al., 2021a; DESAI; RAMASAMY; KIPER, 2021; LIU et al., 2022; XIA; FIELDER; SIRAGUSA, 2013). Por exemplo, no trabalho de Kovanovic *et al.* (2016) foram utilizados caracteríticas como o Coh-Metrix, LIWC, LSA e entidades nomeadas para classificar postagens de acordo com a Presença Cognitiva. Em Neto *et al.* (2018) foi utilizada a mesma abordagem proposta por Kovanovic *et al.* (2016) para classificar postagens, escritas em Português, também de acordo com a Presença Cognitiva. Zou *et al.* (2021) conduziram experimentos com algoritmos de aprendizado profundo (BERT, RNN com a camada de atenção) e tradicionais (*Random Forest* e *Naive Bayes*) para classificar postagens em Inglês de acordo com a Presença Social. Gasevic *et al.* (2019) apresentaram uma abordagem que combina técnicas de análise de redes sociais e epistêmicas para analisar discussões com base nas presenças Social e Cognitiva.

2.4.1 Limitações da área de PLN para a língua portuguesa

Além das questões relevantes para o contexto educacional, este trabalho também aborda um ponto técnico relevante que é a limitação de recursos computacionais para análise de texto em Português.

Segundo a revisão sistemática realizada por Pereira *et al.* (2021), nos últimos anos tem crescido o interesse pela criação e validação de ferramentas direcionadas para o processamento linguístico em Português, como: lematização, análise morfológica, análise semântica e verificação gramatical. Podendo-se citar ferramentas como: TreeTagger, CoGrOO, LX-Center8, nlpnet11, CitiusTagger12, *Natural Language Toolkit* (NLTK) e entre outras. Apesar dos esforços para a evolução da área de PLN para a língua portuguesa, as características linguísticas para o Inglês possuem maior maturidade (ALBUQUERQUE et al., 2023; PEREIRA, 2021). Com o mesmo pensamento, Araújo *et al.* (2020) pontuaram que a maioria das características linguísticas são construídas para o Inglês, sendo grande parte das versões para outras línguas (ex. Português) adaptações lexicais, em alguns casos, sem as devidas validações. Com isso, os referidos autores conduziram experimentos onde foram realizadas traduções de textos de 14 idiomas para o inglês. Os resultados mostraram que a tradução dos textos dos seus idiomas primários para o Inglês foram benéficas. De forma semelhante, Cirqueira *et al.* (2017) apresentaram um estudo comparativo entre características linguísticas em suas versões para Português e Inglês. Corroborando com trabalhos anteriores, os resultados evidenciaram a superioridade das características para a língua inglesa.

Em resumo, em alguns contextos, utilizar mecanismos de tradução como forma de possibilitar a utilização de ferramentas construídas originalmente para o inglês pode ser mais eficaz (PEREIRA, 2021; ARAÚJO; PEREIRA; BENEVENUTO, 2020; CIRQUEIRA et al., 2017).

3 RUMO À ANÁLISE AUTOMÁTICA DE CONTEÚDO DA PRESENÇA SOCIAL EM TRANSCRIÇÕES DE DISCUSSÕES *ONLINE*

No primeiro estudo da série destinada a responder questão de pesquisa **PP1.** (ver Seção 1), buscou-se entender quais as características linguísticas e os algoritmos mais relevantes para a construção de classificadores focados na análise da Presença Social, este estudo se inspirou nas pesquisas já existentes com ênfase na Presença Cognitiva do CoI (artigo completo no Apêndice A). Dessa forma, a primeira questão de pesquisa deste artigo consistiu de:

PP1.1. Até que ponto os métodos de mineração de texto podem codificar automaticamente mensagens de discussão *online* de acordo com as categorias da Presença Social?

Além de abordar a questão de pesquisa acima treinando um algoritmo de AM supervisionado para Presença Social, também estávamos interessados em fornecer informações adicionais sobre as características mais relevantes para cada uma das três categorias de Presença Social. Para tal, exploramos um método semelhante ao aplicado por (KOVANOVIć et al., 2016) e (NETO et al., 2018b). Assim, nossa segunda questão de pesquisa é:

PP1.2. Quais características preveem melhor cada categoria da Presença Social?

Finalmente, estávamos interessados em saber se as mensagens codificadas automaticamente preservam as mesmas propriedades estruturais quando as associações entre Presença Social e tópicos de discussão foram analisadas. Ou seja, estávamos interessados em examinar até que ponto a análise de associações entre mensagens codificadas automaticamente produzia resultados semelhantes à análise de mensagens codificadas manualmente de acordo com as categorias de Presença Social. Portanto, nossa terceira questão de pesquisa é:

PP1.3. As mensagens codificadas automaticamente preservam propriedades estruturais semelhantes na análise de associações entre as categorias de Presença Social e tópicos de discussão aos resultados da análise realizada com mensagens manualmente codificadas manualmente de acordo com as categorias de Presença Social?

3.1 DESIGN EXPERIMENTAL

Este estudo foi realizado considerando uma base de dados, formada por postagens em idioma Inglês, fruto de um curso de mestrado totalmente *online* em Engenharia de Software oferecido por uma Universidade pública no Canadá. O conjunto de dados consiste em um total de 1.747 postagens resultado da interação entre 81 estudantes durante seis ofertas do curso (inverno 2008, outono 2008, verão 2009, outono 2009, inverno 2010, inverno 2011) (GAŠEVIĆ et al., 2015a). O objetivo da discussão *online* foi debater em torno de vídeos sobre trabalhos de pesquisa relacionados a um dos tópicos do curso. A participação na discussão representou 15% da nota final (GAŠEVIĆ et al., 2015a). Durante as duas primeiras ofertas do curso, a participação dos estudantes foi impulsionada principalmente pelos fatores motivacionais extrínsecos (ou seja, nota do curso), com suporte de andaime limitado. Os estudantes das duas primeiras ofertas são chamados de grupo de controle. Após as duas primeiras ofertas de cursos, foi implementado um andaime de participação na discussão por meio de atribuições de papéis e instruções claras (grupo de tratamento).

Dois codificadores especialistas categorizaram o conjunto de dados, considerando os 12 indicadores de Presença Social (KOVANOVIC et al., 2014a). Ou seja, para cada postagem no conjunto de dados, cada indicador recebeu o valor "um" (tem o indicador) ou valor "zero" (não tem o indicador). O percentual de concordância entre os avaliadores foi de 84%, sendo que um terceiro avaliador resolveu os casos discordantes. Seguindo (KOVANOVIC et al., 2014a), três indicadores (Continuando um tópico, Elogiando e Vocativos) foram removidos porque continham um alto número de mensagens.

Finalmente, como o objetivo deste estudo era construir classificadores binários para cada categoria de Presença Social, as categorias foram reorganizadas para ter codificação binária (negativo 0 ou positivo 1). Para que uma mensagem seja classificada em positivo (1), ela deve ter pelo menos um indicador anotado com o valor "um" na respectiva categoria. Por exemplo, se uma mensagem tivesse para a categoria afetiva, os indicadores Emoções = 0, Humor = 0 e Autorrevelação = 1, ela foi codificada como positiva (1). Finalmente, obtivemos o conjunto de dados conforme mostrado na Tabela 2.

Tabela 2 – Distribuição final das fases de presença social

Categoria	**Negativo (0)**	**Positivo (1)**
Afetiva	1217	530
Interativa	717	1030
Coesão de grupo	421	1326

Portanto, o conjunto de dados adotado neste estudo foi dividido em conjuntos de treinamento e conjuntos de teste; a primeira (formação) formada pelas cinco ofertas iniciais do curso (inverno 2008, outono 2008, verão 2009, outono 2009, inverno 2010) e a segunda (teste) para última oferta (inverno 2011) de acordo com as recomendações de (FARROW; MOORE; GAŠE-

VIĆ, 2019). O grupo de treinamento teve 1.510 (86%) postagens e o grupo de teste com 237 (14%) postagens. As classes negativas e positivas apresentaram distribuições aproximadas nas categorias Interativa e Coesão de grupo, com maior diferença na distribuição das classes apenas para a categoria Afetiva.

3.2 RESUMO DOS RESULTADOS

Ao abordar a questão de pesquisa 1.1, a avaliação da classificação automática de Presença Social revelou que a combinação de características tradicionais de mineração de texto e a contagens de palavras extraídas com a LIWC e o Coh-Metrix foram eficazes na classificação de mensagens de discussão *online* em todas as categorias (afetiva, interativa e coesão de grupo). A Cohen κ de 0.49, 0.83 e 0.88, para afetiva, interativa e coesão de grupo, respectivamente, representa uma concordância média a substancial entre avaliadores (LANDIS; KOCH, 1977). Além disso, em duas das três categorias o resultado ficou acima de 0.70, sendo esse resultado comumente usado, em pesquisas envolvendo o CoI, como o limite exigido antes que os resultados da codificação manual sejam considerados válidos.

Ao abordar a questão de pesquisa 1.2, este estudo realizou uma análise detalhada das características utilizadas. Dessa forma, pode-se tirar duas conclusões: (i) para cada categoria, haviam características relacionadas à palavra frequência e as ferramentas LIWC e Coh-Metrix, mostrando a importância de ambos os aspectos; (ii) embora as características relacionadas à frequência de palavras possam levar ao *overfitting* dependendo do domínio, no caso do presente estudo, as palavras com alto ganho de informação foram palavras gerais como esperança, feliz, ouvir, concordo, oi, olá, entre outros. Assim, diminui as chances de *overfitting*, pois essas palavras podem acontecer em mensagens de diferentes domínios.

A análise da importância das características também destacou uma possível correlação entre as principais características identificadas neste estudo e os indicadores considerados mais preditivos da Presença Social (ROURKE et al., 1999). Por exemplo, dentre aqueles selecionados pelo classificador da categoria Afetiva, as características: *hope*, *happi* (happy), liwc.exclam (número de pontos de exclamação) e liwc.negemo (número de emoções negativas) são relacionados à expressão de emoções e ao uso do humor. Enquanto o recurso cm.WRDPRP1s (pontuação de incidência de pronome, primeira pessoa do singular) pode estar associado ao indicador de autorrevelação, uma vez que o estudante demonstra autorrevelação ao apresentar detalhes da vida fora de casa, sala de aula ou vulnerabilidade expressa (ROURKE et al., 1999).

Para a categoria Interativa, as características liwc.you (contagem de palavras na 2ª pessoa) e cm.WRDPRP2 (pontuação de incidência do pronome da segunda pessoa) foram relacionados aos indicadores da categoria Interativa citando e referenciando abertamente outras mensagens ou pessoas na discussão. Além disso, na interação não verbal, ao fazer uma pergunta é comum o uso do sinal de pontuação interrogativo. Assim, o indicador da categoria Interativa denominado "Fazendo Perguntas"é representado pelas características liwc.QMark (número de pontos de interrogação) e liwc.interrog (número de sentenças interrogativas). Outra demonstração de interação (baseada no modelo CoI) são as expressões de concordância das mensagens dos alunos; neste aspecto, as características centrais foram a palavra *agre* (agreement) e o número de acenos por post (liwc.assent).

Por fim, a presença do recurso liwc.we (número de palavras de primeira pessoa do plural) corrobora a relevância do uso de pronomes inclusivos (nós, nosso) como forma de demonstrar a coesão do grupo. Outro indicador da categoria coesão de grupo, a demonstração de saudações, pode ser reconhecido pelas características *hi*, *hello*, liwc.affiliation (número de afiliações) e liwc.social (número de processos sociais).

Ao abordar a questão de pesquisa 1.3, adotamos a ENA para investigar as semelhanças entre as associações entre os tópicos de discussão e as categorias de Presença Social geradas com os dados codificados manual e automaticamente. Obteve-se resultados semelhantes após aplicar a ENA com códigos atribuídos manualmente e automaticamente para as projeções individuais dos estudantes e a rede de subtração dos dois grupos. Além disso, demonstramos que esses resultados estão correlacionados usando o Coeficiente de Correlação de *Pearson*. Assim, concluímos que as análises realizadas com códigos atribuídos automaticamente podem reproduzir os resultados das análises baseadas em dados codificados manualmente em um nível razoável de confiança que pode preservar as propriedades estruturais das associações de Presença Social com outras construções relevantes (MESSICK, 1995). Isso também oferece segurança adicional na validade dos resultados das análises baseadas em mensagens codificadas automaticamente.

3.3 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Este trabalho tem três contribuições. Primeiro, a proposta de três classificadores binários *Random Forest*, utilizando as características linguísticas LIWC, Coh-Metrix e frequência de palavras, para classificar automaticamente mensagens de discussão *online* em categorias de

Presença Social (Afetiva, Interativa e Coesão de grupo). Todas as categorias atingiram os valores de κ de mais de 0.49, uma concordância média a substancial entre avaliadores. Em segundo lugar, os resultados fornecem informações sobre as características psico e sociolinguísticas mais relevantes para cada indicador de Presença Social, vinculando cada uma delas à literatura CoI. Esses resultados esclarecem adicionalmente a natureza de cada indicador de Presença Social, que não havia sido relatado anteriormente na literatura. Finalmente, o uso de mensagens de discussão codificadas automaticamente na análise de associações de Presença Social com outras construções relevantes (por exemplo, tópicos de discussão) produziu resultados quase idênticos às análises realizadas com códigos de Presença Social atribuídos manualmente.

Apesar dos resultados promissores, algumas limitações podem ser identificadas, como o pequeno número de exemplos de mensagens utilizadas no presente estudo (1.747 postagens). Em seguida, os conjuntos de treinamento e teste foram divididos com base em diferentes ofertas do mesmo curso, dificultando a generalização dos resultados apresentados para outros contextos. Por fim, usar a frequência de palavras para compor vetores de características pode significar um forte viés dos modelos de classificação criados no contexto de treinamento.

Trabalhos futuros devem buscar otimizar a abordagem proposta neste estudo para reduzir a dimensionalidade dos vetores de características mantendo os resultados promissores já obtidos, o que é importante para evitar *overfitting*. Além disso, também pretendemos realizar experimentação com a abordagem e possível avaliação com tamanhos de amostra maiores compostos por dados de diferentes domínios.

4 RUMO À CLASSIFICAÇÃO AUTOMÁTICA DE MENSAGENS DE DISCUSSÃO *ONLINE* PARA PRESENÇA SOCIAL

No estudo anterior foram experimentadas características linguísticas, como a LIWC e o Coh-metrix, em conjunto com a discretização dos textos nos vetores de características. Conforme foi possível perceber a partir dos resultados alcançados, e com base na literatura, a alta dimensionalidade dos vetores pode implicar em sobreajuste dos modelos. Ainda com foco na **PP1.** (ver Seção 1), neste estudo foi ampliada a quantidade de características linguísticas avaliadas de modo a priorizar características estruturais em detrimento das oriundas exclusivamente dos sacos de palavras. A estratégia de aumentar o número de recursos teve como objetivo avaliar o maior número de características de modo a identificar as mais preditivas para, assim, obter-se vetores com baixas dimensalidades. Além disso, optou-se por avaliar o desempenho dos algoritmos de árvore de decisão como o XGBoost que foi proposto por Chen e Guestrin (2016), o AdaBoost e do modelo linguístico de aprendizado profundo BERT para detecção automática de Presença Social em discussões *online* (artigo completo no Apêndice B). Em resumo, seguimos a mesma metodologia proposta por trabalhos anteriores, desenvolvendo modelos binários, mas estendendo-os usando os algoritmos baseados em árvores mencionados acima e um novo conjunto de características. Além disso, comparamos os algoritmos baseados em árvores com o BERT. Portanto, a primeira questão de pesquisa respondida pelo presente estudo é:

PP1.4. Qual é o desempenho dos algoritmos baseados em árvores mais recentes em comparação com os convencionais na classificação da Presença Social? Até que ponto os algoritmos de árvore se aproximam de métodos de aprendizagem profunda neste contexto?

Além de avaliar novos algoritmos de árvore para o problema de identificação de Presença Social, também procuramos explorar uma variedade mais ampla de características e sua relevância para cada categoria de Presença Social. Assim, a segunda questão de pesquisa foi formulada como:

PP1.5. Quais características são as melhores preditoras das categorias de Presença Social?

Por fim, (KOVANOVIć et al., 2016) sugeriu que o aumento do número de características na análise de mensagens de discussão *online* poderia aumentar as chances de sobreajuste dos algoritmos de classificação, principalmente quando o *bag-of-words* abordagens são usadas (CORICH; HUNT; HUNT, 2006; MCKLIN, 2004). Em contraste, a adoção de características não relacionadas ao conteúdo (ou seja, LIWC e Coh-Metrix) poderia aumentar a generalização do classificador devido ao fato de que eles medem a estrutura do texto em vez do conteúdo em si. Além disso, a redução do número de características, removendo os características *bag-of-words*, também pode diminuir as chances de *overfitting* (KHALID; KHALIL; NASREEN, 2014). Assim, a última questão de pesquisa a ser respondida é:

PP1.5. Quais combinações de características que não são de conteúdo são as melhores preditoras das categorias de Presença Social?

4.1 DESIGN EXPERIMENTAL

A base de dados utilizada no presente estudo foi a mesma relatada na pesquisa anterior (ver Capítulo 2) com 1.747 mensagens. Tendo, este estudo, como principal diferencial o fato de avaliar diferentes características aplicadas à identificação automática das categorias de Presença Social, são eles: LIWC e Coh-Metrix, as características baseadas em PLN (LSA, NER e análise de sentimentos), as características estruturais (SNA e características RCD) e as abordagens tradicionais de mineração de texto (frequência de palavras).

A classificação de texto tem sido adotada como a principal abordagem para várias aplicações educacionais nos últimos anos (FERREIRA-MELLO et al., 2019). No entanto, recentemente (FARROW; MOORE; GAŠEVIĆ, 2019) sugeriu que, para aumentar a validação dos resultados, os conjuntos de treinamento e teste devem ser divididos em diferentes contextos ou cursos executados em vez de usar a divisão de dados tradicional (75%–25%). Portanto, o conjunto de dados adotado neste estudo foi dividido nas cinco ofertas iniciais do curso (inverno 2008, outono 2008, verão 2009, outono 2009, inverno 2010) e a última oferta (inverno 2011) para os conjuntos de treinamento e teste, respectivamente. Conforme 3, o grupo de treinamento teve 1510 (86%) postagens e o grupo de teste com 237 (14%) postagens. As classes negativas e positivas apresentaram distribuições aproximadas nas categorias Interativa e Coesão de grupo, com maior diferença na distribuição das classes apenas para a categoria Afetiva.

Tabela 3 – Distribuição de Postagens nos Grupos de Treinamento e Teste

	Grupo	negativa (0)	positiva (1)	Total
Afetiva	Train	1038 (69%)	472 (31%)	1510
	Test	179 (76%)	58 (24%)	237
Interativa	Train	620 (41%)	890 (59%)	1510
	Test	97 (41%)	140 (59%)	237
Coesão de grupo	Train	362 (24%)	1148 (76%)	1510
	Test	59 (25%)	178 (75%)	237

Com o objetivo de responder a pergunta de pesquisa 1, duas abordagens para analisar os dados foram seguidas: 1) uma análise por meio de uma *crossvalidation* usando apenas o conjunto de treinamento e cada curso como um *fold*; e 2) a aplicação tradicional de conjunto de treinamento e teste para construir e avaliar o modelo.

Uma análise comparativa dos 179 algoritmos de classificação (FERNÁNDEZ-DELGADO et al., 2014) revelou que *Random Forest* e *Support Vector Machine* (SVM) (com o kernel Gaussiano) alcançam melhores resultados do que outros algoritmos tradicionais de aprendizagem de máquina. Devido a essa vantagem de desempenho e interpretabilidade das características mais relevantes, as aplicações de classificação de texto no domínio educacional geralmente adotam classificadores *Random Forest* (FERREIRA-MELLO et al., 2019). No entanto, a literatura recente introduziu novos algoritmos de árvore de decisão, AdaBoost e XGBoost, que podem alcançar melhores resultados quando comparados ao *Random Forest* (JHAVERI et al., 2019; CHEN; GUESTRIN, 2016).

Apesar das semelhanças, esses algoritmos possuem algumas diferenças. Segundo (HARTMANN, 2021), o *Random Forest* é uma técnica chamada bagging que combina diferentes árvores de decisão geradas a partir da subamostragem dos exemplos presentes no conjunto de treinamento. A previsão final de um *Random Forest* é determinada a partir da média dos resultados individuais. Já os algoritmos AdaBoost e XGBoost utilizam a técnica de *boosting*, que consiste em um processo iterativo onde cada árvore criada foca nos exemplos que foram classificados incorretamente na iteração anterior (CHEN; GUESTRIN, 2016).

4.2 RESUMO DOS RESULTADOS

A questão de pesquisa 1.4 analisou o desempenho da árvore de decisão do estado da arte e dos classificadores de aprendizado profundo para a identificação da Presença Social. O

XGBoost e o AdaBoost alcançaram, na melhor das hipóteses, κ de 0.38, 0.79 e 0.94, para as categorias Afetiva, Interativa e Coesão de grupo, respectivamente. Esses valores representam acordos médios entre os avaliadores (LANDIS; KOCH, 1977). Em duas das três categorias, a concordância foi acima de 0.70, que é o valor comumente usado na pesquisa de CoI como o limite acima do qual os resultados da codificação manual são considerados válidos. Esses valores representaram um aumento de 123%, 27% e 20% para as categorias Afetiva, Interativa e Coesão de grupo, respectivamente, quando comparados aos resultados do *Random Forest*. Uma possível causa para o resultado inferior na classificação da categoria Afetiva é o número de instâncias positivas (mensagens contendo um indicador desta categoria) no conjunto de dados. Enquanto a categoria Afetiva possui 530 instâncias positivas, a Interativa e Coesão de grupo possuem 1030 e 1326, respectivamente.

Além dessa análise, foram avaliados os classificadores, excluindo as características de frequências de palavras. Até onde sabemos, tal análise nunca foi realizada antes no contexto da Presença Social. Os resultados mostraram uma queda no desempenho de 23%, 3% e 27% para as classes Afetiva, Interativa e Coesão de grupo, respectivamente. Evidenciando que a identificação automática das dimensões CoI é dependente do domínio da discussão (GARRISON; ANDERSON; ARCHER, 2010). No entanto, os resultados 0.29 (Afetiva), 0.76 (Interativa) e 0.68 (Coesão de grupo) ainda representam uma concordância média entre avaliadores para as categorias Interativa e Coesão de grupo (LANDIS; KOCH, 1977). É importante destacar que a base analisada possui conjuntos de dados desbalanceados para as categorias Afetiva e Coesão de grupo.

Por fim, os resultados apresentados também mostraram que os melhores classificadores de árvore de decisão superam o modelo BERT neste caso. Isso corrobora que a combinação de características artesanais e classificadores de árvore de decisão são a melhor opção para a identificação automática de presenças de modelos CoI (KOVANOVIć et al., 2016; NETO et al., 2021), e mais especificamente para Presença Social (BARBOSA et al., 2021; FERREIRA et al., 2020a). Existem trabalhos encontrados na literatura que alcançaram melhores resultados usando deep learning (ZOU et al., 2021b; ZOU et al., 2021a). No entanto, esses estudos não exploraram características extraídas por meio do Coh-Metrix e da LIWC para a criação dos modelos *Random Forest*.

Ao abordar a questão de pesquisa 1.5, este estudo realizou uma análise detalhada de vários grupos de características. Dentre as características utilizadas, as características analisadas também estiveram presentes em estudos anteriores (KOVANOVIĆ et al., 2016; FERREIRA et al., 2020a) além de novas características que foram selecionadas por sua relevância para as categorias de Presença Social, como a SNA e a análise de sentimentos.

A primeira análise conduzida demonstrou o potencial dos algoritmos de árvore de decisão para reduzir a dimensionalidade de um vetor de características. Os algoritmos analisados neste estudo (*Random Forest*, XGBoost e AdaBoost) realizam uma análise de importância de características usando a medida *Mean Decrease in Gini* (MDG) para classificar as características de acordo com sua relevância. As características com MDG igual a zero foram retiradas da etapa de treinamento. Até onde sabemos, nenhum trabalho anterior, em ambientes educacionais, realizou essa análise.

O conjunto inicial investigado era composto por 50.644 características. O classificador *AdaBoost* reduziu essas características para apenas 42 na classificação da categoria Interativa. No pior caso, para a categoria Coesão de grupo, o algoritmo *Random Forest* reduziu o conjunto de características para 8.42% do original.

Foi possível também perceber um alto domínio de características de frequência de palavras de acordo com a diminuição da MDG. Trabalhos anteriores também apresentaram uma tendência semelhante (FERREIRA et al., 2020a); no entanto, as principais características neste estudo tiveram uma maior dependência das frequências das palavras. Embora as características relacionadas à frequência de palavras possam levar a um *overfitting* dependendo do domínio, no caso do presente estudo, as palavras com alto ganho de informação foram as gerais, como *hope*, *happy* e *due* (Afetiva); *"agree"*, *"absolutely"* e *"depends"* (interativa); *"hello"*, *"hi"* e *"happy"* (coesão de grupo).

Por fim, a questão de pesquisa 1.5 analisou as características mais significativas para o problema de detecção automática de Presença Social, excluindo as características de frequência de palavras. Entre as características selecionadas pelo classificador das categorias afetivas, as mais importantes foram relacionadas à contagem de palavras (cm.DESWC e LIWC.WC), ao número de pronomes de primeira pessoa (LIWC.i e cm.WRDPRP1) e ao número de pontos de exclamação (LIWC.Exclam), o que está alinhado com a literatura (FERREIRA et al., 2020a). A novidade dessa análise foram as características relacionadas à amizade (LIWC.friend), análise de sentimentos (sa.emotions) e citações a mensagens anteriores (LIWC.Quote). Essas características não foram utilizadas na literatura anterior sobre os classificadores para as presenças Social e Cognitiva.

Para a categoria afetiva as características mais importante foram os indicadores de fazer perguntas, expressar concordância e se referir explicitamente a mensagens de outras pessoas (ver A análise da categoria Coesão de grupo mostrou a importância dos pronomes numéricos de primeira e segunda pessoa e entidades nomeadas (cm.WRDPRP1, cm.WRDPRP2 e NER). Corroborando com o trabalho anterior (FERREIRA et al., 2020a).

4.3 CONSIDERAÇÕES FINAIS DO CAPÍTULO

As implicações práticas deste estudo são três. Primeiro, a Presença Social é comumente associada a várias teorias educacionais na literatura, como autorregulação e autoeficácia (DOO; BONK, 2020), prestígio dos estudantes (ZOU et al., 2021a), e tem sido usado como um fator para identificar a colaboração na discussão *online* (POQUET et al., 2018). Portanto, a classificação automática das categorias de Presença Social é relevante para o estudo desses fenômenos educacionais.

Em segundo lugar, a análise de características proposta neste estudo pode ser usada para avaliar o desempenho do estudante na discussão *online* quando combinada com outros aspectos. Por exemplo, trabalhos anteriores propuseram medidas para avaliar o desempenho dos estudantes em tarefas colaborativas, destacando a relevância da presença Social (ROLIM et al., 2019a).

Finalmente, as categorias da Presença Social podem ser úteis para auxiliar o instrutor na análise do progresso dos estudantes (ou grupo de estudantes) ao longo do tempo em painéis para facilitar o rastreamento das contribuições dos estudantes no contexto de aprendizagem colaborativa assistida por computador (YAMADA; KANEKO; GODA, 2016).

5 CLASSIFICAÇÃO AUTOMÁTICA DA PRESENÇA SOCIAL EM DISCUSSÕES *ONLINE* ESCRITAS EM PORTUGUÊS

Os estudos anteriores, ver os capítulos 2 e 3, focaram na identificação da Presença Social em postagens escritas em Inglês. Neste trabalho, também com foco na **PP1.** (ver Seção 1), buscou-se construir e validar um método que permite a classificação automática das mensagens, trocadas em fóruns *online* de ensino a distância escritas em português brasileiro, de acordo com as categorias (Afetiva, Interativa e Coesão de grupo) da Presença Social (artigo completo no Apêndice C). Para atingir esse objetivo, o método proposto faz uso de um conjunto de 116 características extraídas por meio de técnicas de mineração de texto e contagem de palavras como a LIWC e o Coh-Metrix. Dessa forma, a primeira pergunta de pesquisa deste trabalho é:

PP1.6. Até que ponto os métodos de mineração de texto podem classificar automaticamente as mensagens de discussão, escritas em Português, de acordo com as categorias da Presença Social?

Além dessa questão citada acima, pretende-se também disponibilizar informações sobre quais as características que são mais relevantes para classificar cada uma das três categorias. Para isso, foram utilizados alguns parâmetros aplicados por (KOVANOVIĆ et al., 2016), (NETO et al., 2018a) e (FERREIRA et al., 2020b). Então, a segunda questão de pesquisa é:

PP1.7 Quais características melhor preveem cada categoria da Presença Social?

5.1 DESIGN EXPERIMENTAL

A base de dados utilizada foi gerada por meio de mensagens trocadas em fóruns de discussão de um curso de graduação de Biologia, oferecido totalmente *online* por uma universidade pública brasileira. Foram extraídas 1.500 mensagens produzidas por 215 estudantes durante quatro semanas de curso. O objetivo do fórum era de promover discussões sobre um tema proposto pelo professor, em que a participação representava 20% da nota final do curso.

Dois codificadores anotaram o conjunto de dados levando em consideração os 12 indicadores da Presença Social assim como foi feito em (KOVANOVIC et al., 2014b). Cada mensagem do conjunto de dados para cada indicador recebeu o valor "um" (possui o indicador) ou o valor "zero" (não possui o indicador). Assim como em (KOVANOVIC et al., 2014b), três indicadores (Continuar uma conversa, Expressar apreço/concordância e Vocativos) foram removidos pois continham um grande número de mensagens.

Por fim, como o objetivo neste trabalho foi construir classificadores binários para cada categoria da Presença Social, as categorias foram compostas pelos indicadores anotados. Para que uma mensagem seja classificada como positiva (1), ela deve ter ao menos um indicador da respectiva categoria anotado com o valor "um". Por exemplo, se uma mensagem continha os indicadores A1 = 0, A2 = 0 e A3= 1 , então ela era considerada positiva para a categoria afetiva. Durante os experimentos, o conjunto de dados foi subdividido em 75% para treinamento e 25%.

5.2 RESUMO DOS RESULTADOS

Com o objetivo de responder a questão 1.6, foram realizados experimentos com o classificador *Random Forest*, com os mesmos parâmetros utilizados nos estudos anteriores. Com isso, para todas as categorias os resultados de acurácia foram superiores a 0.97. Com vista na PP1.7, este trabalho também analisou as contribuições das diversas características para o desempenho final do classificador. Apesar de terem sido utilizados os mesmos vetores de características para discriminar as classes (positivas e negativas) das três categorias, cada classificador possui um conjunto de variáveis diferentes consideradas como mais importantes. O algoritmo *Random Forest* usa a para definir o grau de relevância de uma característica. Para a categoria afetiva o conjunto das variáveis mais importantes apresentadas contém dez variáveis de frequência de palavras, quatro LIWC e uma Coh-Metrix. As duas mais importantes foram a palavra "ser"e a variável cm.DESSC.

Para a categoria interativa, temos um conjunto com sete variáveis LIWC, seis frequência de palavras, uma Coh-Metrix e uma do contexto de discussão. A mais importante foi a posição da mensagem dentro da discussão (message.depth). Também vale destacar que a variável liwc.wd6letters (palavras com mais de 6 letras) também atingiu uma pontuação considerável.

Finalmente, as principais variáveis da categoria Coesão de grupo, nas quais sete são frequência de palavras,quatro Coh-Metrix, três LIWC, e uma RCD. Sendo as palavras comumente utilizadas para cumprimentar/saudar obtiveram boas notas ('boa', 'noit' e 'ola'). Dessa forma, as variáveis mais importantes apresentadas, estão alinhadas com a teoria da Presença Social proposta por (GARRISON; ANDERSON; ARCHER, 1999b).

5.3 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Este trabalho abordou o problema da identificação automática das categorias da Presença Social em mensagens de fóruns educacionais, escritas em português. Neste sentido, podemos destacar duas contribuições: a proposição de três classificadores binários, um para cada categoria da presença Social: Afetiva, Interativa e Coesão de grupo; a identificação das quinze variáveis mais importantes para o classificador de cada categoria. Importante destacar que com a utilização desses classificadores, os professores/tutores podem identificar o nível da presença Social do grupo ou de cada estudante e realizar uma intervenção no decorrer do curso para melhorá-lo e consequentemente auxiliar no processo de ensino-aprendizagem.

6 O IMPACTO DA TRADUÇÃO AUTOMÁTICA DE TEXTO NA CLASSIFICAÇÃO DE DISCUSSÕES *ONLINE* PARA PRESENÇAS SOCIAL E COGNITIVA

A adoção de características oriundas de técnicas como a *bag-of-words* pode aumentar a dimensionalidade dos vetores de características (KHALID; KHALIL; NASREEN, 2014). Podendo a alta dimensionalidade impulsionar um super ajuste dos modelos ao conjunto de treinamento impactando, assim, na capacidade de generalização. No âmbito do CoI, características como a LIWC e o Coh-Metrix têm sido utilizadas (ver os Capítulos 2, 3 e 4). Contudo, as características citadas e outras oriundas de importantes bibliotecas de PLN como a NLTK, a TextBlob e o WordNet não possuem a mesma maturidade para o Português que apresentam para a língua inglesa (SCHMITT et al., 2019; FERREIRA; OLIVEIRA; RODRIGUES, 2019).

Portanto, buscando responder a **PP1.** (ver Seção 1), este estudo relata os resultados obtidos com o emprego de métodos de tradução automática de texto na classificação automatizada de mensagens de discussão *online* de acordo com as categorias das presenças Social e Cognitiva (artigo completo no Apêndice D). Assim, nossa primeira questão de pesquisa é:

PP1.8. Qual a eficácia da adoção de métodos de tradução automática de textos para diferentes idiomas no processo de classificação automática de mensagens de discussão *online* para as presenças Social e Cognitiva?

Especificamente, a questão de pesquisa 1.8. teve como objetivo analisar até que ponto a precisão da classificação automática pode ser alcançada quando métodos de tradução de texto são usados. Além da avaliação da precisão do classificador, também nos interessou fornecer informações adicionais sobre as características mais relevantes para os textos originais e traduzidos. Essa análise pode fornecer *insights* sobre o impacto da limitação das características linguísticas para idiomas diferentes do inglês. Assim, a segunda pergunta de pesquisa foi:

PP1.9. Quais recursas melhor preveem presenças Social e Cognitiva para mensagens de transcrição *online* originais e traduzidas?

A questão de pesquisa 1.9. teve como objetivo entender melhor até que ponto os métodos de tradução poderiam superar o problema de características limitadas para idiomas diferentes do Inglês.

6.1 DESIGN EXPERIMENTAL

Os experimentos deste estudo foram conduzidos em duas bases de dados, uma com 1.500 mensagens de discussão em português (ver Capítulo 4) e outra com 1.747 em inglês (ver Capítulo 2 e 3). No nosso caso, uma técnica de tradução de texto foi aplicada para superar a disponibilidade limitada das características linguísticas para diferentes idiomas. Este artigo adotou a amplamente usada API de tradução do Google [1], disponível para a linguagem Python. Portanto, traduzimos os conjuntos de dados do Inglês para o Português e vice-versa.

Seguindo-se a mesma abordagem relatada na literatura relacionada (KOVANOVIć et al., 2016; BARBOSA et al., 2020a), adotamos várias características independentes de linguagem para treinar nossos classificadores automáticos de texto para presenças Social e Cognitiva. Este estudo explora características baseadas em ferramentas linguísticas (LIWC e Coh-Metrix) e informações estruturais (SNA e RCD).

6.2 RESUMO DOS RESULTADOS

Para abordar a questão de pesquisa 1.7, avaliamos a influência da tradução do texto na classificação final das presenças Social e Cognitiva. O desempenho de precisão dos dados originais em Português melhorou quando os classificadores foram treinados no conjunto de dados traduzido para o Inglês. Na primeira análise (tradução do Inglês para o Português) é possível perceber que o desempenho reduziu drasticamente para a Presença Social diminuindo para 57%, 64% e 53% em termos de κ para as categorias afetiva, interativa e Coesão de grupo, respectivamente. Outro exemplo é o classificador de Presença Cognitiva que, usando a tradução para o Inglês, atingiu 0.83 e 0.69 em termos de precisão e κ, respectivamente. Esses resultados são semelhantes ao classificador proposto por (NETO et al., 2018b) 0.83 (precisão) e 0.72 (κ). Nosso resultado é ainda mais relevante porque (FARROW; MOORE; GAŠEVIĆ, 2019) demonstrou que os resultados de (NETO et al., 2018c) podem ter sido contaminados no processo de superamostragem de dados. É importante ressaltar que seguimos as sugestões de (FARROW; MOORE; GAŠEVIĆ, 2019) para evitar o problema de contaminação de dados, são elas: cuidados durante a etapa de processamento dos dados de modo a evitar o vazamento de dados, em especial, ao utilizar técnicas de reamostragem com foco no balanceamento; separação dos conjuntos de treinamento e teste respeitando a temporalidade dos dados (por exemplo, ofertas

[1] https://pypi.org/project/googletrans/

de um curso - treinamento em dados passados e teste em dados futuros).

A análise da tradução do Inglês para o Português resultou em perda de desempenho, exceto para o classificador de Presença Cognitiva que alcançou resultados semelhantes. Esse resultado era esperado, pois haviam mais características linguísticas disponíveis para o Inglês do que para o Português. Portanto, a tradução resultou em uma diminuição no número de características analisadas. Mesmo com esse declínio, os resultados foram comparáveis aos trabalhos anteriores (FERREIRA et al., 2020c; FARROW; MOORE; GAŠEVIĆ, 2019) sem a aplicação de técnicas de *oversampling*. Além disso, este trabalho introduziu medidas SNA como características para classificação de presenças Social e Cognitiva. Os classificadores de presença Social treinados neste estudo usando os dados originais em Inglês mostraram um aumento no desempenho em comparação com a abordagem de última geração (FERREIRA et al., 2020c).

Finalmente, nossa abordagem superou o algoritmo proposto por (BARBOSA et al., 2020a) em termos de classificação da Presença Cognitiva. (BARBOSA et al., 2020a) propôs um classificador multilíngua onde o modelo inicial foi treinado em dados ingleses e então aplicado a um conjunto de dados Português, alcançando 0.53 de κ. Neste estudo, sugerimos a tradução ao invés da abordagem multilínguas, que obteve 0.69 de κ. Embora os resultados não pudessem ser comparados diretamente, pois usaram metodologias diferentes, esses resultados indicam a relevância da abordagem analisada.

A questão de pesquisa 1.8 enfocou a interpretação da importância das características para os diferentes modelos gerados. Para atingir esse objetivo, este estudo replicou a mesma metodologia utilizada nos trabalhos anteriores (ver os Capítulos 2, 3, 4), que adotaram a MDG para estimar as principais características, para cada classificador criado. Embora tenhamos adotado os mesmos vetores de características para discriminar as classes (presenças Social e Cognitiva), cada modelo considerou diferentes variáveis como as mais importantes. Com base nos resultados relatados, podemos tirar duas conclusões: (i) as características mais importantes, em geral, foram relacionadas as características linguísticas (LIWC e Coh-Metrix); (ii) as características extraídas com a LIWC tiveram uma maior dependência do idioma, pois a importância das características da LIWC diminuíram nos conjuntos de dados traduzidos.

Os classificadores da categoria afetiva treinados com os textos originais tiveram um número significativamente maior de características pertencentes a LIWC, seguindo o mesmo resultado relatado por (FERREIRA et al., 2020c), que contém principalmente características de contagem de palavras nas 15 características mais importantes para esta categoria. Os classificadores para a categoria afetiva baseados nas traduções tiveram predominância de características oriundas do Coh-Metrix. Além disso, a característica sna.closeness, nunca antes avaliada para classificação das presenças Social e Cognitiva, teve uma importância significativa para a categoria afetiva. Em ambos os casos, os dados em Inglês e Português, as características SNA, exceto sna.closeness e as características contextuais, não foram selecionadas no top-20.

A categoria Interativa apresentou as discrepâncias mais significativas quando comparada com o trabalho anterior (FERREIRA et al., 2020c). No entanto, as discrepâncias aconteceram porque os conjuntos de características usadas eram muito diferentes. Enquanto Ferreira *et al.* (2020) adotaram características de frequências de palavras, neste estudo foram utilizadas características contextuais, além das extraídas com a SNA, ambas pontuadas como relevantes em nossa análise dos dados em Inglês. Essa constatação, importância do SNA e das características contextuais, está totalmente alinhada aos indicadores da categoria interativa (ROURKE et al., 1999), como continuar um tópico e referir-se explicitamente à mensagem dos outros. Portanto, a adição dessas características representa uma contribuição significativa para a literatura. Em contrapartida, os resultados dos dados escritos em Português não seguiram a mesma tendência. Isso aconteceu devido à natureza das discussões *online* – ou seja, fórum de perguntas e respostas – com base nas quais os dados portugueses foram coletados (NETO et al., 2018c).

Com relação a categoria de Coesão de grupo foi possível perceber a importância das características da LIWC para as postagens em Inglês e a relevância reduzida para as postagens em Português. Sendo a explicação desse contraste atribuída as diferenças dessas características em ambos os idiomas.

Finalmente, o classificador de Presença Cognitiva mostrou um maior grau de similaridade com o previamente relatado na literatura (KOVANOVIć et al., 2016; NETO et al., 2018c; FARROW; MOORE; GAŠEVIĆ, 2019; BARBOSA et al., 2020a) incluindo características relacionadas a: (i) número de palavras (cm.DESWC), (ii) número de pontos de interrogação (liwc.QMark); (iii) semelhança entre as mensagens (dcf.sim.next, dcf.sim.prev); e (iv) pronomes de primeira e segunda pessoa (liwc.i, liwc.you). Este resultado corrobora que é possível usar características não-conteúdos para categorizar as fases da Presença Cognitiva (KOVANOVIć et al., 2016). Novamente, a principal novidade da análise aqui é a inclusão de medidas SNA (sna.closeness, sna.betweenness e sna.degree) como relevantes para este contexto. Isso mostra que a dinâmica e a coesão dos estudantes são importantes para alcançar níveis mais elevados da Presença Cognitiva (KOZAN; RICHARDSON, 2014; ROLIM et al., 2019b).

6.3 CONSIDERAÇÕES FINAIS DO CAPÍTULO

A principal contribuição deste trabalho é a investigação da eficácia da adoção de métodos de tradução automática de textos como forma de mitigar a limitação de características linguísticas disponíveis para análise de diferentes idiomas. Mais especificamente, foi analisado o impacto das traduções de/para inglês e português do Brasil na classificação automática de mensagens de discussão *online* para as presenças Social e Cognitiva. Os resultados obtidos indicaram que o inglês poderia ser utilizado como língua pivô para tal objetivo. Além disso, os classificadores alcançaram resultados comparáveis aos relatados na literatura em todos os casos analisados, o que confirma o potencial do uso de métodos de tradução automática de textos. Mais importante ainda, os achados aqui obtidos sugerem que manifestações específicas das presenças Social e Cognitiva podem ser capturadas com sucesso por classificadores de texto baseados em características extraídas em outras línguas.

Este estudo também ofereceu uma análise detalhada das principais características utilizadas por cada um dos classificadores propostos. Esta investigação adicional fornece mais base sobre as características relevantes que inferem a natureza das categorias de Presença Social e as fases cognitivas. Nesse sentido, a adição de medidas SNA representou a contribuição mais significativa deste estudo. Por fim, a avaliação das características empregadas para os diferentes classificadores explica porque os resultados dos modelos portugueses atingiram resultados inferiores quando comparados com a utilização dos modelos ingleses.

7 NASC: ANÁLISE DE REDE PARA DESCOBRIR O DISCURSO SOCIOCOGNITIVO DOS PAPÉIS DOS ESTUDANTES

Os papéis que os estudantes assumem durante as discussões online são um aspecto importante da experiência educacional. Os papéis podem ser atribuídos aos estudantes e/ou podem surgir espontaneamente por meio da interação estudante-estudante. Embora a pesquisa existente proponha várias abordagens para análise de funções emergentes, existem pesquisas limitadas em métodos analíticos focadas em: i) detectar automaticamente funções emergentes que podem ser interpretadas em termos de construções de colaboração de ordem superior; ii) analisar até que ponto os estudantes cumpriram os papéis roteirizados e como os papéis emergentes se comparam aos roteirizados; e iii) acompanhar a progressão dos papéis na progressão do conhecimento social ao longo do tempo. Para preencher essas lacunas na literatura, e assim responder a **PP2.** (ver Seção 1), este estudo propõe uma abordagem analítica de rede que combina técnicas de análise de cluster e a ENA (artigo completo no Apêndice G). Assim, a primeira questão de pesquisa abordada no presente estudo foi:

PP2.1. Até que ponto os papéis emergentes que os estudantes ocupam em uma discussão *online* assíncrona podem ser identificados por meio da análise automatizada de indicadores das presenças Social e Cognitiva?

Se as funções emergentes puderem ser detectadas automaticamente, seria útil determinar como elas se relacionam com as funções do script para informar as decisões instrucionais e potencialmente fornecer informações para ferramentas baseadas em análise de *feedback* personalizado, como *OnTask* (PARDO et al., 2018). Assim, a segunda questão de pesquisa foi formulada como:

PP2.2. Até que ponto os papéis roteirizados, que visam promover discussões online produtivas, estão associados a papéis emergentes que são identificados por meio de análises automatizadas de indicadores das presenças Social e Cognitiva?

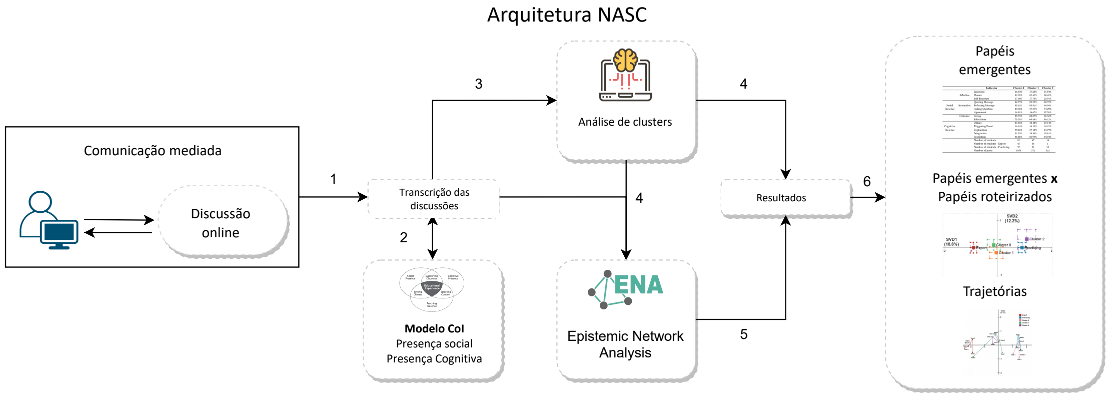
Por fim, é igualmente importante acompanhar a progressão dos aprendizes ao longo do tempo e como eles desempenham determinados papéis e como esses papéis progridem na construção de seu conhecimento social. Novamente, isso pode informar decisões sobre intervenções de ensino e até mesmo ser usado para construir *feedback* personalizado, baseado em análises em escala, para ajudar os estudantes a melhorar sua construção de conhecimento social. Assim, a terceira pergunta de pesquisa foi:

PP2.3. Até que ponto podemos rastrear mudanças nas associações entre presenças Social e Cognitiva, para cada papel, ao longo do tempo?

7.1 ABORDAGEM PROPOSTA

Para responder às questões de pesquisa propostas, propusemos uma abordagem analítica de aprendizagem chamada Análise de Rede do Discurso Sóciocognitivo (NASC). A Figura 3 apresenta as etapas gerais incluídas no NASC. O processo é iniciado (Etapa 1) com a coleta de transcrições de discussão *online* de um ambiente de aprendizagem. Em seguida, os dados são anotados de acordo com as categorias de presenças Social e Cognitiva (Etapa 2). Embora tenhamos usado um conjunto de dados que foi codificado manualmente, a literatura apresenta várias abordagens para codificação automática de mensagens de discussão *online* de acordo com as categorias de presenças Social e Cognitiva (FERREIRA et al., 2020a; BARBOSA et al., 2020b; BARBOSA et al., 2021; NETO et al., 2021). A Etapa 3 envolve o uso de um algoritmo de agrupamento para identificar os papéis emergentes e produzir a atribuição de estudantes aos papéis emergentes relevantes (Etapa 4). E, por fim, atribuições dos estudantes a papéis emergentes e roteirizados usados para construir redes epistêmicas e pré-formar análise de rede epistêmica para comparar papéis emergentes e roteirizados de modo a rastrear mudanças na construção do conhecimento social em cada papel ao longo do tempo (Etapa 5). Ao das etapas citadas, o NASC produz resultados que incluem informações sobre funções emergentes detectadas automaticamente, diferenças entre funções emergentes e com script e os gráficos de trajetória mostrando como as funções mudaram ao longo do tempo.

Figura 3 – Análise de rede do discurso sociocognitivo



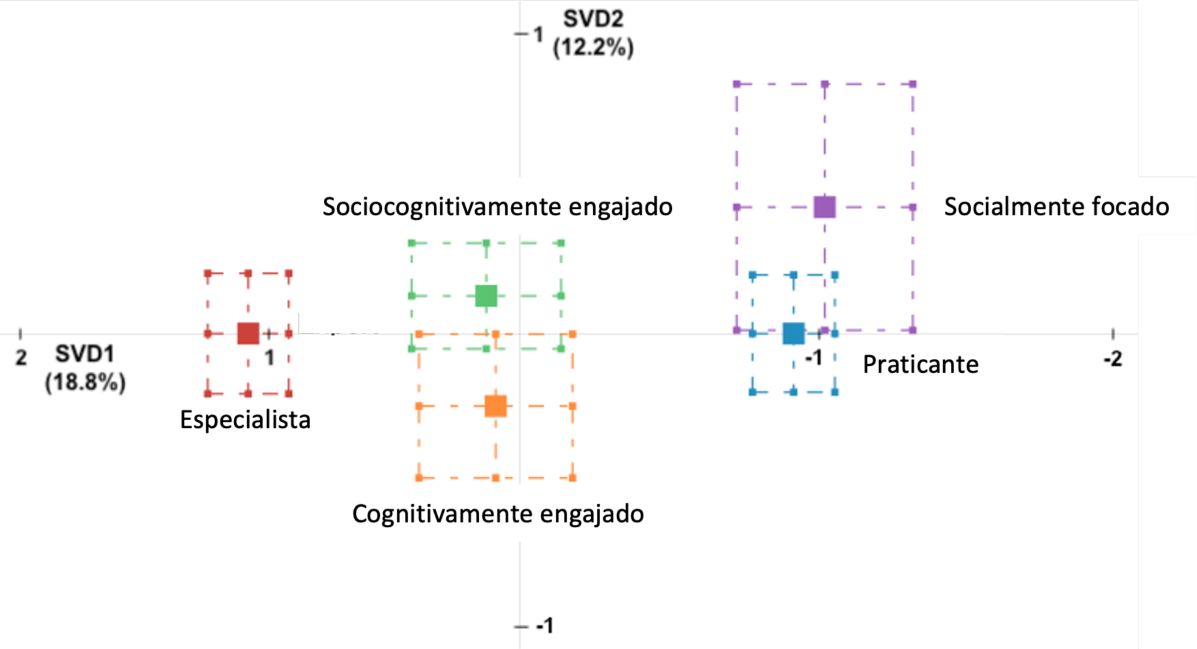
Fonte: o autor (2023)

7.2 RESUMO DOS RESULTADOS

Os resultados relacionados à questão de pesquisa 2.1 mostraram que os indicadores de Presença Social e fases de Presença Cognitiva podem ser adequados para identificar papéis emergentes (definidos como clusters), conforme confirmado com o valor da silhueta de 0.3644 (STARCZEWSKI; KRZYŻAK, 2015). Os resultados alcançados detalham as principais características dos diferentes papéis emergentes, que foram categorizados como sociocognitivamente engajados, cognitivamente engajados e socialmente focados. Conforme mostra a Figura 4, os papéis emergentes (Sociocognitivamente engajado, Cognitivamente engajado e Socialmente focado) demonstraram algumas semelhanças e diferenças com papéis roteirizados (Especialista e Praticante).

Os papéis emergentes identificados com o uso da abordagem NASC tiveram alguns paralelos com a pesquisa existente sobre papéis emergentes e roteirizados. Em primeiro lugar, a literatura mostra que quando um instrutor atribui aos estudantes papéis roteirizados, a tendência geral é que os estudantes participem de acordo com seus papéis roteirizados, especialmente quando a participação em trabalhos de grupo conta para notas finais (STRIJBOS; WEINBERGER, 2010; MAYORDOMO; ONRUBIA, 2015). No entanto, quando o mesmo estudante precisa

Figura 4 – Clusters formados a partir da abordagem NASC



Fonte: o autor (2023)

agir de maneira diferente em discussões *online* distintas, papéis anteriores que o estudantes tenham seguido pode influenciar as atitudes dos estudantes nas futuras atribuições de papéis (POZZI, 2011). Nossos resultados confirmam essa tendência, pois dois dos três papéis emergentes (exceto o papel emergente com foco social) continham estudantes de ambos os papéis roteirizados. Em segundo lugar, é importante destacar que os papéis roteirizados investigados neste estudo estavam relacionados a instruções simples (especialistas – iniciando e mantendo uma discussão e pesquisadores praticantes – fazendo perguntas e postando comentários). No entanto, a comunidade de ACAC geralmente sugere a adoção de papéis roteirizados mais relacionados cognitivamente (por exemplo, papéis argumentativos) (STRIJBOS; WEINBERGER, 2010; FISCHER et al., 2013). Portanto, os papéis emergentes identificados neste estudo mostram uma maior semelhança com trabalhos anteriores da literatura de ACAC e fornecem informações adicionais sobre o comportamento dos estudantes durante a discussão.

A questão de pesquisa 2.2 busca entender a associação entre as presenças Social e Cognitiva para papéis roteirizados e emergentes em discussões on-line assíncronas. Os resultados confirmaram que cada papel emergente tendia a estar relacionado a um dos papéis roteirizados. Enquanto os papéis emergentes sociocognitivamente engajados e cognitivamente engajados estavam mais próximos do papel roteirizado de especialista, os estudantes identificados com papel emergente com foco social tiveram um nível de participação semelhante ao do papel roteirizado de participante praticante. Apesar das semelhanças entre os papéis emergentes sociocognitivamente engajados e cognitivamente engajados e o papel do especialista, os resultados mostraram que o papel do especialista se inclinava para níveis mais altos de Presença Cognitiva. Este resultado é sensato, pois o papel do especialista foi projetado para ser responsável pelas discussões (GAŠEVIĆ et al., 2015b). Além disso, a participação dos estudantes em papéis emergentes sociocognitivamente engajados e cognitivamente engajados foi ligeiramente diferente. Por um lado, o papel sociocognitivamente engajado mostrou predominantemente Presença Social em termos de indicadores interativos; isso pode explicar a forte relação do papel sociocognitivamente engajado com as fases de exploração e integração, que são relatadas na literatura como mais relacionadas à categoria interativa de Presença Social (ROURKE et al., 1999; ROLIM et al., 2019a). Em contraste, os estudantes no papel emergente engajado cognitivamente focaram em mensagens afetivas e fizeram perguntas representativas da fase do evento desencadeador (ROURKE et al., 1999; ROLIM et al., 2019a).

Como o papel emergente com foco social abrangeu um subgrupo de estudantes que assumiram o papel de participantes praticantes, sua relação com o papel de especialista é semelhante à relação entre os papéis de prática e de especialista conforme relatado na literatura (ROLIM et al., 2019a). Em suma, o papel emergente com foco social permaneceu principalmente em baixos níveis de Presença Cognitiva e foi relacionado à saudação e à pergunta indicadores de Presença Social. A comparação direta entre o papel emergente com foco social e o papel do roteiro de pesquisador praticante mostrou que os estudantes no papel emergente com foco social não se envolveram profundamente na discussão, pois suas mensagens estavam mais relacionadas à fase do evento desencadeador e à saudação e fazendo perguntas indicadores de Presença Social. Assim, os estudantes no papel com foco social poderiam estar no grupo de desempenho inferior na discussão (DOWELL; POQUET, 2021).

A última questão de pesquisa (2.3) visava analisar as mudanças na participação de cada função emergente ao longo do tempo por meio do uso da análise de trajetória fornecida pela ENA. Essa análise enfatizou novamente a necessidade de analisar os papéis emergentes, mesmo que o design do curso tenha introduzido anteriormente papéis roteirizados. É claro que os estudantes que atuam como especialistas se concentram em responder perguntas (ou seja, fase de exploração) e encorajar a Coesão do grupo (ou seja, grupo e autorrevelação). No entanto, eles não agiram totalmente como esperado, pois o objetivo dessa função era o início e a conclusão da discussão que poderia estar relacionada às fases de desencadeamento e resolução, respectivamente (GAŠEVIĆ et al., 2015c; ROLIM et al., 2019c). Por outro lado, tanto os papéis sociocognitivamente engajados quanto os cognitivamente engajados estavam intimamente relacionados a esse objetivo.

7.2.1 Considerações finais do capítulo

Os resultados do estudo mostraram que o uso da abordagem NASC é promissor para a análise de papéis emergentes em discussões *online*. A abordagem poderia não apenas identificar automaticamente papéis emergentes teoricamente significativos, mas também compará-los com os papéis do roteiro e acompanhar a progressão ao longo do tempo em termos de construções de ordem superior bem estabelecidas das presenças Social e Cognitiva. Isso pode oferecer informações úteis aos professores para informar suas decisões e melhorar a presença geral do ensino em uma comunidade de investigação (GARRISON, 2015). O NASC pode fornecer aos professores insights sobre a eficácia de suas intenções pedagógicas (ou seja, papéis do roteiro), para que possam tomar medidas relevantes, como mudanças nos roteiros dos papéis. Também pode ajudá-los a acompanhar o progresso do estudante em relação as presenças Social e Cognitiva e criar uma base para o aprimoramento da experiência educacional, fornecendo feedback oportuno e reconhecendo os estudantes que não estão cognitivamente engajados na discussão (ROLIM et al., 2019a). Além disso, o NASC também tem potencial para ser usado na avaliação formativa de habilidades de colaboração (DOWELL; POQUET, 2021) ao enfatizar a natureza desenvolvimentista da avaliação (MARTINEZ-MALDONADO et al., 2021) por meio do uso de gráficos de trajetória da ENA.

Embora a abordagem do NASC tenha mostrado resultados promissores na análise das funções dos alunos, um potencial ainda maior está em sua integração com o corpo de pesquisa existente em análise de aprendizado e colaboração (SCHNEIDER; DOWELL; THOMPSON, 2021). A abordagem pode efetivamente ser usada em conjunto com as abordagens existentes para classificação automática de mensagens de discussão *online* com base nas presenças Social e Cognitiva (KOVANOVIć et al., 2016; BARBOSA et al., 2021; NETO et al., 2021). Isso, por sua vez, facilita a adoção potencial da abordagem analítica proposta sem a necessidade de os usuários codificarem manualmente as mensagens, como foi feito no estudo atual. No entanto, seria uma direção de pesquisa promissora investigar como a abordagem NASC pode ser complementada com o trabalho existente em análise de colaboração que faz uso da análise de redes sociais para investigar ainda mais a dimensão social da colaboração (GAŠEVIĆ et al., 2019; POQUET; JOVANOVIC, 2020) e métodos automáticos para análise de outros constructos da dimensão cognitiva da colaboração (JOKSIMOVIĆ et al., 2020).

8 CONSIDERAÇÕES FINAIS

A importância dos fóruns de dicussão no contexto da Educação *online* é amplamente defendida por professores e pesquisadores (XIA; FIELDER; SIRAGUSA, 2013; KOVANOVIć et al., 2014; ROLIM; FERREIRA; COSTA, 2016). Também é consenso a dificuldade que os professores e os tutores podem enfrentar para acompanhar os fóruns a medida em que o número de postagens aumenta. Dessa forma, esta tese propôs uma abordagem capaz de possibilitar o acompanhamento dos papéis roteirizados e emergentes, sob a óptica das presenças Social e Cognitiva do CoI, ao passo em que as discussões evoluem.

Para construir a referida abordagem a primeira preocupação foi em ter uma análise automatizada das postagens inseridas nos fóruns sob a óptica das presenças Social e Cognitiva do CoI. Nesse sentido, a partir de uma revisão da literatura foram identificados trabalhos focados em classificar automaticamente postagens de acordo com a Presença Cognitiva, sendo mapeados 7 para o idioma Inglês e 1 para o idioma Português. Além de um trabalho com foco na Presença Social do CoI para a língua inglesa. Após os achados, e considerando o crescimento da EAD no Brasil, definiu-se: 1) mapear e utilizar as boas práticas utilizadas em pesquisas anteriores no contexto da Presença Cognitiva; 2) construir uma abordagem de classificação focada na Presença Social. Durante o processo de construção do classificador social, e com o embasamento da literatura, foi possível perceber uma diferença de maturidade das ferramentas de PLN produzidas para o Inglês em relação as construídas para outros idiomas (ex. Português). Com isso, optou-se por propor uma abordagem de classificação multilingua (Português e Inglês), para as presenças Social e Cognitiva do CoI, baseada em tradução de textos, que possibilitasse o uso das ferramentas elaboradas para o Inglês mesmo em postagens escritas em Português. Os resultados obtidos evidenciaram que a tradução das postagens do idioma Português para o idioma Inglês potencializou o desempenho dos classificadores. Tendo como premissa a possibilidade de analisar automaticamente postagens de fóruns com base as presenças Social e Cognitiva do CoI, o passo seguinte consistiu em construir uma abordagem de análise visual capaz de utilizar as saídas dos classificadores para gerar *insights* visuais sobre as relações entre as presenças Social e Cognitiva do CoI e os papéis roteirizados/emergentes nos fóruns.

8.1 LIMITAÇÕES

Esta tese utilizou dois conjuntos de dados que quando somados resultam em um total de 3247 postagens de fóruns de discussão. Um dos conjuntos é oriundo de discussões ocorridas em um curso de graduação de Biologia, promovido por uma Universidade pública brasileira, enquanto o outro foi originado a partir de um curso de mestrado, em Engenharia de Software, ofertado por uma Universidade canadense.

Outro ponto importante a ser considerado são os designs instrucionais por traz de cada discussão que originou as bases de dados adotadas. A base em Português foi populada a partir de interações de 215 estudantes ocorridas em um fórum que representou 20% da nota final do curso. A base em Inglês, por sua vez, consistiu de interações de 81 estudantes em um fórum que representou 15% da nota final. É importante enfatizar esses pontos, pois motivações extrínsecas (notas) podem influenciar o comportamento dos estudantes durante os debates. Além disso, no caso da base em Inglês, existiu a atribuição de papéis roteirizados. Portanto, as generalizações a partir dos resultados obtidos devem considerar as especifidades dos cursos considerados nesta tese.

Também é relevante ponderar as decisões de cunho experimental como remover três indicadores da Presença Social (Continuar uma conversa, Expressar apreciação e Vocativos) do processo de construção da abordagem proposta. Além da adoção do *Google Translate* para traduzir as postagens gerando, assim, um viés em razão de eventuais erros de tradução.

8.2 TRABALHOS FUTUROS

- Avaliar o desempenho da abordagem proposta nesta tese em postagens oriundas de fóruns de discussões de diferentes áreas do conhecimento;
- Investigar o uso de abordagens para rotulação de dados de forma semiautomatizada para lidar com a dificuldade de obtenção de grandes massas de dados no contexto educacional;
- Experimentar abordagens de transferência de conhecimento de modo a permitir a utilização de algoritmos de Aprendizado profundo nas etapas de classificação da NASC;
- Mensurar o impacto da NASC para o acompanhamento dos níveis de interações sociais e da produção de conhecimento significativo, nos fóruns de discussões, por parte dos professores/tutores;

- Mensurar o impacto da NASC para o reconhecimento dos pápeis roteirizados/emergentes com base no CoI, nos fóruns de discussões, por parte dos professores/tutores;
- Compreender o nível de conhecimento necessário aos professores/tutores para a correta interpretação das análises geradas pela NASC;

REFERÊNCIAS

AKYOL, Z.; GARRISON, D. R. Understanding cognitive presence in an online and blended community of inquiry: Assessing outcomes and processes for deep approaches to learning. *British Journal of Educational Technology*, Wiley Online Library, v. 42, n. 2, p. 233–250, 2011.

ALBLAWI, A. S.; ALHAMED, A. A. Big data and learning analytics in higher education: Demystifying variety, acquisition, storage, nlp and analytics. In: IEEE. *2017 IEEE conference on big data and analytics (ICBDA)*. [S.l.], 2017. p. 124–129.

ALBUQUERQUE, H. O.; SOUZA, E.; GOMES, C.; PINTO, M. H. d. C.; FILHO, P. R.; COSTA, R.; LOPES, V. T. d. M.; SILVA, N. F. da; CARVALHO, A. C. de; OLIVEIRA, A. L. Named entity recognition: a survey for the portuguese language. *Procesamiento del Lenguaje Natural*, v. 70, p. 171–185, 2023.

ALENCAR, M.; SANTOS, E.; NETTO, J. F. Identifying students with evasion risk using data mining. In: CARLINER, S.; FULFORD, C.; OSTASHEWSKI, N. (Ed.). *Proceedings of EdMedia + Innovate Learning 2015*. Montreal, Quebec, Canada: Association for the Advancement of Computing in Education (AACE), 2015. p. 611–616. Disponível em: <https://www.learntechlib.org/p/151328>.

ALMATRAFI, O.; JOHRI, A. Systematic review of discussion forums in massive open online courses (moocs). *IEEE Transactions on Learning Technologies*, IEEE, v. 12, n. 3, p. 413–428, 2018.

ALMATRAFI, O.; JOHRI, A.; RANGWALA, H. Needle in a haystack: Identifying learner posts that require urgent response in mooc discussion forums. *Computers & Education*, Elsevier, 2017.

ALMATRAFI, O.; JOHRI, A.; RANGWALA, H. Needle in a haystack: Identifying learner posts that require urgent response in mooc discussion forums. *Computers & Education*, Elsevier, v. 118, p. 1–9, 2018.

ALRUSHIEDAT, N.; OLFMAN, L. Facilitating collaboration and peer learning through anchored asynchronous online discussions. In: *Proceedings of the Nineteenth Americas Conference on Information Systems*. [S.l.: s.n.], 2013. v. 19, n. 1, p. 1–10.

ANDERSON, T.; KANUKA, H. On-line forums: New platforms for professional development and group collaboration. *Journal of Computer-Mediated Communication*, Wiley Online Library, v. 3, n. 3, p. 0–0, 1997.

ARAÚJO, M.; PEREIRA, A.; BENEVENUTO, F. A comparative study of machine translation for multilingual sentence-level sentiment analysis. *Information Sciences*, Elsevier, v. 512, p. 1078–1102, 2020.

ARBAUGH, J. B. Does academic discipline moderate coi-course outcomes relationships in online mba courses? *The Internet and Higher Education*, Elsevier, v. 17, p. 16–28, 2013.

ARBAUGH, J. B.; BANGERT, A.; CLEVELAND-INNES, M. Subject matter effects and the community of inquiry (coi) framework: An exploratory study. *The internet and higher education*, Elsevier, v. 13, n. 1-2, p. 37–44, 2010.

BARBOSA, A.; FERREIRA, M.; MELLO, R. F.; LINS, R. D.; GAŠEVIĆ, D. The impact of automatic text translation on classification of online discussions for social and cognitive presences. In: *LAK21: 11th International Learning Analytics and Knowledge Conference*. [S.l.: s.n.], 2021. p. 77–87.

BARBOSA, G.; CAMELO, R.; CAVALCANTI, A. P.; MIRANDA, P.; MELLO, R. F.; KOVANOVIĆ, V.; GAŠEVIĆ, D. Towards automatic cross-language classification of cognitive presence in online discussions. In: *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. [S.l.: s.n.], 2020. p. 605–614.

BARBOSA, G. et al. Towards automatic cross-language classification of cognitive presence in online discussions. In: *Proceedings of the tenth international conference on learning analytics & knowledge*. Frankfurt, Germany: [s.n.], 2020. p. 605–614. doi: 10.1145/3375462.3375496.

BRI, D.; GARCÍA, M.; COLL, H.; LLORET, J. A study of virtual learning environments. *Wseas transactions on advances in engineering education*, v. 6, n. 1, p. 33–43, 2009.

BURGESS, M. L.; SLATE, J. R.; ROJAS-LEBOUEF, A.; LAPRAIRIE, K. Teaching and learning in second life: Using the community of inquiry (coi) model to support online instruction with graduate students in instructional technology. *The Internet and Higher Education*, Elsevier, v. 13, n. 1-2, p. 84–88, 2010.

CARLON, S.; BENNETT-WOODS, D.; BERG, B.; CLAYWELL, L.; LEDUC, K.; MARCISZ, N.; MULHALL, M.; NOTEBOOM, T.; SNEDDEN, T.; WHALEN, K. et al. The community of inquiry instrument: Validation and results in online health care disciplines. *Computers & Education*, Elsevier, v. 59, n. 2, p. 215–221, 2012.

CHEN, B.; CHANG, Y.-H.; OUYANG, F.; ZHOU, W. Fostering student engagement in online discussion through social learning analytics. *The Internet and Higher Education*, Elsevier, v. 37, p. 21–30, 2018.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. [S.l.: s.n.], 2016. p. 785–794.

CIRQUEIRA, D.; JACOB, A.; LOBATO, F.; SANTANA, A. L. de; PINHEIRO, M. Performance evaluation of sentiment analysis methods for brazilian portuguese. In: SPRINGER. *Business Information Systems Workshops: BIS 2016 International Workshops, Leipzig, Germany, July 6-8, 2016, Revised Papers 19*. [S.l.], 2017. p. 245–251.

CLOW, D. An overview of learning analytics. *Teaching in Higher Education*, Taylor & Francis, v. 18, n. 6, p. 683–695, 2013.

CORICH, S.; HUNT, K.; HUNT, L. Computerised content analysis for measuring critical thinking within discussion forums. *Journal of e-learning and knowledge society*, Italian e-Learning Association, v. 2, n. 1, 2006.

DESAI, U.; RAMASAMY, V.; KIPER, J. Evaluation of student collaboration on canvas lms using educational data mining techniques. In: *Proceedings of the 2021 ACM southeast conference*. [S.l.: s.n.], 2021. p. 55–62.

DOO, M. Y.; BONK, C. J. The effects of self-efficacy, self-regulation and social presence on learning engagement in a large university class using flipped learning. *Journal of Computer Assisted Learning*, Wiley Online Library, v. 36, n. 6, p. 997–1010, 2020. doi: 10.1111/jcal.12455.

DOWELL, N. M.; POQUET, O. Scip: Combining group communication and interpersonal positioning to identify emergent roles in scaled digital environments. *Computers in Human Behavior*, Elsevier, p. 106709, 2021.

DYCKHOFF, A. L.; ZIELKE, D.; BÜLTMANN, M.; CHATTI, M. A.; SCHROEDER, U. Design and implementation of a learning analytics toolkit for teachers. *Journal of Educational Technology & Society*, JSTOR, v. 15, n. 3, p. 58, 2012.

ELIAS, T. Learning analytics: Definitions, processes and potential. *Learning*, p. 1–22, 2011.

FARROW, E.; MOORE, J.; GAŠEVIĆ, D. Analysing discussion forum data: a replication study avoiding data contamination. In: *LAK' 2019*. [S.l.: s.n.], 2019. p. 170–179.

FERGUSON, R. *The construction of shared knowledge through asynchronous dialogue*. Tese (Doutorado) — The Open University, 2009. Disponível em: <https://oro.open.ac.uk/19908/>.

FERNÁNDEZ-DELGADO, M.; CERNADAS, E.; BARRO, S.; AMORIM, D. Do we need hundreds of classifiers to solve real world classification problems? *The journal of machine learning research*, JMLR. org, v. 15, n. 1, p. 3133–3181, 2014. doi: 10.5555/2627435.2697065.

FERREIRA, J.; OLIVEIRA, H. G.; RODRIGUES, R. Improving nltk for processing portuguese. In: SCHLOSS DAGSTUHL-LEIBNIZ-ZENTRUM FUER INFORMATIK. *8th Symposium on Languages, Applications and Technologies (SLATE 2019)*. [S.l.], 2019.

FERREIRA, M.; MELLO, R. F.; LINS, R. D.; GAŠEVIĆ, D. Analytics of emerging and scripted roles in online discussions: An epistemic network analysis approach. In: SPRINGER. *International Conference on Artificial Intelligence in Education*. [S.l.], 2021. p. 156–161.

FERREIRA, M.; PINHEIRO, A.; ROLIM, V.; LINS, R. D.; FALCãO, T. P.; MELLO, R. F. Adopting Learning Analytics to Promote Collaboration in Online Discussions Written in Portuguese. *Computer-Based Learning in Context*, v. 4, n. 1, p. 1–15, jun. 2021. Disponível em: <https://doi.org/10.5281/zenodo.4936893>.

FERREIRA, M.; ROLIM, V.; MELLO, R. F.; LINS, R. D.; CHEN, G.; GAŠEVIĆ, D. Towards automatic content analysis of social presence in transcripts of online discussions. In: *Proceedings of the tenth international conference on learning analytics & knowledge*. Frankfurt, Germany: [s.n.], 2020. p. 141–150. doi: 10.1145/3375462.3375495.

FERREIRA, M.; ROLIM, V.; MELLO, R. F.; LINS, R. D.; CHEN, G.; GAŠEVIĆ, D. Towards automatic content analysis of social presence in transcripts of online discussions. In: *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. [S.l.: s.n.], 2020. p. 141–150.

FERREIRA, M.; ROLIM, V.; MELLO, R. F.; LINS, R. D.; CHEN, G.; GAsEVIć, D. Towards automatic content analysis of social presence in transcripts of online discussions. In: *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. New York, NY, USA: Association for Computing Machinery, 2020. p. 141–150. ISBN 9781450377126. Disponível em: <https://doi.org/10.1145/3375462.3375495>.

FERREIRA, M. A. D.; MELLO, R. F.; KOVANOVIC, V.; NASCIMENTO, A.; LINS, R.; GAŠEVIĆ, D. Nasc: Network analytics to uncover socio-cognitive discourse of student roles. In: *LAK22: 12th International Learning Analytics and Knowledge Conference*. [S.l.: s.n.], 2022. p. 415–425.

FERREIRA, M. A. D.; MELLO, R. F.; NASCIMENTO, A.; LINS, R. D.; GAŠEVIĆ, D. Toward automatic classification of online discussion messages for social presence. *IEEE Transactions on Learning Technologies*, IEEE, v. 14, n. 6, p. 802–816, 2021.

FERREIRA-MELLO, R.; ANDRÉ, M.; PINHEIRO, A.; COSTA, E.; ROMERO, C. Text mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, p. e1332, 2019.

FISCHER, F.; KOLLAR, I.; MANDL, H.; HAAKE, J. M. *Scripting computer-supported collaborative learning: Cognitive, computational and educational perspectives*. [S.l.]: Springer Science & Business Media, 2007. v. 6.

FISCHER, F.; KOLLAR, I.; STEGMANN, K.; WECKER, C. Toward a script theory of guidance in computer-supported collaborative learning. *Educational psychologist*, Taylor & Francis, v. 48, n. 1, p. 56–66, 2013.

FREITAS, S. de; GIBSON, D.; ALVAREZ, V.; IRVING, L.; STAR, K.; CHARLEER, S.; VERBERT, K. How to use gamified dashboards and learning analytics for providing immediate student feedback and performance tracking in higher education. In: INTERNATIONAL WORLD WIDE WEB CONFERENCES STEERING COMMITTEE. *Proceedings of the 26th International Conference on World Wide Web Companion*. [S.l.], 2017. p. 429–434.

GABA, A. K.; LI, W. Growth and development of distance education in india and china: A study on policy perspectives. *Open Praxis*, v. 7, n. 4, p. 311–323, 2015.

GARRISON, D. R. *Thinking collaboratively: Learning in a community of inquiry*. [S.l.]: Routledge, 2015.

GARRISON, D. R.; ANDERSON, T.; ARCHER, W. Critical Inquiry in a Text-Based Environment: Computer Conferencing in Higher Education. *The Internet and Higher Education*, v. 2, n. 2-3, p. 87–105, 1999.

GARRISON, D. R.; ANDERSON, T.; ARCHER, W. Critical inquiry in a text-based environment: Computer conferencing in higher education. *The internet and higher education*, Elsevier, v. 2, n. 2-3, p. 87–105, 1999.

GARRISON, D. R.; ANDERSON, T.; ARCHER, W. Critical thinking, cognitive presence, and computer conferencing in distance education. *American Journal of distance education*, Taylor & Francis, v. 15, n. 1, p. 7–23, 2001. Disponível em: <https://doi.org/10.1080/08923640109527071>.

GARRISON, D. R.; ANDERSON, T.; ARCHER, W. The first decade of the community of inquiry framework: A retrospective. *The Internet and Higher Education*, v. 13, n. 1-2, p. 5–9, 2010.

GARRISON, D. R.; ARBAUGH, J. B. Researching the community of inquiry framework: Review, issues, and future directions. *The Internet and Higher Education*, Elsevier, v. 10, n. 3, p. 157–172, 2007.

GAŠEVIĆ, D.; ADESOPE, O.; JOKSIMOVIĆ, S.; KOVANOVIĆ, V. Externally-facilitated regulation scaffolding and role assignment to develop cognitive presence in asynchronous online discussions. *The internet and higher education*, Elsevier, v. 24, p. 53–65, 2015.

GAŠEVIĆ, D.; ADESOPE, O.; JOKSIMOVIĆ, S.; KOVANOVIĆ, V. Externally-facilitated regulation scaffolding and role assignment to develop cognitive presence in asynchronous online discussions. *The internet and higher education*, Elsevier, v. 24, p. 53–65, 2015.

GAŠEVIĆ, D.; ADESOPE, O.; JOKSIMOVIć, S.; KOVANOVIć, V. Externally-facilitated regulation scaffolding and role assignment to develop cognitive presence in asynchronous online discussions. *The Internet and Higher Education*, v. 24, p. 53–65, 2015.

GAŠEVIĆ, D.; DAWSON, S.; SIEMENS, G. Let's not forget: Learning analytics are about learning. *TechTrends*, Springer, v. 59, n. 1, p. 64–71, 2015.

GAŠEVIĆ, D.; JOKSIMOVIĆ, S.; EAGAN, B. R.; SHAFFER, D. W. Sens: Network analytics to combine social and cognitive perspectives of collaborative learning. *Computers in Human Behavior*, Elsevier, v. 92, p. 562–577, 2019.

GAYTAN, J. Comparing faculty and student perceptions regarding factors that affect student retention in online education. *American Journal of Distance Education*, Routledge, v. 29, n. 1, p. 56–66, 2015.

GOKHALE, A. A. Collaborative learning enhances critical thinking. Digital Library and Archives of the Virginia Tech University Libraries, 1995.

GUIMARÃES, I. C.; CAÇÃO, O.; COUTINHO, V. Da interação à colaboração em comunidades e fóruns de discussão. *Internet Latent Corpus Journal*, v. 3, n. 1, p. 49–64, 2013.

GUTIÉRREZ-SANTIUSTE, E.; RODRÍGUEZ-SABIOTE, C.; GALLEGO-ARRUFAT, M.-J. Cognitive presence through social and teaching presence in communities of inquiry: A correlational–predictive study. *Australasian Journal of Educational Technology*, v. 31, n. 3, 2015.

HARTMANN, J. Classification using decision tree ensembles. In: *Machine Age Customer Insight*. Bingley, UK: Emerald, 2021. p. 103–117. doi: 10.1108/978-1-83909-694-520211011.

HUGHES, S. C.; WICKERSHAM, L.; RYAN-JONES, D. L.; SMITH, S. A. Overcoming social and psychological barriers to effective on-line collaboration. *Educational Technology & Society*, JSTOR, v. 5, n. 1, p. 86–92, 2002.

INEP. *Ensino a distância cresce 474% em uma década*. 2022. Disponível em: <https://www.gov.br/inep/pt-br/assuntos/noticias/censo-da-educacao-superior/ensino-a-distancia-cresce-474-em-uma-decada>. Acesso em: 01 set. 2022.

JAIN, P. Virtual learning environment. *International Journal in IT & Engineering*, International Journals of Multi-Dimensional Research, v. 3, n. 5, p. 75–84, 2015.

JELODAR, H.; WANG, Y.; ORJI, R.; HUANG, S. Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: Nlp using lstm recurrent neural network approach. *IEEE Journal of Biomedical and Health Informatics*, IEEE, v. 24, n. 10, p. 2733–2742, 2020.

JHAVERI, S.; KHEDKAR, I.; KANTHARIA, Y.; JASWAL, S. Success Prediction using Random Forest, CatBoost, XGBoost and AdaBoost for Kickstarter Campaigns. In: *IEEE 3rd International Conference Computing Methodologies & Communication (ICCMC)*. Erode, India: [s.n.], 2019. p. 1170–1173. doi: 10.1109/ICCMC.2019.8819828.

JOKSIMOVIĆ, S.; JOVANOVIĆ, J.; KOVANOVIĆ, V.; GAŠEVIĆ, D.; MILIKIĆ, N.; ZOUAQ, A.; STAALDUINEN, J. P. V. Comprehensive analysis of discussion forum participation: from speech acts to discussion dynamics and course outcomes. *IEEE Transactions on Learning Technologies*, IEEE, v. 13, n. 1, p. 38–51, 2020.

KARA, M.; ERDOGDU, F.; KOKOÇ, M.; CAGILTAY, K. Challenges faced by adult learners in online distance education: A literature review. *Open Praxis*, v. 11, n. 1, p. 5–22, 2019.

KHALID, S.; KHALIL, T.; NASREEN, S. *A survey of feature selection and feature extraction techniques in machine learning*. 2014. 372–378 p. doi: 10.1109/SAI.2014.6918213.

KOVANOVIC, V.; JOKSIMOVIC, S.; GAŠEVIĆ, D.; HATALA, M. What is the source of social capital? the association between social network position and social presence in communities of inquiry. In: EDM. *Workshop at Educational Data Mining Conference*. [S.l.], 2014.

KOVANOVIC, V.; JOKSIMOVIC, S.; GAŠEVIĆ, D.; HATALA, M. What is the source of social capital? the association between social network position and social presence in communities of inquiry. In: *Proceedings of the Workshops held at Educational Data Mining 2014 co-located with 7th International Conference on Educational Data Mining (EDM 2014)*. [S.l.]: Citeseer, 2014.

KOVANOVIĆ, V.; JOKSIMOVIĆ, S.; WATERS, Z.; GAŠEVIĆ, D.; KITTO, K.; HATALA, M.; SIEMENS, G. Towards automated content analysis of discussion transcripts: A cognitive presence case. In: *Proceedings of the sixth international conference on learning analytics & knowledge*. [S.l.: s.n.], 2016. p. 15–24.

KOVANOVIć, V.; JOKSIMOVIć, S.; WATERS, Z.; GAsEVIć, D.; KITTO, K.; HATALA, M.; SIEMENS, G. Towards automated content analysis of discussion transcripts: A cognitive presence case. In: *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*. New York, NY, USA: Association for Computing Machinery, 2016. (LAK '16), p. 15–24. ISBN 9781450341905. Disponível em: <https://doi.org/10.1145/2883851.2883950>.

KOVANOVIĆ, V. et al. Towards automated content analysis of discussion transcripts: A cognitive presence case. In: *Proceedings 6th International Conference Learning Analytics & Knowledge (LAK'16)*. Edinburgh, UK: [s.n.], 2016. p. 15–24. doi: 10.1145/2883851.2883950.

KOVANOVIć, V.; JOKSIMOVIć, S.; GAŠEVIĆ, D.; HATALA, M. Automated cognitive presence detection in online discussion transcripts. In: *LAK'14*. Indianapolis, IN: [s.n.], 2014.

KOVANOVIć, V.; JOKSIMOVIć, S.; WATERS, Z.; GAŠEVIĆ, D.; KITTO, K.; HATALA, M.; SIEMENS, G. Towards automated content analysis of discussion transcripts: A cognitive presence case. In: *LAK'16*. New York, NY, USA: ACM, 2016. p. 15–24.

KOZAN, K.; RICHARDSON, J. C. Interrelationships between and among social, teaching, and cognitive presence. *The Internet and higher education*, Elsevier, v. 21, p. 68–73, 2014.

LANDAUER, T. K.; FOLTZ, P. W.; LAHAM, D. An introduction to latent semantic analysis. *Discourse processes*, Taylor & Francis, v. 25, n. 2–3, p. 259–284, 1998. doi: 10.1080/01638539809545028.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. *biometrics*, JSTOR, p. 159–174, 1977.

LANG, C.; SIEMENS, G.; WISE, A.; GAŠEVIĆ, D. *Handbook of Learning Analytics*. [S.l.]: SOLAR, Society for Learning Analytics and Research, 2017.

LEITNER, P.; KHALIL, M.; EBNER, M. Learning analytics in higher education—a literature review. In: *Learning Analytics: Fundaments, Applications, and Trends*. [S.l.]: Springer, 2017. p. 1–23.

LIN, F.-R.; HSIEH, L.-S.; CHUANG, F.-T. Discovering genres of online discussion threads via text mining. *Computers & Education*, Elsevier, v. 52, n. 2, p. 481–495, 2009.

LIN, G.-Y. Scripts and mastery goal orientation in face-to-face versus computer-mediated collaborative learning: Influence on performance, affective and motivational outcomes, and social ability. *Computers & Education*, Elsevier, v. 143, p. 103691, 2020.

LIU, S.; LIU, S.; LIU, Z.; PENG, X.; YANG, Z. Automated detection of emotional and cognitive engagement in mooc discussions to predict learning achievement. *Computers & Education*, Elsevier, v. 181, p. 104461, 2022.

MANNING, C.; SCHUTZE, H. Foundations of Statistical Natural Language Processing. *Computing Linguistics*, MIT press, v. 26, n. 2, p. 277–279, 1999. doi: 10.1162/coli.2000.26.2.277.

MARCOS-GARCÍA, J.-A.; MARTÍNEZ-MONÉS, A.; DIMITRIADIS, Y. Despro: A method based on roles to provide collaboration analysis support adapted to the participants in cscl situations. *Computers & Education*, Elsevier, v. 82, p. 335–353, 2015.

MARTINEZ-MALDONADO, R.; GAŠEVIC, D.; ECHEVERRIA, V.; NIETO, G. F.; SWIECKI, Z.; SHUM, S. B. What do you mean by collaboration analytics? a conceptual model. *Journal of Learning Analytics*, ERIC, v. 8, n. 1, p. 126–153, 2021.

MAYORDOMO, R. M.; ONRUBIA, J. Work coordination and collaborative knowledge construction in a small group collaborative virtual task. *The Internet and Higher Education*, Elsevier, v. 25, p. 96–104, 2015.

MCKLIN, T. E. *Analyzing Cognitive Presence in Online Courses Using an Artificial Neural Network*. Tese (Doutorado), Atlanta, GA, USA, 2004. AAI3190967.

MCNAMARA, D. S.; GRAESSER, A. C.; MCCARTHY, P. M.; CAI, Z. *Automated evaluation of text and discourse with Coh-Metrix*. [S.l.]: Cambridge University Press, 2014.

MEDHAT, W.; HASSAN, A.; KORASHY, H. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, Elsevier, v. 5, n. 4, p. 1093–1113, 2014. doi: 10.1016/j.asej.2014.04.011.

MESSICK, S. Standards of validity and the validity of standards in performance asessment. *Educational measurement: Issues and practice*, Wiley Online Library, v. 14, n. 4, p. 5–8, 1995.

MINHOTO, P.; MEIRINHOS, M. As redes sociais na promoção da aprendizagem colaborativa: um estudo no ensino secundário. *Educação, Formação & Tecnologias-ISSN 1646-933X*, v. 4, n. 2, p. 25–34, 2012.

MURPHY, E. Recognising and promoting collaboration in an online asynchronous discussion. *British Journal of Educational Technology*, Wiley Online Library, v. 35, n. 4, p. 421–431, 2004.

NETO, V.; ROLIM, V.; CAVALCANTI, A. P.; LINS LINS, R. D.; GAŠEVIĆ, D.; MELLO, R. F. Automatic content analysis of online discussions for cognitive presence: A study of the generalizability across educational contexts. *IEEE Transactions on Learning Technologies*, IEEE, 2021. doi: 10.1109/TLT.2021.3083178.

NETO, V.; ROLIM, V.; FERREIRA, R.; KOVANOVIĆ, V.; GAŠEVIĆ, D.; LINS, R. D.; LINS, R. Automated analysis of cognitive presence in online discussions written in portuguese. In: SPRINGER. *European conference on technology enhanced learning*. [S.l.], 2018. p. 245–261.

NETO, V.; ROLIM, V.; FERREIRA, R.; KOVANOVIĆ, V.; GAŠEVIĆ, D.; LINS, R. D.; LINS, R. Automated analysis of cognitive presence in online discussions written in portuguese. In: SPRINGER. *European Conference on Technology Enhanced Learning*. [S.l.], 2018. p. 245–261.

NETO, V.; ROLIM, V.; FERREIRA, R.; KOVANOVIĆ, V.; GAŠEVIĆ, D.; LINS, R. D.; LINS, R. Automated analysis of cognitive presence in online discussions written in portuguese. In: SPRINGER. *European Conference on Technology Enhanced Learning*. Springer International Publishing, 2018. p. 245–261. ISBN 978-3-319-98572-5. Disponível em: <https://doi.org/10.1007/978-3-319-98572-5_19>.

NOURI, J.; SAQR, M.; FORS, U. Predicting performance of students in a flipped classroom using machine learning: towards automated data-driven formative feedback. In: *10th International conference on education, training and informatics (ICETI 2019)*. [S.l.: s.n.], 2019. v. 17, n. 4, p. 17–21.

PANIGRAHI, R.; SRIVASTAVA, P. R.; SHARMA, D. Online learning: Adoption, continuance, and learning outcome—a review of literature. *International Journal of Information Management*, Elsevier, v. 43, p. 1–14, 2018.

PANITZ, T. The case for student centered instruction via collaborative learning paradigms. ERIC, 1999.

PARDO, A.; BARTIMOTE, K.; SHUM, S. B.; DAWSON, S.; GAO, J.; GAŠEVIĆ, D.; LEICHTWEIS, S.; LIU, D.; MARTÍNEZ-MALDONADO, R.; MIRRIAHI, N. et al. Ontask: Delivering data-informed, personalized learning support actions. *Journal of Learning Analytics*, v. 5, n. 3, p. 235–249, 2018.

PEACOCK, S.; COWAN, J.; IRVINE, L.; WILLIAMS, J. An exploration into the importance of a sense of belonging for online learners. *International Review of Research in Open and Distributed Learning*, Érudit, v. 21, n. 2, p. 18–35, 2020.

PENG, X.; XU, Q.; GAN, W. Sbtm: A joint sentiment and behaviour topic model for online course discussion forums. *Journal of Information Science*, SAGE Publications Sage UK: London, England, v. 47, n. 4, p. 517–532, 2021.

PEREIRA, D. A. A survey of sentiment analysis in the portuguese language. *Artificial Intelligence Review*, Springer, v. 54, n. 2, p. 1087–1115, 2021.

PETERSON, A. T.; ROSETH, C. J. Effects of four cscl strategies for enhancing online discussion forums: Social interdependence, summarizing, scripts, and synchronicity. *International Journal of Educational Research*, Elsevier, v. 76, p. 147–161, 2016.

POQUET, O.; JOVANOVIC, J. Intergroup and interpersonal forum positioning in shared-thread and post-reply networks. In: *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. [S.l.: s.n.], 2020. p. 187–196.

POQUET, O.; KOVANOVIĆ, V.; VRIES, P. de; HENNIS, T.; JOKSIMOVIĆ, S.; GAŠEVIĆ, D.; DAWSON, S. Social presence in massive open online courses. *International Review of Research in Open and Distributed Learning*, v. 19, n. 3, 2018.

POZZI, F. The impact of scripted roles on online collaborative learning processes. *International Journal of Computer-Supported Collaborative Learning*, Springer, v. 6, n. 3, p. 471–484, 2011.

QAYYUM, A.; ZAWACKI-RICHTER, O. The state of open and distance education. In: *Open and distance education in Asia, Africa and the Middle East*. [S.l.]: Springer, 2019. p. 125–140.

QU, Z.; LIU, Y. Sentence dependency tagging in online question answering forums. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. [S.l.], 2012. p. 554–562.

RADKOWITSCH, A.; VOGEL, F.; FISCHER, F. Good for learning, bad for motivation? a meta-analysis on the effects of computer-supported collaboration scripts. *International Journal of Computer-Supported Collaborative Learning*, Springer, v. 15, n. 1, p. 5–47, 2020.

RAPANTA, C.; BOTTURI, L.; GOODYEAR, P.; GUÀRDIA, L.; KOOLE, M. Online university teaching during and after the covid-19 crisis: Refocusing teacher presence and learning activity. *Postdigital science and education*, Springer, v. 2, n. 3, p. 923–945, 2020.

REIS, F. L. dos. A importância dos fóruns de debate na comunicação e interação no ensino online. *Revista de Estudos da Comunicação*, v. 9, n. 19, 2008.

ROLIM, V.; FERREIRA, R.; COSTA, E. Identificação automática de dúvidas em fóruns educacionais. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2016. v. 27, n. 1, p. 936.

ROLIM, V.; FERREIRA, R.; LINS, R. D.; GăSEVIĆ, D. A network-based analytic approach to uncovering the relationship between social and cognitive presences in communities of inquiry. *Internet & Higher Education*, Elsevier, v. 42, p. 53–65, 2019. doi: 10.1016/j.iheduc.2019.05.001.

ROLIM, V.; FERREIRA, R.; LINS, R. D.; GăSEVIĆ, D. A network-based analytic approach to uncovering the relationship between social and cognitive presences in communities of inquiry. *The Internet and Higher Education*, Elsevier, v. 42, p. 53–65, 2019.

ROLIM, V.; MELLO, R. F. L. de; KOVANOVIC, V.; GAŠEVIC, D. Analysing social presence in online discussions through network and text analytics. In: IEEE. *IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*. [S.l.], 2019. v. 2161, p. 163–167.

ROMERO, C.; VENTURA, S. Educational data mining and learning analytics: An updated survey. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, Wiley Online Library, v. 10, n. 3, p. e1355, 2020.

ROURKE, L.; ANDERSON, T.; GARRISON, D. R.; ARCHER, W. Assessing Social Presence In Asynchronous Text-based Computer Conferencing. *The Journal of Distance Education*, v. 14, n. 2, p. 50–71, 1999.

SAQR, M.; VIBERG, O. Using diffusion network analytics to examine and support knowledge construction in cscl settings. In: *In Proceedings of the 15th European Conference on Technology Enhanced Learning*. Cham, Switzerland: Springer, 2020. p. 158–172.

SCHELLENS, T.; KEER, H. V.; WEVER, B. D.; VALCKE, M. Scripting by assigning roles: Does it improve knowledge construction in asynchronous discussion groups? *International Journal of Computer-Supported Collaborative Learning*, Springer, v. 2, p. 225–246, 2007.

SCHMITT, X.; KUBLER, S.; ROBERT, J.; PAPADAKIS, M.; LETRAON, Y. A replicable comparison study of ner software: Stanfordnlp, nltk, opennlp, spacy, gate. In: IEEE. *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. [S.l.], 2019. p. 338–343.

SCHNEIDER, B.; DOWELL, N.; THOMPSON, K. Collaboration analytics—current state and potential futures. *Journal of Learning Analytics*, v. 8, n. 1, p. 1–12, 2021.

SCOTT, J. Social network analysis. *Sociology*, v. 22, n. 1, p. 109–127, 1988. doi: 10.1177/0038038588022001007.

SHEA, P.; BIDJERANO, T. Cognitive presence and online learner engagement: A cluster analysis of the community of inquiry framework. *Journal of Computing in Higher Education*, Springer, v. 21, n. 3, p. 199, 2009.

SHELAR, H.; KAUR, G.; HEDA, N.; AGRAWAL, P. Named entity recognition approaches and their comparison for custom NER model. *Science & Technology Libraries*, Taylor & Francis, v. 39, n. 3, p. 324–337, 2020. doi: 10.1080/0194262X.2020.1759479.

SIEMENS, G.; GAŠEVIĆ, D.; HAYTHORNTHWAITE, C.; DAWSON, S.; SHUM, S. B.; FERGUSON, R.; DUVAL, E.; VERBERT, K.; BAKER, R. *Open Learning Analytics: an integrated & modularized platform*. Tese (Doutorado) — Open University Press Doctoral dissertation, 2011.

STARCZEWSKI, A.; KRZYŻAK, A. Performance evaluation of the silhouette index. In: SPRINGER. *International Conference on Artificial Intelligence and Soft Computing*. [S.l.], 2015. p. 49–58.

STENBOM, S. A systematic review of the community of inquiry survey. *The internet and higher education*, Elsevier, v. 39, p. 22–32, 2018.

STOYTCHEVA, M. Developing a sense of belonging in a collaborative distance learning course: Breaking isolation in online learning. In: AIP PUBLISHING. *AIP Conference Proceedings*. [S.l.], 2021. v. 2333, n. 1.

STRIJBOS, J.-W.; WEINBERGER, A. Emerging and scripted roles in computer-supported collaborative learning. *Computers in Human Behavior*, Elsevier, v. 26, n. 4, p. 491–494, 2010.

TAUSCZIK, Y. R.; PENNEBAKER, J. W. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, Sage Publications Sage CA: Los Angeles, CA, v. 29, n. 1, p. 24–54, 2010.

TEIXEIRA, J. B.; COSTA, E. de B.; DIONÍSIO, M.; NASCIMENTO, A. C.; MELLO, R. F. L. de. Classificação automática da presença social em discussões online escritas em português. In: SBC. *Anais do XXXI Simpósio Brasileiro de Informática na Educação*. [S.l.], 2020. p. 942–951.

TENÓRIO, A.; JUNIOR, J. F.; TENÓRIO, T. A visão de tutores sobre o uso de fóruns em cursos a distância. *Revista Brasileira de Aprendizagem Aberta e a Distância*, v. 14, 2015.

THOMAS, L.; HERBERT, J.; TERAS, M. A sense of belonging to enhance participation, success and retention in online programs. 2014.

VYTASEK, J. M.; WISE, A. F.; WOLOSHEN, S. Topic models to support instructors in mooc forums. In: *Proceedings of the seventh international learning analytics & knowledge conference*. [S.l.: s.n.], 2017. p. 610–611.

WANG, Q.; ROSE, C. P.; MA, N.; JIANG, S.; BAO, H.; LI, Y. Design and application of automatic feedback scaffolding in forums to promote learning. *IEEE Transactions on Learning Technologies*, IEEE, v. 15, n. 2, p. 150–166, 2022.

WATERS, Z.; KOVANOVIć, V.; KITTO, K.; GAŠEVIĆ, D. Structure matters: Adoption of structured classification approach in the context of cognitive presence classification. In: *Information Retrieval Technology*. [S.l.: s.n.], 2015. p. 227–238.

WEVER, B. D.; KEER, H. V.; SCHELLENS, T.; VALCKE, M. Applying multilevel modelling to content analysis data: Methodological issues in the study of role assignment in asynchronous discussion groups. *Learning and instruction*, Elsevier, v. 17, n. 4, p. 436–447, 2007.

WISE, A. F.; SCHWARZ, B. B. Visions of cscl: Eight provocations for the future of the field. *International Journal of Computer-Supported Collaborative Learning*, Springer, v. 12, n. 4, p. 423–467, 2017.

XIA, C.; FIELDER, J.; SIRAGUSA, L. Achieving better peer interaction in online discussion forums: A reflective practitioner case study. *Issues in Educational Research*, v. 23, n. 1, p. 97–113, 2013.

YAMADA, M.; KANEKO, K.; GODA, Y. Social presence visualizer: Development of the collaboration facilitation module on CSCL. In: *International Conference Collaboration Technologies*. Kanazawa, Japan: [s.n.], 2016. p. 174–189. doi: 10.1007/978-981-10-2618-8$_14$.

ZARRA, T.; CHIHEB, R.; FAIZI, R.; AFIA, A. E. Student interactions in online discussion forums: Visual analysis with lda topic models. In: *Proceedings of the International Conference on Learning and Optimization Algorithms: Theory and Applications*. [S.l.: s.n.], 2018. p. 1–5.

ZOU, W.; HU, X.; PAN, Z.; LI, C.; CAI, Y.; LIU, M. Exploring the relationship between social presence and learners' prestige in mooc discussion forums using automated content analysis and social network analysis. *Computers in Human Behavior*, Elsevier, v. 115, p. 106582, 2021.

ZOU, W.; PAN, Z.; LI, C.; LIU, M. Does social presence play a role in learners' positions in MOOC learner network? A machine learning approach to analyze social presence in discussion forums. In: SPRINGER. *Proceedings 3rd International Conference Quantitative Ethnography (ICQE'21)*. Malibu, CA, USA, 2021. p. 248–264. doi: 10.1007/978-3-030-67788-6$_1$7.

APÊNDICE A – TOWARDS AUTOMATIC CONTENT ANALYSIS OF SOCIAL PRESENCE IN TRANSCRIPTS OF ONLINE DISCUSSIONS

Towards Automatic Content Analysis of Social Presence in Transcripts of Online Discussions

Máverick Ferreira
Centro de Informática
Universidade Federal de Pernambuco
Recife, PE, Brazil
madf@cin.ufpe.br

Vitor Rolim
Centro de Informática
Universidade Federal de Pernambuco
Recife, PE, Brazil
vbr@cin.ufpe.br

Rafael Ferreira Mello
Centro de Informática
UFRPE
Recife, Pernambuco, Brasil
rafael.mello@ufrpe.br

Rafael Dueire Lins
Centro de Informática
UFRPE
Recife, Pernambuco, Brasil
rdl@cin.ufpe.br

Guanliang Chen
Faculty of Information Technology
Monash University
Melbourne, VIC, Australia
Guanliang.Chen@monash.edu

Dragan Gašević
Faculty of Information Technology
Monash University
Melbourne, VIC, Australia
dragan.gasevic@monash.edu

ABSTRACT

This paper presents an approach to automatic labeling of the content of messages in online discussion according to the categories of social presence. To achieve this goal, the proposed approach is based on a combination of traditional text mining features and word counts extracted with the use of established linguistic frameworks (i.e., LIWC and Coh-metrix). The best performing classifier obtained 0.95 and 0.88 for accuracy and Cohen's kappa, respectively. This paper also provides some theoretical insights into the nature of social presence by looking at the classification features that were most relevant for distinguishing between the different categories. Finally, this study adopted epistemic network analysis to investigate the structural construct validity of the automatic classification approach. Namely, the analysis showed that the epistemic networks produced based on messages manually and automatically coded produced nearly identical results. This finding thus produced evidence of the structural validity of the automatic approach.

CCS CONCEPTS

• **Information systems → Clustering and classification**; • **Applied computing → E-learning**; **Distance learning**;

KEYWORDS

Community of Inquiry Model, Content Analytics, Online Discussion, Text Classification, Epistemic Network Analysis

ACM Reference format:
Máverick Ferreira, Vitor Rolim, Rafael Ferreira Mello, Rafael Dueire Lins, Guanliang Chen, and Dragan Gašević. 2020. Towards Automatic Content Analysis of Social Presence in Transcripts of Online Discussions. In *Proceedings of Proceedings of the 10th International Conference on Learning Analytics and Knowledge, Frankfurt, Germany, March 23–27, 2020 (LAK '20),* 10 pages.
https://doi.org/10.1145/3375462.3375495

1 INTRODUCTION

Online learning enables students and instructors to participate in the teaching and learning process without being in the same physical space [34]. Due to this characteristic, it has promoted access to education for people located in regions that are difficult to reach or distant from educational institutions. One of the critical challenges facing instructors of online courses is creating a supportive and productive environment for student communication and collaboration through technology [11]. According to the literature, asynchronous online discussion is a resource with high potential for promoting collaboration in online education [46] supporting students' social interactions and social-constructivist pedagogies [1], which encourage the engagement of learners [4].

Within this context, a social constructivist model called Community of Inquiry (CoI) [12] is a frameworks developed to support instructors in online learning environments. The study of CoI is heavily depended on the analysis of messages exchanged in online discussions. The most commonly used approach to this analysis is based on *Quantitative Content Analysis* (QCA) [9, 22] of the transcripts of asynchronous online discussions. Krippendorff [27] states that content analysis is *"a research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use"*[p.18]. QCA methods can use predefined coding schemes to analyze text artifacts (i.e., messages in online discussions) with respect to the defined research goals and objectives. The CoI model defines the QCA coding schemes for each of the three presences that can be applied to analyze online discussion messages. As widely done in the social sciences, research of CoI primarly uses QCA for retrospection and research after online discussions are over, without much impact on the actual student learning and outcomes in real-time [43].

However, it is possible to adopt automated methods for text analysis commonly used within learning analytics [17] to perform

LAK '20, March 23–27, 2020, Frankfurt, Germany
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7712-6/20/03…$15.00
https://doi.org/10.1145/3375462.3375495

automatic and real-time analysis of online discussion messages according to the CoI presences [23].

Existing literature reports several approaches to automating the process of content analysis of online discussion according to the coding scheme for cognitive presence as proposed in the CoI model [10, 24, 31], even for languages different than English, like Portuguese [35]. However, methods for automatic content analysis of online discussions for indicators of social presence are not commonly found in the literature.

This paper describes a method that combines several text analytics techniques for automatic content analysis of social presence from online discussion transcripts. The study combines traditional text mining features based on content words with tools that extract different psychological processes indicators, and measures of text coherence and complexity [32, 44]. We developed three classifiers, one for each category of social presence, that use different feature sets and achieved up to 0.95 and 0.88 for accuracy and Cohen's Kappa, respectively. Besides, we proposed a network analytic approach to evaluate the structural validity of our proposal in practice. The results and their implications are also discussed in the paper.

2 BACKGROUND WORK

2.1 The Community of Inquiry Model

Several models of and approaches to understanding students' interactions in online environments have been proposed. Among them, the Community of Inquiry (CoI) model is one of the most researched structures when the objective is to describe the essential facets of social interactions and knowledge construction in online and blended education [14]. CoI proposes three dimensions that explain the processes of social knowledge construction of learners and instructors with the goal of describing promoting effective educational experience. Garrison et al. [12] distinguishes between the three key dimensions of CoI, known as presences, as follows: (i) **Social presence** measures the ability to humanize the relationships among participants in a discussion. It focuses on social interactions and tries to model the social climate within a group of learners (i.e., cohesion, affectivity, and open communication) [41]; (ii) **Cognitive presence** is highly related to the development of learning outcomes. It aims to capture the progress of interactions in students' cognitive processes that support the development of critical thinking, knowledge construction, and problem-solving [13]; (iii) **Teaching presence** concerns teaching role before (i.e., course design) and during (i.e., facilitation and direct instruction) the course [2].

Given the focus of the current study, social presence is further unpacked. Garrison et al. [12] define social presence as *"the ability of participants in a community of inquiry to project themselves socially and emotionally, as 'real' people (i.e., their full personality), through the medium of communication being used"* (p. 94). Besides, in contrast to face-to-face interaction, in online discussions, it is essential to textually express such abilities in order to establish a socio-emotional communication [15]. Social presence, as defined in the CoI model, includes three categories: (i) **Affective**: This category analyses the translation of real emotions into text. It encompasses emotion, feelings, and mood expressions; (ii) **Interactive**: This category focuses on the interactivity of the messages exchanged among participants. The main goal of this category is to enhance open communication among students; (iii) **Group Cohesion**: This category investigates the sense of union and group commitment among students.

Each of the three categories of social presence has several indicators related, as described in Table 1. These indicators are a roadmap to interpret the interactions though the social presence concept.

Table 1: Indicators social presence [41].

Category	Indicator	Label
Affective	1. Expression of emotions	Emotions
	2. Use of humor	Humor
	3. Self-disclosure	Self_disclosure
Interactive	4. Continuing a thread	Cont_Thread
	5. Quoting from others' messages	Quoting_Mess
	6. Referring explicitly to others' message	Referring_Mess
	7. Asking questions	Asking_Q
	8. Complimenting, expressing appreciation	Complimenting
	9. Expressing agreement	Agreement
Cohesive	10. Vocatives	Vocatives
	11. Addresses of refers to the group using inclusive pronouns	Group
	12. Phatics, salutations	Salutations

2.2 Analysis of the CoI

The published literature presented two methods for the analysis of the three presences within the CoI perspective through the use of questionnaires and the adoption of content analysis.

Several questionnaires have been proposed and validated in the context of CoI to examine the perception of students about their experience online interactions. The most broadly adopted is the instrument proposed by Arbaugh et al. [3], in which a 34-item survey measures the perception of the students regarding the three presences using a five-point Likert scale (1 = strongly disagree to 5 = strongly agree). This form is adopted by several studies to analyze individual presences [38] and relationship among them [26].

The second approach to the analysis of the CoI presences is content analysis of online discussion transcripts. Rourke et al. [41] and Garrison et al. [13] defined coding schemes to analyze social and cognitive presences. These schemes have widely been adopted for manual content analysis of CoI. For instance, Gašević et al. [16] adopted the cognitive presence scheme to evaluate the improvement of asynchronous online discussions after an instructional intervention. Following a similar idea, Kovanovic et al. [22] used the manual codding to evaluate the association between social presence and social network position.

Initial proposals to automate content analysis according to the coding schemes of the CoI model primarily relied upon features traditionally used in text mining such as word and phrase counts. For instance, Mcklin [31] an artificial neural network based on word frequency features to classify online discussion messages according to their cognitive presence. The classifier reached 0.31 Cohen's κ.

Recent studies examined the use of other features and classifiers. Kovanović et al. [24] examined the use of a combination of bag-of-words (n-gram) and Part-of-Speech (POS) N-gram features for classifying cognitive presence using the Support Vector Machines (SVMs) classifier, achieving 0.41 Cohen's κ. Kovanović et al. [25] and Neto et al. [35] adopted features based on Coh-metrix [32], LIWC [44], latent semantic analysis (LSA) similarity, named entities, and discussion context [45], to identify phases of cognitive presence for messages written in English (0.63 Cohen's κ) and Portuguese (0.72 Cohen's κ). Besides, the authors applied a random forest classifier [6], which also allowed for the analysis of the influence of the different features on the final classification results.

Although there are studies to extract the phases of cognitive presence automatically, to our knowledge, there is no publication that looked at the automatic content analysis of social presence.

3 RESEARCH QUESTIONS

As discussed in Section 2.1, social presence has a key role in the CoI, influencing the development of cognitive presence in online learning environments. It enhances personal relationships and promotes the sense of community among students. Although several studies demonstrated its importance [21], there is no automatic method to code online discussion messages according to the categories of social presence(affective, interactive and cohesive). Hence, our first research question is:

> **Research Question 1 (RQ1):**
> *To what extent can accurately text mining methods automatically code online discussion messages according to the categories of social presence?*

In addition to addressing the above research question by training a supervised machine learning algorithm (i.e., classifier) for social presence, we were also interested in providing additional insights into the features that were more relevant to each of the three categories of social presence. To do so, we explored a method similar to the one applied by Kovanović et al. [25] and Neto et al. [35]. As such, our second research question is:

> **Research Question 2 (RQ2):**
> *Which features do best predict each category of social presence?*

Finally, we were interested in whether the automatically coded messages preserve the same structural properties when associations between social presence and discussion topics were analyzed. That is, we were interested in examining the extent to which the analysis of associations between automatically coded messages produced results similar to the analysis of manually coded messages according to the categories of social presence. Therefore, our third research question is:

> **Research Question 3 (RQ3):**
> *Do automatically coded messages preserve similar structural properties in the analysis of associations between the categories of social presence and discussion topics to the results of the analysis performed with manual manually coded messages according to the categories of social presence?*

4 METHOD

4.1 Data and course design

The dataset used in this study was taken from a fully online master's degree course in software engineering offered by a public university in Canada. The dataset consists of a total of 1.747 posts from the interaction between 81 students during six offers of the course (winter 2008, fall 2008, summer 2009, fall 2009, winter 2010, winter 2011) [16]. The goal of the online discussion was to debate around videos about research papers related to one of the course topics. The participation in the discussion accounted for 15% of the final grade [16].

Table 2: Distribution of social presence categories.

	Messages					
Category	Control		Treatment		*Total*	
Affective	266	14.27%	264	13.84%	530	30.34%
Interactive	825	44.28%	878	46.04%	1,703	97.48%
Cohesive	772	41.45%	765	40.11%	1,535	87.98%
Total	1,863	100%	1,907	100%	3,770	100.00%

During the first two offerings of the course, the participation of the students was primarily driven by the extrinsic motivational factors (i.e., course grade), with limited scaffolding support. The students from the first two offerings are referred to as the *control group*. After the first two course offerings, a scaffolding of discussion participation through role assignments and clear instructions was implemented (*treatment group*). Table 2 shows the number of messages accounting for control and treatment groups. It is important to remark that the same message could have more them one category of social presence.

Two expert coders categorized the dataset, considering the 12 indicators of social presence (see Table 1) [22]. That is, for each post in the dataset, each indicator received the value "one" (has the indicator) or value "zero" (does not have the indicator). The percentage of agreement between the evaluators was 84%, and a third evaluator resolved the cases with disagreements. Following Kovanovic et al. [22], three indicators (Continuing a thread, Complimenting, and Vocatives) were removed because they had a high number of messages.

Finally, as the objective of this study was to construct binary classifiers for each category of social presence, the categories were reorganized to have binary coding (negative 0 or positive 1). For a message to be classified in positive (1), it must have at least one indicator annotated with the value "one" in the respective category. For instance, if a message had for the affective category, the indicators Emotions = 0, Humor = 0, and Self_disclosure = 1, it was coded as positive (1). Finally, we obtained the dataset as shown in Table 3.

Table 3: Final distribution of social presence phases

Category	Negative (0)	Positive (1)
Affective	1217	530
Interactive	717	1030
Cohesive	421	1326

4.2 Training and test data preparation

The classification of texts has been the target of several educational works over the last years. In the systematic review of the literature presented in [11], 343 studies that applied text mining techniques in educational problems were selected, from which 109 (31.77%) studies focus on text classification. These studies applied machine learning algorithms that used a previously labeled training set to generate a model capable of predicting the correct labels of examples whose labels were unknown (future cases). Therefore, the data set adopted in this study was divided into training and test sets; the first (training) one formed by the five initial offerings of the course (winter 2008, fall 2008, summer 2009, fall 2009, winter 2010) and the second (test) for last offer (winter 2011) according to the recommendations by Farrow et al. [10]. As shown in Table 4, the training group had 1,510 (86%) posts and the test group with 237 (14%) posts. The negative and positive classes presented approximate distributions in the Interactive and Cohesive categories, with a greater difference in class distribution only for the Affective category.

Table 4: Distribution of posts in training and testing groups

	Group	negative (0)	positive (1)	Total
Affective	Train	1038 (0.69%)	472 (0.31%)	1510
	Test	179 (76%)	58 (24%)	237
Interactive	Train	620 (0.41%)	890 (0.59%)	1510
	Test	97 (41%)	140 (59%)	237
Cohesive	Train	362 (24%)	1148 (76%)	1510
	Test	59 (25%)	178 (75%)	237

4.3 Feature extraction

This work combines traditional text mining features, like word frequencies, with the linguistic tools LIWC and Coh-Metrix to extract indications of social presence from textual contributions. These tools are widely validated in the literature as suitable extractors of cohesive, psychological, and social aspects of texts [37]. The remainder of the subsection provides an overview of these features as well as justifications for their use to automate the identification of social presence in asynchronous online discussions.

4.3.1 LIWC features. Linguistic Inquiry Word Count (LIWC) is a linguistic text analysis resource that extracts 93 features divided into the categories as follows: summary of language variables, linguistic dimensions, grammar, and psychological processes. The last category has words that express Affective Processes, Positive Emotion, Negative Emotion, Anger, Sadness, Social Processes, and others. By relating the definition of social presence (students' ability to demonstrate that they are real people [37]), and the social indicators proposed in the CoI [12] model, we hypothesized the use of this language resource might contribute to the construction of classifiers capable of correctly discriminating messages with or without evidence of social presence. LIWC was already used in a previous study focusing on automating the identification of cognitive presence, reaching high levels of accuracy [25, 35]. Thus, in this study, the LIWC 2015 version was used to extract features from the messages in our dataset.

4.3.2 Coh-Metrix features. According to [19], Coh-Metrix uses lexicons, POS Tagger, LSA, among other natural language processing (NLP) techniques to analyze cohesion and textual coherence. Several studies have reported good results when using Coh-Metrix to generate text cohesion indicators [18, 30]. Therefore, we hypothesized that the existence/absence of social presence indicators could be related to the index of textual cohesion and complexity proposed in Coh-Metrix. For instance, in the message "You got it right!" there is a hint of social interaction through the pronominal "You" cohesion. Hence, when considering the use of cohesive words as a way of demonstrating group projection, Coh-Metrix was also adopted as a component of the analysis in the process of automatic identification of social presence.

4.3.3 Word frequency. Finally, we adopt a bag-of-words vector, a traditional text mining technique, as the last set of features to extract social presence. More specifically, we adopt the method which transforms the textual documents (in this case, online discussion message) into an array consisting of the terms count. A problem found in this type of technique is the high dimensionality of the generated vector of features because as the text itself is used as discriminant the size of the vocabulary of the documents will correspond to the size of the matrix. Therefore, three techniques were used to reduce the dimensionality of the term count matrix. The first was a spelling correction of the texts. The second was the removal of stopwords, which consisted of removing words of little significance in a text such as articles, conjunctions, and prepositions [29]. Finally, the last technique was stemming, which seeks to reduce words to their respective radicals [36]. For example, the words "engineer" and "engineering" become "engine".

4.4 Data preprocessing

The main focus of machine learning is the creation of inductors based on past data (training set) and with the ability to generalize learned patterns to future examples [6]. One of the challenges in machine learning is dealing with datasets that have unbalanced class distributions [20]. According to He and Garcia [20], in these cases, the generated inductors usually prioritize the majority class. As presented in section 4.2, the negative and positive classes in all categories (Affective, Interactive, and Cohesive) were unbalanced. The Cohesive category of social presence was particularly highly unbalanced, where the negative class presented approximately 25% of the data and the positive class approximately 75%, suggesting that the classifier could prioritize the positive class. According to Chawla et al. [7], there are basically two approaches to solving the data imbalance problem: (i) cost-sensitive classification, i.e., penalizing the majority and minority prediction errors in different ways in order to force the algorithm prioritizing classes with fewer examples; and (ii) resampling of the data, with the options of undersampling the majority class in order to balance the number of examples which has the negative point of data loss or oversampling of the minority class. Thus, we decided to use the oversampling technique in the data of the training sets of each category. For this, we adopted the SMOTE algorithm, which is used in several works to create artificial data of the minority class (oversampling) [20].

4.5 Model Selection and Evaluation

To address research question 1, we trained three machine learning classifiers – one for each category of social presence. Recent studies show that combining machine learning classifiers tends to yield better results compared to those obtained by individual classifiers [6]. Ensembles can be performed by combining several distinct algorithms or by using only one algorithm with different training sets. Random Forest, one of the most widely used ensembles in the literature, combines decision trees using a technique called bagging which randomly samples characteristics. In other words, each tree is trained with different views (feature sets) of the same problem. Finally, all decisions are combined using the majority vote decision rule [6]. Random Forest is also often used to estimate the importance of an individual feature where it considers metric Mean decrease gini impurity index (MDG) [6], categorizing this technique as a white-box algorithm. Thus, Random Forest was the algorithm chosen for this study.

In educational research, to measure the performance of a supervised machine learning algorithm, the accuracy and Cohen's kappa[8] metrics are used [22, 35]. Hence, these metrics were also used in this work.

As highlighted in [6], the main parameters of the Random Forest algorithm are the number of input variables randomly chosen from each division (max_features) and the number of trees in the forest (n_estimators). To optimize the final performance, we performed a tunning in the parameters (max_features and n_estimators) of the Random Forest classifier through executions using the cross-validation technique for each training set (Affective, Interactive, and Cohesive). Each validation fold consisted of one of the course offerings in the respective training set (winter 2008, fall 2008, summer 2009, fall 2009, winter 2010). The first parameter to be adjusted was *max_features*. The literature considers that the Random Forest performance stabilizes after a certain number of trees [25]. Therefore, we set the number of trees at 1500 to ensure the convergence of the algorithm and, consequently, choosing the best *max_features* parameter. To obtain a deterministic behavior during parameter setting, one seed (five) was established. In each cross-validation execution, a total of 140 values were verified for the *max_features* parameter, which was set at random, respecting the maximum number of possible characteristics (6418) and without repetition. At the end of the executions, the average performances and the standard deviations obtained for each parameter were reported. In Figure 1, it is shown that the average accuracy obtained in each of the three categories (Affective, Interactive, and Cohesive) began to stabilize when the number of characteristics was about 2000.

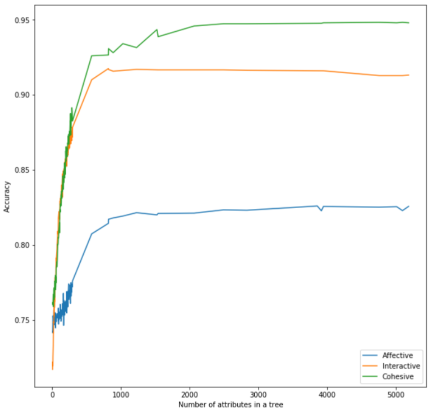


Figure 1: Random forest parameter tuning results

After setting the *max_features parameter*, the next step was to estimate the appropriate number of trees *n_estimators* for the problem. For this, the (Out-of-Bag) OOB error was used and it was calculated in a set of observations that were not used to build the current tree. As shown in Figure 2, it was not possible to notice drastic differences when changing the n_estimators parameter, so we set the number of trees to 800.

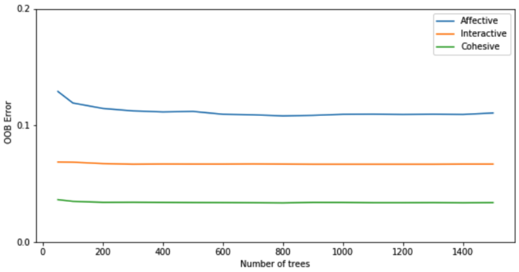


Figure 2: Best random forest configuration performance

Besides, the development of the classifier, we also provide additional insights on which features are more relevant for each category. One popular measure to calculate the feature importance from a Randon Forest is Mean Decrease Gini (MDG) which is based on the reduction in Gini impurity measure [6]. In this paper, we adopted MDG address the second research question, the evaluation of the relevance of different features to the outcome of our classifiers.

4.6 Epistemic Network Analysis

We applied Epistemic Network Analysis (ENA) [42] to address research question 3 and provide insights into the validity of the results produced by the proposed classifier. ENA is a graph-based technique used for analysis of associations between different concepts (called *codes*) used for coding textual datasets. Within ENA, a network of relationships among different *codes* is created for each *unit of analysis* (e.g., student). Two codes are considered related if they co-occur in the same chunk of text, called *stanza* (or *conversation*).

In this work, we reproduced the experiment proposed by Rolim et al. [39] using the automatic classifier to generate the social presence codes. Rolim et al. [39] adopted LDA to extract topics from the dataset (15 topics in total). Then, the authors used the students as *units of analysis*, social presence categories and course topics as *codes*, and individual students' discussion messages as *stanza*. Again, the goal of this analysis is to compare the graphs plotted based on the humans' annotation, and those automatically generated by the proposed classifier.

ENA mainly provides three graphical outcomes: (1) Projection graph; (2) Epistemic Network; and (3) Subtraction network. A relevant characteristic of ENA is that all these graphical outcomes are represented in the same bi-dimensional space, called *analytic space*, composed by X and Y axes. So it is possible to analyze different aspects at the same time. Each graph produced by ENA has elements to be analyzed, as follows: (i) **Projection graph** presents the units of analysis (i.e., the students in the current study) distributed in the analytic space. In this graph, each unit of analysis is represented as a pair (x,y) that are related to their position in the axes X and Y. (ii) **Network graph** is undirected graph and contains three important elements: the *size* of the nodes represents the frequency of occurrence of the nodes; the *nearness* of the node represents the similarity among them; and the *strength* of the code relationship represents the frequency of their co-occurrence; and (iii) **Subtraction network** captures the differences between two ENA networks and only shows edges that are different between the two networks. Moreover, it presents the same aspects of the regular ENA networks.

The graphs produced by ENA provide a promising approach to assessing the validity of the results of the classifier, mainly for allowing us to compare the results with our previous work [39], which provides insights into how the students' social presence was related to the course topics; moreover, all relationships developed by the students can be quantified. Several other studies have already applied ENA in the context of CoI. For example, Rolim et al. [40] presents an approach that uses ENA to understand the relationships between cognitive and social presences and uncover how students progress over time in their social inquiry. In another paper, Rolim et al. [39] analyze, using ENA, the relationships between course topics and indicators of social presence categories.

Aiming to quantify the comparison of the proposed classifier output with the manual codded data using ENA, we used the Pearson correlation coefficient (PCC), which measures the linear relationship between two variables [5]. In our study, we measured the correlation between the projections, the pair (x,y) of each student, comparing the data generated with human-annotated and automatically generated data. Thus, we wanted to measure the extent to which the epistemic networks with automatically and manually assigned codes were similar.

In order to measure the PCC between the subtraction network of the automatically generated and manually coded data we used the following three variables: (i) two variables related to the nodes positions, divided into the position in relation to axes X and Y; and (ii) the strength of the links between the nodes. In this case, we report three values of PCC.

5 RESULTS

5.1 Model training and evaluation – RQ1

Initially, we evaluated the influence of parameter tuning in the final classification. Table 5 shows the average results reported of the performance of the Random Forest classifiers with the default parameters and tuned parameters using the training set and the cross-validation approach. The results, in terms of accuracy, increased by 12.5%, 22.6% and 21.79% for the affective, interactive and cohesive categories, respectively. Regarding Cohen's kappa, it achieved even higher improvements of 145%, 66% and 114.6% for the same three categories. These results demonstrate the importance of fine-tuning the algorithm parameters.

Table 5: Random forest parameter tuning results

Category	Optimization	Accuracy	Kappa
Affective	Default parameters	0.72 (0.05)	0.20 (0.10)
	Tuned parameters	0.81 (0.04)	0.49 (0.08)
Interactive	Default parameters	0.75 (0.03)	0.50 (0.05)
	Tuned parameters	0.92 (0.02)	0.83 (0.05)
Cohesive	Default parameters	0.78 (0.06)	0.41 (0.11)
	Tuned parameters	0.95 (0.03)	0.88 (0.06)

After parameter optimization, the Random Forest classifier was ran ten times for each social presence category. In each execution, the training set examples (1.510 posts – initial five runs of the courses in our data) were used to generate a binary classifier, and its generalization capacity was verified in the respective test sets (237 posts – the last run of the courses in our data). Thus, the following results were obtained for the test sets of each category: Affective – accuracy = 0.80 (0.01) and Cohen's kappa 0.34 (0.02); Interactive – accuracy = 0.92 (0.0) and Cohen's kappa 0.85 (0.0); and Cohesive – accuracy = 0.97 (0.0) and Cohen's kappa 0.93 (0.0). Despite the high dimensionality of the feature vectors and fine parameter adjustments (max_features and n_estimators), the results achieved in the test sets approximate the average accuracy obtained in the validation sets (cross-validation). Therefore, it shows that there were no overfitting in the training set. Instead, the models demonstrated good generalizability for unknown examples.

Table 6 shows the confusion matrix generated for each category. Corroborating with the average values of accuracy and Cohen's kappa presented, it is possible to notice that the highest occurrences of false positives occurred in the Affective class achieving 37 examples. On the other hand, there were only 9% occurrences of false positives for the Interactive category and 2% for Cohesive.

Table 6: Confusion matrix for the best performing models

	Affective		Interactive		Cohesive	
	neg*	pos*	neg	pos*	neg	pos*
neg*	170	9	91	6	56	3
pos*	37	21	12	128	3	175

* pos = positive and neg = negative

5.2 Feature importance analysis – RQ2

Although the same feature vectors were used to discriminate the classes (positive and negative) of the three categories, each classifier considered different variables as the most important. The Random Forest uses the Mean Decrease Gini impurity index (MDG) measure to define the degree of relevance of a feature. Tables 7, 8 and 9 present the top-15 features for the classifier of each category (Affective, Interactive and Cohesive).

The most important set of variables for the Affective category shows six from the word frequency, eight LIWC, and one Coh-Metrix features. Besides, the two most important variables are the words "hope" and "happi" (without applying the stemming technique, "happy"), reaching 11.68 and 10.33 and MDG, respectively. It is also noteworthy that two characteristic cm.WRDPRP1s concerning the number of first-person pronouns (e.g., I, me, mine) is in the top-15 set of features.

The most predictive features of the Interactive category were divided into word frequency (three), LIWC (eight), and Coh-Metrix (four). Table 8 shows that the most important was the number of Question marks which achieved MDG of 44.2. The presence of the word "agre" (stemmed from the word agreement) and the liwc.assent which measures the agreement was also noticeable.

Finally, Table 9 presents the main features of the Cohesive class being 10 from word frequency, five from LIWC, and none from Coh-Metrix. It may be highlighted that the presence of words commonly used to greet ("hi" - MDG of 54.18 and "hello" MDG of 3.64) and of socially biased variables like liwc.affiliations (e.g., ally, friend, and social) and Liwc.social (e.g., mate, talk, and they). Therefore, the most predictive feature listings for each category demonstrate the importance of the three language resources used in this study.

Table 7: Fifteen most important variables for the Affective category according to MDG

Variable	Description	MDG	Negative	Positive
hope	Word frequency	11.68	0.02 (0.14)	0.24 (0.46)
happi	Word frequency	10.33	0.0 (0.05)	0.22 (0.51)
liwc.i	1st pers singular	3.88	2.3 (2.04)	3.39 (2.19)
cm.WRDPRP1s	Incidence score of pronouns, first person, single form	3.31	22.86 (19.94)	33.53 (21.68)
liwc.Apostro	Number of Apostrophes	2.07	0.7 (1.25)	1.08 (1.4)
liwc.Exclam	Number of Exclamation marks	1.18	0.21 (0.84)	0.66 (3.58)
liwc.negemo	Number of negative emotion	1.18	0.63 (1.04)	0.83 (1.03)
work	Word frequency	1.14	0.25 (0.64)	0.54 (0.98)
bb	Word frequency	1.14	0.35 (0.48)	0.55 (0.5)
experi	Word frequency	1.07	0.09 (0.37)	0.26 (0.78)
develop	Word frequency	0.83	0.53 (1.1)	0.79 (1.43)
liwc.power	words with power idea	0.79	1.47 (1.59)	1.55 (1.33)
liwc.hear	hear	0.72	0.29 (0.66)	0.41 (0.79)
liwc.negate	Number of negations	0.7	0.98 (1.16)	1.35 (1.17)
liwc.we	1st pers plural	0.62	0.21 (0.65)	0.37 (0.81)

Table 8: Fifteen most important variables for the Interactive category according to MDG

Variable	Description	MDG	Negative	Positive
liwc.QMark	Number of question marks	44.2	0.07 (0.4)	1.5 (1.38)
agre	Word frequency	9.21	0.01 (0.1)	0.25 (0.5)
present	Word frequency	5.98	0.54 (0.9)	1.29 (1.23)
liwc.assent	Number of assent	1.77	0.26 (1.43)	0.3 (0.71)
liwc.auxverb	Auxiliary verbs	0.81	7.58 (4.2)	8.82 (2.77)
liwc.you	2nd person	0.8	1.57 (2.91)	2.25 (1.88)
liwc.Period	Number of periods	0.79	7.64 (5.2)	5.51 (2.54)
liwc.AllPunc	Number all punctuation	0.78	17.22 (14.72)	13.3 (10.48)
cm.WRDPOLc	Number of senses (core meanings) of a word	0.73	3.72 (0.83)	3.98 (0.57)
cm.WRDPRP2	Incidence score of pronouns, second person	0.69	15.56 (28.84)	22.26 (18.65)
liwc.Dic	Dictionary words	0.65	75.83 (14.12)	80.5 (7.4)
cm.DESWLltd	Mean number of letters in the words within the text	0.64	3.34 (1.75)	2.98 (0.81)
cm.SYNSTRUTt	Proportion of intersection tree nodes between all sentences	0.51	0.07 (0.06)	0.06 (0.03)
did	Word frequency	0.49	0.0 (0.04)	0.15 (0.42)
liwc.function.	Total function words	0.47	44.29 (12.67)	48.66 (6.74)

Table 9: Fifteen most important variables for the Cohesive category according to MDG

Variable	Description	MDG	Negative	Positive
hi	Word frequency	54.18	0.0 (0.07)	0.86 (0.35)
liwc.affiliation	Number of affiliations	10.48	0.98 (2.46)	1.97 (2.17)
regard	Word frequency	5.02	0.05 (0.23)	0.29 (0.52)
hello	Word frequency	3.64	0.0 (0.0)	0.05 (0.21)
cheer	Word frequency	1.12	0.0 (0.05)	0.07 (0.26)
bb	Word frequency	1.01	0.52 (0.5)	0.38 (0.49)
liwc.social	Social processes	0.96	5.49 (5.34)	7.12 (3.94)
thank	Word frequency	0.76	0.6 (0.64)	0.77 (0.68)
grant	Word frequency	0.55	0.1 (0.3)	0.03 (0.18)
liwc.Clout	Number of clout-related words	0.5	48.4 (20.16)	58.07 (19.09)
present	Word frequency	0.5	0.76 (1.08)	1.05 (1.19)
hey	Word frequency	0.32	0.0 (0.0)	0.01 (0.1)
liwc.Apostro	Number of Apostrophes	0.31	1.23 (1.77)	0.68 (1.1)
liwc.we	1st pers plural	0.3	0.19 (0.68)	0.28 (0.72)
sy	Word frequency	0.3	0.04 (0.19)	0.01 (0.11)

5.3 Epistemic Network Analysis – RQ3

We reproduced the work by Rolim et al. [39] in order to evaluate the impact on the validity of the automatic classification in producing codes that are used in ENA. Figure 3 presents the projection graph for both manually and automatically coded datasets. In this graph, each node represents a student, and the squares are the mean values for the two groups; control and treatment group are represented as red and blue nodes, respectively. Although the first two SVD dimensions (i.e., x- and y-axes) small differences in explained variance, it is possible to visually identify a high level of similarity between the two graphics (Fig. 3a and 3b); for instance, the positions of the group centroids (red and blue squares), and the distribution of the units (red and blue points) along the axis. The Mann-Whitney test showed that along the X-axis the students from the control and treatment groups were statistically significantly different at the alpha=0.05 level for both manually (U=1497.00, p=0.00, r=0.76) and automatically coded (U=164.00, p=0.00, r=0.81) data with very similar effect sizes as reflected by the r values.

Moreover, we calculated the PCC between the same students with the data manually coded and the automatically generated. As it student is projected as a pair (x,y) we analyzed the PCC variables separately for each dimension (X and Y axes). The final values of PCC reached 0.93 and 0.84 for axes X and Y, respectively, which demonstrate a high correlation between the distribution of the students.

Figure 4 shows the subtraction network between the control (red) and treatment(blue) groups for both manually and automatically coded data. Visually analyzing, we can highlight the arrangement of the codes, where the course topics are mostly in the bottom-left corner of both networks, whereas the social presence categories are plotted in the upper-central part o both networks. Also, the connections between codes have a similar strength among the codes. Figure 4b also reveals a slight change in the sizes of "Interactive", "Cohesive" and "const.meth" codes. Another difference is the position of the three social presence categories with the most significant one for the "Affective" code, which before was positioned closer to the course topics, and in Figure 4b it is closer of the other two categories; closeness in the project graphs means higher similarity.

We also evaluated the PCC between links strength and the node positions in the subtraction network. In this case, we did not evaluate the individual students from the projection graph (as reported previously in this section), but the position of the codes in Figure 4. The results for the positions of the nodes achieved 0.96 and 0.89 for the axes X and Y, respectively. Regarding the PCC of the strength between codes the results reached 0.93. The high PCC values indicate the convergence of the two networks.

6 DISCUSSION

In addressing research question 1, the evaluation of the automatic classification of social presence revealed that the combination of traditional text mining features and word counts extracted from LIWC and Coh-Metrix were effective in classifying online discussion messages message in all categories (affective, interactive and cohesive). Cohen's κ of 0.49, 0.83 and 0.88, for affective, interactive and cohesive, respectively, represent a medium to substantial interrater agreement [28], and in two out of three categories it is above 0.70, which is the CoI research commonly used as the threshold limit required before manual coding results are considered valid. The optimization of the max_features (i.e., the number of attributes used in each tree of the forest) and n_estimators (i.e., the number of trees used in each iteration) parameters improved the final result in all cases (Table 5).

Although we did not find any other related work which performed a similar analysis of social presence to compare to, it is important to mention that the approach presented here reached accuracy results better than the classifiers of cognitive presence developed for English [24, 25, 45].

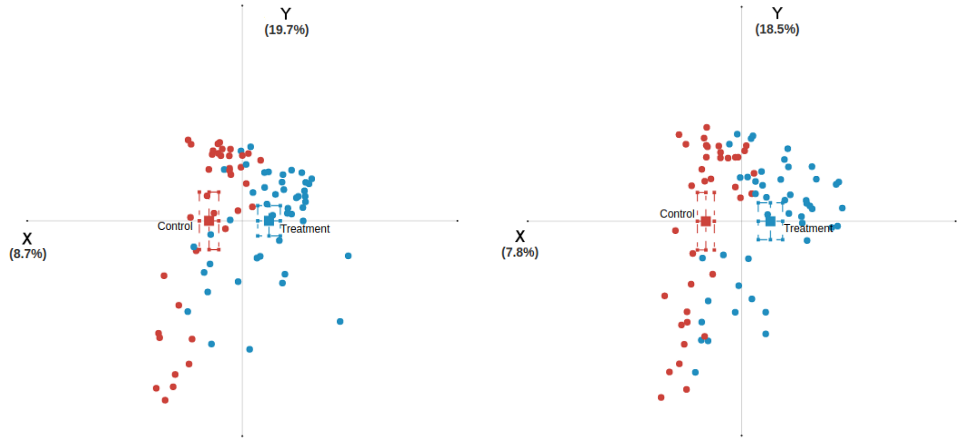
In addressing research question 2, this study conducted a detailed analysis of the features used. By analyzing the features provided in tables 7, 8 and 9, we can draw two conclusions: (i) for every category, there were feature related to word frequency and the tools LIWC and Coh-Metrix, showing the importance of both aspects; (ii) although the features related to word frequency could lead to overfitting depending on the domain, in the case of the current study, the words with high information gain were general ones like hope, happy, hear, agree, hi, hello, among others. Thus, it decreases the chances of overfitting because these words can happen in messages of different domains.

The analysis of the feature importance also highlighted a possible correlation between the main features identified in this study and the indicators considered the most predictive of social presence [41]. For instance, among those selected by the Affective category classifier, the features: *hope*, *happi* (happy), liwc.exclam (number of exclamation points) and liwc.negemo (number of negative emotions) are related to the expression of emotions and the use of humor. While the feature cm.WRDPRP1s (pronoun incidence score, first-person singular) may be associated with the self-disclosure indicator since the student demonstrates self-disclosure when presenting details of life outside the home, classroom, or express vulnerability [41].

For the Interactive category, the features liwc.you (2nd person word count) and cm.WRDPRP2 (second person pronoun incidence score) were related to the indicators of the Interactive category (see Table 1) by citing and referencing openly other messages or people in the discussion. Moreover, in nonverbal interaction, when asking a question it is common to use the question punctuation mark. Thus, the Interactive category indicator named Asking Questions is represented in the list of most essential features by the features liwc.QMark (number of question marks) and liwc.interrog (number of interrogative sentences). Another demonstration of interaction (based on the CoI model) is expressions of agreement students' messages; in this respect, the central features were the word *agre* (agree) and the number of nods per post (liwc.assent).

Finally, the presence of the feature liwc.we (number of first-person plural words) in table 9 corroborate the relevance of using inclusive pronouns (us, ours) as a way of demonstrating group cohesion. Another indicator of the cohesive category, the demonstration of salutations, can be recognized by the characteristics *hi*, *hello*, liwc.affiliation (number of affiliations) and liwc.social (number of social processes).

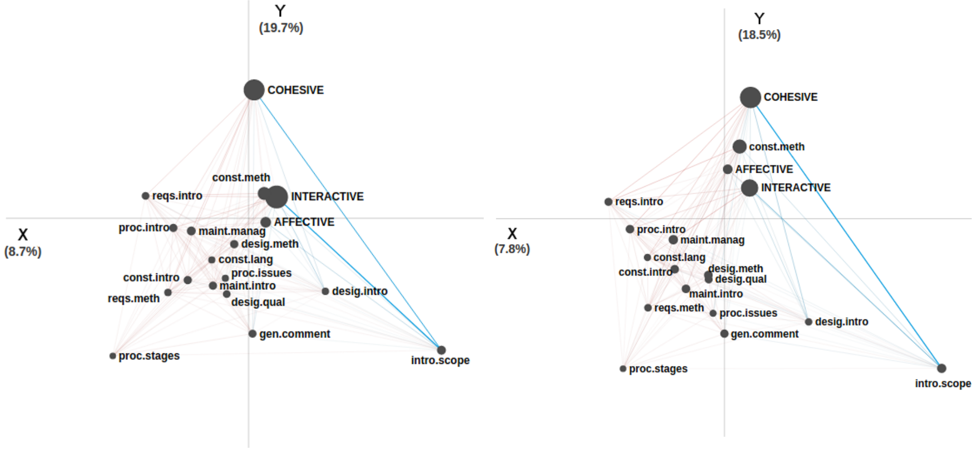
In addressing research question 3, we adopted ENA to investigate the similarities between associations between discussion topics and social presence categories as generated with the manually and automatically coded data. As can be seen in section 5.3, we obtained similar results after performing ENA with manually and automatically assigned codes for both individual projections of students and the subtraction network of the two groups. Also, we demonstrated that these outcomes are correlated using Pearson Correlation Coefficient. Thus, we conclude that analyses performed with automatically assigned codes can reproduce the results of analyses based on manually coded data on a reasonable level of confidence that can preserve structural properties of the associations of social presence with other relevant constructs [33]. This also offers additional reassurance in the validity of the results of analyses that are based on automatically coded messages.



(a) Using manual codded labels.

(b) Using automatic generated labels.

Figure 3: ENA projection of the networks of the students related to social presences and course topics between control (red) and treatment (blue) groups.



(a) Using manual codded labels.

(b) Using automatic generated labels.

Figure 4: Subtraction mean network between control (red) and treatment (blue) groups.

7 CONCLUSIONS

This paper has three contributions. First, the proposal of three binary Random Forests classifiers, using the LIWC, Coh-Metrix and word frequency linguistic resources, to automatically classify online discussion messages into social presence categories (Affective, Interactive and Cohesive). Every category reached Cohen's kappa values of more than 0.49, a medium to substantial inter-rater agreement. Second, the results provide insights into the psycho- and socio-linguistic features that are more relevant for each social presence indicator, linking each of them with the CoI literature. These results additionally clarify the nature of each social presence indicator, which have not been previously reported in the literature. Finally, the use of automatically coded discussion messages in analysis of associations of social presence with other relevant constructs (e.g., discussion topics) produced nearly identical results to the analyses performed with manually assigned codes of social presence.

Despite promising results, some limitations can be identified, such as the small number of message examples used in the current study (1.747 posts). Next, the training and test sets were divided based on different offerings of the same course, making it difficult to generalize the results presented to other contexts. Finally, using word frequency to compose feature vectors can mean a strong bias of the classification models created in the training context.

Future work should seek to optimize the approach proposed in this study to reduce the dimensionality of the feature vectors while maintaining the promising results already obtained, which is important to avoid overfitting. Besides, we also aim to conduct experimentation with the approach and possible evaluation with larger sample sizes composed of data from different domains.

ACKNOWLEDGEMENT

This research was partially supported by the fellowship 310060/2018-4 of the National Council for Scientific and Technological Development – CNPq.

REFERENCES

[1] Terry Anderson and Jon Dron. 2010. Three generations of distance education pedagogy. *The International Review of Research in Open and Distance Learning* 12, 3 (2010), 80–97.

[2] Terry Anderson, Liam Rourke, D. Randy Garrison, and Walter Archer. 2001. Assessing Teaching Presence in a Computer Conferencing Context. *Journal of Asynchronous Learning Networks* 5 (2001), 1–17.

[3] J Ben Arbaugh, Martha Cleveland-Innes, Sebastian R Diaz, D Randy Garrison, Philip Ice, Jennifer C Richardson, and Karen P Swan. 2008. Developing a community of inquiry instrument: Testing a measure of the community of inquiry framework using a multi-institutional sample. *The internet and higher education* 11, 3-4 (2008), 133–136.

[4] Jason J Barr. 2016. Developing a Positive Classroom Climate. *IDEA Center, Inc.* (2016).

[5] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*. Springer, 1–4.

[6] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[7] Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz. 2004. Special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter* 6, 1 (2004), 1–6.

[8] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[9] Roisin Donnelly and John Gardner. 2011. Content analysis of computer conferencing transcripts. *Interactive learning environments* 19, 4 (2011), 303–315.

[10] Elaine Farrow, Johanna Moore, and Dragan Gašević. 2019. Analysing discussion forum data: a replication study avoiding data contamination. In *LAK' 2019*. 170–179.

[11] Rafael Ferreira-Mello, Máverick André, Anderson Pinheiro, Evandro Costa, and Cristobal Romero. 2019. Text mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (2019), e1332.

[12] D. Randy Garrison, Terry Anderson, and Walter Archer. 1999. Critical Inquiry in a Text-Based Environment: Computer Conferencing in Higher Education. *The Internet and Higher Education* 2, 2-3 (1999), 87–105.

[13] D. Randy Garrison, Terry Anderson, and Walter Archer. 2001. Critical Thinking, Cognitive Presence, and Computer Conferencing in Distance Education. *American Journal of Distance Education* 15, 1 (2001), 7–23.

[14] D. Randy Garrison, Terry Anderson, and Walter Archer. 2010. The first decade of the community of inquiry framework: A retrospective. *The Internet and Higher Education* 13, 1-2 (2010), 5–9.

[15] D Randy Garrison and J Ben Arbaugh. 2007. Researching the community of inquiry framework: Review, issues, and future directions. *The Internet and Higher Education* 10, 3 (2007), 157–172.

[16] Dragan Gašević, Olusola Adesope, Srećko Joksimović, and Vitomir Kovanović. 2015. Externally-facilitated regulation scaffolding and role assignment to develop cognitive presence in asynchronous online discussions. *The internet and higher education* 24 (2015), 53–65.

[17] Dragan Gašević, Vitomir Kovanović, and Srećko Joksimović. 2017. Piecing the learning analytics puzzle: a consolidated model of a field of research and practice. *Learning: Research and Practice* 3, 1 (2017), 63–78.

[18] Arthur C Graesser, Danielle S McNamara, and Jonna M Kulikowich. 2011. Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational researcher* 40, 5 (2011), 223–234.

[19] Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers* 36, 2 (2004), 193–202.

[20] Haibo He and Edwardo A Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering* 21, 9 (2009), 1263–1284.

[21] Srećko Joksimović, Dragan Gašević, Vitomir Kovanović, Bernhard E Riecke, and Marek Hatala. 2015. Social presence in online discussions as a process predictor of academic performance. *Journal of Computer Assisted Learning* 31, 6 (2015), 638–654.

[22] Vitomir Kovanovic, Srecko Joksimovic, Dragan Gasevic, and Marek Hatala. 2014. What is the source of social capital? The association between social network position and social presence in communities of inquiry. In *Workshop at Educational Data Mining Conference*. EDM.

[23] Vitomir Kovanović, Dragan Gašević, and Marek Hatala. 2014. Learning analytics for communities of inquiry. *Journal of Learning Analytics* 1, 3 (2014), 195–198.

[24] Vitomir Kovanović, Srećko Joksimović, Dragan Gašević, and Marek Hatala. 2014. Automated cognitive presence detection in online discussion transcripts. In *LAK'14*. Indianapolis, IN.

[25] Vitomir Kovanović, Srećko Joksimović, Zak Waters, Dragan Gašević, Kirsty Kitto, Marek Hatala, and George Siemens. 2016. Towards automated content analysis of discussion transcripts: A cognitive presence case. In *LAK'16*. ACM, New York, NY, USA, 15–24.

[26] Kadir Kozan and Jennifer C Richardson. 2014. Interrelationships between and among social, teaching, and cognitive presence. *The Internet and higher education* 21 (2014), 68–73.

[27] Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

[28] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[29] Rachel Tsz-Wai Lo, Ben He, and Iadh Ounis. 2005. Automatically building a stopword list for an information retrieval system. In *Journal on Digital Information Management*, Vol. 5. 17–24.

[30] Philip M McCarthy, Gwyneth A Lewis, David F Dufty, and Danielle S McNamara. 2006. Analyzing Writing Styles with Coh-Metrix.. In *FLAIRS Conference*. 764–769.

[31] Thomas E. Mcklin. 2004. *Analyzing Cognitive Presence in Online Courses Using an Artificial Neural Network*. Ph.D. Dissertation. Atlanta, GA, USA. Advisor(s) Harmon, Stephen W. AAI3190967.

[32] Danielle S McNamara, Arthur C Graesser, Philip M McCarthy, and Zhiqiang Cai. 2014. *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

[33] Samuel Messick. 1995. Standards of validity and the validity of standards in performance asessment. *Educational measurement: Issues and practice* 14, 4 (1995), 5–8.

[34] Natalie B Milman. 2015. Distance education. (2015).

[35] Valter Neto, Vitor Rolim, Rafael Ferreira, Vitomir Kovanović, Dragan Gašević, Rafael Dueire Lins, and Rodrigo Lins. 2018. Automated analysis of cognitive presence in online discussions written in portuguese. In *European Conference on Technology Enhanced Learning*. Springer, 245–261.

[36] Viviane Moreira Orengo and Christian Huyck. 2001. A stemming algorithm for the portuguese language. In *Proceedings. Eighth International Symposium on*. IEEE, 186–193.

[37] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. Technical Report.

[38] Oleksandra Poquet, Vitomir Kovanović, Pieter de Vries, Thieme Hennis, Srećko Joksimović, Dragan Gašević, and Shane Dawson. 2018. Social presence in massive open online courses. *International Review of Research in Open and Distributed Learning* 19, 3 (2018).

[39] Vitor Rolim, Rafael Ferreira Leite de Mello, Vitomir Kovanovic, and Dragan Gašević. 2019. Analysing Social Presence in Online Discussions Through Network and Text Analytics. In *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*, Vol. 2161. IEEE, 163–167.

[40] Vitor Rolim, Rafael Ferreira, Rafael Dueire Lins, and Dragan Găsević. 2019. A network-based analytic approach to uncovering the relationship between social and cognitive presences in communities of inquiry. *The Internet and Higher Education* 42 (2019), 53–65.

[41] Liam Rourke, Terry Anderson, D. Randy Garrison, and Walter Archer. 1999. Assessing Social Presence In Asynchronous Text-based Computer Conferencing. *The Journal of Distance Education* 14, 2 (1999), 50–71.

[42] David Williamson Shaffer, David Hatfield, Gina Navoa Svarovsky, Padraig Nash, Aran Nulty, Elizabeth Bagley, Ken Frank, André A. Rupp, and Robert Mislevy. 2009. Epistemic Network Analysis: A Prototype for 21st-Century Assessment of Learning. *International Journal of Learning and Media* 1, 2 (2009), 33–53.

[43] Jan-Willem Strijbos. 2011. Assessment of (computer-supported) collaborative learning. *IEEE transactions on learning technologies* 4, 1 (2011), 59–73.

[44] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.

[45] Zak Waters, Vitomir Kovanović, Kirsty Kitto, and Dragan Gašević. 2015. Structure matters: Adoption of structured classification approach in the context of cognitive presence classification. In *Information Retrieval Technology*. 227–238.

[46] C Xia, John Fielder, and Lou Siragusa. 2013. Achieving better peer interaction in online discussion forums: A reflective practitioner case study. *Issues in Educational Research* 23, 1 (2013), 97–113.

APÊNDICE B – TOWARD AUTOMATIC CLASSIFICATION OF ONLINE DISCUSSION MESSAGES FOR SOCIAL PRESENCE

Toward Automatic Classification of Online Discussion Messages for Social Presence

Máverick André, Rafael Ferreira Mello, André Nascimento, Rafael Dueire Lins, and Dragan Gašević

***Abstract*—Social Presence is an essential construct of the well-known Community of Inquiry (CoI) model, which is created to support design, facilitation, and analysis of asynchronous online discussions. Social Presence focuses on the extent to which participants of online discussions can see each other as "real persons" in computer-mediated communication. In the CoI model, Social Presence is looked at through the Affective, Interactive, and Cohesive categories. Previous research has obtained good results in the automatic identification of these three categories in the analysis of transcripts of asynchronous online discussions using the random forest algorithms benefiting from traditional text mining features in combination with structural features such as Coh-Metrix and Linguistic Inquiry and Word Count (LIWC). In this context, this study evaluated the performance of the state-of-the-art decision tree algorithms and the deep learning linguistic model BERT for automatic detection of Social Presence in online discussions. The results revealed that XGBoost (eXtreme Gradient Boosting) and AdaBoost (Adaptive Boosting) outperformed the traditional random forest classifier that was commonly used in the previous works on automatic analysis of Social Presence reaching .38, .79, and .94 for the Affective, Interactive, and Cohesive categories, respectively. Moreover, the proposed classifiers also reached better result when compared to BERT. Finally, this study explored a broad range of features used in the automatic classification of online discussion messages according to the categories of Social Presence. The results showed the importance of the features provided by the well-known linguistic framework LIWC and features calculated by techniques such as social network analysis (SNA) and sentiment analysis, that had never been reported previously in the literature for the automatic detection of Social Presence.**

***Index Terms*—Community of Inquiry Model, Social Presence, Online Discussion, Content Analytics, Text Classification.**

I. Introduction

Back in 1976, Short, Williams, and Christie developed the initial theory of Social Presence [1] to explain how the communication medium affects the way people communicate. For them, Social Presence measures the degree of salience between two people using a communication medium. The Internet made possible online courses where asynchronous computer-mediated communication (CMC) is the major form of discourse. According to Lowenthal [2], as the popularity of CMC courses grew, communication researchers began to apply the theory of Social Presence to better understand the interaction between teachers and students and among students.

Today, the majority of learning management systems (LMS) focuses on providing tools to create and share information such as learning materials and activities. The interactions among course participants, which is also a relevant factor for the success of online courses, are not adequately approached in LMSs [3]. For instance, asynchronous online discussions, a resource with a high potential for strengthening the collaboration in online education, does not fully support students' social interactions and engagement if it is not adequately managed.

Another critical challenge that instructors of online courses face is the creation of a supportive and productive environment for student communication and collaboration through technology [4], encouraging the formation of social ties and using social-constructivist pedagogies [5]. Many studies have confirmed the importance of social interactions to generate collaboration and social ties, which can also increase academic performance in online courses [6].

In this context, the Community of Inquiry (CoI) [7] is a model that explores different facets of interactions in online discussions. The CoI model has three dimensions: Social Presence, Cognitive Presence, and Teaching Presence. The three presences examine the engagement of students and learning outcomes, unpack interactions in computer-supported collaborative scenarios, and analyze how different technology-use profiles of students are associated with educational experience [8]. Social Presence investigates the students' abilities to establish the socio-emotional communication in online environments [9], which is essential to promote student retention and to develop Cognitive processes [9].

The quantitative content analysis (QCA) is a traditional method used to analyze the transcripts of asynchronous online discussions [10]. QCA uses predefined coding schemes to analyze text artifacts (i.e., messages in online discussions) and it has been widely applied in the CoI research [11]. The literature reports numerous approaches to automate the QCA of online discussions according to the coding scheme for the Cognitive Presence as proposed in the CoI model [12]–[14], and Social Presence [11], [15]–[17].

However, those methods focus on using traditional machine learning algorithms (i.e., Random Forrest classifiers) and a limited variety of features (in general based on word counts

Manuscript received November 24, 2020; revised August 30, 2021 and December 10, 2021; accepted TBD. Date of publication TBD; date of current version November XX, 2021. This work was supported by the Brazilian National Council for Scientific and Technological Development (CNPq) under Fellowship 310060/2018-4. (*Corresponding author: Rafael Ferreira Mello*).

M. André is with the Informatics Center, Federal University of Pernambuco, Recife, Pernambuco 50670-901, Brazil (e-mail: madf@cin.ufpe.br).

R. F. Mello, A. Nascimento, and R. D. Lins are with the Department of Computer Science, Federal Rural University of Pernambuco, Recife, Pernambuco 52171-900, Brazil (e-mail: {rafael.mello, andre.camara, rafael.dueirelins}@ufrpe.br).

D. Gašević is with the Faculty of Information Technology, Monash University, Clayton, VIC 3800, Australia; with the School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, U.K.; and with the Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia (e-mail: dragan.gasevic@monash.edu).

and resources such as Linguistic Inquiry and Word Count (LIWC) and Coh-Metrix). Although the promising results of these approaches, there are state-of-the-art decision tree algorithms for classification [18] and categories of features (i.e., sentiment analysis and social network analysis (SNA)) that have not been explored for this problem before.

Therefore, this paper evaluates the efficacy of XGBoost (eXtreme Gradient Boosting) and AdaBoost (Adaptive Boosting) decision tree algorithms for classification [18], given that most of the existing research on the automatic detection of social and Cognitive Presence in CoI has used random forest. Furthermore, this study examined a wide range of features applied to the problem of automatic identification of Social Presence in online discussion transcripts. This paper reports on the findings of a study that investigated some features extracted by using 1) traditional text mining approaches based on the analysis of words used in the content of messages exchanged in online discussions [19]; 2) pretrained models (i.e., sentiment analysis and latent semantic analysis) [20]; 3) social network analysis [21]; 4) indicators of different psychological processes, and measures of text coherence and complexity [22], [23]. Finally, this study also evaluates the performance of BERT classifier in order to compare the decision tree algorithms with a deep learning algorithm.

II. Theoretical Background

A. The Community of Inquiry Model

The CoI model provides three dimensions (called presences) that aim to explain the interactions between students and instructors in online learning environments. The CoI model describes the behaviors of students and instructors in order to create an effective educational experience through online discussions [24]. These dimensions are [7]:

- **Social Presence** looks at the ability to humanize the relationships among participants in a discussion. It focuses on social interactions and aims to highlight the importance of the social climate within a group of learners [25].
- **Cognitive Presence** aims to investigate and support the students' problem-solving and critical thinking skills. It evaluates the progress of interactions in the students' Cognitive process, such as knowledge construction [24].
- **Teaching Presence** mainly concerns about instructors roles before (i.e., course design) and through social interactions (i.e., facilitation and direct instruction) within the course [7].

The central value of Social Presence in online discussions, in contrast to face-to-face interactions, is to enable establishing a socio-emotional interaction through text-based communication [9]. Social Presence is analyzed through three categories and twelve indicators, as shown in I. The **Affective** category examines how the translation of real emotions and feelings map onto the text of the messages exchanged. The indicators of this category include the expression of emotions, the sense of humor, and self-disclosure. The **Interactive** category is characterized by indicators such as quoting messages, asking questions, and expressing agreement. It aims to understand and enhance the open communication among participants. In general, the Interactive category is the one that most commonly occurs in discussions within a CoI. Finally, the **group cohesion** category investigates the sense of union and group engagement. The indicators in this category include vocatives and addressing or referring to the group using inclusive pronouns.

TABLE I
Indicators Social Presence [25]

Category	Indicator
Affective	1. Expression of emotions 2. Use of humor 3. Self-disclosure
Interactive	4. Continuing a thread 5. Quoting from others' messages 6. Referring explicitly to others' message 7. Asking questions 8. Complimenting, expressing appreciation 9. Expressing agreement
Cohesive	10. Vocatives 11. Addresses of refers to the group using inclusive pronouns 12. Phatics, salutations

Cognitive Presence has a central role in the CoI setting because it captures how "*participants move deliberately from understanding the problem or issue through to exploration, integration, and application.*" [9, p. 162]. It develops through a cycle of four phases: 1) triggering event, which initiates the cycle by introducing problems or questions proposed by the instructor and students; 2) exploration, where the participants are encouraged to propose solutions to the problem, involving brainstorming and the exchange of the findings; 3) integration, during this phase, the students starts to construct new concepts from the information collected; and 4) resolution, where the students propose hypothesis testing or real-world applications based on the problem/dilemma that triggered the learning cycle.

In spite of the importance of Cognitive Presence, previous studies highlighted that the applications of CoI should be committed to the framework as a whole and not only to individual presences [9], [24], [26]. Within this context, topics such as the prediction of student outcomes [24], the impact of Social Presence on the development of Cognitive Presence [9], [24], and the analysis on the students' progress across social and Cognitive Presences overtime [26] have been largely studied.

The recent methods to investigate the use of social and Cognitive Presences are based on natural language processing and machine learning techniques. The main approaches to code automatically the online discussion messages according to categories of social and Cognitive Presence found in the literature are presented in the next section. It is important to highlight that the automatic analysis of Teaching Presence has not received much attention as the other two CoI dimensions.

B. Extraction of the CoI Dimensions from Text

Content analysis of transcripts of online discussion messages is a commonly accepted method for analysis of the CoI presences. In this context, Garrison *et al*. [25] established coding schemes to examine social and Cognitive Presences, based on the categories explained in the previous section. These schemes are extensively used for the manual and automatic content analysis of CoI. The manual coding has been adopted for evaluating the impact of an instructional intervention.

However, manual coding is labor intensive and requires experienced coders which reduces the viability of this approach.

In terms of automatic analysis, the literature presents several papers dedicated to automatic detection of Cognitive Presence, which is addressed as the task of text classification [12]–[14], [27]–[29]. Early studies based their method on a combination of word and phrase counts and traditional machine learning algorithms (i.e., support-vector machine (SVM) and neural network) [12], [27] reaching results up to .31 Cohen's κ. While the results suggest a fair accuracy, the use of black-box classifiers and content-based features reduces the explainability and generalizability of the final models, respectively [28].

Recent studies examined the use of different features and classifiers. For instance, Kovanović *et al.* [28] developed an approach that relies on features based on Coh-Metrix [23], LIWC [22], LSA similarity, named entities, and discussion context [30] instead of word counts used by previous works. Moreover, they applied a random forest algorithm to classify the messages according to the categories of Cognitive Presence. The approach was evaluated with 10-fold cross-validation and a dataset composed of 1747 messages in English, manually categorized by two experts (Cohen's κ = .974) into five categories (Other, Triggering Event, Exploration, Integration, Resolution); the proposed method reached 70.3% for accuracy and .63 for Cohen's κ.

Neto *et al.* [29] followed a similar approach to identify Cognitive Presence phases in Portuguese. The authors evaluated the model developed using a dataset containing 1500 discussion messages (Cohen's κ = .86) divided into the phases of Cognitive Presence. The best model proposed achieved 83% of accuracy and .72 Cohen's κ. The same authors also proposed an analysis related to the generalization of the model to categorize Cognitive Presence [31]. In particular, the authors applied the same feature set and the Random Forrest classifier, but the evaluation was different. In this case, the classifiers were evaluated in two datasets extract from different disciplines (Biology and Technology). The Biology and Technology data sets encompassed 1500 and 734 messages, respectively. This study reported results for the evaluation on the entire data, which reached 76% of accuracy and .55 Cohen's κ, and for the process of training on one dataset and test on the other. The models reached .20 of Cohen's κ when trained on the biology data and evaluated on the technology one, and .38 for the other way around (DatasetTech for training and DatasetBio test).

An automatic approach to performing content analysis according to the categories of Social Presence has been proposed recently [11]. The approach combines the Coh-Metrix and LIWC indicators and standard text mining features (i.e., bag-of-words) to identify the categories of the Social Presence. This approach is based on cascading of three binary random forest classifiers, one for each category. The authors evaluated the performance of the model using the same data used by Kovanović *et al.* [28], but with the coding scheme focused on Social Presence. The final results reached .34, .85, and .93 of Cohen's κ for Affective, Interactive, and Cohesive categories of Social Presence, respectively.

Adopting the same set of features and dataset (of [28] and [29]) Barbosa *et al.* [15] presented an approach that use text translation to categorize Social Presence in different languages. The results reported the efficiency of using English text as a training set to classify Portuguese messages. In terms of Social Presence, the proposed approach reached .49, .92, and .44 of Cohen's κ for the Affective, Interactive, and Cohesive categories, respectively.

Recently, a few studies had reported the use of deep learning methods to identify Social Presence [16], [17]. Zou *et al.* [16] proposed an assessment of different deep learning approaches (BERT, One-way RNN with Attention layer, One-way RNN with Mish activation, and Attention layer, One-way RNN, Bidirectional RNN) and traditional algorithms (random forest and naive bayes) for the problem of identifying Social Presence. For the evaluation approach, the authors initially coded a dataset of 1000 messages, then generalized these categories to other messages using linguistic markers (e.g., verbs, adjectives, adverbs, punctuations, Etc.) in order to increase the size of the dataset. This was done to ensure sufficient data for training to support deep learning models training, which demands large quantities of training data. BERT obtained the best results (0.85 of accuracy and 0.83 of F1). Likewise, Zou *et al.* [17] compared the performance of BERT and random forest on a dataset of 3500 sentences containing the Social Presence indicators. Different from the previous paper [16], the authors evaluated different features for the classifiers. In the end, the combination of BERT and NER features reached the best results (0.83 of accuracy and 0.81 of F1). Although the benefits of using BERT, these studies [16], [17] also reported the benefits of using Coh-Metrix and LIWC features for this problem.

The existing studies have demonstrated the value of using random forest classifier as a white-box algorithm and different groups of features. For instance, Coh-Metrix and LIWC are high predictors of Cognitive Presence while general content words (i.e., hello, hi, and cheer) are important to the classification of Social Presence [11], [14], [15]. However, most recently, other decision tree approaches (i.e., AdaBoost and XGBoost) reached better performance when compared to random forest [18]. Therefore, this paper evaluates the performance of AdaBoost and XGBoost for the problem of Social Presence identification in online discussion messages. Furthermore, the results of the decision tree algorithms were also compared with BERT, the state-of-the-art deep learning model for linguistic analysis [32]. Finally, we also investigate the importance of the whole set of features previously proposed and extend it by including sentiment analysis (to extract students' social and emotional projections in a community [9]) and social network analysis (measures related to the dynamics of a discussion [33]) for this problem.

III. Research Questions

As discussed in Section II.B, the automatic content analysis methods applied for the identification of social and Cognitive Presence have not explored new classification algorithms. All the previous work found in the literature made use of traditional machine learning algorithms, especially random forest. Furthermore, there is a limited number of features evaluated in the context of Social Presence automatic identification. In this study, we followed the same methodology proposed by previous works, developing binary models, but extending it by using state-of-the-art white-box algorithms and a new feature set. Moreover, we compared the white-box algorithms with BERT, a state-of-the-art deep learning model for computational linguistics application, including text classification. Therefore, the first research question answered by the current study is:

> **Research Question 1 (RQ1):**
> *What is the performance of state-of-the-art white-box algorithms in comparison to the conventional ones in the classification of the Social Presence? To which extent do the white-box algorithms over perform deep learning methods in this context?*

In addition to evaluating new white-box algorithms for the problem of Social Presence identification, we were also keen to explore a more extensive variety of features and their relevance for each category of Social Presence. Hence, the second research question was formulated as:

> **Research Question 2 (RQ2):**
> *Which features are the best predictors of the categories of Social Presence?*

Finally, Kovanović *et al.* [28] suggested that increasing the number of features in the analysis of online discussion messages could increase the chances overfitting the classification algorithms, mainly when bag-of-words approaches are used [12], [27]. In contrast, the adoption of non content features (i.e., LIWC and Coh-Metrix) could increase the generalization of the classifier due to the fact that they measure the structure of the text instead of the content itself. Moreover, the reduction of the number of features, by removing the bag-of-words features, can also decrease the chances of overfitting [34]. Thus, the last research question to be answered is:

> **Research Question 3 (RQ3):**
> *What combinations of non content features are best predictors of the categories of Social Presence ?*

IV. Method

A. The Data and Course Design

The data used in the present study originates from six course offerings (Winter 2008, Fall 2008, Summer 2009, Fall 2009, Winter 2010, Winter 2011) of a master level research-intensive degree in software engineering offered entirely online, through the Moodle LMS, at a Canadian public university between 2008 and 2011. In those six offerings, a total of 81 students posted 1747 messages. The course encompassed six modules that covered fourteen different topics related to software engineering. The goal of the online discussions was to debate issues arisen around videos prerecorded by the students about research papers related to one of the course topics. The participation in the discussions accounted for 15% of the final grade [35].

Two expert coders categorized the dataset, considering the twelve indicators of Social Presence (see I) [36]. That is, for each post in the dataset, each indicator received the value "one" (has the indicator) or value "zero" (does not have the indicator). The percentage of agreement between the evaluators was 84%, which means that the coders had different opinion in only 16% of the messages. A third evaluator resolved the cases with disagreements [36]. Following Kovanovic *et al.* [36], three indicators (Continuing a thread, Complimenting, and Vocatives) were removed because they had a high number of messages. This removal was particularly necessary for this study because we are exploring the categories of Social Presence (Affective, Interactive, and Cohesive) and not the indicators (shown in I). Thus, the use of the aforementioned indicators would create extremely unbalanced data (i.e., all messages would be labeled with the Interactive category of Social Presence).

Finally, as the objective of this study is to construct a one binary classifiers for each category of Social Presence, the classes were reorganized to have binary coding (negative 0 or positive 1). For a message to be classified in positive (1), it must have at least one indicator annotated with the value "one" in the respective category. For instance, if a message had for the Affective category, the indicators Emotions = 0, Humor = 0, and Self_disclosure = 1, it was coded as positive (1). Finally, the dataset was tagged as shown in II.

TABLE II
Distribution of the Categories of the Social Presence

	Messages			
Category	Positive		Negative	
Affective	530	33.33%	1217	66.67%
Interactive	1030	58.95%	717	41.05%
Cohesive	1326	75.90%	421	24.10%

B. Feature Extraction

This work evaluates different features applied to the automatic identification of the categories of Social Presence. The range of features includes linguistic resources (LIWC and Coh-Metrix), natural language processing-based features (LSA, NER, and sentiment analysis), structural features (SNA and DCF), and traditional text mining features (word frequency). The following sections provide an overview of those features.

1) LIWC features: The Linguistic Inquiry Word Count (LIWC) is a resource that allows the extraction of 93 characteristics from a textual document considering four categories: Summary of language variables, linguistic dimensions, grammar, and psychological processes. These features are commonly used for automatic classification of social and Cognitive Presences of the CoI model [11], [28], [29]. For instance, the psychological processes category includes indicators, such as Affective processes, positive/negative emotion, social processes, and Cognitive processes [37]. LIWC was particularly important for the identification of Affective and Interactive categories in previous work [11]. This study employed the LIWC 2015 version to extract features from the online discussion messages in our dataset. In this study, we evaluated all the LIWC features available.

2) Coh-Metrix features: In addition to the psychological indicators provided by LIWC, the Coh-Metrix linguistic resource [23] was also applied. It allows the extraction of 108 features related to textual cohesion, coherence, linguistic complexity, text readability, and lexical category [23]. Several studies have reported the efficiency of Coh-Metrix to analyze aspects relevant to Social Presence such as cohesion and coherence of texts [38], [39]. Ferreira *et al.* [11] showed that Coh-Metrix had 30% of the features that are most predictive of Interactive categories, highlighting its importance to this problem. Moreover, the recent literature on the identification of CoI dimensions indicates the importance of this resource [11], [14], [28], [29]. Similarly to LIWC, we used all the Coh-Metrix features available for this study.

3) Latent semantic analysis similarity: According to [40], Latent Semantic Analysis (LSA) is a natural language processing technique used for verifying the contextual meaning of words or documents. LSA is widely employed to compute semantic similarity in the identification of concepts and their context. This measure reached relevant results in the identification of the Cognitive Presence, but it has not been applied to the automatic detection of Social Presence before [14]. We hypothesized that LSA is relevant for identifying indicators of the Interactive category, such as quoting from others' messages, expressing agreement, and referring explicitly to others' message. In this paper, we followed the approach proposed by Kovanovic *et al.* [28] to capture the degree of coherence of each message using the LSA semantic space.

4) Number of Named Entities: Named Entity Recognition (NER) is a natural language processing technique that seeks to identify named entities (e.g., named objects such as people, organizations, and geographical locations) mentioned in a piece of text (e.g., a discussion message) [41]. As shown in the previous studies, the number of named entities is highly indicative of the different phases of the Cognitive Presence [14], [28]. Here one hypothesized that the Cohesive category of the Social Presence should contain more named entities (e.g., mentions to the names of the participants in the discussion).

5) Sentiment analysis: Social Presence is related to the students' social and emotional projection in a community [7], [9]. Therefore, sentiment analysis (SA) could contribute to the construction of a Social Presence classifier. SA is the field dedicated to the automatic analysis of feelings, opinions, attitudes, and emotions from text documents [20].

According to Medhat, Hassan, and Korashy [20], one method of SA is the adoption of sentiment-enriched dictionaries. Therefore, this study adopted four dictionaries to extract the feelings expressed in the online discussion messages: TextBlob, SentiWordNet, AFINN [42], and Emotions Ekphrasis [43].

It is important to note that one document can contain phrases that express different feelings. There are methods of SA that perform the analysis at the document level assigning a single sentiment (e.g., Positive, Neutral, or Negative) to the entire text [44], [45]. Thus, using the dictionaries aforementioned, this study inferred whether a given post expresses a sentiment considering the text at a sentence or document level. The presence or absence of sentiment was extracted:

- In a sentence using the Textblob.
- In a sentence using AFFIN.
- In a sentence using SentiWordNet.
- In the entire message using Textblob.
- In the entire message using AFFIN.
- In the entire message using SentiWordNet.
- In the entire message using Emotions Ekphrasis.

6) Discussion Context Features (DCF): In addition to the adoption of the linguistic resources and natural language processing features, several discussion context features (DCF) initially proposed by Kovanovic *et al.* [28] were extracted, including:

- *Message Depth*: The numeric position of the message within the discussion thread.
- *Number of replies*: The number of responses received by each discussion message.
- *Cosine similarity to previous/next message*: The cosine similarity of the message text to the previous and next message within the discussion.
- *Start/end indicators*: A binary value that indicates whether the message is starting or ending the discussion.

Since the theory of Social Presence suggests that its categories develop overtime [24], the DCF could assist the proposed approach for automatic predicting the indicators presented in I. For instance, Cohesive messages are more likely to occur later on in the discussion process, as the students take time to know each other before recognizing themselves as a community.

7) Social Network Analysis (SNA): Social network analysis (SNA) offers numerical and visual insights into the types of relationships and interactions that occur between individuals (e.g., students and instructors), groups, and communities within a large number of social interactions [46]. With SNA, it is possible to understand how social interactions are occurring and which individuals are more or less communicative [33]. The literature shows several studies that have adopted SNA to understand the students' interactions in asynchronous online discussion. In this study, the measures that are widely used in the indication of connections around a specific node [47] were investigated:

- *Closeness centrality* evaluates how close a student in the network is to all other students. According to Yusof *et al.* [33], students with higher closeness degree tend to be effective in spreading information over the network.

- *Betweenness centrality* investigates the importance of a student's intervention for the interactions of other students. Therefore, a high level of betweenness indicates the leadership of a student in relation to their peers.
- *Degree centrality* seeks to analyze the level of interaction of the student with the other students present in the network. Thus, a higher degree centrality of a student means more influence in the network.

In this study, the networks were created based on the online discussion graph, which means that an explicit reply to a specific message creates a link in the network. To our knowledge, the SNA measures have not been explored in the context of the automatic identification of the CoI presences. The hypothesis raised here is that the SNA features reveal tendencies towards more interconnected and stronger social networks that support the identification of indicators of the Cohesive category of Social Presence.

8) Word frequency: A bag-of-words vector, a traditional text mining technique, was adopted as the last set of features to extract Social Presence. This method performs a transformation from the textual document (e.g., online discussion messages) into an array consisting of the terms weights [48]. More specifically, this study adopted the TF–IDF (term frequency–inverse document frequency) technique [48], which was calculated in three steps, as shown in Equations 1, 2, and 3.

$$\text{TF} = \frac{\#occurrencesofaterminadocument}{total\#termsinthedocument} \tag{1}$$

$$\text{IDF} = 1 + log_e \frac{total\#documents}{\#documentsthathaveacertainterm} \tag{2}$$

$$\text{TF–IDF} = TF \times IDF \tag{3}$$

A limitation of TF–IDF is the high dimensionality of the final vector generated. This happens because the TF–IDF uses the entire vocabulary of the documents analyzed in order to produce the final output. To minimize the effects of this problem, several preprocessing methods were applied:

- *Removal of stopwords*: it removes the words with little significance in a text such as articles, conjunctions, and prepositions.
- *Stemming*: it reduces the words contained in the corpus to their respective radical, decreasing the variability of the vocabulary of the dataset. For example, the words "engineer" and "engineering" become "engine."
- *Lemmatization*: It replaces the words contained in the corpus with their root forms known as lemma. For instance, the verbs "is" and "been" are converted to "be."
- *Restriction of grammatical classes*: It applies a POS Tagger to analyze each word or term contained in a sentence and then assigns each item a grammatical class. In this study, only adjectives, adverbs, nouns, and verbs were considered relevant.
- *Word disambiguation*: It standardizes the vocabulary of a corpus according to the meaning of the words to minimize problems resulting from the adoption of synonymy. Lesk, a word-sense disambiguation algorithm available in the NLTK library, was adopted for this step.

9) Summary of the analyzed features: III summarizes the features evaluated in this study including the linguistic resources (LIWC and Coh-Metrix), the natural language processing-based features (LSA, NER, and sentiment analysis), the structural features (SNA and DCF), and the traditional text mining approaches (Word frequency). It shows the number of features extracted in each group and their IDs used in the rest of this paper.

TABLE III
List of Linguistic Resources

ID	Gourp of features	Size
1	LIWC	93
2	Coh-Metrix	108
3	LSA	1
4	DCF	4
5	NER	1
6	SNA	3
7	Sentiment Analysis	7
8	Word Frequency	9976
9	Word Frequency–(stopWords)	9942
10	Word Frequency–(stopWords + stemming)	6436
11	Word Frequency–(word desambiguation)	9472
12	Word Frequency–(stopWords + POS–adjective, adverb, noun, verb)	5672
13	Word Frequency–(stopWords + lemmatization)	8929
	Total	**50 644**

This study explored all the features from the suggested resources, even the features that could be similar. For instance, both LIWC and Coh-Metrix have features related to word counts. However, the values for these features are not always the same. It is important to highlight that we used decision tree algorithms that perform internally feature selection processes. So, the use of all features will not create overfitting in the proposed models.

C. Data Processing

Text classification has been adopted as the primary approach to several educational applications over the last years [4]. However, recently Farrow *et al.* [13] suggested that, in order to increase the validation of the results, the training and test sets should be divided into different contexts or courses run instead of using the traditional data split (75%–25%). Therefore, the data set adopted in this study was divided into the five initial offerings of the course (winter 2008, fall 2008, summer 2009, fall 2009, winter 2010) and the last offer (winter 2011) for the training and testing sets, respectively. As shown in IV, the training group had 1510 (86%) posts and the test group with 237 (14%) posts. The negative and positive classes presented approximate distributions in the Interactive and Cohesive categories, with a greater difference in class distribution only for the Affective category.

TABLE IV
Distribution of Posts in Training and Testing Groups

	Group	negative (0)	positive (1)	Total
Affective	Train	1038 (69%)	472 (31%)	1510
	Test	179 (76%)	58 (24%)	237
Interactive	Train	620 (41%)	890 (59%)	1510
	Test	97 (41%)	140 (59%)	237
Cohesive	Train	362 (24%)	1148 (76%)	1510
	Test	59 (25%)	178 (75%)	237

Aiming to answer research question 1, two approaches to analyze the data were followed: 1) a cross-validation analysis using only the training set and each course as one fold; and 2) the traditional application of training and testing set to build and evaluate the model.

D. Model Selection and Evaluation

A comparative analysis of the 179 classification algorithms [49] revealed that Random forests and Gaussian kernel SVMs reach better results than other traditional Machine Learning Algorithms. Because of this performance advantage and the white-box nature, the applications of text classification in the educational domain usually adopts random forest classifiers [4]. However, the recent literature introduced new decision tree algorithms, AdaBoost and XGBoost, that can achieve better results when compared to random forest [50].

Despite the similarities, these algorithms have some differences. According to [51], random forest is a technique called bagging that combines different decision trees generated from the sub-sampling of the examples present in the training set. The final forecast of a random forest is determined from the average of the results of the individual ones. On the other hand, the AdaBoost and XGBoost algorithms use the boosting technique, which consists of an iterative process where each tree created focuses on the examples that have incorrectly been classified in the previous iteration [18]. Finally, we also evaluated the contextual embedding BERT for this text classification task [32]. This study compared the random forest, AdaBoost, XGBoost, and BERT algorithms to address research question 1.

The use of decision tree algorithms provides two crucial features of relevance for answering research question two and three: 1) the estimation of individual features importance through the analysis of the metric mean decrease in Gini impurity index (MDG) [52]; 2) the internal feature selection mechanism that highlights the main resources for each classifier [53]. The approach followed in this study analyzed the outcomes of the classifiers using all the features proposed in Section IV-B and excluding the Word Frequency features to answer research questions two and three, respectively.

Moreover, in order to compare the results with the state-of-the-art deep learning method, we adopted the BERT classifier. BERT (Bidirectional Encoder Representations from Transformers) is a word embedding approach that takes the context of each word into consideration, which increases the performance in several natural language processing applications [32]. We decide to use BERT in this study due to the adoption of this model in previous papers on the analysis of Social Presence [16], [17]. It is important to mention that we have not done any data extension and preparation in our dataset before use BERT.

Finally, in educational data mining and learning analytics [29], [36], to measure the performance of a supervised machine learning algorithm, Cohen's κ [54] is commonly used and was also adopted in this study. Moreover, we also evaluated the results of traditional machine learning measures: Precision (P), Recall (R), and F1-score.

V. Results

A. Research Question 1: Performance of the Algorithms

The first research question aimed to compare the results of state-of-the-art decision tree and deep learning algorithms to the random forest classifier. Initially, we focused on the analysis of random forest, AdaBoost and XGBoost. V shows the best results achieved by each algorithm using cross-validation in the training set (as described in IV-D), with and without considering the word frequency features, respectively. We performed this comparison because previous research suggests that word frequency features could reduce the generalizability of the final classifier [28]. The outcomes showed that the XGBoost algorithm outperformed AdaBoost and random forest in the three categories of Social Presence and with all measures. It is also important to highlight that the classifiers that used the word frequency features outperformed the classifiers without them. The results, in terms of Cohen's κ, decreased at least 36%, 10%, and 10% for the Affective, Interactive, and Cohesive categories, respectively. The F1-score follows a similar trend as the results decreased for the models that did not considered the word frequency features. Such results demonstrate the importance of the Word Frequency features in the training step.

After the initial evaluation, all classifiers were applied to the training set examples (1510 posts–initial five runs of the courses in our data) to generate a binary classifier, and their generalization capacities were verified in the respective testing sets (237 posts–the last run of the courses in our data). VI presents the application of those models for each classifier to the test set. In general, the models demonstrated good generalizability for unknown examples, as the lowest values for precision and recall were 0.74 and 0.75, respectively. Similar to the previous analysis, the classifiers reached better results when the word frequency features were used. However, in this analysis with the different testing dataset, the AdaBoost classifier outperformed XGBoost in the Affective category.

Finally, VII compares the best results of the decision tree models with BERT on the test set. XGBoost and AdaBoost outperformed the BERT model for all Social Presence category and in for all measures analyzed.

B. Research Question 2: Best Features for the Problem

Research question two intended to highlight the most relevant features for the automatic identification of Social Presence. This study applied the same feature vectors to

TABLE V
RESULTS FOR SOCIAL PRESENCE CLASSIFICATION CONSIDERING WORD FREQUENCY FEATURES–CROSS-VALIDATION

Category	Algorithms	With Word Frequency				Without Word Frequency			
		κ	P	R	F1	κ	P	R	F1
Affective	XGBoost	.49 (.05)	0.80 (0.05)	0.81 (0.04)	0.80 (0.05)	.27 (.06)	0.71 (0.05)	0.73 (0.04)	0.71 (0.05)
	AdaBoost	.44 (.04)	0.76 (0.04)	0.77 (0.04)	0.77 (0.04)	.25 (.05)	0.69 (0.04)	0.70 (0.04)	0.69 (0.04)
	Random Forest	.31 (.04)	0.71 (0.04)	0.72 (0.04)	0.71 (0.04)	.20 (.03)	0.71 (0.04)	0.73 (0.04)	0.72 (0.04)
Interactive	XGBoost	.83 (.05)	0.92 (0.01)	0.92 (0.02)	0.92 (0.02)	.74 (.06)	0.88 (0.02)	0.87 (0.02)	0.87 (0.02)
	AdaBoost	.80 (.04)	0.89 (0.03)	0.89 (0.03)	0.89 (0.03)	.70 (.04)	0.86 (0.01)	0.85 (0.02)	0.85 (0.02)
	Random Forest	.75 (.04)	0.81 (0.03)	0.80 (0.03)	0.80 (0.03)	.68 (.07)	0.83 (0.04)	0.82 (0.05)	0.83 (0.05)
Cohesive	XGBoost	.84 (.05)	0.93 (0.02)	0.93 (0.03)	0.93 (0.03)	.56 (.06)	0.85 (0.05)	0.84 (0.06)	0.84 (0.06)
	AdaBoost	.85 (.04)	0.94 (0.03)	0.93 (0.03)	0.93 (0.03)	.51 (.10)	0.82 (0.06)	0.84 (0.06)	0.84 (0.07)
	Random Forest	.60 (.04)	0.83 (0.04)	0.80 (0.08)	0.81 (0.09)	.54 (.06)	0.85 (0.04)	0.84 (0.06)	0.84 (0.07)

* K = Kappa; P = Precision; R = Recall

TABLE VI
RESULTS FOR SOCIAL PRESENCE CLASSIFICATION CONSIDERING WORD FREQUENCY FEATURES–TEST SET

Category	Algorithms	With Word Frequency				Without Word Frequency			
		κ	P	R	F1	κ	P	R	F1
Affective	XGBoost	.33	0.79	0.80	0.77	.28	0.74	0.77	0.74
	AdaBoost	.38	0.75	0.77	0.75	.29	0.74	0.75	0.74
	Random Forest	.17	0.79	0.78	0.72	.21	0.76	0.78	0.72
Interactive	XGBoost	.79	0.91	0.91	0.91	.76	0.88	0.88	0.88
	AdaBoost	.77	0.89	0.89	0.89	.65	0.83	0.83	0.83
	Random Forest	.62	0.81	0.81	0.81	.65	0.82	0.82	0.82
Cohesive	XGBoost	.94	0.97	0.97	0.97	.68	0.89	0.89	0.89
	AdaBoost	.89	0.96	0.96	0.96	.64	0.87	0.87	0.87
	Random Forest	.74	0.83	0.81	0.82	.66	0.89	0.88	0.88

* K = Kappa; P = Precision; R = Recall

TABLE VII
COMPARING THE BEST DECISION TREE ALGORITHM WITH BERT–TEST SET

Category	Algorithms	κ	P	R	F1
Affective	XGBoost without Word Frequency	.28	0.74	0.77	0.74
	AdaBoost with Word Frequency	.38	0.75	0.77	0.75
	BERT	.17	0.73	0.76	0.74
Interactive	XGBoost without Word Frequency	.76	0.88	0.88	0.88
	XGBoost with Word Frequency	.79	0.91	0.91	0.91
	BERT	.19	0.61	0.62	0.61
Cohesive	XGBoost without Word Frequency	.68	0.89	0.89	0.89
	XGBoost with Word Frequency	.94	0.97	0.97	0.97
	BERT	.06	0.75	0.75	0.75

* K = Kappa; P = Precision; R = Recall

discriminate the classes (positive and negative) of the three categories of Social Presence. Although we have used the whole feature set (Table III) as input for all classifiers, the internal feature selection mechanism of the three decision tree algorithms considered different features as the most predictive for the Affective, Interactive, and Cohesive categories.

Table VIII shows the number of features selected by each algorithm according to the different Social Presence categories (with the results presented on Tables V–VII). The study revealed that the most relevant features for the Affective and Interactive categories were LIWC and word frequency, with the exception of the Interactive category with the random forest classifier, which included features of sentiment analysis. All the classifiers used the word frequency features for the detection of the Cohesive category.

We performed an additional analysis of feature importance,

TABLE VIII
NUMBER OF FEATURES PER LINGUISTIC RESOURCE USED BY THE ALGORITHMS CONSIDERING WORD FREQUENCY

	XGBoost			**Adaboost**			**RandomForest**		
Group of features ID	Affective	Interactive	Cohesive	Affective	Interactive	Cohesive	Affective	Interactive	Cohesive
1	80	77	–	13	17	–	91	92	–
2	–	–	–	–	–	–	–	–	–
3	–	–	–	–	–	–	–	–	–
4	–	–	–	–	–	–	–	–	–
5	–	–	–	–	–	–	–	–	–
6	–	–	–	–	–	–	–	–	–
7	–	–	–	–	–	–	–	3	–
8–13	169	99	117	32	25	43	4176	4035	2532
Total	**249**	**176**	**117**	**45**	**42**	**43**	**4267**	**4130**	**2532**

TABLE IX
TEN MOST IMPORTANT VARIABLES FOR THE AFFECTIVE CATEGORY WITH WORD FREQUENCY FEATURES

Variable	**Description**	**MDG**	**Negative**	**Positive**
happy	Word frequency	3.440	0.000 (0.003)	0.021 (0.058)
hope	Word frequency	2.450	0.001 (0.009)	0.015 (0.033)
due	Word frequency	2.053	0.002 (0.017)	0.006 (0.026)
designed	Word frequency	1.867	0.001 (0.012)	0.003 (0.016)
programming	Word frequency	1.523	0.010 (0.046)	0.010 (0.038)
suggest	Word frequency	1.522	0.003 (0.018)	0.003 (0.017)
job	Word frequency	1.420	0.004 (0.022)	0.006 (0.026)
mining	Word frequency	1.386	0.003 (0.028)	0.004 (0.031)
happens	Word frequency	1.341	0.001 (0.012)	0.002 (0.018)
sorry	Word frequency	1.337	0.001 (0.009)	0.006 (0.027)

* MDG = mean decrease in gini

TABLE X
TEN MOST IMPORTANT VARIABLES FOR THE INTERACTIVE CATEGORY WITH WORD FREQUENCY FEATURES

Variable	**Description**	**MDG**	**Negative**	**Positive**
LIWC.QMark	Number of question marks	10.392	0.065 (0.397)	1.495 (1.378)
agree	Word frequency	5.692	0.000 (0.005)	0.016 (0.036)
absolutely	Word frequency	3.282	0.000 (0.005)	0.002 (0.017)
depends	Word frequency	2.953	0.001 (0.012)	0.002 (0.014)
late	Word frequency	2.581	0.001 (0.010)	0.001 (0.014)
independent	Word frequency	2.537	0.000 (0.004)	0.001 (0.012)
say	Word frequency	2.265	0.006 (0.025)	0.007 (0.026)
finding	Word frequency	1.925	0.001 (0.007)	0.003 (0.019)
guess	Word frequency	1.652	0.003 (0.022)	0.004 (0.021)
right	Word frequency	1.496	0.005 (0.024)	0.005 (0.021)

* MDG = mean decrease in gini

TABLE XI
TEN MOST IMPORTANT VARIABLES FOR THE COHESIVE CATEGORY WITH WORD FREQUENCY FEATURES

Variable	**Description**	**MDG**	**Negative**	**Positive**
hello	Word frequency	23.743	0.000 (0.004)	0.045 (0.030)
hi	Word frequency	7.125	0.000 (0.007)	0.005 (0.025)
bhi	Word frequency	5.054	0.000 (0.000)	0.002 (0.026)
cheer	Word frequency	3.919	0.000 (0.001)	0.011 (0.046)
regard	Word frequency	2.904	0.002 (0.014)	0.022 (0.050)
hey	Word frequency	2.861	0.000 (0.000)	0.002 (0.024)
techniqu	Word frequency	1.693	0.005 (0.027)	0.008 (0.038)
internet	Word frequency	1.564	0.004 (0.032)	0.002 (0.021)
sorri	Word frequency	1.461	0.002 (0.021)	0.002 (0.021)
gregg	Word frequency	1.427	0.005 (0.052)	0.004 (0.026)

* MDG = mean decrease in gini

using mean decrease in Gini impurity index (MDG), for each category using the XGBoost classifier as it reached better results when compared to random forest and AdaBoost. Tables IX, X, and XI present the top-ten features for the XGBoost classifier of each category (Affective, Interactive, and Cohesive). The Tables show a predominance of the word frequency features among the top-ten.

The most important set of variables for the Affective category shows only features related to word frequency. Although the word frequency features represent content words, the majority of the words presented in the top-ten are general ones, such as "hope," "happy," "due," and "sprry." The most predictive features for the Interactive category were nine from the word frequency, and one from the LIWC features. Table X shows that the number of Question marks (LIWC feature) was the most important one reaching MDG of 10.39. The presence of the word "agree" (stemmed from the word agreement), "absolutely," and "depends," which represents agreement, was also noticeable.

Finally, Table XI presents the main features of the Cohesive category of Social Presence where all top-ten were from the word frequency features. Again, general words, such as "hello," "hi," "cheer," and "regard" were relevant in this context. On the other hand, domain-related words like "techniqu" (stemmed from the word technique) and "internet" were also valuable.

C. Research Question 3: Best non Content Features

The last research question presents a similar analysis, as RQ2, but excluding the Word Frequency features. This section presents the most relevant features considering the following resources: LIWC, Coh-Metrix, LSA, DSF, NER, SNA, and sentiment analysis. Table XII details the features for the Social Presence categories used by each classifier (with the results presented on Tables V–VII). The results show the importance of investigating distinct aspects of the text, as all linguistic resources were necessary, to detect Social Presence in online discussions. Table XII also highlights the importance of LIWC and Coh-Metrix for this classification tasks and also highlights the relevance of the DCF, LSA, and NER features.

As in the previous section, Tables XIII, XIV, and XV introduce the top-ten features for each category adopting MDG for the XGBoost classifier. Differently from the previous RQ, the main features were divided between LIWC and Coh-Metrix.

The top-fifteen features for the Affective category include ten, four, and one from LIWC, Coh-Metrix, and sentiment analysis, respectively. Among the LIWC features, the relevant features were the number of first-person singular pronouns (LIWC.i) with MDG of 1.76, the number of words associated with friendship (LIWC.friend) with MDG of 1.38, and the number of words (LIWC.WC) with MDG of 1.76. For Coh-Metrix, the number of words, and the number of first-person pronouns were the most significant. Finally, the sentiment analysis using the Emotions resource was also significant in this case.

The LIWC was also predominant for the Interactive category (nine features), including the number of question marks, number of agreement words, and number of nonfluent words. Coh-Metrix had five relevant features, including sentence count and the number of senses of words. The LSA similarity was also important for this category of Social Presence.

Finally, Table XV presents the relevant features for the Cohesive category, which includes the appearance of features associated with linguistic expressions commonly used by students to make references to the other participants in a discussion. For instance, the presence of words related to affiliation/proximity and number of agreement words both from LIWC were important for classification of messages in the Cohesive category of Social Presence. The features of the number of second-person pronouns, number of connectives, and the number of named entities of Coh-Metrix were also significant for the Cohesive category of Social Presence.

VI. Discussion

A. Research Question 1: Performance of the Algorithms

The first research question analyzed the performance of state-of-the-art decision tree and deep learning classifiers to the Social Presence identification. XGBoost and AdaBoost achieved, in the best case scenario, Cohen's κ of .38, .79, and .94, for the Affective, Interactive, and Cohesive categories of Social Presence, respectively. These values represent medium to substantial inter-rater agreements [55]. In two out of the three categories, the agreement was above .70, which is the value commonly used in CoI research as the threshold above which manual coding results are considered valid. These values represented an increase of 123%, 27%, and 20% for the Affective, Interactive, and Cohesive, respectively when compared to the random forest results. One possible cause for the lower result on the classification of Affective category is the number of positive instances (messages containing an indicator of this category) in the dataset. While the Affective category has 530 positive instances, the Interactive and Cohesive has 1030 and 1326, respectively.

The literature reported a similar result .34 (Affective), .85 (Interactive), and .93 (Cohesive) using a random forest classifier, but after data balancing and parameter optimization [11]. The results obtained indicate the potential of using XGBoost and AdaBoost. Moreover, the proposed approach reached the values of κ which are better than those of the classifiers of Cognitive Presence developed for English [28], [30].

In addition to this analysis, the classifiers, excluding the word frequencies features were evaluated. To the best of our knowledge, such an analysis has never been performed before in the context of Social Presence. The results showed a decrease in the performance of 23%, 3%, and 27% for Affective, Interactive and Cohesive, respectively. It revealed that the automatic identification of the CoI dimensions is dependent on the domain of the discussion [24]. However, the results .29 (Affective), .76 (Interactive), and .68 (Cohesive) still represent a medium inter-rater agreement for the Interactive and Cohesive categories [55]. It is important to highlight that the analyzed corpus have unbalanced sets of data for Affective and Cohesive categories. However, the results for the Cohesive category reached better results due to the nature of the features analyzed (i.e., Coh-Metrix and LIWC) [**?**], [15].

TABLE XII
Number of Features Per Linguistic Resource Used by the Algorithms Without Word Frequency

	XGBoost			**Adaboost**			**RandomForest**		
Group of features ID	Affective	Interactive	Cohesive	Affective	Interactive	Cohesive	Affective	Interactive	Cohesive
1	82	75	77	25	20	21	91	91	91
2	91	87	91	15	14	21	100	100	100
3	1	1	1	1	1	1	1	1	1
4	4	4	3	1	1	1	4	4	4
5	1	1	1	–	–	–	1	1	1
6	2	3	2	1	–	1	3	3	3
7	1	–	1	2	–	–	4	5	4
Total	**182**	**171**	**99**	**45**	**36**	**45**	**204**	**205**	**204**

TABLE XIII
Fifteen Most Important Variables For the Affective Category Without Word Frequency Features

Variable	**Description**	**MDG**	**Negative**	**Positive**
cm.DESWC	Number of words–Coh-Metrix	5.882	122.154 (88.464)	171.291 (140.519)
LIWC.i	Number of first person singular pronouns	1.996	2.303 (2.044)	3.386 (2.191)
LIWC.WC	Number of words–LIWC	1.760	123.254 (89.249)	174.602 (141.709)
cm.WRDPRP1	Number of first person pronouns	1.546	22.855 (19.939)	33.533 (21.676)
LIWC.friend	Number of words associated with friendship	1.380	0.021 (0.144)	0.044 (0.206)
LIWC.Quote	Number of Quotation marks	1.267	1.499 (3.222)	1.689 (2.990)
LIWC.netspeak	Number of internet common words	1.229	0.482 (1.991)	0.489 (1.495)
cm.LDVOCD	Lexical diversity, VOCD	1.212	50.833 (53.392)	72.393 (51.689)
sa.emotions	Setiment Analysis using Emotions	1.165	0.071 (0.256)	0.208 (0.406)
LIWC.Parenth	Number of parentheses	1.086	0.999 (1.701)	1.225 (1.661)
LIWC.Exclam	Number of exclamation marks	1.020	0.205 (0.842)	0.660 (3.577)
LIWC.WPS	Number of words per sentence	1.019	15.900 (6.888)	16.736 (6.708)
LIWC.bio	Number of biological processes words	1.001	0.188 (0.505)	0.267 (0.520)
LIWC.male	Number of male references	0.971	0.042 (0.343)	0.067 (0.266)
cm.PCNARp	Measure of narrativity	0.957	40.542 (24.358)	51.684 (23.727)

Finally, the presented results also showed that the best decision tree classifiers outperform the BERT model in this case. It corroborates the combination of handcrafted features and decision tree classifiers are the best option for the automatic identification of CoI model presences [28], [31], and more specifically for Social Presence [11], [15]. There are papers found in the literature that reached better results using deep learning [16], [17]. However, these studies have not explored features extracted from Coh-Metrix and LIWC in the creation of the random forest models.

B. Research Question 2: Best Features for the Problem

In addressing research question 2, this study conducted a detailed analysis of several groups of features. Among the resources used, features analyzed were also present in previous studies [11], [56] in addition to new resources which were selected given their relevance for the categories of Social Presence, such as SNA, and sentiment analysis.

The first analysis presented in Table VIII demonstrates the potential of the decision tree algorithms to reduce the dimensionality of a feature vector. The algorithms analyzed in this study (Random Forrest, XGBoost and AdaBoost) perform a feature importance analysis using the MDG measure to rank the features according to their relevance. The features with MDG equal to zero were removed from the training step. To our knowledge, no previous work, in educational settings, have performed this analysis.

The initial set investigated was composed of 50 644 of features. The AdaBoost classifier shrunk these features to only 42 in the identification of the Interactive category. In the worst case, for the Cohesive category, random forest reduced the feature set to 8.42% of the original. Table VIII also highlighted that LIWC combined with word frequencies were the only features used by all classifiers created in Section V-A. The only exception was the Interactive category using the random forest classifier, which also considered three features of sentiment analysis.

Tables IX, X, and XI revealed a high dominance of word frequency features in the top-fifteen according to the mean decrease in Gini impurity index (MDG). Previous work also presented a similar trend [11]; however, top features in this study had a higher dependence of the word frequencies. Although the features related to word frequency could lead to overfitting depending on the domain, in the case of the current study, the words with high information gain were general ones such as *hope*, *happy*, and *due* (Affective); agree, absolutely, and depends (Interactive); *hello*, *hi*, and *cheer* (Cohesive). Such general words are suggestive of Social Presence indicators (i.e., “addresses to the students in the group using pronouns” and “expression of emotions” [25]) and are aligned

TABLE XIV
FIFTEEN MOST IMPORTANT VARIABLES FOR THE INTERACTIVE CATEGORY WITHOUT WORD FREQUENCY FEATURES

Variable	**Description**	**MDG**	**Negative**	**Positive**
LIWC.QMark	Number of question marks	15.063	0.065 (0.397)	1.495 (1.378)
cm.WRDPOLc	number of senses (core meanings) of a word	3.600	3.719 (0.826)	3.982 (0.571)
LIWC.assent	Number of agreement words	2.642	0.261 (1.431)	0.301 (0.711)
LIWC.nonflu	Number of non fluent words	2.343	0.173 (0.689)	0.277 (0.589)
cm.DESSC	Sentence count	1.511	7.025 (4.720)	9.181 (6.342)
LIWC.OtherP	Number of unusual punctuation	1.259	2.783 (6.688)	0.994 (2.763)
LIWC.i	Number of first person singular pronouns	1.257	2.521 (2.465)	2.708 (1.893)
cm.CNCADC	Number of adversative/contrastive connectives	1.146	13.982 (14.053)	15.784 (12.849)
cm.LSAGN	Average givenness of each sentence	1.141	0.160 (0.082)	0.188 (0.068)
LIWC.Colon	Number of colon	1.127	0.620 (1.435)	0.400 (0.828)
LIWC.see	Perceptual processes: seeing (e.g., view, saw, seen)	1.063	1.024 (2.002)	0.822 (1.073)
LIWC.ppron	Number of Personal pronouns	1.046	5.091 (4.121)	5.893 (2.884)
cm.CRFCWO1	Average content word overlap	0.967	0.076 (0.055)	0.089 (0.046)
LIWC.Dash	Number of dashes	0.959	1.908 (3.908)	1.167 (7.581)
lsa.similarity	Average LSA similarity between sentences	0.957	0.500 (0.287)	0.559 (0.206)

TABLE XV
FIFTEEN MOST IMPORTANT VARIABLES FOR THE COHESIVE CATEGORY WITHOUT WORD FREQUENCY FEATURES

Variable	**Description**	**MDG**	**Negative**	**Positive**
LIWC.affiliation	Presence of words related to affiliation/proximity	5.959	0.979 (2.457)	1.967 (2.171)
cm.WRDPRP2	Number of second person pronouns	2.62	16.64 (29.592)	20.422 (21.276)
LIWC.assent	Number of agreement words	2.415	0.375 (1.45)	0.256 (0.911)
cm.PCDCz	Measure of cohesion	2.248	0.000 (1.903)	0.065 (1.498)
cm.CRFAO1	Average argument overlap	2.194	0.393 (0.286)	0.408 (0.241)
LIWC.SemiC	Number of Semicolons	1.788	0.209 (1.198)	0.121 (0.485)
cm.CNCAll	Number of connectives	1.732	72.075 (35.011)	75.247 (27.632)
LIWC.anger	Number of anger related words	1.565	0.099 (0.363)	0.055 (0.239)
LIWC.bio	Number of biological processes words	1.425	0.221 (0.544)	0.209 (0.500)
LIWC.health	Number of health related words	1.389	0.118 (0.384)	0.100 (0.362)
cm.DESWC	Number of words	1.22	127.95 (107.462)	139.953 (109.754)
cm.WRDPRP1p	Number of first person pronouns	1.219	1.918 (6.726)	2.662 (6.892)
similarity.previous.post	Average similarity between sentences	1.123	0.200 (0.206)	0.229 (0.235)
LIWC.social	Number of social processes	1.063	1.271 (1.399)	1.269 (1.249)
ner	Number of Named Entities	1.062	30.805 (23.667)	33.986 (24.105)

with previous work [11].

C. Research Question 3: Best non Content Features

Finally, the last research question analyzed the most significant features to the problem of automatic detection of Social Presence, excluding the word frequency features. Table XII demonstrated the importance of the groups of features investigated. For instance, the random forest classifier employed at least one feature from each group in the classification process. Table XII also showed the same trend of VIII in terms of feature reduction, where AdaBoost performed the high compression followed by XGBoost and random forest. This reduction in the number of features could represent a better computational performance when adopted in real applications. Therefore, in order to deploy those classifiers in a practical scenario, the project should consider not only the final results but also the number of features. Tables XIII, XIV, and XV highlight the importance of LIWC and Coh-Metrix for the automation of the extraction of CoI presences, which is aligned with the previous studies that developed classifiers for both Cognitive and Social Presences with high importance of LIWC and Coh-Metrix features [11], [14], [29], [56], [57].

Among the features selected by the Affective category classifier, the most important ones were related to the word count (cm.DESWC and LIWC.WC), the number of first-person pronouns (LIWC.i and cm.WRDPRP1), and number of exclamation marks (LIWC.Exclam), which is aligned with the literature [11]. The novelty of this analysis were the features related to friendship (LIWC.friend), sentiment analysis (sa.emotions), and quotations to previous messages (LIWC.Quote). Those features have not been used in the previous literature on the classifiers for Social Presence (nor for Cognitive Presence), but they proved relevant in the research that performed manual content analysis, especially research that aimed to identify the Affective (LIWC.friend and sa.emotions) and Interactive (LIWC.Quote) categories.

The top-fifteen features of Table XIV included the number of question marks (LIWC.QMark), agreement words (LIWC.assent), and adversative connectives (cm.CNCADC) which are directly aligned to the conceptualisation of the Interactive category [25]. More specifically, these features are related to the indicators of Asking questions, Expressing agreement, and Referring explicitly to other people messages (see Table I). The average givenness of each sentence (cm.LSAGN) and average LSA similarity between sentences are indicative of the continuity in a thread that is extremely important for the Interactive category. The main difference from the previous work on the automatic identification of Social Presence in online discussion messages is the lack of features related to the second person pronoun incidence score, which is indicative of citing and referencing openly other messages or people in the discussion [11]. However, in the proposed approach, the number of Personal pronouns (LIWC.ppron) could also be a characteristic of the same indicator.

The analysis of the Cohesive category shows the importance of the number first and second-person pronouns and named entities (cm.WRDPRP1, cm.WRDPRP2, and NER). It corroborate with the work by Ferreira *et al.* [11] where the authors showed the relevance of using inclusive pronouns (us, ours) and direct mentioning of other students' names (the main NER extracted were related to names) as a way of demonstrating group cohesion. Another indicator of the Cohesive category, the demonstration of salutations, the number of affiliations (LIWC.affiliation), and social processes (LIWC.social) [11]. However, the most notable feature selected for this category is the Coh-Metrix measure of cohesion (cm.PCDCz) which is directly related to the goal of this category of Social Presence.

VII. Final Remarks

Practical implications of this study are three fold. *First*, the Social Presence is commonly associated with several educational theories in the literature, such as self-regulation, and self-efficacy [58], learners' prestige [17], and has been used as a factor to identify collaboration in online discussion [59]. Therefore, the automatic classification of Social Presence categories is relevant to studying these educational phenomena.

Second, the feature analysis proposed in this study could be used to evaluate the student performance in online discussion when combined with other aspects. For instance, previous works proposed measures to evaluate student performance in collaborative tasks, highlighting the Social Presence relevance [26].

Finally, the categories of Social Presence could be useful to assist the instructor in the analysis of the progress of the students (or group of students) over time in dashboards to facilitate the tracking of students contributions in computer-supported collaborative learning context [60].

A total of eight groups of features were analyzed including linguistic resources (LIWC and Coh-Metrix), natural language processing-based features (LSA, NER, and sentiment analysis), structural features (SNA and DCF), and traditional text mining approaches (Word frequency). The results confirmed that the word frequency features increased the final performance of the models. On the other hand, both XGBoost and random forest algorithms selected the newly introduced features (sentiment analysis and SNA) as important features for the classification process. These results indicate the potential of these features (sentiment analysis and SNA) in the automatic identification of the categories of Social Presence.

The limitations of the present study include: 1) the size and the nature of the dataset used. Although the dataset was based on the coding process that reached a high level of agreement between the annotators and the dataset was used in several other studies, it encompasses a relatively small number of instances (1747 messages). 2) The degree of generalizability of the presented results needs to be further investigated as all the data used was from only one course. However, the RQ3 explore non content features which minimize such a limitation. For instance, [14] demonstrated the generalization of this approach by proposing a cross-language classifier, for texts in English and Portuguese, using non content features. For further work, the authors plan to extend the current work by analyzing different data sources, including different languages (i.e., Portuguese), to overcome the aforementioned limitations. Moreover, we also intend to perform more experiments to understand why the feature reduction for the algorithms analyzed was different and explore the potential of using optimization methods and oversampling or undersampling algorithms to improve the outcomes.

References

[1] J. Short, E. Williams, and B. Christie, *The social psychology of telecommunications*. London, U.K.: Wiley, 1976, doi: 10.2307/2065899.

[2] P. Lowenthal, "Social presence," *Encyclopedia of distance learn.*, pp. 1900–1906, 2009, doi: 10.4018/978-1-60566-198-8.ch280.

[3] J. J. Barr, "Developing a positive classroom climate," Manhattan, KS, USA, IDEA Paper 61, Oct. 2016, [Online]. Available: https://www.ideaedu.org/idea_papers/developing-a-positive-classroom-climate/.

[4] R. Ferreira Mello, M. André, A. Pinheiro, E. Costa, and C. Romero, "Text mining in education," *Wiley Interdisciplinary Rev.: Data Mining & Knowl. Discovery*, vol. 9, no. 6, p. e1332, 2019, doi: 10.1002/widm.1332.

[5] T. Anderson and J. Dron, "Three generations of distance education pedagogy," *Int. Rev. Res. Open & Distrib. Learn.*, vol. 12, no. 3, pp. 80–97, 2010, doi: 10.19173/irrodl.v12i3.890.

[6] I. Galikyan and W. Admiraal, "Students' engagement in asynchronous online discussion: The relationship between cognitive presence, learner prominence, and academic performance," *Internet & Higher Educ.*, vol. 43, p. 100692, 2019, doi: 10.1016/j.iheduc.2019.100692.

[7] D. R. Garrison, T. Anderson, and W. Archer, "Critical inquiry in a text-based environment: Computer conferencing in higher Education," *Internet & Higher Educ.*, vol. 2, no. 2–3, pp. 87–105, 1999, doi: 10.1016/S1096-7516(00)00016-6.

[8] V. Kovanović, D. Gašević, S. Joksimović, M. Hatala, and O. Adesope, "Analytics of communities of inquiry: Effects of learning technology use on cognitive presence in asynchronous online discussions," *Internet & Higher Educ.*, vol. 27, pp. 74–89, 2015, doi: 10.1016/j.iheduc.2015.06.002.

[9] D. R. Garrison and J. B. Arbaugh, "Researching the community of inquiry framework: Review, issues, and future directions," *Internet & Higher Educ.*, vol. 10, no. 3, pp. 157–172, 2007, doi: 10.1016/j.iheduc.2007.04.001.

[10] R. Donnelly and J. Gardner, "Content analysis of computer conferencing transcripts," *Interactive Learn. Environ.*, vol. 19, no. 4, pp. 303–315, 2011, doi: 10.1080/10494820903075722.

[11] M. Ferreira, V. Rolim, R. F. Mello, R. D. Lins, G. Chen, and D. Gašević, "Towards automatic content analysis of social presence in transcripts of online discussions," in *Proc. 10th Int. Conf. Learning Analytics & Knowledge (LAK'20)*, Frankfurt, Germany, Mar. 23–27, 2020, pp. 141–150, doi: 10.1145/3375462.3375495.

[12] T. E. Mcklin, "Analyzing cognitive presence in online courses using an artificial neural network," Ph.D. dissertation, Dept. Middle-Secondary Educ. Instructional Technol., Georgia State Univ., Atlanta, GA, USA, 2004, [Online]. Available: https://scholarworks.gsu.edu/msit_diss/1.

[13] E. Farrow, J. Moore, and D. Gašević, "Analysing discussion forum data: A replication study avoiding data contamination," in *Proc. 9th Int. Conf. Learning Analytics & Knowledge*, Tempe, Arizona, Mar. 4–8, 2019, pp. 170–179, doi: 10.1145/3303772.3303779.

[14] G. Barbosa *et al.*, "Towards automatic cross-language classification of cognitive presence in online discussions," in *Proc. 10th Int. Conf. Learning Analytics & Knowledge (LAK'20)*, Frankfurt, Germany, Mar. 23–27, 2020, pp. 605–614, doi: 10.1145/3375462.3375496.

[15] A. Barbosa, M. Ferreira, R. Ferreira Mello, R. Lins, Rafael Dueire, and D. Gašević, "The impact of automatic text translation on classification of online discussions for social and cognitive presences," in *Proc. 11th Int. Conf. Learning Analytics & Knowledge (LAK'21)*, Irvine, USA, Apr. 12–16, 2021, pp. 77–87, doi: 10.1145/3448139.3448147.

[16] W. Zou, Z. Pan, C. Li, and M. Liu, "Does social presence play a role in learners' positions in MOOC learner network? A machine learning approach to analyze social presence in discussion forums," in *Proc. 3rd Int. Conf. Quantitative Ethnography (ICQE'21).* Malibu, CA, USA: Springer, Nov. 6–9, 2021, pp. 248–264, doi: 10.1007/978-3-030-67788-6_17.

[17] W. Zou, X. Hu, Z. Pan, C. Li, Y. Cai, and M. Liu, "Exploring the relationship between social presence and learners' prestige in MOOC discussion forums using automated content analysis and social network analysis," *Comput. Human Behav.*, vol. 115, p. 106582, 2021, doi: 10.1016/j.chb.2020.106582.

[18] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD'16)*, California, USA, Aug. 13–17, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[19] R. T.-W. Lo, B. He, and I. Ounis, "Automatically building a stopword list for an information retrieval system," in *J. Digital Information Management: Special Issue 5th Dutch-Belgian Information Retrieval Workshop (DIR)*, vol. 5, Utrecht, Netherlands, Jan. 10–11, 2005, pp. 17–24, doi: 10.1.1.111.3041.

[20] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Eng. journal*, vol. 5, no. 4, pp. 1093–1113, 2014, doi: 10.1016/j.asej.2014.04.011.

[21] S. Wasserman and K. Faust, *Social network analysis: Methods and applications*. Cambridge University Press, 1994, doi: 10.1017/CBO9780511815478.

[22] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *J. Lang. & social Psychol.*, vol. 29, no. 1, pp. 24–54, 2010, doi: 10.1177/0261927X09351676.

[23] D. S. McNamara, A. C. Graesser, P. M. McCarthy, and Z. Cai, *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press, 2014, doi: 10.1017/CBO9780511894664.

[24] D. R. Garrison, T. Anderson, and W. Archer, "The first decade of the community of inquiry framework: A retrospective," *Internet & Higher Educ.*, vol. 13, no. 1–2, pp. 5–9, 2010, doi: 10.1016/j.iheduc.2009.10.003.

[25] L. Rourke, T. Anderson, D. R. Garrison, and W. Archer, "Assessing social presence in asynchronous text-based computer conferencing," *J. Distance Educ.*, vol. 14, no. 2, pp. 50–71, 1999, available: http://hdl.handle.net/2149/732.

[26] V. Rolim, R. Ferreira, R. D. Lins, and D. Gǎsević, "A network-based analytic approach to uncovering the relationship between social and cognitive presences in communities of inquiry," *Internet & Higher Educ.*, vol. 42, pp. 53–65, 2019, doi: 10.1016/j.iheduc.2019.05.001.

[27] S. Corich, K. Hunt, and L. Hunt, "Computerised content analysis for measuring critical thinking within discussion forums," *J. e-Learn. & Knowl. Soc.*, vol. 2, no. 1, 2006, doi: 10.20368/1971-8829/700.

[28] V. Kovanović *et al.*, "Towards automated content analysis of discussion transcripts: A cognitive presence case," in *Proc. 6th Int. Conf. Learning Analytics & Knowledge (LAK'16).* New York, NY, USA: ACM, Apr. 25–29, 2016, pp. 15–24, doi: 10.1145/2883851.2883950.

[29] V. Neto *et al.*, "Automated analysis of cognitive presence in online discussions written in portuguese," in *European Conf. Technology Enhanced Learning*, Leeds, UK, Sep. 3–6, 2018, pp. 245–261, doi: 10.1007/978-3-319-98572-5_19.

[30] Z. Waters, V. Kovanović, K. Kitto, and D. Gašević, "Structure matters: Adoption of structured classification approach in the context of cognitive presence classification," in *11th Asia Information Retrieval Symposium Information Retrieval Technology*, Brisbane, Australia, Dec. 2–4, 2015, pp. 227–238, doi: 10.1007/978-3-319-28940-3_18.

[31] V. Neto, V. Rolim, A. P. Cavalcanti, R. D. Lins, Lins, D. Gašević, and R. Ferreira Mello, "Automatic content analysis of online discussions for cognitive presence: A study of the generalizability across educational contexts," *IEEE Trans. Learn. Technol.*, 2021, doi: 10.1109/TLT.2021.3083178.

[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[33] Erlin, N. Yusof, and A. Abdul, Rahman, "Students' interactions in online asynchronous discussion forum: A social network analysis," in *IEEE Int. Conf. Education Technology & Computer (ETC'2009*, Singapore, Malaysia, Apr. 17–20, 2009, pp. 25–29, doi: 10.1109/ICETC.2009.48.

[34] S. Khalid, T. Khalil, and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning," pp. 372–378, 2014, doi: 10.1109/SAI.2014.6918213.

[35] D. Gašević, O. Adesope, S. Joksimović, and V. Kovanović, "Externally-facilitated regulation scaffolding and role assignment to develop cognitive presence in asynchronous online discussions," *Internet & Higher Educ.*, vol. 24, pp. 53–65, 2015, doi: 10.1016/j.iheduc.2014.09.006.

[36] V. Kovanovic, S. Joksimovic, D. Gašević, and M. Hatala, "What is the source of social capital? the association between social network position and social presence in communities of inquiry," in *CEUR Workshop Proc.*, vol. 1183, Rome, Italy, 07 2014.

[37] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of LIWC2015," Tech. Rep., 2015, doi: 10.15781/T29G6Z.

[38] P. M. McCarthy, G. A. Lewis, D. F. Dufty, and D. S. McNamara, "Analyzing Writing Styles with Coh-Metrix," in *Proc. Int. Conf. Florida Artificial Intelligence Research Society Conference (FLAIRS'2006)*, Melbourne Beach, USA, May. 11–13, 2006, pp. 764–769, doi: 10.3758/BF03195564.

[39] A. C. Graesser, D. S. McNamara, and J. M. Kulikowich, "Coh-Metrix: Providing multilevel analyses of text characteristics," *Educ. researcher*, vol. 40, no. 5, pp. 223–234, 2011, doi: 10.3102/0013189X11413260.

[40] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse processes*, vol. 25, no. 2–3, pp. 259–284, 1998, doi: 10.1080/01638539809545028.

[41] H. Shelar, G. Kaur, N. Heda, and P. Agrawal, "Named entity recognition approaches and their comparison for custom NER model," *Science & Technology Libraries*, vol. 39, no. 3, pp. 324–337, 2020, doi: 10.1080/0194262X.2020.1759479.

[42] F. Å. Nielsen, "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs," 2011, *arXiv:1103.2903*.

[43] C. Baziotis, N. Pelekis, and C. Doulkeridis, "Datastories at semeval-2017 task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis," in *Proc. 11th Int. workshop semantic evaluation (SemEval'2017)*, 2017, pp. 747–754, doi: 10.18653/v1/S17-2126.

[44] A. Yessenalina, Y. Yue, and C. Cardie, "Multi-level structured models for document-level sentiment classification," in *Proc. Conf. Empirical Methods Natural Language Processing (EMNLP'2010)*, Cambridge, MA, Oct. 9–11, 2010, pp. 1046–1056, doi: 10.5555/1870658.1870760.

[45] N. Farra, E. Challita, R. A. Assi, and H. Hajj, "Sentence-level and document-level sentiment mining for arabic texts," in *IEEE Proc. Int. Conf. Data Mining Workshops (ICDM'2010)*, Sydney, Australia, Dec. 13–13, 2010, pp. 1114–1119, doi: 10.1109/ICDMW.2010.95.

[46] J. Scott, "Social network analysis," *Sociology*, vol. 22, no. 1, pp. 109–127, 1988, doi: 10.1177/0038038588022001007.

[47] P. J. Carrington, J. Scott, and S. Wasserman, *Models and methods in social network analysis*. Cambridge university press, 2005, vol. 28, doi: 10.1017/CBO9780511811395.

[48] C. Manning and H. Schutze, "Foundations of Statistical Natural Language Processing," *Comput. Linguistics*, vol. 26, no. 2, pp. 277–279, 1999, doi: 10.1162/coli.2000.26.2.277.

[49] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3133–3181, 2014, doi: 10.5555/2627435.2697065.

[50] S. Jhaveri, I. Khedkar, Y. Kantharia, and S. Jaswal, "Success Prediction using Random Forest, CatBoost, XGBoost and AdaBoost for Kickstarter Campaigns," in *IEEE 3rd Int. Conf. Computing Methodologies & Communication (ICCMC)*, Erode, India, Mar. 27–29, 2019, pp. 1170–1173, doi: 10.1109/ICCMC.2019.8819828.

[51] J. Hartmann, "Classification using decision tree ensembles," in *Machine Age Customer Insight*. Bingley, UK: Emerald, 2021, pp. 103–117, doi: 10.1108/978-1-83909-694-520211011.

[52] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[53] V. Sugumaran, V. Muralidharan, and K. Ramachandran, “Feature selection using decision tree and classification through proximal support vector machine for fault diagnostics of roller bearing,” *Mech. Syst. & signal Process.*, vol. 21, no. 2, pp. 930–942, 2007, doi: 10.1016/j.ymssp.2006.05.004.

[54] J. Cohen, “A coefficient of agreement for nominal scales,” *Educ. & Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960, doi: 10.1177/001316446002000104.

[55] J. R. Landis and G. G. Koch, “The measurement of observer agreement for categorical data,” *Biometrics*, pp. 159–174, 1977, doi: 10.2307/2529310.

[56] V. Kovanović *et al.*, “Towards automated content analysis of discussion transcripts: A cognitive presence case,” in *Proc. 6th Int. Conf. Learning Analytics & Knowledge (LAK’16)*, Edinburgh, UK, Apr. 25–29, 2016, pp. 15–24, doi: 10.1145/2883851.2883950.

[57] R. Ferreira, V. Kovanović, D. Gašević, and V. Rolim, “Towards combined network and text analytics of student discourse in online discussions,” in *Int. Conf. Artificial Intelligence Education*. London, UK: Springer, Jun. 27–30, 2018, pp. 111–126, doi: 10.1007/978-3-319-93843-1_9.

[58] M. Y. Doo and C. J. Bonk, “The effects of self-efficacy, self-regulation and social presence on learning engagement in a large university class using flipped learning,” *J. Comput. Assisted Learn.*, vol. 36, no. 6, pp. 997–1010, doi: 10.1111/jcal.12455.

[59] O. Poquet *et al.*, “Social presence in massive open online courses,” *Int. Rev. Res. Open & Distrib. Learn.*, vol. 19, no. 3, pp. 63–68, 2018, doi: 10.19173/irrodl.v19i3.3370.

[60] M. Yamada, K. Kaneko, and Y. Goda, “Social presence visualizer: Development of the collaboration facilitation module on CSCL,” in *Int. Conf. Collaboration Technologies*, Kanazawa, Japan, Sep. 14–16, 2016, pp. 174–189, doi: 10.1007/978-981-10-2618-8_14.

Máverick André received a Master’s degree in applied computing from the Federal Rural University of Pernambuco (UFRPE), Recife, Brazil. He is part of the AIBox Lab (https://aiboxlab.org/) at UFRPE, where he has been involved in research on the application of text mining techniques in virtual learning environments. His recent work applies natural language processing to examine student collaboration in online discussions.

Rafael Ferreira Mello holds a permanent faculty position at the Federal Rural University of Pernambuco, Recife, Brazil, where he is one of the coordinators of the AIBox Lab. Dr. Mello has worked on several multinational research projects involving institutional and organizational partners in Europe, Australia, and Latin America. He has coauthored 84 publications, including one book, four book chapters, 18 journal articles, and 61 refereed conference papers. A key theme of his recent research has been the use of natural language processing in applied fields like education, and for analyzing the content of a range of documents through text-summarization and topic-modeling algorithms.

André Nascimento holds a permanent faculty position at the Federal Rural University of Pernambuco, Recife, Brazil, where he is one of the coordinators of the AIBox Lab. Dr. Nascimento’s main areas of expertise include supervised machine learning, recommendation systems, and text mining. He has coauthored 18 publications, including book chapters, journal articles, and conference papers.

Rafael Dueire Lins has been a full professor at Federal University of Pernambuco, Recife, Brazil, since 2010, and at the Federal Rural University of Pernambuco, Recife, Brazil, since 2016. Prof. Lins coauthored the bestselling book *Garbage Collection: Algorithms for Dynamic Memory Management* (John Wiley Sons, 1996; also published in Chinese by ChinaPub in 2004). His pioneering contributions encompass the creation of the lambda calculus with explicit substitution, and the first general and efficient solution to cyclic reference counting in sequential, parallel, and distributed architectures. He has published 45 articles in refereed journals and over 200 papers in international conference proceedings. Prof. Lins served as vice-chair of the International Association on Pattern Recognition’s Technical Committee 10 (Graphics Recognition) from 2011 to 2015 and as chair of the same committee from 2015 to 2017.

Dragan Gašević is a distinguished professor of learning analytics in the Faculty of Information Technology and director of the Centre for Learning Analytics at Monash University, Melbourne, Australia. He is also an honorary professor with the University of Edinburgh, Edinburgh, U.K., where he was previously a professor and the Sir Tim O’Shea Chair in Learning Analytics and Informatics in the Moray House School of Education and the School of Informatics. Additionally, he is a [POSITION] with the Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia. Prof. Gašević is one of the cofounders of the Society for Learning Analytics Research (SoLAR) and served as its president from 2015 to 2017. He had the pleasure of serving as a founding program chair of SoLAR’s International Conference on Learning Analytics Knowledge (LAK) in 2011 and 2012, general chair of LAK in 2016, founding program chair of the Learning Analytics Summer Institute (LASI) in 2013 and 2014, and a founding editor of the *Journal of Learning Analytics*. He is currently an editor-in-chief of *Computers Education: Artificial Intelligence*, an associate editor of the IEEE transactions on learning technologies, and serves on the editorial boards of several other journals. A computer scientist by formal education, Prof. Gašević considers himself a learning analyst who develops computational methods that can shape next-generation learning technologies and advance our understanding of self-regulated and collaborative learning. He is a frequent keynote speaker and a (co)author of numerous research papers and books.

APÊNDICE C – CLASSIFICAÇÃO AUTOMÁTICA DA PRESENÇA SOCIAL EM DISCUSSÕES ONLINE ESCRITAS EM PORTUGUÊS

IX Congresso Brasileiro de Informática na Educação (CBIE 2020)
Anais do XXXI Simpósio Brasileiro de Informática na Educação (SBIE 2020)

Classificação Automática da Presença Social em Discussões Online Escritas em Português

Jean B. Teixeira[1], **Evandro de B. Costa**[1], **Rafel F. Mello**[2], **Máverick Ferreira**[3], **André N. Camara**[2]

[1]Instituto de Computação – Universidade Federal de Alagoas (UFAL)

[2]Centro de Informática - Universidade Federal Rural de Pernambuco (UFRPE)

[3]Centro de Informática – Universidade Federal de Pernambuco (UFPE)

{jbt,evandro}@ic.ufal.br, {rafel.mello,andre.camara}@ufrpe.br,

madf@cin.ufpe.br

Abstract. *This work presents a method that allows the automatic classification of messages exchanged in online distance learning forums written in Brazilian Portuguese according to categories (Affective, Interactive and Cohesive) of social presence. To achieve this goal, the adopted method uses a set of 116 resources extracted from text mining and word counting techniques, such as LIWC and Coh-Metrix. The classifier with the best performance presented 0,97 % and 0,95 % for precision and cohen kappa, respectively. This work also provides an analysis of the nature of social presence, looking at the most relevant classification characteristics to distinguish the three categories of social presence.*

Resumo. *Este trabalho apresenta um método que permite a classificação automática das mensagens trocadas em fóruns online de ensino a distância escritas em português brasileiro de acordo com as categorias (Afetiva, Interativa e Coesiva) da presença social. Para atingir esse objetivo, o método proposto faz uso de um conjunto de 116 características extraídas de técnicas de mineração de texto e contagem de palavras como o LIWC e Coh-Metrix. O classificador com melhor desempenho obteve 0,97% e 0,95% para acurácia e cohen kappa, respectivamente. Este trabalho também fornece uma análise da natureza da presença social, observando as características de classificação que foram mais relevantes para distinguir cada uma das três categorias.*

1. Introdução

O Ambiente Virtual de Aprendizagem (AVA) é geralmente utilizado como um mecanismo para facilitar a interação entre professores/tutores e estudantes em cursos de educação *online*. Os AVAs apresentam diversos recursos que proporcionam essa interação, dentre os quais, um dos mais populares é o fórum de discussão [Soares et al. 2016]. Os fóruns de discussão são ferramentas assíncronas que proporcionam uma enorme interatividade entre alunos e professores [Barros 2011], permitindo postagem de dúvidas, cometários sobre o conteúdo da disciplina, postagens de materiais extras, entre outros. Uma pesquisa em aprendizagem *online* e educação a distância aponta que o envolvimento em fóruns de discussão assíncronos acarreta em uma melhora dos resultados acadêmicos

DOI: 10.5753/cbie.sbie.2020.942

[Suhang et al. 2014]. Portanto, analisar as interações entre os alunos e professores torna-se bastante relevante para o processo pedagógico [Garrison et al. 1999].

Na perspectiva de interação mencionada, o presente estudo foca no modelo da Comunidade de Investigação (do inglês *Community of Inquiry (CoI)*. CoI é um dos modelos pedagógicos desenvolvido para apoiar os professores/tutores no desenvolvimento de experiências de aprendizagem *online* moderna [Garrison et al. 1999]. O CoI estabelece três elementos, conhecidos como presenças, que modelam o aprendizado *online* dos alunos, quais sejam: presença social, presença de ensino e presença cognitiva. Dentre elas, a presença social demonstra ser um elemento importante para o sucesso da experiência educacional [Garrison et al. 1999] e também é considerada relevante na observação da forma com que os alunos se lançam nas interações e em sua manutenção nos cursos a distância [Palloff and Pratt 2004].

Dado o contexto descrito anteriormente, este trabalho aborda o problema da identificação automática das categorias de presença social, em mensagens de fóruns educacionais escritas em português brasileiro, propondo uma solução via mineração de texto para desenvolver três classificadores. A principal contribuição é a utilização de características linguísticas, LIWC e Coh-Metrix, que serão detalhadas na Seção 3.4, além de apontar quais características são as mais relevantes para a predição das categorias. Os resultados são analisados sob a perspectiva de aprendizagem colaborativa na área de CSCL (do inglês *Computer-Supported Collaborative Learning*).

2. Fundamentação Teórica e Trabalhos Relacionados

2.1. O Modelo de Comunidade de Investigação (CoI)

O Modelo de Comunidade de Investigação (CoI), é um modelo conceitual proposto por [Garrison et al. 1999] com foco no processo social de construção conjunta e colaborativa do conhecimento em ambientes de comunicação assíncrona baseada em texto. Ele é bastante utilizado para orientar a pesquisa e a prática da aprendizagem *online*, na qual uma comunidade de investigação é constituída por meio de três dimensões, também conhecidas como presenças, essenciais para que haja uma experiência educacional de sucesso. Segundo [Garrison et al. 1999], são elas: (i) A presença social que mede a capacidade de humanizar os relacionamentos entre os participantes de uma discussão; (ii) A presença cognitiva que esta fortemente relacionada ao desenvolvimento dos resultados da aprendizagem; (iii) A presença de ensino que refere-se ao papel do professor antes (isto é, da concepção do curso) e durante o curso.

Este trabalho tem como foco na dimensão da presença social, definida por [Garrison et al. 1999] como a capacidade dos participantes de uma comunidade de investigação de se projetarem social e emocionalmente, como pessoas reais, ou seja, sua personalidade completa, através do meio de comunicação em uso. A presença social, uma das três dimensões do modelo CoI, possui três categorias: (i) Afetiva: categoria que analisa a tradução de emoções reais no texto. Contempla emoções, sentimentos e expressões de humor; (ii) Interativa: tem como foco a interatividade das mensagens trocadas entre os participantes. Seu principal objetivo é melhorar a comunicação aberta entre os alunos; (iii) Coesão de grupo: investiga o senso de união e compromisso de grupo entre os alunos.

2.2. Análises Automáticas do CoI

A Análise de Conteúdo Quantitativo (QCA), é apontada como um método largamente adotado no contexto das três presenças de CoI [Strijbos et al. 2006] com o objetivo de medir/avaliar os processos relacionados a construção do conhecimento nas discussões *online* e fornecer suposições válidas e confiáveis a partir da análise de dados textuais [Bauer 2007]. Em [Garrison et al. 2001], os autores definiram esquemas de codificação para analisar as presenças sociais e cognitivas, os quais têm sido amplamente adotados para a análise de conteúdo manual de CoI.

Em [Kovanovic et al. 2014], os autores utilizaram a codificação manual para avaliar a associação entre presença social e a posição na rede social. As primeiras propostas para automatizar a análise de conteúdo de acordo com os esquemas de codificação do modelo CoI baseavam-se principalmente em recursos tradicionais utilizados em mineração de texto, por exemplo contagem de palavras e frases. Um estudo mais recente utilizou recursos baseados no LIWC e Coh-Metrix para identificar fases da presença social em mensagens escritas em inglês no qual o melhor classificador atingiu 0,95 e 0,88 em acurácia e kappa, respectivamente [Ferreira et al. 2020].

Apesar de existirem estudos para extrair de forma automática as fases do desenvolvimento da presença cognitiva e categorias da presença social em inglês, até o momento não foram encontradas publicações que abordem a classificação automática da presença social em mensagens assíncronas de ambientes *online* de discussão em português.

3. Metodologia

3.1. Questões de Pesquisa

De acordo com o que foi abordado na Seção 2.1, a presença social tem sua importância no CoI pois influencia o desenvolvimento da presença cognitiva nos ambientes virtuais de aprendizagem. Dessa forma, a primeira pergunta de pesquisa deste trabalho é:

Questão de Pesquisa 1 (QP01): *Até que ponto os métodos de mineração de texto podem classificar automaticamente as mensagens de discussão online de acordo com as categorias da presença social?*

Além dessa questão citada acima, pretende-se também disponibilizar informações sobre quais as características que são mais relevantes para classificar cada uma das três categorias. Para isso, foram utilizados alguns parâmetros aplicados por [Kovanović et al. 2016], [Neto et al. 2018] e [Ferreira et al. 2020] com essa mesma finalidade. Então, a segunda questão de pesquisa é:

Questão de Pesquisa 2 (QP02): *Quais características melhor preveem cada categoria da presença social?*

3.2. Descrição do *Corpus*

O *corpus* utilizado foi gerado através de mensagens trocadas em fóruns de discussão de um curso de graduação de Biologia, oferecido totalmente *online* por uma universidade pública brasileira. Foram extraídas 1.500 mensagens produzidas por 215 alunos durante quatro semanas de curso. O objetivo do fórum era de promover discussões sobre um tema proposto pelo professor, em que a participação representava 20% da nota final do curso. Basicamente as discussões foram do tipo pergunta e resposta, ou seja, os fóruns eram iniciados com uma pergunta pelo professor e os alunos deveriam responder deixando suas contribuições.

Dois codificadores anotaram o conjunto de dados levando em consideração os 12 indicadores da presença social assim como foi feito em [Kovanovic et al. 2014]. Cada mensagem do conjunto de dados para cada indicador recebeu o valor “um” (possui o indicador) ou o valor “zero” (não possui o indicador). Assim como em [Kovanovic et al. 2014], três indicadores (Continuar uma conversa, Expressar apreço/concordância e Vocativos) foram removidos pois continham um grande número de mensagens.

Por fim, como o objetivo neste trabalho foi construir classificadores binários para cada categoria da presença social, as categorias foram compostas pelos indicadores anotados. Para que uma mensagem seja classificada como positiva (1), ela deve ter ao menos um indicador da respectiva categoria anotado com o valor “um”. Por exemplo, se uma mensagem continha os indicadores A1 = 0, A2 = 0 e A3= 1 , então ela era considerada positiva para a categoria afetiva.

3.3. Preparação dos Dados de Treinamento e Teste

De acordo com a revisão sistemática da literatura apresentada em [Ferreira-Mello et al. 2019], pode-se constatar que os estudos que se concentram na classificação de texto aplicam algoritmos de aprendizagem de máquina em um conjunto de dados divididos em dois subconjuntos, treinamento e teste, em que o subconjunto de treinamento previamente rotulado é utilizado para gerar um modelo que seja capaz de prever casos futuros, ou seja, prever exemplos da base de teste dos quais os rótulos são desconhecidos. Com base nisso o conjunto de dados foi subdividido em 75% para treinamento e 25% para teste conforme a Tabela 1.

Tabela 1. Distribuição das mensagens entre os grupos de treinamento e teste

Categoria	Grupo	Classe negativa (0)	Classe positiva (1)	Total
Afetiva	Treino	1088 (97%)	37 (3%)	1125
	Teste	367 (98%)	8 (2%)	375
	Total	**1455**	**45**	**1500**
Interativa	Treino	486 (43%	639 (57%)	1125
	Teste	169 (45%	206 (55%)	375
	Total	**655**	**845**	**1500**
Coesiva	Treino	988 (88%)	137 (12%)	1125
	Teste	341 (91%)	34 (9%)	375
	Total	**1329**	**171**	**1500**

3.4. Extração de Características

Além das características tradicionais de mineração de texto neste trabalho utilizou-se as ferramentas linguísticas LIWC e Coh-Metrix para extrair indicativos da presença social nos textos. Para esse trabalho foram utilizadas um total de 116 características extraídas através de ferramentas apresentadas a seguir.

Linguistic Inquiry and Word Count (LIWC): É uma ferramenta desenvolvida por [Pennebaker et al. 2001] com o intuito de fornecer um método eficiente para estudos sobre fatores emocionais, psicológicos, cognitivos entre outros, presentes em trechos de falas verbais e escritas de indivíduos. Em sua versão original em inglês o dicionário possui aproximadamente 6.540 palavras e cada uma associada a uma ou mais categorias dentre as 73 disponíveis. Neste trabalho foi utilizada a versão em português que possui cerca de 127.149 palavras que estão assinaladas a uma ou mais das 64 categorias (social, afetiva, concordância dentre outras) disponíveis [Balage Filho et al. 2013]. Ao relacionar a definição de presença social bem como seus indicadores, hipoteticamente sua utilização pode contribuir para o desenvolvimento dos classificadores capazes de diferenciar de forma correta mensagens com ou sem evidência da presença social. Em [Ferreira et al. 2020], os autores utilizaram a versão em inglês da ferramenta.

Coh-Metrix: Para este trabalho foi utilizada a versão da ferramenta para o português, o Coh-Metrix-PT [1], que possui 48 medidas implementadas de nível léxico, sintático em nível de sintagmas nominais, semântico, e discursivo [Scarton et al. 2010]. Sendo assim, supomos que o índice de coesão e complexidade textual proposto na ferramenta pode relacionar-se com a existência/ausência de indicadores da presença social.

Características de Contexto da Discussão: Com o objetivo de incorporar mais informações de contexto ao conjunto de características neste trabalho, foram incluídas quatro características de contexto utilizadas em [Kovanović et al. 2016]. São elas: (i) Número de respostas, (ii) Profundidade da mensagem, (iii) Similaridade de Cosseno para a mensagem anterior e (iv) Similaridade de Cosseno para a mensagem seguinte.

Frequência de palavras Também foi utilizada uma técnica tradicional de mineração de texto conhecida como *Bag of Words* (BoW), que é uma representação do texto através de uma matriz composta pela quantidade de ocorrência de cada palavra. Ao utilizar essa técnica é comum ocorrer um problema de alta dimensionalidade do vetor devido ao vasto vocabulário presente no texto. Para resolver esse problema e diminuir a dimensionalidade da matriz de ocorrência de palavras foram utilizadas algumas técnicas para limpar o texto removendo URL's, normalizando o texto, removendo *stopwords* que são palavras de pouca importância no texto. Por fim, também foi utilizada uma técnica que visa a redução das palavras aos seus radicais [Orengo and Huyck 2001], por exemplo, as palavras "concordamos" e "concordo" se tornam "concord".

3.5. Pré-processamento dos Dados

Conforme apresentado na Tabela 1, as classes negativas e positivas em todas as categorias apresentam desbalanceamento. Segundo [He and Garcia 2009], lidar com conjuntos de dados que possuem distribuições de classe desequilibradas é um grande desafio pois geralmente as classes majoritárias são priorizadas pelos indutores. Por exemplo, no conjunto de dados utilizados neste trabalho, a categoria afetiva possui 3% de ocorrências da classe positiva e 97% da classe negativa, sugerindo que o classificador pode priorizar a classe negativa. Para resolver esse problema utilizou-se o algoritmo SMOTE que permite a criação de dados artificiais da classe minoritária (*oversampling*) [Chawla et al. 2002].

[1] http://143.107.183.175:22680/

3.6. Seleção e Avaliação do Modelo

Existem diversos algoritmos de aprendizado de máquina para construção de modelos supervisionados. Em [Fernández-Delgado et al. 2014] foi realizada uma análise comparativa utilizando-se de 179 algoritmos de classificação de propósito geral em 121 conjuntos de dados diferentes. Nesse estudo, o algoritmo *Random Forest* foi um dos que apresentaram melhor desempenho. Este trabalho utilizou esse algoritmo não apenas pelo seu ótimo desempenho mas também por ser um algoritmo caixa branca podendo assim identificar quais as características que mais influenciaram na classificação das categorias da presença social. A medida mais utilizada para realizar essa avaliação da importância das características é o *Mean Decrease Gini* (MDG), que explica a separabilidade de uma determinada característica em relação as categorias [Breiman 2001]

Para o algoritmo *Random Forest*, foram estabelecidos dois parâmetros: (i) *n_estimators*: o número de árvores geradas pelo algoritmo; e (ii) *max_features*: o número de características aleatórias selecionadas por cada árvore. Os valores para cada um deles foram baseados na etapa de otimização de parâmetros realizada por [Ferreira et al. 2020] onde constatou-se que os valores de acurácia e kappa se estabilizavam ao utilizar em *max_features* e *n_estimator*, os valores 2.000 e 800 respectivamente.

Nesta etapa, foram utilizados os recursos da biblioteca em *python* denominada *scikit-learn* pois ela possui uma série de funções para serem aplicadas no pré-processamento, treinamento e validação de algoritmos de aprendizado de máquina que supriram as necessidades neste trabalho.

4. Resultados e Discussões

4.1. Modelo de Treinamento e Avaliação - QP01

Como mencionado anteriormente, os valores dos dois parâmetros (*max_features* e *n_estimator*) utilizados no classificador *Random Forest* foram baseados no trabalho de [Ferreira et al. 2020] onde encontrou-se bons valores para eles. Sendo assim foram utilizados o conjunto de dados de treinamento (1.125 mensagens) para gerar um classificador binário de cada categoria, e sua capacidade de generalização foi verificada no conjunto de teste (375 mensagens), os resultados para cada categoria podem ser observados na Tabela 2.

Tabela 2. Resultados dos classificadores por categoria

Categoria	Acurácia	Kappa
Afetiva	0,9786	0,1931
Interativa	0,976	0,9516
Coesiva	0,9786	0,8706

A Tabela 3 mostra a matriz de confusão gerada para cada categoria. É possível notar que as ocorrências mais altas de falsos positivos foram na classe Afetiva com 7 exemplos em um universo de 8 instâncias positivas enquanto as demais categorias tiveram menos de 2% de falsos positivos. Vale ressaltar que a categoria afetiva no conjunto de dados de teste possuía a menor quantidade de instâncias positivas, tornando difícil para o classificador aprender efetivamente como reconhecer as mensagens da categoria.

Tabela 3. Matriz de confusão por categoria

	Afetiva		**Interativa**		**Coesiva**	
	neg*	pos*	neg*	pos*	neg*	pos*
neg*	366	1	167	2	337	4
pos*	7	1	7	199	4	30

*pos = classe postiva e neg = classe negativa

4.2. Análise das Características Importantes - QP02

Este trabalho também analisou as contribuições das diversas características para o desempenho final do classificador. Apesar de terem sido utilizados os mesmos vetores de características para discriminar as classes (positivas e negativas) das três categorias, cada classificador possui um conjunto de variáveis diferentes consideradas como mais importantes. O algoritmo *Random Forest* usa a medida do índice de impureza média de diminuição de gini (MDG) para definir o grau de relevância de uma característica. As Tabelas 4, 5 e 6 apresentam as 15 variáveis mais importantes para o classificador de cada categoria (Afetiva, Interativa e Coesiva).

Para a categoria afetiva o conjunto das variáveis mais importantes apresentadas na Tabela 4 contém dez variáveis de frequência de palavras, quatro LIWC e uma Coh-Metrix. As duas mais importantes foram a palavra “ser” e a variável cm.DESSC, atingindo valores de MDG 12,91 e 9,33 respectivamente.

Tabela 4. Quinze variáveis mais importantes para a categoria afetiva

Variável	**Descrição**	**MDG**
ser	Frequência de palavras	12,91
cm.DESSC	Contagem de frases, número de frases	9,33
liwc.see	Número de palavras que fazem referência a visão	4,47
ter	Frequência de palavras	2,90
morr	Frequência de palavras	2,40
acredt	Frequência de palavras	2,24
brasil	Frequência de palavras	1,94
nov	Frequência de palavras	1,79
def	Frequência de palavras	1,79
vinic	Frequência de palavras	1,71
liwc.friend	Número de palavras que fazem referência a amizade	1,53
profes	Frequência de palavras	1,50
liwc.i	Primeira pessoa do singular	1,42
liwc.hear	Número de palavras que fazem referência a audição	1,29
individu	Frequência de palavras	1,22

Para a categoria interativa, conforme a Tabela 5, temos um conjunto com sete variáveis LIWC, seis frequência de palavras, uma Coh-Metrix e uma do contexto de discussão. A mais importante foi a posição da mensagem dentro da discussão (message.depth) com MDG de 64,88. Também vale destacar que a variável liwc.wd6letters (palavras com mais de 6 letras) também atingiu uma pontuação considerável de 10,58.

Tabela 5. Quinze variáveis mais importantes para a categoria interativa

Variável	Descrição	MDG
message.depth	Posição da mensagem dentro da discussão	64,88
liwc.wd6letters	Palavras com mais de 6 letras	10,58
cm.DESWC	Contagem de palavras, número de palavras	6,94
liwc.words	Quantidade de palavras	3,52
liwc.incl	Número de palavras que fazem referência a inclusão	1,15
liwc.article	Número de artigos	0,9
liwc.preps	Número de preposições	0,56
celul	Frequência de palavras	0,45
liwc.funct	Quantidade de palavras funcionais	0,42
opc	Frequência de palavras	0,35
liwc.cogmech	Número de palavras que fazem referência a cognição	0,33
fer	Frequência de palavras	0,33
dign	Frequência de palavras	0,22
boa	Frequência de palavras	0,20
permit	Frequência de palavras	0,19

Tabela 6. Quinze variáveis mais importantes para a categoria coesiva

Variável	Descrição	MDG
boa	Frequência de palavras	35,99
noit	Frequência de palavras	14,99
ola	Frequência de palavras	8,71
cm.DESPC	Número de parágrafos	4,99
bom	Frequência de palavras	4,00
tarde	Frequência de palavras	3,78
dia	Frequência de palavras	2,29
message.depth	Posição da mensagem dentro da discussão	1,78
obrig	Frequência de palavras	1,70
cm.DESSC	Número de sentenças	1,47
m.DESPL	Número médio de frases em cada parágrafo do texto	1,42
liwc.wd6letters	Palavras com mais de 6 letras	1,08
liwc.swear	Número de palavras que fazem referência a xingamento	0,78
liwc.cogmech	Número de palavras que fazem referência a cognição	0,70
cm.DESSL	Duração da frase, número de palavras, média	0,69

Finalmente, a Tabela 6, aponta as principais variáveis da categoria coesiva, nas quais sete são frequência de palavras,quatro Coh-Metrix, três LIWC, e uma do contexto de discussão. Percebe-se que palavras comumente utilizadas para cumprimentar / saudar obtiveram boas notas ('boa' - MDG de 35,99, 'noit' - MDG de 14,99 e 'ola' - MDG de 8,71). Dessa forma, as variáveis mais importantes apresentadas nas tabelas acima, estão alinhadas com a teoria da presença social proposta por [Garrison et al. 1999], por exemplo, podemos destacar mensagens que possuem: i) palavras que expressam emoções positivas; ii) quantidade de pronomes em segunda pessoa; iii) palavras que expressam concordância; iv) palavras que fazem referência a inclusão; v) palavras que indicam saudações.

5. Conclusão

Este trabalho abordou o problema da identificação automática das categorias de presença social em mensagens de fóruns educacionais, escritas em português. Neste sentido, podemos destacar duas contribuições: a modelagem do problema sob a ótica de aprendizagem de máquina supervisionada, no qual foram propostos três classificadores binários, um para cada categoria da presença social: afetiva, interativa e coesiva. Para isso utilizou-se o algoritmo *Random Forest* e os recursos linguísticos LIWC, Coh-Metrix, contexto da discussão e frequência de palavras (BoW). Com exceção da categorias afetiva, as demais atingiram os valores de *Cohen's kappa* acima de 0,80, que é um acordo entre avaliadores muito bom. A outra contribuição foi a identificação das quinze variáveis mais importantes para o classificador de cada categoria. Importante destacar que com a utilização desses classificadores, os professores/tutores podem identificar o nível da presença social do grupo ou de cada aluno e realizar uma intervenção no decorrer do curso para melhorá-lo e consequentemente auxiliar no processo de ensino-aprendizagem.

Vale ressaltar que a pesquisa apresentou algumas limitações, podemos destacar o problema com o tamanho pequeno da base de dados utilizadas e as categorias desbalanceadas, apesar de que essa situação reflita com as encontradas na literatura, podem afetar o desempenho do classificador. Para trabalhos futuros pretende-se utilizar uma amostra maior de dados composta por diferentes domínios, bem como realizar uma etapa de otimização buscando ajustar os parâmetros utilizados para melhorar os resultados obtidos principalmente na categoria afetiva. Também está sendo desenvolvida uma versão mais completa da ferramenta Coh-Metrix para o português brasileiro.

Referências

Balage Filho, P., Pardo, T. A. S., and Aluísio, S. (2013). An evaluation of the brazilian portuguese liwc dictionary for sentiment analysis. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*.

Barros, Maria das Graças e Carvalho, A. B. G. (2011). As concepções de interatividade nos ambientes virtuais de aprendizagem. *Campina Grande: EDUEPB*.

Bauer, M. W. (2007). Content analysis. an introduction to its methodology–by klaus krippendorff from words to numbers. narrative, data and social science–by roberto franzosi. *The British Journal of Sociology*, 58(2):329–331.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, 15(1):3133–3181.

Ferreira, M., Rolim, V., Mello, R. F., Lins, R. D., Chen, G., and Gašević, D. (2020). Towards automatic content analysis of social presence in transcripts of online discussions. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, pages 141–150.

Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., and Romero, C. (2019). Text mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(6):e1332.

Garrison, D. R., Anderson, T., and Archer, W. (1999). Critical inquiry in a text-based environment: Computer conferencing in higher education. *The internet and higher education*, 2(2-3):87–105.

Garrison, D. R., Anderson, T., and Archer, W. (2001). Critical thinking, cognitive presence, and computer conferencing in distance education. *American Journal of distance education*, 15(1):7–23.

He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284.

Kovanovic, V., Joksimovic, S., Gasevic, D., and Hatala, M. (2014). What is the source of social capital? the association between social network position and social presence in communities of inquiry. In *Proceedings of the Workshops held at Educational Data Mining 2014 co-located with 7th International Conference on Educational Data Mining (EDM 2014)*. Citeseer.

Kovanović, V., Joksimović, S., Waters, Z., Gašević, D., Kitto, K., Hatala, M., and Siemens, G. (2016). Towards automated content analysis of discussion transcripts: A cognitive presence case. In *Proceedings of the sixth international conference on learning analytics & knowledge*, pages 15–24.

Neto, V., Rolim, V., Ferreira, R., Kovanović, V., Gašević, D., Lins, R. D., and Lins, R. (2018). Automated analysis of cognitive presence in online discussions written in portuguese. In *European conference on technology enhanced learning*, pages 245–261. Springer.

Orengo, V. M. and Huyck, C. R. (2001). A stemming algorithmm for the portuguese language. In *spire*, volume 8, pages 186–193.

Palloff, R. M. and Pratt, K. (2004). *O aluno virtual-um guia para trabalhar com estudantes on-line*. Penso Editora.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

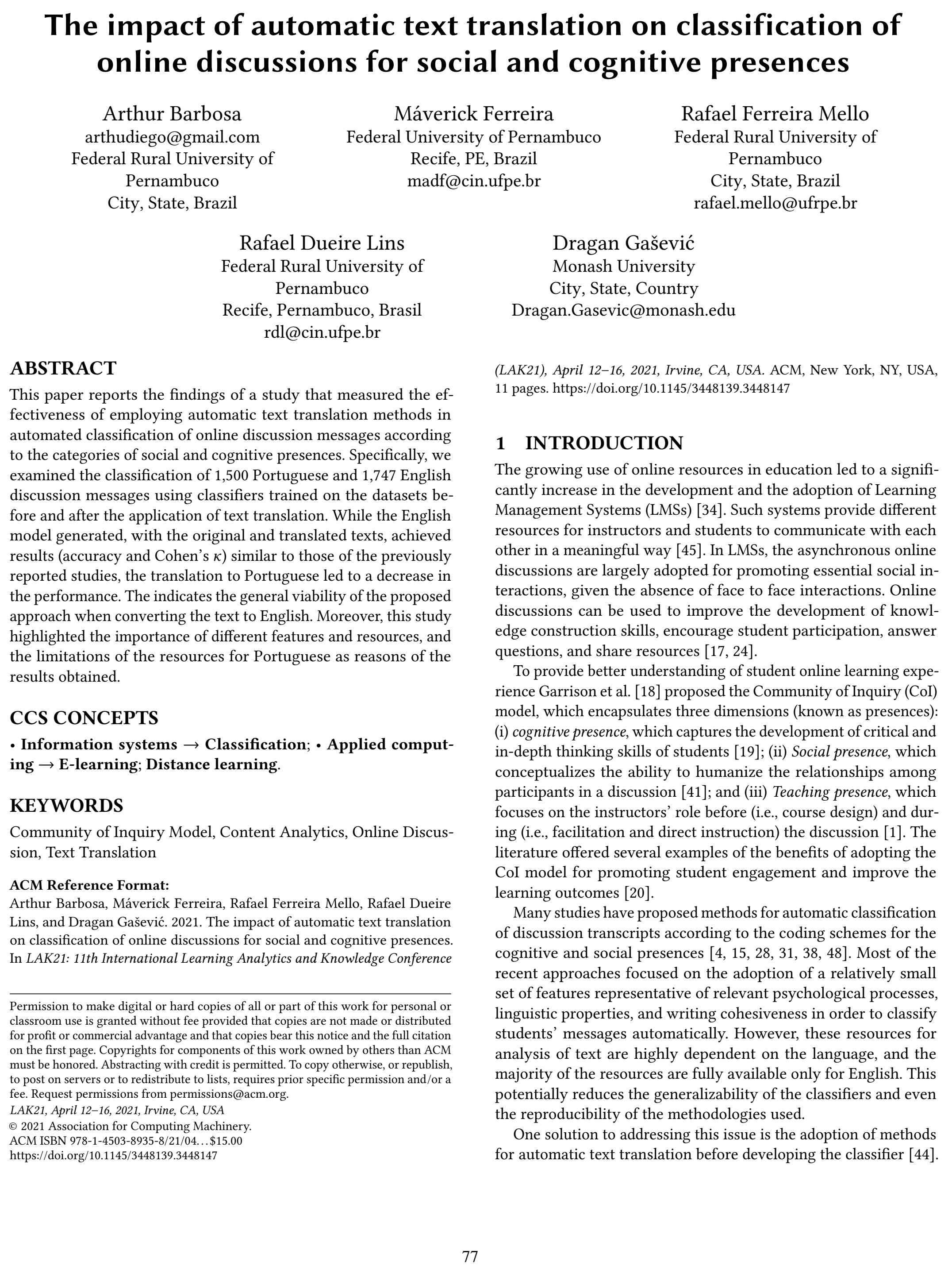
Scarton, C., Gasperin, C., and Aluisio, S. (2010). Revisiting the readability assessment of texts in portuguese. In *Ibero-American Conference on Artificial Intelligence*, pages 306–315. Springer.

Soares, F. B. M., Machado, C. J. R., Diniz, D., and Maciel, A. M. A. (2016). Educational data mining to support distance learning students with difficulties in the portuguese grammar. In *Anais do XXVII Simpósio Brasileiro de Informática na Educação (SBIE 2016)*, pages 956–965, Brasil.

Strijbos, J.-W., Martens, R. L., Prins, F. J., and Jochems, W. M. (2006). Content analysis: What are they talking about? *Computers & Education*, 46(1):29–48.

Suhang, J., Williams, A., Schenke, K., Warschauer, M., and Odowd, D. (2014). Predicting mooc performance with week 1 behavior. *Educational Data Mining*.

APÊNDICE D – THE IMPACT OF AUTOMATIC TEXT TRANSLATION ON CLASSIFICATION OF ONLINE DISCUSSIONS FOR SOCIAL AND COGNITIVA PRESENTES



The impact of automatic text translation on classification of online discussions for social and cognitive presences

Arthur Barbosa
arthudiego@gmail.com
Federal Rural University of Pernambuco
City, State, Brazil

Máverick Ferreira
Federal University of Pernambuco
Recife, PE, Brazil
madf@cin.ufpe.br

Rafael Ferreira Mello
Federal Rural University of Pernambuco
City, State, Brazil
rafael.mello@ufrpe.br

Rafael Dueire Lins
Federal Rural University of Pernambuco
Recife, Pernambuco, Brasil
rdl@cin.ufpe.br

Dragan Gašević
Monash University
City, State, Country
Dragan.Gasevic@monash.edu

ABSTRACT

This paper reports the findings of a study that measured the effectiveness of employing automatic text translation methods in automated classification of online discussion messages according to the categories of social and cognitive presences. Specifically, we examined the classification of 1,500 Portuguese and 1,747 English discussion messages using classifiers trained on the datasets before and after the application of text translation. While the English model generated, with the original and translated texts, achieved results (accuracy and Cohen's κ) similar to those of the previously reported studies, the translation to Portuguese led to a decrease in the performance. The indicates the general viability of the proposed approach when converting the text to English. Moreover, this study highlighted the importance of different features and resources, and the limitations of the resources for Portuguese as reasons of the results obtained.

CCS CONCEPTS

• **Information systems → Classification**; • **Applied computing → E-learning**; **Distance learning**.

KEYWORDS

Community of Inquiry Model, Content Analytics, Online Discussion, Text Translation

ACM Reference Format:
Arthur Barbosa, Máverick Ferreira, Rafael Ferreira Mello, Rafael Dueire Lins, and Dragan Gašević. 2021. The impact of automatic text translation on classification of online discussions for social and cognitive presences. In *LAK21: 11th International Learning Analytics and Knowledge Conference (LAK21), April 12–16, 2021, Irvine, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3448139.3448147

1 INTRODUCTION

The growing use of online resources in education led to a significantly increase in the development and the adoption of Learning Management Systems (LMSs) [34]. Such systems provide different resources for instructors and students to communicate with each other in a meaningful way [45]. In LMSs, the asynchronous online discussions are largely adopted for promoting essential social interactions, given the absence of face to face interactions. Online discussions can be used to improve the development of knowledge construction skills, encourage student participation, answer questions, and share resources [17, 24].

To provide better understanding of student online learning experience Garrison et al. [18] proposed the Community of Inquiry (CoI) model, which encapsulates three dimensions (known as presences): (i) *cognitive presence*, which captures the development of critical and in-depth thinking skills of students [19]; (ii) *Social presence*, which conceptualizes the ability to humanize the relationships among participants in a discussion [41]; and (iii) *Teaching presence*, which focuses on the instructors' role before (i.e., course design) and during (i.e., facilitation and direct instruction) the discussion [1]. The literature offered several examples of the benefits of adopting the CoI model for promoting student engagement and improve the learning outcomes [20].

Many studies have proposed methods for automatic classification of discussion transcripts according to the coding schemes for the cognitive and social presences [4, 15, 28, 31, 38, 48]. Most of the recent approaches focused on the adoption of a relatively small set of features representative of relevant psychological processes, linguistic properties, and writing cohesiveness in order to classify students' messages automatically. However, these resources for analysis of text are highly dependent on the language, and the majority of the resources are fully available only for English. This potentially reduces the generalizability of the classifiers and even the reproducibility of the methodologies used.

One solution to addressing this issue is the adoption of methods for automatic text translation before developing the classifier [44].

LAK21, April 12–16, 2021, Irvine, CA, USA
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8935-8/21/04…$15.00
https://doi.org/10.1145/3448139.3448147

This approach has been adopted in applications using standard bag-of-words features [49] and word counts based on psychological processes[12]. Given the use of psychologically sound measures of text for human cognition and linguistic indicators of text complexity and coherence, English-based features may perform comparably well on non-English datasets, such as for example, dual-language analysis [5, 26].

This paper investigates the feasibility of applying techniques for automatic text translation to overcome the limitation of the resources used for text analysis in automatic classification of the CoI constructs in languages other than English. This study also examines how the different psychological and linguistic features impact on the classification of the original and the translated texts. To the best of our knowledge, no previous work had evaluated the potential of using text translation to overcome the limitation of the linguistic resources for different languages. The results obtained revealed that using English as the pivot language before the categorization of cognitive and social presences improves the quality of the classifiers developed.

2 THEORETICAL BACKGROUND

2.1 The Community of Inquiry Model (CoI)

The Community of Inquiry (CoI) model aims to describe the essential facets of social interactions and knowledge construction in online and blended education [20]. CoI proposes three dimensions applied to analyze these interactions. Among them, social and cognitive presences explain how educational experience is happening during online discussions [18].

Social presence is *"the ability of participants in a community of inquiry to project themselves socially and emotionally, as 'real' people (i.e., their full personality), through the medium of communication being used"* [18, p. 94]. As shown in Table 1, social presence comprises several indicators divided into the following categories: (i) **Affective**: analyzes the translation of real emotions into text; (ii) **Interactive**: focuses on the interactivity of the messages exchanged among participants; (iii) **Group Cohesion**: investigates the sense of union and group commitment among students. In general, the interpretation of this presentation is provided based on these categories and indicators.

Cognitive presence focuses on *"the extent to which the participants in any particular configuration of a community of inquiry are able to construct meaning through sustained communication"* [18, p. 89]. There are four phases defined by this presence: (i) The **triggering event** phase starts the cycle of critical inquiry with a problem statement, challenge or dilemma usually introduced by the instructor; (ii) The **exploration** phase includes exploration of ideas, reflection, brainstorming, and exchange of findings, information, and ideas based on the initial provocation. The main goal is to explore the nature of the problem; (iii) The **integration** phase is characterized by the connection of relevant ideas, information, and findings in other to construct meaning from the previous phases; and (iv) **Resolution** is the phase in which the participants employ the newly-constructed knowledge to real-word applications and hypothesis testing. The output of this last phase could trigger another inquiry cycle.

2.2 Analysis of the Cognitive and Social Presences

The literature reports two main approaches to analyzing the CoI presences. The first one is using a questionnaire developed for this goal [11, 50]. Arbaugh et al. [2] present one of the most popular questionnaires, which examines the perception of students about their experience of online interactions using five-point Likert scales (1 = strongly disagree to 5 = strongly agree). The main limitation of this approach is the fact that it can be used only after the interactions are completed and it is typically applied after the end of a course .

To overcome this limitations, many studies adopt a *Quantitative Content Analysis (QCA)* method to assess the CoI presences in the analysis of online discussion transcripts [6]. QCA benefits from predetermined coding schemes to code (manually and automatically) the discussion messages of the students according to the categories of the three presences [19, 43]. As the only input of this approach is student message, it could be used to analyze the discussion in real-time with the ultimate goal of using the CoI model to drive instructional interventions and enhance student learning outcomes [30].

Within the automatic methods for QCA, the initial studies adopted machine learning algorithms that were trained on traditional word and phrase counts features as input for different classifiers. For instance, Mcklin [35] developed a neural network classifier based on word frequency features to categorize online discussion messages according to the phrases of cognitive presence. The classifier reached accuracy of 0.31 of Cohen's κ. Kovanović et al. [31] applied a combination of bag-of-words (n-gram) and Part-of-Speech (POS) N-gram features in combination with a Support Vector Machines (SVMs) classifier, obtaining accuracy of 0.41 for Cohen's κ.

Although approaches that used word/phrase counts for text analysis reached good values of accuracy, they had two main limitations: (i) the use of word counts, in general, implicates that classification models are extremely dependent on the domain (i.e., the issue of limited generalizability is pronounced); and (ii) the application black-box algorithms (such as neural network and SVM) do not provide information on why a specific class was chosen for each message. Therefore, the recent literature proposes the adoption of a combination of (i) textual features reflective of psychological processes, writing cohesiveness and discussion structure, and (ii) decision tree classifiers [7], which allow, due to their white box nature, for the analysis of the influence of the different features on the final classification results. Kovanović et al. [28] and Neto et al. [38] adopted random forest classifiers and features based on Coh-metrix [37], LIWC [46], latent semantic analysis (LSA)-based similarity, named entities, and discussion context [48], to identify phases of cognitive presence for messages written in English (0.63 Cohen's κ) and Portuguese (0.72 Cohen's κ). Ferreira et al. [15] proposed a similar approach, using the same features and classifiers as used in [28, 38], to develop three binary classifiers for automatic coding of a discussion message based on the three categories of social presence. These classifiers reached 0.49, 0.83, and 0.88 of κ for the affective, interactive, and cohesive categories of social presence, respectively. Finally, Barbosa et al. [4] proposed a cross-language classifier for cognitive presence with the use of the random forest

Table 1: Indicators social presence [43].

Category	Indicator	Label
Affective	1. Expression of emotions	Emotions
	2. Use of humor	Humor
	3. Self-disclosure	Self_disclosure
Interactive	4. Continuing a thread	Cont_Thread
	5. Quoting from others' messages	Quoting_Mess
	6. Referring explicitly to others' message	Referring_Mess
	7. Asking questions	Asking_Q
	8. Complimenting, expressing appreciation	Complimenting
	9. Expressing agreement	Agreement
Cohesive	10. Vocatives	Vocatives
	11. Addresses of refers to the group using inclusive pronouns	Group
	12. Phatics, salutations	Salutations

algorithm and by making use of the the same features applied all previously mentioned studies. The main difference from the previous work is the optimization method used for balancing the dataset. The final result reached κ of 0.53.

This section has summarized different studies that coded discussion messages according to the phases and categories of cognitive and social presence, respectively and these studies coded discussions even in different languages. However, to our knowledge, no publication evaluated the limitation of the linguistic resources for different languages and the impact of automatic text translation with the goal of automatic coding of online discussion messages according to cognitive and social presence.

3 RESEARCH QUESTIONS

As presented in Section 2.2, there are several methods for automatic coding of online discussion messages according to the categories of social presence (affective, interactive, and cohesive) and the phases of cognitive presence (triggering event, exploration, integration, and resolution) for different languages. However, some languages do not have the linguistic resources necessary to apply the methodology proposed by the state-of-the-art classifiers. Hence, our first research question is:

> **Research Question 1 (RQ1):**
> *What is the effectiveness of adopting methods for automatic text translation for different languages in the process of automatic classification of online discussion messages for social and cognitive presence?*

Specifically, research question 1 aimed to analyze the extent to which accuracy of automatic classification can be achieved when text translation methods are used. In addition to the evaluation of the classifier accuracy, we were also interested in providing additional information on the most relevant features for the original and translated texts. This analysis could provide insights into the impact of the limitation of the linguistic resources for languages different from English. Therefore, the second research question was:

> **Research Question 2 (RQ2):**
> *Which features do best predict social and cognitive presences for original and translated online transcript messages?*

Research question 2 aimed to better understand the extent to which translation methods could overcome the problem of limited resources for non-English languages.

4 METHOD

4.1 Data Description

This study made use of the datasets from the previous two studies that investigated the classification process of discussion messages for the levels of cognitive presence. The English and Portuguese language datasets from the studies by Kovanović et al. [29] and Neto et al. [39] were used. The work described in reference Barbosa et al. [5] also applied the same dataset to propose a multi-language classifier. Some details on both datasets follow.

4.1.1 English dataset. The English dataset contained 1,747 messages extracted from six offerings of a master level research-intensive course in software engineering offered entirely online at a Canadian public university. During the six offerings of the course (winter 2008, fall 2008, summer 2009, fall 2009, winter 2010, winter 2011), 81 students produced 1,747 messages. The students recorded and engaged in asynchronous online discussions about video presentations about research papers related to one of the course topics (i.e., software requirements, design, and maintenance). This activity accounted for 15% of the course final mark [for further details on the course design, see 21].

The annotation process for both cognitive and social presence followed a similar approach. Using the coding scheme proposed by Garrison et al. [19], two expert coders annotated all online discussion messages according to the indicators and phases of social and cognitive presence, respectively. In terms of social presence, the messages were categorized using “one” (have the indicator) or value "zero" (not have the indicator) At the end of the annotations, the coders reached 84% of agreement. In terms of cognitive presence, the inter-rater agreement between the coders was excellent (Percent agreement=98.1% and κ=0.974), with only 32 disagreements. In both cases, the differences were resolved through discussion among coders.

Finally, following [16, 27], the annotations regarding some of the social presence indicators (Continuing a topic, Praises, and Vocatives) were removed because almost every message contained them. Moreover, our analysis focused on the category level, not on indicator one. Thus, this paper assigned the final label if the message has at least one indicator related to the category. Tables 2 and 3 shows the final distribution of messages according to social and cognitive presences, respectively.

4.1.2 Portuguese dataset. The Portuguese dataset encompassed 1,500 discussion messages produced by 215 students. It was generated during four weeks of a single offering of an entirely online undergraduate level-biology course offered at a Brazilian public university. In this case, the online forum focused on the discussion about a question/topic raised by the instructor, where the participation accounted for 20% of the final course mark.

The annotation processed followed the same approach used in the English dataset. Two coders categorized each message based on indicators of the social presence indicators and the phases of cognitive presence. The social presence categorization used the same binary approach for indicators which led to the final numbers of category instances (affective, interactive, and cohesive) . The coders reached an agreement percentage of 90.41%. In terms of cognitive presence, the coders obtained a high inter-rater agreement (Percent agreement=91.4% and κ=0.86). For both cases, a third coder resolved the disagreements. Tables 2 and 3 present the final distribution of the Portuguese data in terms of social and cognitive presences, respectively.

4.2 Automatic Text Translation

The main objective of this research was to verify the impact of text translation on the construction of classifiers for social and cognitive presences of the CoI model. This analysis is important due to the lack of resources for languages such as Portuguese. Previous work had used the translation approach to improving different natural language processing applications such as text classification [44], text summarization [8], and information extraction [25]. In our case, a text translation technique was applied to overcome the imitated availability of the linguistic resources for different languages. This paper adopted the widely used Google translate API [1] available for the Python language. Therefore, we translated the datasets described in section 4.1 from English to Portuguese and vice-versa. In the reminder of the paper, these datasets are referenced as the *source* (original dataset) and the *translation* (text obtained after the use of the Google translate API) datasets.

4.3 Features Extraction

Following the same approach as reported in the related literature [4, 28], we adopted several language-independent features to train our automatic text classifiers for social and cognitive presences. This study explores features based on linguistic resources (LIWC and Coh-Metrix) and structural information (SNA and Discussion Context Features). The following sections provide an overview of these features.

4.3.1 LIWC features. The Linguistic Inquiry Word Count (LIWC) is a dictionary that allows the extraction of grammatical, psychological, and social characteristics from a textual document [46]. It has been widely used in previous studies focused on the construction of classifiers associated with the CoI [15, 28, 38] model. This study benefited from the LIWC 2015 [3] and LIWC 2007 [9] versions for English and Portuguese, respectively. It is important to mention that the English version had 29 features that were not available for Portuguese. For instance, features related to Cognitive Processes and Perpetual Processes are not included in the Portuguese dictionary.

4.3.2 Coh-Metrix features. In addition to the indicators implemented by LIWC, we also evaluated the usability of the indicators provided by the Coh-Metrix computational linguistics tool [37]. Coh-Metrix allows for the extraction of features associated with [23, 37]: text cohesion (argument overlap), linguistic complexity (based on syntactic tree structures), text readability (Flesch reading ease), and lexical diversity (type-token ratio). In this study, we adopted the recent web tool that contains 94 features for each language.

4.3.3 Discussion context features (DCF). According to [20] the discussion structure could indicate aspects of the social and cognitive presences. For instance, given that the cognitive presence is developed over time through discourse and reflection, integration and resolution phases tend to happen in later messages. Such an information can be captured by those features. We made use of indicators that capture the context and structure of online discussions, as initially proposed by Kovanović et al. [28]: i) *Depth of the message (dcf.depth)*, refers to the numerical position of the message within the discussion; ii) *Number of replies (dcf.numReplies)* , consists of the count of responses that each message in the database received; iii) *Cosine similarity with the previous/next message (dcf.sim.prev and dcf.sim.next)*, refers to the cosine similarity of the text of each message with the message that precedes and succeeds it; iv) *Start and end the discussion (dcf.first and dcf.last)*, a binary value that indicates whether a message is starting or ending a discussion.

4.3.4 Social Network Analysis (SNA). Social network analysis (SNA) has been widely used in research that seeks to measure the level of social interaction among participants in interaction networks such as online discussions [22, 33, 40]. SNA allows for identifying which participants are most influential and/or intermediaries within an interaction network. Therefore, the last classification features in our list were the most commonly considered SNA measures: i) *Closeness centrality (sna.closeness)* assesses how close a student in the network is to all other students. According to [51], students with a greater degree of proximity tend to be effective in disseminating information over the network; ii) *Betweenness centrality (sna.betweenness)* investigates the importance of a student's intervention in interactions among other students. Therefore, a high level of intermediation indicates the student's leadership among peers; *Degree centrality (sna.degree)* seeks to analyze the level of student interaction with other students present in the network. Thus, a higher degree of centrality of a student means that the student is more influential in the network.

[1] https://pypi.org/project/googletrans/

Table 2: Distribution of social presence for both dataset.

Category	English data					Portuguese data				
	Positive	%	Negative	%	Total	Positive	%	Negative	%	Total
Affective	530	33.33%	1,217	66.67%	1,747	45	3.0%	1,455	97.0%	1,500
Interactive	1,030	58.95%	741	41.05%	1,747	845	56.33%	655	43.67%	1,500
Cohesive	1,326	75.90%	421	24.10%	1,747	171	11.40%	1,329	88.60%	1,500

Table 3: Distribution of cognitive presence for both dataset

		English data		Portuguese data	
ID	Phase	Messages	%	Messages	%
0	*Other*	140	8.01%	196	13.07%
1	Triggering event	308	17.63%	235	15.67%
2	Exploration	684	39.15%	871	58.07%
3	Integration	508	29.08%	154	10.27%
4	Resolution	107	6.13%	44	2.92%
	Total	1,747	100.00%	1,500	100.00%

4.4 Data processing and evaluation

Previous studies had addressed several issues in the automatic classification of online discussion messages according to social and cognitive presences of the CoI model. Particularly, data imbalance [5] and contamination in the evaluation [13] are highlighted. The current study did not concentrate on any of these problems, but we considered the recommendations suggested by Farrow et al. [13] to perform an adequate analysis. As the goal of this research is the investigation of the impact of automatic text translation in the automatic classification of online discussion messages for social and cognitive presences, we did not evaluated different classification algorithms and their parameters.

Following the method proposed by Kovanović et al. [28], and its extensive use in the literature [4, 13, 15, 38], we adopted the Random Forest algorithm [7] to implement our classifiers. Another reason for making such a decision is the fact that Random Forest usually reaches good performance when applied to different domains Fernández-Delgado et al. [14]. The Random Forest algorithm is an ensemble classification method which internally involves the creation of a large number (hundreds or even thousands) of decision trees and then selecting a majority vote for its final output. Each tree is built on a bootstrap sample of the original training data (i.e., a sample of the same size as the original data, with repetition) and evaluated on the data that did not enter the bootstrap sample (approximately one-third of the training dataset). In this manner, random forests can achieve a low variance of their prediction without increasing their bias Breiman [7]. Moreover, following the studies by Kovanović et al. [28] and Ferreira et al. [15], we trained three binary classifiers for each social presence category and one multi-class classifier for the cognitive presence problem.

For the data separation in training and test sets, we adopted a stratified 10-fold cross-validation technique taking into consideration the recommendations proposed by [13]. Four different scenarios were evaluated:

- English Source (SourceEN): Experiments using the original English data;
- Portuguese Translation (TranslationPT): Experiments using the original English data translated to Portuguese;
- English Source (SourcePT): Experiments using the original Portuguese data;
- English Translation (TranslationEN): Experiments using the original Portuguese data translated to English;

To evaluate the classifiers, we adopted accuracy and Cohen's $kappa$[10] because they are widely used in learning analytics and educational data mining and in the related works [27, 38]. For each case, the classifier was performed 30 times in order to create different samples to run statistical tests in order to answer *research question 1*. Then, we applied the U Mann-Whitney test [36] to check if the translation resulted in significantly different results.

Finally, to answer *research question 2*, we presented the top-20 features for each scenario/classifiers according to the Mean Decrease Gini (MDG), which is based on the reduction in Gini impurity measure [7]. We also compared the outcome of MDG and the resources available for each language.

5 RESULTS

5.1 Research Question 1

Initially, we evaluated the influence of text translation in the final classification of social and cognitive presences. Table 4 presents the average accuracy and Cohen's κ results obtained, from cycles of 30 executions, by each one of the proposed classifiers. The results revealed that the classifiers reached better results for English texts, even when the translation was applied.

In the first analysis (translation from English to Portuguese) it is possible to see that the preformance reduced drastically for social presence decreasing for 57%, 64% and 53% in terms of κ for affective, interactive and cohesive categories, respectively. The cognitive presence retained similar outcomes achieving 0.38 of κ before and after the translation. On the other hand, the translation from Portuguese to English led to improvements in all classifier. For instance, the cohesive category increased by 120%, in terms of κ, after the translation. The interactive category had the smallest gain among the classifiers, i.e., κ increased only by 1%.

Finally, Table **??** presents the results of the application of Mann-Whitney U test for each classifier in order to compare the quality of the translation results. For social presence, the tests showed that the differences among the classifiers were significant (α=0.05), which means that the classifiers built using the English data performed better than the Portuguese one, even for the data that was originally written in Portuguese. In terms of cognitive presence, the test revealed that the translation from Portuguese to English improved the results while the classifier transcribed from English to Portuguese maintained the same results.

5.2 Research Question 2

This study adopted the same feature set to create all the four classifiers, following the same methodology as used in the related literature [5, 16]. However, each classifier considered different variables as the most important. We ranked the features based on the Mean Decrease Gini impurity index (MDG) measure. Tables 6, 7, 8, and 9 present the top-20 features for the classifier of each social presence categories (Affective, Interactive and Cohesive) and cognitive presence phases [2]. The items in boldface highlight the features that appeared before and after the translation for each database.

The most important set of variables for the Affective category were based on LIWC and Coh-Metrix, followed by four SNA features and one DCF feature. It is noteworthy that liwc.i (number of first noun pronouns) and sna.closeness were the most relevant features for English and Portuguese, respectively, before and after the translation. Moreover, while both English classifiers highlighted the importance of punctuation, the two Portuguese classifiers emphasized the Coh-Metrix word information category (e.g., cm.WRDFAMc – rating of how familiar a word seems to an adult, cm.WRDMEAc – meaningfulness rating of a word, cm.WRDAOAc – the age of acquisition norms from MRC[3], cm.WRDIMGc – how easy it is to construct a mental image of the word is also provided in the merged ratings of the MRC, and cm.WRDCNCc – how concrete or non-abstract a word is).

In contrast to the Affective category, the most predictive features of the Interactive category included SNA and DCF features. This resulted in a large intersection of the features between the original and translated classifiers for the English data, as shown in Table 7. On the other hand, the classifiers trained on the Portuguese data did not reach the same level of correspondence among the features, but the most predominant feature was the sna.closeness achieving the MDG value five times higher than those of the other features.

Similar to the Affective category, the classifier for the Cohesive category of social presence showed (see Table 8) the dominance of LIWC and Coh-Metrix features with only four SNA and one DCF measures again. However, the most relevant information in this table is the importance of the LIWC features. While they accounted for 45% of the English classifiers, they were not relevant for the Portuguese ones, which were dominated by the Coh-Metrix features (95%). Also, the sna.closeness was an important feature in all cases.

Table 9 reports the most relevant features used for automatic classification of cognitive presence. As the nature of the model was different from those used for socal presence (multi-class instead of binary classifiers), the number of features per resources were more balanced. The higher concordance in the English data happened because of the non-linguistic features. Moreover, liwc.QMark (the number of question marks) remained relevant for all classifiers, reaching the first or second place in terms of MDG.

Finally, this study also intended to evaluate the importance of the resources applied in the proposed classifiers. Tables 10 and 11 show the most commonly used resources by each classifier in the top-20 features according to the MDG. The most significant change was the importance of LIWC and Coh-Metrix before and after the translation. In both cases, the LIWC indexes obtained higher relevance for the original text. Another discrepancy is the importance of the DCF features when comparing the models built on the English data in relation to the Portuguese one, where there were no occurrences. SNA kept similar importance across all cases.

6 DISCUSSION

6.1 RQ1: Effects of automatic text translation

To address research question 1, we evaluated several random forest classifiers trained based on original and translated transcripts of online discussions. The accuracy performance of the original Portuguese data improved when the classifiers were trained on the dataset translated to English. For instance, the cognitive presence classifier, using the English translation, reached 0.83 and 0.69 in terms of accuracy and Cohen's κ, respectively. These results are similar to the classifier proposed by Neto et al. [39] 0.83 (accuracy) and 0.72(κ). Our result is even more relevant because Farrow et al. [13] demonstrated that the results of Neto et al. [38] could have been contaminated in the data oversampling process. It is important to stress that we followed the suggestions by Farrow et al. [13] in order to avoid the problem of data contamination.

The analysis of the translation from English to Portuguese resulted in a loss of performance, except for the cognitive presence classifier which reached similar results. This outcome was expected as there were more linguistic resources available for English than for Portuguese. Therefore, the translation resulted in a decrease in the number of features analyzed. Even with this decline, the results were comparable to the previous work [13, 15] without the application of oversampling techniques. Moreover, this work introduced SNA measures as features for classification of social and cognitive presences. The classifiers for social presence trained in this study using the original English data showed a rise in the performance comparing to the state-of-the-art approach [15].

Finally, our approach outperformed the algorithm proposed by Barbosa et al. [4] in terms of cognitive presence classification. Barbosa et al. [4] proposed a cross-language classifier where the initial

Table 4: Results obtained for social and cognitive presence.

	Affective		Interactive		Cohesive		Cognitive	
Dataset	Accuracy	κ	Accuracy	κ	Accuracy	κ	Accuracy	κ
SourceEN	0.72 (0.00)	0.26 (0.01)	0.88 (0.00)	0.75 (0.01)	0.87 (0.00)	0.60 (0.01)	0.58 (0.00)	0.38 (0.01)
TranslationPT	0.68 (0.00)	0.15 (0.01)	0.75 (0.00)	0.48 (0.01)	0.79 (0.00)	0.32 (0.01)	0.57 (0.00)	0.38 (0.00)
SourcePT	0.90 (0.00)	0.40 (0.02)	0.95 (0.00)	0.91 (0.00)	0.85 (0.00)	0.20 (0.02)	0.73 (0.00)	0.52 (0.01)
TranslationEN	0.91 (0.00)	0.49 (0.01)	0.96 (0.00)	0.92 (0.00)	0.88 (0.00)	0.44 (0.01)	0.83 (0.00)	0.69 (0.00)

[2] For more details about the features description see section 4.3 (for sna and dcf), http://cohmetrix.com/ (for coh-metrix) and https://liwc.wpengine.com/compare-versions/ (for LIWC)

[3] https://websites.psychology.uwa.edu.au/school/MRCDatabase/uwa_mrc.htm

Table 5: Features selected by resources - Portuguese Data

	SourcePT				TranslationEN			
	LIWC	**Coh-metrix**	**SNA**	**DCF**	**LIWC**	**Coh-metrix**	**SNA**	**DCF**
Affective	13	5	2	0	3	16	1	0
Interactive	10	9	1	0	6	11	3	0
Cohesive	8	11	1	0	0	19	1	0
Cognitive	11	8	1	0	9	10	1	0
Total	42	37	5	0	18	56	6	0

Table 6: Twenty most important variables for the affective category according to MDG

English Data				Portuguese Data			
SourceEN		**TranslationPT**		**SourcePT**		**TranslationEN**	
Features	**MDG**	**Features**	**MDG**	**Features**	**MDG**	**Features**	**MDG**
liwc.i	5.62	**liwc.i**	2.43	**sna.closeness**	10.92	**sna.closeness**	11.76
liwc.WC	2.97	cm.SYNMEDpos	2.03	cm.DESWLlt	2.86	liwc.period	9.10
cm.DESWC	1.92	cm.WRDFAMc	1.87	liwc.relativ	2.56	**cm.WRDFAMc**	2.55
liwc.Exclam	1.56	cm.LDMTLDa	1.75	cm.WRDIMGc	2.36	**cm.WRDCNCc**	2.54
cm.WRDPRP1s	1.52	cm.WRDMEAc	1.63	**cm.WRDFAMc**	2.30	liwc.comma	2.09
cm.LDVOCD	1.35	cm.SYNLE	1.55	cm.WRDMEAc	2.29	liwc.WPS	2.08
liwc.Parenth	0.89	**cm.LDVOCD**	1.52	cm.DESWLltd	2.09	liwc.friend	1.95
liwc.insight	0.87	sna.closeness	1.52	**cm.DESWLsyd**	2.03	liwc.allPunc	1.86
cm.DESWLlt	0.83	liwc.present	1.49	**cm.RDFKGL**	1.90	liwc.informal	1.57
liwc.Apostro	0.80	cm.WRDAOAc	1.45	cm.SMCAUSwn	1.82	liwc.exclam	1.25
liwc.Authentic	0.78	cm.WRDIMGc	1.38	cm.WRDADJ	1.81	liwc.affiliation	1.23
dcf.sim.prev	0.77	cm.WRDFRQmc	1.38	cm.WRDVERB	1.77	**cm.DESWLsyd**	1.20
liwc.negate	0.74	cm.WRDCNCc	1.29	cm.DESWC	1.77	liwc.present	1.20
cm.WRDHYPv	0.73	liwc.relativ	1.26	liwc.cogmech	1.75	**cm.RDFKGL**	1.17
cm.SMCAUSwn	0.73	cm.RDFKGL	1.22	cm.WRDAOAc	1.67	sna.betweenness	1.01
liwc.time	0.71	cm.DESSLd	1.22	cm.LDMTLDa	1.51	liwc.function	0.99
liwc.we	0.71	cm.DESWLsyd	1.22	cm.DESWLsy	1.48	liwc.clout	0.95
cm.WRDHYPn	0.69	cm.WRDFRQa	1.19	liwc.funct	1.47	liwc.sixltr	0.93
cm.CNCTemp	0.68	cm.SMCAUSlsa	1.18	**cm.WRDCNCc**	1.47	liwc.adverb	0.93
liwc.article	0.68	cm.WRDFRQc	1.17	cm.WRDNOUN	1.37	cm.DESSL	0.93

model was trained on English data then applied to a Portuguese dataset, reaching 0.53 of Cohen's κ. In this study, we suggest the translation instead of the cross-language approach, which obtained 0.69 of Cohen's κ. Although the outcomes could not be directly compared, as they used different methodologies, these results indicate the relevance of the approach analyzed in this paper.

6.2 RQ2: Analysis of features and resources

Research question 2 focused on the interpretation of the feature importance for the different models generated. To achieve this goal, this study replicated the same methodology proposed by Kovanović et al. [28], which adopted the MDG measure to estimate the top-20 features for each classifier created. Although we adopted the same feature vectors to discriminate the classes (social and cognitive presences), each model considered different variables as the most important. Based on the results reported in Section 5.2 we can draw two conclusions: (i) the most important features, in general, were related to the linguistic resources (LIWC and Coh-Metrix); (ii) the LIWC resource had a higher dependency on the language as the importance of LIWC features decreased on the translated datasets.

The affective category classifiers trained using the original texts had a significantly higher number of features that belongs to the LIWC resource, following the same result reported by [15], which mainly contains word counts features in the top-15 most important features for this category. The classifiers for affective category based on the translations had a predominance of Coh-Metrix features. Moreover, the sna.closeness feature, never evaluated for classification of cognitive and social presences before, had a significant importance for the affective category. In both cases, the English and Portuguese data, the SNA features, except for sna.closeness, and DCF features were not selected in the top-20.

The Interactive category had the most significant discrepancies when compared to the previous work [15]. However, the discrepancies happened because the sets of features used were very different. While Ferreira et al. [15] adopted word frequencies features, they did not evaluate the SNA and DCF features, which were among the most relevant features in our analysis of the English data. This finding, importance of SNA and DCF, is fully aligned to the indicators of the interactive category [43], such as continuing a thread and referring explicitly to others' message. Therefore, the addition of these features represents a significant contribution to the literature.

Table 7: Twenty most important variables for the interactive category according to MDG

English Data				**Portuguese Data**			
SourceEN		**TranslationPT**		**SourcePT**		**TranslationEN**	
Features	**MDG**	**Features**	**MDG**	**Features**	**MDG**	**Features**	**MDG**
liwc.QMark	42.05	**dcf.sim.next**	8.15	**sna.closeness**	54.68	**sna.closeness**	47.54
liwc.assent	6.79	**sna.betweenness**	4.71	**cm.DESWC**	4.02	cm.WRDPRO	11.11
dcf.sim.next	3.46	**cm.DESWLltd**	4.29	liwc.funct	4.00	**cm.DESWC**	4.88
sna.betweenness	2.28	**dcf.depth**	3.82	liwc.cogmech	3.06	liwc.WC	3.91
liwc.AllPunc	1.41	**dcf.sim.prev**	3.44	cm.WRDNOUN	2.43	cm.WRDPRP1p	3.29
sna.degree	1.39	**sna.degree**	2.66	cm.WRDADJ	2.16	liwc.WPS	1.75
dcf.depth	1.20	cm.DESWLsyd	2.46	liwc.article	1.68	liwc.QMark	1.71
liwc.informal	1.18	cm.RDFKGL	1.38	liwc.incl	1.64	liwc.AllPunc	1.27
dcf.sim.prev	1.15	cm.DESSL	1.30	cm.LDVOCDa	1.31	cm.DRPVAL	1.16
liwc.dic	0.97	cm.SMCAUSlsa	1.28	**cm.LDTTRa**	1.07	liwc.Exclam	0.84
liwc.function	0.88	cm.SYNLE	1.27	cm.WRDVERB	0.94	liwc.you	0.80
liwc.otherP	0.73	liwc.humans	1.24	liwc.relativ	0.74	cm.CRFSOa	0.64
cm.DESWLltd	0.68	cm.DESWLsy	1.24	sna.degree	0.72	**cm.LDTTRa**	0.61
liwc.Period	0.55	**sna.closeness**	1.17	cm.WRDPRP2	0.69	cm.LDTTRc	0.56
dcf.numReplies	0.54	**cm.LSAGN**	1.16	liwc.preps	0.68	liwc.pronoun	0.44
liwc.auxverb	0.50	cm.WRDMEAc	1.12	cm.DESWLltd	0.57	liwc.ppron	0.42
sna.closeness	0.44	cm.DESSLd	1.09	cm.WRDMEAc	0.52	cm.LSAGNd	0.41
liwc.i	0.44	cm.CRFCWO1	1.04	sna.betweenness	0.48	cm.WRDPRP1s	0.38
cm.LSAGN	0.43	cm.CRFCWO1d	1.02	cm.CNCPos	0.48	liwc.insight	0.33
liwc.interrog	0.43	cm.LSASSpd	1.02	cm.DESWLsy	0.48	liwc.friend	0.32

Table 8: Twenty most important variables for the cohesive category according to MDG

English Data				**Portuguese Data**			
SourceEN		**TranslationPT**		**SourcePT**		**TranslationEN**	
Features	**MDG**	**Features**	**MDG**	**Features**	**MDG**	**Features**	**MDG**
liwc.affiliation	34.66	cm.SMCAUSlsa	9.70	**sna.closeness**	5.65	liwc.Exclam	10.28
liwc.social	4.16	cm.LSASSpd	2.28	**cm.LSAGNd**	4.93	**cm.DESSLd**	3.84
liwc.Apostro	2.20	**sna.closeness**	2.17	**cm.DESSLd**	2.77	liwc.reward	3.71
liwc.Dic	1.15	cm.WRDVERB	2.13	cm.SYNMEDpos	2.42	liwc.affiliation	3.20
liwc.Clout	0.82	cm.WRDFRQc	1.91	cm.CRFCWOa	2.27	**sna.closeness**	2.74
liwc.function	0.75	cm.WRDFRQa	1.87	cm.WRDFAMc	2.14	cm.LSASS1	2.08
cm.PCREFz	0.72	**cm.SYNLE**	1.79	cm.RDFKGL	2.05	**cm.LSAGNd**	1.72
liwc.drives	0.69	cm.DESWLlt	1.69	**cm.WRDIMGc**	1.99	cm.SYNLE	1.56
cm.PCREFp	0.63	cm.LSASS1	1.63	cm.RDFRE	1.98	liwc.drives	1.50
cm.SYNMEDpos	0.61	cm.DESWLltd	1.63	cm.SMCAUSlsa	1.95	liwc.leisure	1.15
liwc.certain	0.60	cm.WRDFRQmc	1.46	cm.WRDAOAc	1.94	cm.CNCAdd	1.11
sna.closeness	0.60	cm.LDMTLDa	1.41	cm.WRDMEAc	1.91	cm.LSASS1d	1.06
cm.WRDHYPn	0.60	cm.WRDAOAc	1.39	cm.DESWLltd	1.87	cm.LDMTLDa	1.04
cm.WRDNOUN	0.57	cm.CRFCWO1	1.30	cm.DESWLsyd	1.79	liwc.time	1.02
cm.SYNLE	0.55	cm.DESWLsy	1.29	cm.WRDCNCc	F 1.79	liwc.verb	F 1.01
cm.WRDPOLc	F 0.54	F **cm.SYNMEDpos**	F 1.26	F cm.LSAGN	F 1.76	F cm.DESWC	F 1.01
liwc.tentat	F 0.52	F cm.LSASSp	F 1.24	F cm.DESSL	F 1.64	F liwc.you	F 1.00
dcf.sim.prev	F 0.51	F cm.RDFKGL	F 1.23	F cm.DESWLsy	F 1.61	F cm.SYNSTRUTt	F 0.97
cm.PCTEMPz	F 0.51	F cm.DESWLsyd	F 1.18	F cm.DESWLlt	F 1.58	F cm.LSASSp	F 0.92
cm.SMCAUSr	F 0.49	F cm.SYNMEDwrd	F 1.16	F cm.SYNMEDlem	F 1.46	F **cm.WRDIMGc**	F 0.90

Table 9: Twenty most important variables for the cognitive presence according to MDG

English Data				Portuguese Data			
SourceEN		TrasnlationPT		SourcePT		TrasnlationEN	
Features	MDG	Features	MDG	Features	MDG	Features	MDG
cm.DESWC	5.69	**liwc.QMark**	3.04	liwc.funct	5.88	**liwc.QMark**	12.27
liwc.QMark	2.81	**sna.betweenness**	2.06	**liwc.QMark**	4.71	liwc.WPS	8.32
dcf.depth	2.02	liwc.funct	1.98	**sna.closeness**	4.69	**cm.DESWC**	5.24
cm.LDVOCD	1.75	**sna.closeness**	1.84	cm.WRDNOUN	3.32	cm.WRDPRO	4.58
cm.DESSL	1.49	liwc.space	1.78	**cm.DESWC**	3.05	liwc.you	3.68
dcf.first	1.46	**cm.DESWC**	1.77	liwc.preps	2.97	liwc.WC	2.89
sna.closeness	1.21	**dcf.depth**	1.68	liwc.relativ	2.35	**sna.closeness**	2.88
sna.betweenness	1.21	liwc.incl	1.36	liwc.incl	2.29	cm.DRPVAL	2.78
sna.degree	1.18	**dcf.sim.next**	1.33	liwc.social	1.74	liwc.interrog	2.24
liwc.i	1.06	cm.SMCAUSr	1.32	cm.WRDPRP2	1.64	cm.LDTTRa	1.78
liwc.you	0.90	**dcf.sim.prev**	1.31	**cm.DESSL**	1.60	liwc.Exclam	1.56
liwc.AllPunc	0.83	cm.DESSLd	1.27	cm.RDFKGL	1.56	cm.WRDPRP1p	1.49
liwc.focuspast	0.83	cm.SYNLE	1.23	liwc.pronoun	1.51	liwc.assent	1.23
liwc.cause	0.79	liwc.conj	1.21	liwc.article	1.50	**cm.DESSL**	1.16
dcf.sim.next	0.78	cm.WRDNOUN	1.20	cm.DESWLsyd	1.47	liwc.Period	1.09
cm.SMCAUSlsa	0.77	cm.DESWLsyd	1.14	cm.DESWLltd	1.45	liwc.Comma	1.01
liwc.posemo	0.77	liwc.preps	1.12	cm.DESWLsy	1.40	liwc.AllPunc	0.99
liwc.compare	0.75	liwc.social	1.08	cm.RDFRE	1.35	cm.WRDPRP1s	0.96
liwc.ppron	0.74	liwc.ipron	1.06	**cm.LDMTLDa**	1.30	**cm.LDMTLDa**	0.92
dcf.sim.prev	0.72	cm.DESWLsy	1.03	liwc.ipron	1.29	liwc.conj	0.75

Table 10: Features selected by resources - English Data

	SourceEN				TranslationPT			
	LIWC	Coh-metrix	SNA	DCF	LIWC	Coh-metrix	SNA	DCF
Affective	11	8	0	1	3	16	1	0
Interactive	11	2	3	4	1	13	3	3
Cohesive	9	9	1	1	0	19	1	0
Cognitive	9	4	3	0	8	7	2	0
Total	40	23	7	6	12	55	7	3

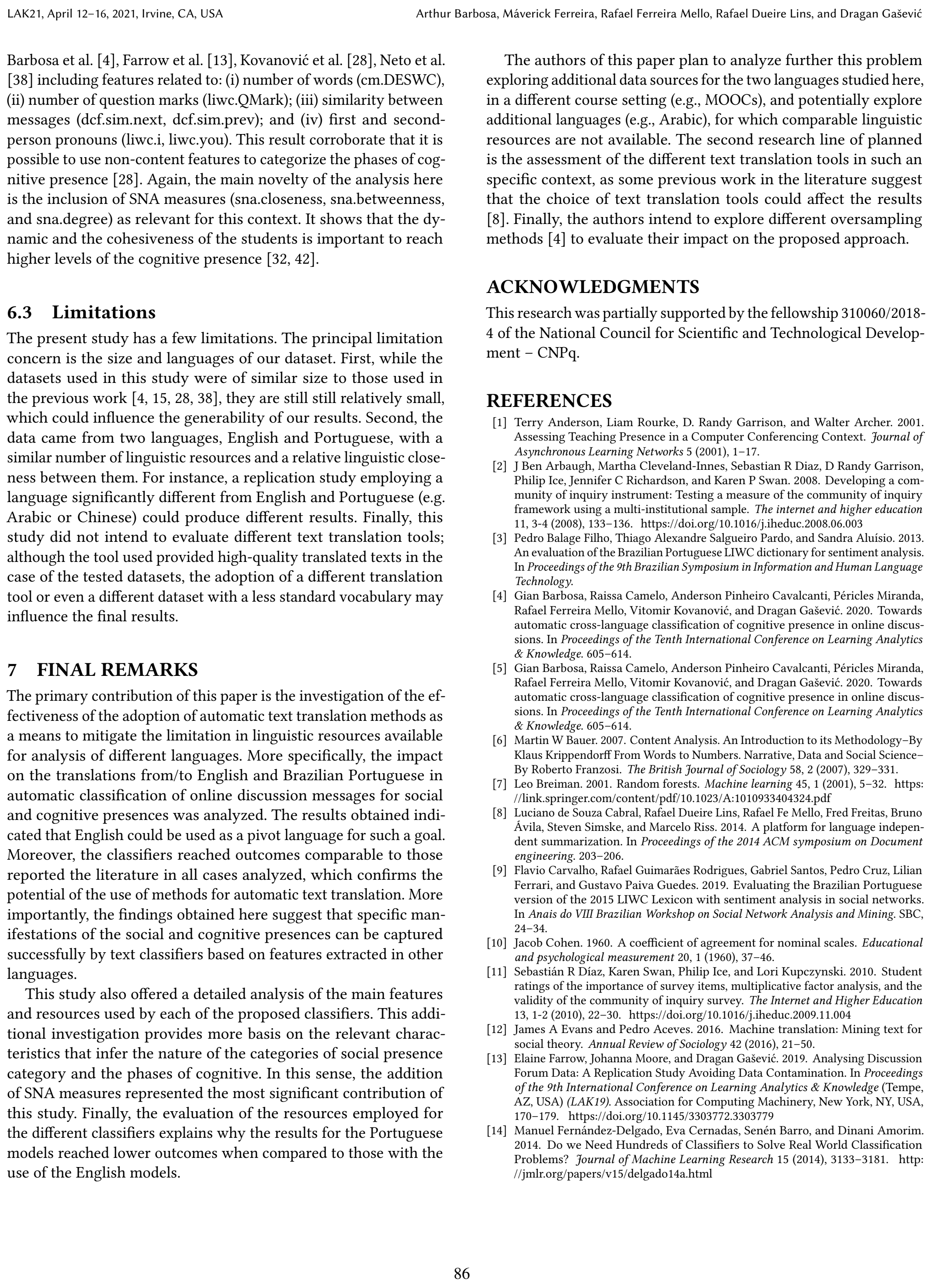
Table 11: Features selected by resources - Portuguese Data

	SourcePT				TranslationEN			
	LIWC	Coh-metrix	SNA	DCF	LIWC	Coh-metrix	SNA	DCF
Affective	13	5	2	0	3	16	1	0
Interactive	10	9	1	0	6	11	3	0
Cohesive	8	11	1	0	0	19	1	0
Cognitive	11	8	1	0	9	10	1	0
Total	42	37	5	0	18	56	6	0

In contrast to this, the results on the Portuguese data did not follow the same trend. This happened due to the nature of the online discussions – i.e., question-answer forum – based on which the Portuguese data was collected [38].

The presence of the LIWC features in Table 8 highlights the relevance of this resource for the identification of the cohesive category in the English data. For instance, the demonstration of salutations, that can be recognized by the characteristics as liwc.affiliation (number of affiliations) and liwc.social (number of social processes), is very indicative of this category Ferreira et al. [15]. On the other hand, the limited number of features of the LIWC resource for Portuguese, as mentioned in section 4.3.1, could explain the lack of these features in the top-20 and the difference between the performance of the English trained models and the Portuguese ones reported in table 4. Moreover, the analysis made here revealed the importance of the sna.closeness measure to demonstrate group cohesion. The results obtained indicated that a higher value of the sna.closeness could mean stronger ties among students which leads to a cohesive and participative network [47].

Finally, the cognitive presence classifier has shown a higher degree of similarity with the previously reported in the literature



Barbosa et al. [4], Farrow et al. [13], Kovanović et al. [28], Neto et al. [38] including features related to: (i) number of words (cm.DESWC), (ii) number of question marks (liwc.QMark); (iii) similarity between messages (dcf.sim.next, dcf.sim.prev); and (iv) first and second-person pronouns (liwc.i, liwc.you). This result corroborate that it is possible to use non-content features to categorize the phases of cognitive presence [28]. Again, the main novelty of the analysis here is the inclusion of SNA measures (sna.closeness, sna.betweenness, and sna.degree) as relevant for this context. It shows that the dynamic and the cohesiveness of the students is important to reach higher levels of the cognitive presence [32, 42].

6.3 Limitations

The present study has a few limitations. The principal limitation concern is the size and languages of our dataset. First, while the datasets used in this study were of similar size to those used in the previous work [4, 15, 28, 38], they are still still relatively small, which could influence the generability of our results. Second, the data came from two languages, English and Portuguese, with a similar number of linguistic resources and a relative linguistic closeness between them. For instance, a replication study employing a language significantly different from English and Portuguese (e.g. Arabic or Chinese) could produce different results. Finally, this study did not intend to evaluate different text translation tools; although the tool used provided high-quality translated texts in the case of the tested datasets, the adoption of a different translation tool or even a different dataset with a less standard vocabulary may influence the final results.

7 FINAL REMARKS

The primary contribution of this paper is the investigation of the effectiveness of the adoption of automatic text translation methods as a means to mitigate the limitation in linguistic resources available for analysis of different languages. More specifically, the impact on the translations from/to English and Brazilian Portuguese in automatic classification of online discussion messages for social and cognitive presences was analyzed. The results obtained indicated that English could be used as a pivot language for such a goal. Moreover, the classifiers reached outcomes comparable to those reported the literature in all cases analyzed, which confirms the potential of the use of methods for automatic text translation. More importantly, the findings obtained here suggest that specific manifestations of the social and cognitive presences can be captured successfully by text classifiers based on features extracted in other languages.

This study also offered a detailed analysis of the main features and resources used by each of the proposed classifiers. This additional investigation provides more basis on the relevant characteristics that infer the nature of the categories of social presence category and the phases of cognitive. In this sense, the addition of SNA measures represented the most significant contribution of this study. Finally, the evaluation of the resources employed for the different classifiers explains why the results for the Portuguese models reached lower outcomes when compared to those with the use of the English models.

The authors of this paper plan to analyze further this problem exploring additional data sources for the two languages studied here, in a different course setting (e.g., MOOCs), and potentially explore additional languages (e.g., Arabic), for which comparable linguistic resources are not available. The second research line of planned is the assessment of the different text translation tools in such an specific context, as some previous work in the literature suggest that the choice of text translation tools could affect the results [8]. Finally, the authors intend to explore different oversampling methods [4] to evaluate their impact on the proposed approach.

ACKNOWLEDGMENTS

This research was partially supported by the fellowship 310060/2018-4 of the National Council for Scientific and Technological Development – CNPq.

REFERENCES

[1] Terry Anderson, Liam Rourke, D. Randy Garrison, and Walter Archer. 2001. Assessing Teaching Presence in a Computer Conferencing Context. *Journal of Asynchronous Learning Networks* 5 (2001), 1–17.

[2] J Ben Arbaugh, Martha Cleveland-Innes, Sebastian R Diaz, D Randy Garrison, Philip Ice, Jennifer C Richardson, and Karen P Swan. 2008. Developing a community of inquiry instrument: Testing a measure of the community of inquiry framework using a multi-institutional sample. *The internet and higher education* 11, 3-4 (2008), 133–136. https://doi.org/10.1016/j.iheduc.2008.06.003

[3] Pedro Balage Filho, Thiago Alexandre Salgueiro Pardo, and Sandra Aluísio. 2013. An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology.*

[4] Gian Barbosa, Raissa Camelo, Anderson Pinheiro Cavalcanti, Péricles Miranda, Rafael Ferreira Mello, Vitomir Kovanović, and Dragan Gašević. 2020. Towards automatic cross-language classification of cognitive presence in online discussions. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge.* 605–614.

[5] Gian Barbosa, Raissa Camelo, Anderson Pinheiro Cavalcanti, Péricles Miranda, Rafael Ferreira Mello, Vitomir Kovanović, and Dragan Gašević. 2020. Towards automatic cross-language classification of cognitive presence in online discussions. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge.* 605–614.

[6] Martin W Bauer. 2007. Content Analysis. An Introduction to its Methodology–By Klaus Krippendorff From Words to Numbers. Narrative, Data and Social Science–By Roberto Franzosi. *The British Journal of Sociology* 58, 2 (2007), 329–331.

[7] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32. https://link.springer.com/content/pdf/10.1023/A:1010933404324.pdf

[8] Luciano de Souza Cabral, Rafael Dueire Lins, Rafael Fe Mello, Fred Freitas, Bruno Ávila, Steven Simske, and Marcelo Riss. 2014. A platform for language independent summarization. In *Proceedings of the 2014 ACM symposium on Document engineering.* 203–206.

[9] Flavio Carvalho, Rafael Guimarães Rodrigues, Gabriel Santos, Pedro Cruz, Lilian Ferrari, and Gustavo Paiva Guedes. 2019. Evaluating the Brazilian Portuguese version of the 2015 LIWC Lexicon with sentiment analysis in social networks. In *Anais do VIII Brazilian Workshop on Social Network Analysis and Mining.* SBC, 24–34.

[10] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[11] Sebastián R Díaz, Karen Swan, Philip Ice, and Lori Kupczynski. 2010. Student ratings of the importance of survey items, multiplicative factor analysis, and the validity of the community of inquiry survey. *The Internet and Higher Education* 13, 1-2 (2010), 22–30. https://doi.org/10.1016/j.iheduc.2009.11.004

[12] James A Evans and Pedro Aceves. 2016. Machine translation: Mining text for social theory. *Annual Review of Sociology* 42 (2016), 21–50.

[13] Elaine Farrow, Johanna Moore, and Dragan Gašević. 2019. Analysing Discussion Forum Data: A Replication Study Avoiding Data Contamination. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (Tempe, AZ, USA) *(LAK19).* Association for Computing Machinery, New York, NY, USA, 170–179. https://doi.org/10.1145/3303772.3303779

[14] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. 2014. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research* 15 (2014), 3133–3181. http://jmlr.org/papers/v15/delgado14a.html

[15] Máverick Ferreira, Vitor Rolim, Rafael Ferreira Mello, Rafael Dueire Lins, Guanliang Chen, and Dragan Gašević. 2020. Towards Automatic Content Analysis of Social Presence in Transcripts of Online Discussions. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. Association for Computing Machinery, New York, NY, USA, 141–150. https://doi.org/10.1145/3375462.3375495

[16] Máverick Ferreira, Vitor Rolim, Rafael Ferreira Mello, Rafael Dueire Lins, Guanliang Chen, and Dragan Gašević. 2020. Towards Automatic Content Analysis of Social Presence in Transcripts of Online Discussions. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. Association for Computing Machinery, New York, NY, USA, 141–150. https://doi.org/10.1145/3375462.3375495

[17] Rafael Ferreira-Mello, Máverick André, Anderson Pinheiro, Evandro Costa, and Cristobal Romero. 2019. Text mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (2019), e1332. https://doi.org/10.1002/widm.1332

[18] D.Randy Garrison, Terry Anderson, and Walter Archer. 1999. Critical Inquiry in a Text-Based Environment: Computer Conferencing in Higher Education. *The Internet and Higher Education* 2, 2 (1999), 87 – 105. https://doi.org/10.1016/S1096-7516(00)00016-6

[19] D. Randy Garrison, Terry Anderson, and Walter Archer. 2001. Critical thinking, cognitive presence, and computer conferencing in distance education. *American Journal of Distance Education* 15, 1 (2001), 7–23. https://doi.org/10.1080/08923640109527071

[20] D. Randy Garrison, Terry Anderson, and Walter Archer. 2010. The first decade of the community of inquiry framework: A retrospective. *The Internet and Higher Education* 13, 1 (2010), 5 – 9. https://doi.org/10.1016/j.iheduc.2009.10.003

[21] Dragan Gašević, Olusola Adesope, Srećko Joksimović, and Vitomir Kovanović. 2015. Externally-facilitated regulation scaffolding and role assignment to develop cognitive presence in asynchronous online discussions. *The internet and higher education* 24 (2015), 53–65.

[22] Dragan Gašević, Srećko Joksimović, Brendan R Eagan, and David Williamson Shaffer. 2018. SENS: Network analytics to combine social and cognitive perspectives of collaborative learning. *Computers in Human Behavior* (2018). https://doi.org/10.1016/j.chb.2018.07.003

[23] Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers* 36, 2 (2004), 193–202.

[24] Khe Foon Hew and Wing Sum Cheung. 2008. Attracting student participation in asynchronous online discussions: A case study of peer facilitation. *Computers & Education* 51, 3 (2008), 1111–1124.

[25] Emna Hkiri, Souheyl Mallat, and Mounir Zrigui. 2017. Arabic-English text translation leveraging hybrid NER. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*. 124–131.

[26] Srećko Joksimović, Areti Manataki, Dragan Gašević, Shane Dawson, Vitomir Kovanović, and Inés Friss de Kereki. 2016. Translating network position into performance: Importance of centrality in different network configurations. In *Proceedings of the 6th International Conference on Learning Analytics & Knowledge (LAK '16)*. ACM, New York, USA, 314–323. https://doi.org/10.1145/2883851.2883928

[27] Vitomir Kovanovic, Srecko Joksimovic, Dragan Gasevic, and Marek Hatala. 2014. What is the source of social capital? The association between social network position and social presence in communities of inquiry. In *Workshop at Educational Data Mining Conference*. EDM.

[28] Vitomir Kovanović, Srećko Joksimović, Zak Waters, Dragan Gašević, Kirsty Kitto, Marek Hatala, and George Siemens. 2016. Towards Automated Content Analysis of Discussion Transcripts: A Cognitive Presence Case. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (Edinburgh, United Kingdom) *(LAK '16)*. Association for Computing Machinery, New York, NY, USA, 15–24. https://doi.org/10.1145/2883851.2883950

[29] Vitomir Kovanović, Srećko Joksimović, Zak Waters, Dragan Gašević, Kirsty Kitto, Marek Hatala, and George Siemens. 2016. Towards Automated Content Analysis of Discussion Transcripts: A Cognitive Presence Case. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (Edinburgh, United Kingdom) *(LAK '16)*. Association for Computing Machinery, New York, NY, USA, 15–24. https://doi.org/10.1145/2883851.2883950

[30] Vitomir Kovanović, Dragan Gašević, and Marek Hatala. 2014. Learning analytics for communities of inquiry. *Journal of Learning Analytics* 1, 3 (2014), 195–198. https://doi.org/10.18608/jla.2014.13.21

[31] Vitomir Kovanović, Srećko Joksimović, Dragan Gašević, and Marek Hatala. 2014. Automated cognitive presence detection in online discussion transcripts. In *LAK'14*. Indianapolis, IN.

[32] Kadir Kozan and Jennifer C Richardson. 2014. Interrelationships between and among social, teaching, and cognitive presence. *The Internet and higher education* 21 (2014), 68–73. https://doi.org/10.1016/j.iheduc.2013.10.007

[33] Zhi Liu., Lingyun Kang., Monika Domanska., Sannyuya Liu., Jianwen Sun., and Changli Fang. 2018. Social Network Characteristics of Learners in a Course Forum and Their Relationship to Learning Outcomes. In *Proceedings of the 10th International Conference on Computer Supported Education - Volume 2: CSEDU,*. INSTICC, SciTePress, 15–21. https://doi.org/10.5220/0006647600150021

[34] Tanya J McGill and Jane E Klobas. 2009. A task technology fit view of learning management system impact. *Computers & Education* 52, 2 (2009), 496–508.

[35] Thomas E. Mcklin. 2004. *Analyzing Cognitive Presence in Online Courses Using an Artificial Neural Network*. Ph.D. Dissertation. Atlanta, GA, USA. Advisor(s) Harmon, Stephen W. https://scholarworks.gsu.edu/msit_diss/1/ AAI3190967.

[36] Patrick E McKnight and Julius Najab. 2010. Mann-Whitney U Test. *The Corsini encyclopedia of psychology* (2010), 1–1.

[37] Danielle S McNamara, Arthur C Graesser, Philip M McCarthy, and Zhiqiang Cai. 2014. *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

[38] Valter Neto, Vitor Rolim, Rafael Ferreira, Vitomir Kovanović, Dragan Gašević, Rafael Dueire Lins, and Rodrigo Lins. 2018. Automated analysis of cognitive presence in online discussions written in portuguese. In *European Conference on Technology Enhanced Learning*. Springer, Springer International Publishing, 245–261. https://doi.org/10.1007/978-3-319-98572-5_19

[39] Valter Neto, Vitor Rolim, Rafael Ferreira, Vitomir Kovanović, Dragan Gašević, Rafael Dueire Lins, and Rodrigo Lins. 2018. Automated analysis of cognitive presence in online discussions written in portuguese. In *European Conference on Technology Enhanced Learning*. Springer, Springer International Publishing, 245–261. https://doi.org/10.1007/978-3-319-98572-5_19

[40] Daniel Olivares, Rafael Ferreira Leite de Mello, Olusola Adesope, Vitor Rolim, Dragan Gaševic, and Christopher Hundhausen. 2019. Using social network analysis to measure the effect of learning analytics in computing education. In *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*, Vol. 2161. IEEE, 145–149. https://doi.org/10.1109/ICALT.2019.00044

[41] Jennifer Richardson, Karen Swan, Patrick Lowenthal, and Phil Ice. 2016. Social presence in online learning: Past, present, and future. In *Global Learn*. Association for the Advancement of Computing in Education (AACE), 477–483.

[42] Vitor Rolim, Rafael Ferreira, Rafael Dueire Lins, and Dragan Gǎsević. 2019. A network-based analytic approach to uncovering the relationship between social and cognitive presences in communities of inquiry. *The Internet and Higher Education* 42 (2019), 53–65.

[43] Liam Rourke, Terry Anderson, D. Randy Garrison, and Walter Archer. 1999. Assessing Social Presence In Asynchronous Text-based Computer Conferencing. *The Journal of Distance Education* 14, 2 (1999), 50–71. https://www.learntechlib.org/p/92000/

[44] Lei Shi, Rada Mihalcea, and Mingjun Tian. 2010. Cross language text classification by model translation and semi-supervised learning. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. 1057–1067.

[45] George Siemens, Dragan Gašević, and Shane Dawson (Eds.). 2015. *Preparing for the digital university: a review of the history and current state of distance, blended, and online learning*. Athabasca University, Edmonton, AB. http://linkresearchlab.org/PreparingDigitalUniversity.pdf

[46] Yla R. Tausczik and James W. Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology* 29, 1 (2010), 24–54. https://doi.org/10.1177/0261927X09351676

[47] Marco Tortoriello, Ray Reagans, and Bill McEvily. 2012. Bridging the knowledge gap: The influence of strong ties, network cohesion, and network range on the transfer of knowledge between organizational units. *Organization science* 23, 4 (2012), 1024–1039.

[48] Zak Waters, Vitomir Kovanović, Kirsty Kitto, and Dragan Gašević. 2015. Structure matters: Adoption of structured classification approach in the context of cognitive presence classification. In *Information Retrieval Technology*. Springer International Publishing, Cham, 227–238. https://doi.org/10.1007/978-3-319-28940-3_18

[49] Ruochen Xu, Yiming Yang, Hanxiao Liu, and Andrew Hsi. 2016. Cross-lingual text classification via model translation with limited dictionaries. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. 95–104.

[50] Cherng-Jyh Yen and Chih-Hsiung Tu. 2008. Online social presence: A study of score validity of the computer-mediated communication questionnaire. *Quarterly Review of Distance Education* 9, 3 (2008), 297. https://digitalcommons.odu.edu/efl_fac_pubs/28/

[51] Norazah Yusof, Azizah Abdul Rahman, et al. 2009. Students' interactions in online asynchronous discussion forum: A Social Network Analysis. In *2009 International Conference on Education Technology and Computer*. IEEE, 25–29. https://doi.org/10.1109/ICETC.2009.48

APÊNDICE E – ANALYTICS OF EMERGING AND SCRIPTED ROLES IN ONLINE DISCUSSIONS: AN EPISTEMIC NETWORK ANALYSIS APPROACH

Analytics of Emerging and Scripted Roles in Online Discussions: An Epistemic Network Analysis Approach

Máverick Ferreira[1], Rafael Ferreira Mello[2]( ), Rafael Dueire Lins[2],
and Dragan Gašević[3]

[1] Informatics Center, Universidade Federal de Pernambuco, Recife, Brazil
[2] Department of Computing, Universidade Federal Rural de Pernambuco,
Recife, Brazil
{rafael.mello,rafael.dueirelins}@ufrpe.br
[3] Faculty of Information Technology, Monash University, Melbourne, Australia
dragan.gasevic@monash.edu

Abstract. This paper investigates emerging roles in the context of the community of inquiry model. The paper reports the results of a study that demonstrated the application of epistemic network and clustering analyses to reveal the roles that different students assumed during an asynchronous course with online discussions. The proposed method highlights the differences and similarities between emerging and scripted roles based on the development of social and cognitive presences, two key constructs of the model of communities of inquiry.

Keywords: Emerging roles · Epistemic network analysis · CSCL

1 Introduction

Asynchronous online discussions are the most widely adopted resource to promote social interactions in computer-supported collaborative learning settings [4]. There is a rising need to understand asynchronous online discussions under different perspectives like engagement, knowledge (co-)construction and the roles of each learner within the discussion [3,19,24]. The Community of Inquiry (CoI) model highlights the social nature of interactions in learning by describing three different constructs [5,6,8]: the social, teaching, and cognitive presences. They measure the student social interactions and their cognitive development, unpack interactions in computer-supported collaborative scenarios [12], demonstrate the importance of teaching presence for the development of an effective community of learners [9,10], and measure the learners' participation in online communities [13,14].

Another relevant pedagogical aspect to observe in the asynchronous online communication process is the roles that learners assume during discussions. It could influence their contribution and interaction patterns with other group

I. Roll et al. (Eds.): AIED 2021, LNAI 12749, pp. 156–161, 2021.
https://doi.org/10.1007/978-3-030-78270-2_28

members [1]. In general, there are two categories of student roles [23]: i) emerging, when the student assumes a role intrinsically during the discussion and (ii) scripted, when the moderator pre-defines the role for each student. These role categories have a high impact on the social knowledge construction, group collaboration and cohesion during online discussions. While the predefined roles have been widely studied in the literature [17,18,23,25], fewer efforts have been devoted towards the analysis of the emerging roles [3]. The study reported in this paper unpacks the relationship between emerging and scripted roles and categories of social and cognitive presences in a CoI. To reach this goal, this study performed a cluster analysis in combination with an epistemic network analysis [21] to shed light on the scripted and emerging roles assumed by the students in asynchronous online discussions. More specifically, the aim of the stud was to answer the following research question: "*To what extent can a clustering algorithm using indicators of the social and cognitive presences predict emerging roles in an asynchronous online discussion?*"

1.1 Data and Course Design

The data used here encompass six offerings of a fully-online master level research-intensive course, with a total of 82 students that posted 1,747 messages, which accounted for 15% of the final mark. During the online discussions, there were two scripted roles defined by the instructors: (i) **experts**, the students that initiated the discussion by posting the video about the research paper they presented, (ii) **practicing researchers**, students who were requested to watch the video and interact asking questions and posting comments for the experts[1]. Not all students played both roles.

Two experts labeled the dataset using the coding scheme proposed by [7]. The coding unit of analysis was the entire message. However, each message could have more that one social presence indicator, but just one cognitive presence phase. The coders reached 84% and 98.1% of the agreement for social presence and cognitive presences, respectively. The differences were resolved through discussions among them. Table 1 presents the distribution of messages in the dataset.

Table 1. Distribution of social and cognitive presences.

Social messages	#	%	Cognitive messages	#	%
Affective positive	530	33.33%	*Other*	140	8.01%
Affective negative	1,217	66.67%	Triggering event	308	17.63%
Interactive positive	1,030	58.95%	Exploration	684	39.15%
Interactive negative	741	41.05%	Integration	508	29.08%
Cohesive negative	1,326	75.90%	Resolution	107	6.13%
Cohesive positive	421	24.10%	*Total*	1,747	100.00%

[1] For further details on the course design [10].

1.2 Clustering Algorithm

Clustering algorithms aim to create sets of high intra-group and low inter-group similarities between elements [22]. The messages posted by the students were clustered using the k-means algorithm [26] to analyze their emerging roles. The total number of instances evaluated was 163 equivalent to all pairs of student/role. The social presence indicators and the phases of cognitive presence were the features for the clustering algorithm. Such indicators from all student/role instance contributions were encompassed into a single line used as input to the k-means algorithm. The silhouette approach was used to identify the ideal number of clusters [11].

1.3 Epistemic Network Analysis (ENA)

ENA provides a mechanism to compare the differences between different groups of analysis units – such as between the emerging and scripted roles in the present study [20]. Each message was coded capturing the presence/absence of the social presence indicators and the cognitive presence phases. Both units of analysis and stanzas were students (i.e., all student messages) within different roles (emerging and scripted). The use of students as units of analysis here enabled us to see the connections between phases of cognitive presence and the indicators of social presence for each student. We removed the social presence indicators (Continuing a thread, Complementing, and Vocatives) that could generate a dominant code to enhance the ENA results [16].

2 Results

This study identified that the best number of groups was 3 (silhouette = 0.3644). The k-means algorithm was applied to identify the three clusters, considered the *emerging roles*. Table 2 shows the percentage of indicators for each cluster identified and the scripted roles (expert and practicing research). Cluster 0 presented the greater values for the social presence indicators, showing the concern in creating discussions that could reach a social climax, similar to the expert role. Cluster 1 had slightly higher values for the affective category indicators, but in general, the values for the social presence were similar to those of clusters 0 and 2. Both clusters 0 and 1 presented similar values of cognitive presence. Cluster 2 had a similar trend with the practicing researcher role when compared the social presence indicators; however, it also presented a lower development of the cognitive presence.

Figure 1 presents the projection of the average students' networks with relationships between the social and cognitive presences. The rectangles represent group-average networks (95% CI are also outlined) for students in expert (red), practicing researcher (blue), cluster 0 (green), cluster 1 (orange) and cluster 2 (purple) roles. The visualization was done using 1 and 2, which accounted for 18.8 and 12.2% of variability between students' network models, respectively.

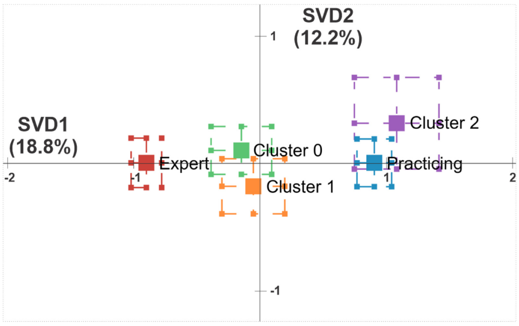


Fig. 1. ENA projections of the means values of group networks of the students who assumed different scripted and emerging roles. Each colored square (95% CI are outlined around the group means) represent a mean network of one role (emerging or scripted). (Color figure online)

Table 2. Indicators for emerging and scripted (Expert and Practicing Researcher)

		Indicator	Cluster 0	Cluster 1	Cluster 2	Expert	Practicing
	Affective	Emotions	16.43%	17.20%	15.84%	22.12%	11.15%
		Humor	02.28%	02.42%	00.42%	02.94%	02.12%
		Self discosure	17.00%	17.74%	23.91%	17.06%	19.73%
Social Presence	Interactive	Continuing thread	95.73%	90.32%	99.38%	90.24%	100.00%
		Quoting message	04.75%	03.23%	00.93%	06.47%	01.11%
		Referring message	05.32%	05.91%	04.04%	05.65%	04.79%
		Asking question	40.36%	37.37%	73.29%	14.47%	75.47%
		Complimenting	81.67%	64.25%	90.68%	70.94%	87.85%
		Agreement	14.81%	16.67%	07.76%	17.18%	10.81%
	Cohesive	Vocatives	84.14%	66.94%	92.55%	77.41%	86.40%
		Group	08.55%	08.87%	06.52%	09.41%	07.13%
		Salutations	75.78%	60.48%	80.12%	72.00%	74.58%
Cognitive presence		Others	07.41%	10.48%	07.14%	07.29%	08.70%
		Triggering event	16.14%	16.13%	24.22%	17.65%	17.61%
		Exploration	38.46%	37.10%	43.79%	39.76%	38.57%
		Integration	31.53%	29.30%	20.81%	29.18%	28.99%
		Resolution	06.46%	06.99%	04.04%	06.12%	06.13%
		Number of students/role	82	67	14	06.12%	06.13%
		Number of posts	1053	372	322	850	897

3 Conclusions and Lines for Further Work

The cluster analysis revealed that the social presence indicators and cognitive presence phases were effective in dividing the students into groups as the silhouette value reached 0.3644 [22]. Table 2 shows that cluster 1 and cluster 2 erre the most diverse groups in relation to the social and cognitive presences, respectively. The literature shows that when the instructor assigns students to scripted roles, the students play their scripted roles (especially when the participation in group work counts towards final marks) [15,23]. Cluster 0 and cluster 2 showed high similarities with the expert and practicing researcher, respectively, corroborating the literature [15,23]. However, cluster 2 had only 14 students in the practicing researcher role, and cluster 0, included students from both scripted roles. It is possible to say that the definition of the scripted roles impacted on the students' participation in group activities, but not every student acted according to their scripted role. Cluster 1, the less active students (an average of 5 posts per student), did not present a high degree of similarity with any scripted role.

The approach proposed here has some limitations that must be acknowledged. First, this study was based on the data from six-course offerings, from a single course and institution, which can affect the generalizability of the results obtained. Second, the findings of the present study may be limited, given the features of the course design and the scripted roles. The authors intend to apply the same analytic approach with other datasets form different course settings and with online discussions in a different language to address those problems. Finally, cluster analysis involves making many decisions, such as on the number of clusters and the features used. Different algorithms and parameters may yields different results. The authors intend to evaluate different clustering algorithms and incorporate other features as suggested in [2,3,18].

References

1. De Laat, M., Lally, V.: It's not so easy: researching the complexity of emergent participant roles and awareness in asynchronous networked learning discussions. J. Comput. Assist. Learn. **20**(3), 165–171 (2004)
2. De Wever, B., Van Keer, H., Schellens, T., Valcke, M.: Roles as a structuring tool in online discussion groups: the differential impact of different roles on social knowledge construction. Comput. Hum. Behav. **26**(4), 516–523 (2010)
3. Dowell, N.M., Poquet, O.: SCIP: combining group communication and interpersonal positioning to identify emergent roles in scaled digital environments. Comput. Hum. Behav. 106709 (2021)
4. Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., Romero, C.: Text mining in education. Wiley Interdiscip. Rev.: Data Mining Knowl. Disc. **9**(6), e1332(2019)
5. Garrison, D.R.: Thinking Collaboratively: Learning in a Community of Inquiry. Routledge, New York (2016)
6. Garrison, D.R., Anderson, T., Archer, W.: Critical inquiry in a text-based environment: computer conferencing in higher education. Internet High. Educ. **2**(2–3), 87–105 (2000). https://doi.org/10.1016/S1096-7516(00)00016-6
7. Garrison, D.R., Anderson, T., Archer, W.: Critical thinking, cognitive presence, and computer conferencing in distance education. Am. J. Distance Educ. **15**(1), 7–23 (2001). https://doi.org/10.1080/08923640109527071
8. Garrison, D.R., Anderson, T., Archer, W.: The first decade of the community of inquiry framework: a retrospective. Internet High. Educ. **13**(1–2), 5–9 (2010)
9. Garrison, D.R., Cleveland-Innes, M.: Facilitating cognitive presence in online learning: interaction is not enough. Am. J. Distance Educ. **19**(3), 133–148 (2005)
10. Gašević, D., Adesope, O., Joksimović, S., Kovanović, V.: Externally-facilitated regulation scaffolding and role assignment to develop cognitive presence in asynchronous online discussions. Internet High. Educ. **24**, 53–65 (2015)

11. Hamerly, G., Elkan, C.: Learning the k in k-means. Adv. Neural Inf. Process. Syst. **16**, 281–288 (2004)
12. Joksimović, S., Gašević, D., Kovanović, V., Adesope, O., Hatala, M.: Psychological characteristics in cognitive presence of communities of inquiry: a linguistic analysis of online discussions. Internet High. Educ. **22**, 1–10 (2014)
13. Kovanović, V., Gašević, D., Joksimović, S., Hatala, M., Adesope, O.: Analytics of communities of inquiry: effects of learning technology use on cognitive presence in asynchronous online discussions. Internet High. Educ. **27**, 74–89 (2015)
14. Kovanović, V., et al.: Examining communities of inquiry in massive open online courses: the role of study strategies. Internet High. Educ. **40**, 20–43 (2019)
15. Mayordomo, R.M., Onrubia, J.: Work coordination and collaborative knowledge construction in a small group collaborative virtual task. Internet High. Educ. **25**, 96–104 (2015)
16. Ferreira Mello, R., Gašević, D.: What is the effect of a dominant code in an epistemic network analysis? In: Eagan, B., Misfeldt, M., Siebert-Evenstone, A. (eds.) ICQE 2019. CCIS, vol. 1112, pp. 66–76. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-33232-7_6
17. Näykki, P., Isohätälä, J., Järvelä, S., Pöysä-Tarhonen, J., Häkkinen, P.: Facilitating socio-cognitive and socio-emotional monitoring in collaborative learning with a regulation macro script-an exploratory study. Int. J. Comput.-Supported Collab. Learn. **12**(3), 251–279 (2017)
18. Pozzi, F.: The impact of scripted roles on online collaborative learning processes. Int. J. Comput.-Supported Collab. Learn. **6**(3), 471–484 (2011)
19. Rolim, V., Ferreira, R., Lins, R.D., Găsević, D.: A network-based analytic approach to uncovering the relationship between social and cognitive presences in communities of inquiry. Internet High. Educ. **42**, 53–65 (2019)
20. Shaffer, D.W.: Epistemic Frames and islands of expertise: learning from infusion experiences. In: Proceedings of the 6th International Conference on Learning Sciences, ICLS 2004, pp. 473–480. International Society of the Learning Sciences, Santa Monica, California (2004). http://dl.acm.org/citation.cfm?id=1149126.1149184
21. Shaffer, D.W., et al.: Epistemic network analysis: a prototype for 21st-century assessment of learning. Int. J. Learn. Med. **1**(2) (2009)
22. Starczewski, A., Krzyżak, A.: Performance evaluation of the silhouette index. In: Rutkowski, L., Korytkowski, M., Scherer, R., Tadeusiewicz, R., Zadeh, L.A., Zurada, J.M. (eds.) ICAISC 2015. LNCS (LNAI), vol. 9120, pp. 49–58. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-19369-4_5
23. Strijbos, J.W., Weinberger, A.: Emerging and scripted roles in computer-supported collaborative learning. Comput. Hum. Behav. **26**(4), 491–494 (2010)
24. Thomas, J.: Exploring the use of asynchronous online discussion in health care education: a literature review. Comput. Educ. **69**, 199–215 (2013)
25. Wise, A.F., Chiu, M.M.: The impact of rotating summarizing roles in online discussions: effects on learners' listening behaviors during and subsequent to role assignment. Comput. Hum. Behav. **38**, 261–271 (2014)
26. Xu, R., Wunsch, D.: Clustering, vol. 10. John Wiley (2008)

APÊNDICE F – NASC: NETWORK ANALYTICS TO UNCOVER SOCIO-COGNITIVE DISCOURSE OF STUDENT ROLES

NASC: Network analytics to uncover socio-cognitive discourse of student roles

Roles that learners assume during online discussions are an important aspect of educational experience. The roles can be assigned to learners and/or can spontaneously emerge through student-student interaction. While existing research proposed several approaches for analytics of emerging roles, there is limited research in analytic methods that can i) automatically detect emerging roles that can be interpreted in terms of higher-order constructs of collaboration; ii) analyse the extent to which students complied to scripted roles and how emerging roles compare to scripted ones; and iii) track progression of roles in social knowledge progression over time. To address these gaps in the literature, this paper propose a network-analytic approach that combines techniques of cluster analysis and epistemic network analysis. The method was validated in an empirical study discovered emerging roles that were found meaningful in terms of social and cognitive dimensions of the well-known model of communities of inquiry. The study also revealed similarities and differences between emerging and script roles played by learners and identified different progression trajectories in social knowledge construction between emerging and scripted roles. The proposed analytic approach and the study results have implications that can inform teaching practice and development techniques for collaboration analytics.

CCS Concepts: • **Applied computing** → **Education**; **Distance learning**; • **Social and professional topics** → **Computing/technology policy**.

Additional Key Words and Phrases: Emerging Roles; Epistemic Network Analysis; CSCL; Clustering Analysis

ACM Reference Format:
. 2022. NASC: Network analytics to uncover socio-cognitive discourse of student roles. In *LAK '22: Learning Analytics and Knowledge Conference, April 11–15, 2021, Newport Beach, CA, US.* ACM, New York, NY, USA, 16 pages.

1 INTRODUCTION

Asynchronous online discussions have been widely used to in computer-supported collaborative learning (CSCL) settings [15]. They promote social interactions and knowledge construction, which are essential aspects of online learning environments. In this line of research, the literature emphasizes a need to provide automatic methods to understand the effectiveness of pedagogies that seek to promote social knowledge construction through online discussions [3, 13, 57]. This understanding can help instructors inform future decisions about pedagogical approaches that aim to support large numbers of students in online discussions [1, 12].

Roles that learners assume during discussions are an important aspect of educational experience and have attracted much attention in the literature [13, 16, 23, 34, 56, 60]. Roles in educational settings can be defined as groups of students with similar behaviour exhibited during their participation in a group activity [10]. In the case of online discussions, there are two categories of student roles [56]: (i) *emerging*, when the student assumes a role intrinsically during an online discussion, and (ii) *scripted*, when the moderator pre-defines the role for each student. Such roles could influence student participation and consequently impact the social knowledge construction, group collaboration and cohesion during an online discussion [50].

Analytics of roles students take in online discussions can offer useful insights that can inform teaching practice [13, 23, 34, 60]. However, to effectively provide analytic insights for instructors, the interpretation of students' roles should be connected to relevant learning designs and/or educational theory [23]. Also, the literature on student roles suggests that social and cognitive aspects of discussion should be considered [13, 23, 56, 58]. In this context, the Community of Inquiry (CoI) model is a well-established framework that outlines how asynchronous online communication shapes student learning and their cognitive development [19]. The CoI model proposes three different constructs (known as presences) to explain educational experience through social interactions [17]: social presence, teaching presence, and cognitive presence. These presences explain the process of student inquiry through social interactions in an online community of learners [17], recognise the role of social context to create a productive community of learners [22, 45], emphasise the importance of teaching for the development of an effective community of learners [1, 21], and could provide information on how students participation is associated with the development of online communities [29, 32].

There is a large body of research that examined associations of the CoI model [17] and script theory of guidance [16] with student engagement, knowledge construction, and learning outcomes [3, 13, 20, 25]. Previous research has explored how scripted roles could mould social and cognitive presences in a CoI [1, 3]. Nevertheless, there is limited research that has investigated the impact of *emerging roles* on the development of social and cognitive presences in a CoI. A further investigation on this topic could (i) reveal new knowledge that advances the understanding how presences develop in a CoI and (ii) provide a analytic approach to inform teaching decisions.

This paper proposes a network analytic approach called Network Analytics for Socio-Cognitive (NASC) for analysis of online discussions based on social and cognitive presences of the CoI model. The NASC approach is based on epistemic network analysis [54] and allows for automatic detection of emerging roles, comparison of emerging roles with scripted roles, and tracking development of social knowledge construction over time. The approach was validated in a study that used scripted roles for guiding students to develop social and cognitive presence in asynchronous online discussions.

2 THEORETICAL BACKGROUND

2.1 The Community of Inquiry Model

The Community of Inquiry (CoI) model aims to describe the processes of social knowledge construction in asynchronous online discussions [20]. The CoI model defines three interdependent dimensions (called presences) that together shape students' online learning experience.

The central construct of the model is *cognitive presence*, that describes students' process of critical thinking and inquiry-based learning. It is formally defined as "the extent to which the participants in any particular configuration of a community of inquiry are able to construct meaning through sustained communication" [18, p. 89]. Operationalised within the practical inquiry model [17], cognitive presence consists of four interconnected phases of inquiry-based learning cycle:

- **Triggering event** phase starts the learning cycle and is initiated by a problem or dilemma that initiates practical inquiry cycle;
- **Exploration** phase includes information sharing and investigation and debating of different ideas and potential solutions;

- **Integration** phase, where students connect and synthesise relevant ideas and information to construct new knowledge; and finally,
- **Resolution** phase, which includes the application of new knowledge to the originating problem, typically triggering a new learning cycle in the process.

Another critical construct within the CoI model is *social presence*, that captures social climate and interpersonal relationships within the course [47]. It is formally defined as "the ability of participants in a community of inquiry to project themselves socially and emotionally, as 'real' people (i.e., their full personality), through the medium of communication used" [18, p. 94]. Social presence consists of three sub-dimensions that capture different aspects of social interactions. The *affective* dimension capture the expression of students' real emotions and feelings map within the learning community. The *interactive* dimensions captures the open and interactive nature of student communication. Finally, *group cohesion* captures the sense of union and group commitment among students in the learning community. A meta-analysis reported in [45] showed a strong positive correlation of social presence with student satisfaction and perceived learning, while social presence has also also found to be positively associated with academic performance with role scripting as a moderator of the association [26].

Finally, *teaching presence* relates to the instructors' role before and during the course, and is formally defined as "the design, facilitation, and direction of cognitive and social processes for the purpose of realizing personally meaningful and educationally worthwhile learning outcomes" [7, p. 5]. It consists of the three components, namely, *design and facilitation* of online learning experience, discourse *facilitation* and *direct instruction*, when necessary. Teaching presence is crucial for the development of both cognitive and social presence. Similarly, social presence mediates the relationship between teaching and the cognitive presences [22, 31], providing necessary environment for providing foundation for the effective inquiry-based learning experience. The use of script roles is also used as a form of teaching presence in communities of inquiry[1].

2.2 Emerging and Scripted Roles in Online Discussions

Asynchronous discussion is a key element to support collaboration in online environments. However, a spontaneous interaction of students in online discussions does not necessarily lead to a high level of social knowledge construction, social presence, and cognitive development [16, 17]. Instead effective approaches to guiding learners in productive online discussions are necessary. In this context, the script theory of guidance [16] gained significant attention in the literature. From the perspective of the CSCL research, the script theory of guidance posits that productive student–student discussions can be supported and high-level of social knowledge construction achieved through scripting and assigning roles to students [16, 56]. According to Strijbos and Weinberger [56, p. 491] roles are "more or less stated functions or responsibilities that guide individual behavior and regulate group interaction," while *scripted roles* are intentionally "designed to improve both learning processes and outcomes" [p. 492]. Role assignment is a type of collaboration script central to the script theory, and has been extensively studied in the CSCL literature as an effective approach to increasing the level of knowledge construction, cognitive processing, and argumentation [3, 11, 50]. Role assignment and scripting are also used in research on CoI to as a way to guide students to achieve high levels of social and cognitive presence [1].

In contrast to scripted roles, emerging roles arise spontaneously from interpersonal exchanges without prior guidance, capture aspects that are not predefined at the beginning of the group work [56]. Emerging roles are also important to promote learner agency given that scripted roles can be seen to promote compliance [62] and even be demotivating [43], in spite of being beneficial for learning [43]. Developments in learning analytics enabled for automated detection of emerging roles from the analysis of digital traces recorded in online discussion forums [35, 51]. Existing approaches are based on natural language processing, social network analysis, clustering algorithms and epistemic network analysis [3, 13, 23, 34, 49, 60]. For instance, Dowell and Poquet [13] proposed an approach that combines automated content analysis and social network analysis to identify emerging roles in online discussion. Gašević et al. [23] suggested an approach that combines techniques of social and epistemic network analysis with techniques for automated content and cluster analysis to study emerging roles through the lenses of social and cognitive dimensions of online collaboration.

2.3 Research Questions

The automatic detection of emerging roles can provide insights into the extent to which roles that the instructors envisioned in their learning designs indeed hold [13, 23]. As such, analysis of emerging roles can inform decisions about future changes in learning design. However, at this stage there has not been sufficient research that looked specifically at the emergence of roles in online discussions based on high-order constructs such as cognitive and social presence. Analysis of emerging roles have so far been done based on lower-level text analysis in combination with cluster analysis [13], or measures of social networks analysis [23, 34] and manually coded messages for higher-order constructs [10, 56] but without focusing on analytic methods that aim to identify patters of participation through automated analysis of higher-order constructs of online group activity to identify emerging roles. In this context, we posit that social and cognitive presences of CoI, as higher order constructs of online participation, could provide a solid socio-cognitive conceptual foundation for analysis of emerging roles in online discussions. As such, the first research question addressed in the current study was:

> **RESEARCH QUESTION 1:** *To what extent can emerging roles students occupy in an asynchronous online discussion be identified through automated analysis of indicators of social and cognitive presences?*

If emerging roles can be automatically detected, it would be useful to determine how they relate to scripted roles to inform instructional decisions and potentially provide input for personalised feedback analytics-based tools such as OnTask [40]. While attempts for identification of emerging roles [3, 13, 23, 34, 49, 60] and analysis of the extent to which students comply to the functions of scripted roles [59, 61] are noted, there has been limited literature that proposed data analytic methods to compare emerging and scripted roles especially on the level of higher-order constructs such as relationships between cognitive and social presence. Hence, the second research question was formulated as:

> **RESEARCH QUESTION 2:** *To what extent are scripted roles, which aim at promoting productive online discussions, associated with emerging roles that are identified through automated analysis of indicators of social and cognitive presence?*

Finally, it is equally important to track the progression of learners over time and how they perform certain roles and how these roles progress in their social knowledge construction. Again, this can inform decisions about teaching interventions and even be used to construct analytics-based personalised feedback at scale with tools to help students improve their social knowledge construction. There has been a shortage of work that looked at how emerging and scripted roles progress over time and data analytic methods that can provide analysis of progression on the level of higher-order constructs such as relations between social and cognitive presence. Thus, the third research question was:

> **RESEARCH QUESTION 3:** *To what extent can we track changes in the associations between social and cognitive presences, for each role, over time?*

3 METHOD

To answer the proposed research questions, we proposed a learning analytic approach called Network Analytics of Socio-Cognitive discourse (NASC). Figure 1 presents the overall steps that are included in NASC. The process is initiated (Step 1) with the collection of online discussion transcripts from a learning environment. Then, the data is annotated according to the categories of social and cognitive presence (Step 2). Although we used a dataset that was manually coded, the literature presents several approaches for automatic coding of online discussion messages according to the categories of social and cognitive presences [6, 8, 9, 37]. Step 3 entails the use of a clustering algorithm to identify the emerging roles and produces the assignment of students to the relevant emerging roles (Step 4). Student assignments to emerging and scripted roles used to construct epistemic networks and preform epistemic network analysis to compare emerging and scripted roles and to track changes in social knowledge construction for each role over time (Step 5). Finally, the NASC produces results that include information about emerging roles detected automatically, differences between emerging and scripted roles, and the trajectory graphs showing how the roles changed over time.

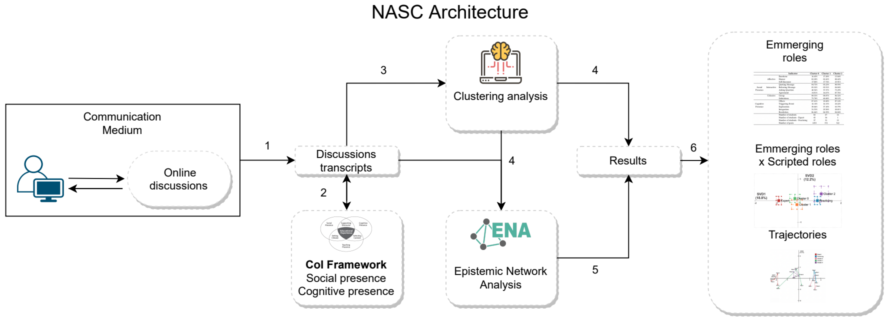


Fig. 1. The description of the proposed analytic approach NASC – network analytic of socio-cognitive discourse

3.1 Discussion Transcript and CoI annotation

The dataset used in the study comprised six offerings (Winter 2008, Fall 2008, Summer 2009, Fall 2009, Winter 2010, Winter 2011) of a fully online master-level research-intensive course in software engineering, with 82 students that posted 1,747 messages. As part of the course assessment (15% of the final mark), the students were instructed to record a video presentation about one research paper related to one of the course topics, in which the other students would engage in online discussions around their presentation. Within each discussion, the students could assume two scripted roles defined by the instructors: (i) **experts**, the students that posted the video presentation that triggered the discussion; they also had the responsibility to answer questions about the video and moderate the discussion; (ii) **practising researchers**, students who watched the video in order to interact with comments and questions for the experts [1]. It is important to mention that the students were assigned a specific role for each discussion. It means that the same student could assume the expert or practising roles in different discussions. Not all students played both roles. The discussions

[1] For further details on the course design [1].

Table 1. Distribution of cognitive presence.

Phase	Messages	%
Triggering event	308	17.63%
Exploration	684	39.15%
Integration	508	29.08%
Resolution	107	6.13%
Other	140	8.01%
Total	1,747	100.00%

happened over the four weeks of the course (Weeks 2-5) and the students could assume the expert role in only one discussions while in all other discussions participated as practising researchers.

As already mentioned, the data analysed in this study was manually annotated. Two experts in CSCL and software engineering codded the dataset for both cognitive and social presence using the coding scheme proposed by Garrison et al. [19], Rourke et al. [46]. All online discussion messages were coded according to the indicators and phases of the social and cognitive presences, respectively. The coding unit of analysis was entire message. Each message could have more than one social presence indicator [44], but just one cognitive presence phase [19]. The coders achieved an excellent level of agreement for cognitive presences, reaching a percentage of agreement = 98.1% and Cohen's k = 0.974. The differences (32 messages) were resolved through discussion among coders. Table 1 shows the number of messages in each phase of cognitive presence in the dataset.

For social presence, the final number of indicator codes was 3,770 instead of 1,747 (the total number of messages) due to multiple codes assigned to one message. The coders reached the agreement of 84% of agreement for each indicator of social presence. Table 2 presents the distribution of the indicators of social presence after the coding process. We omitted some of the social presence indicators (i.e., Continuing a thread, Complementing, and Vocatives) in our analysis as they had a disproportionately large number of messages with such codes. It follows the guideline used in previous work [30] and avoids the problem of a dominant code in epistemic network analysis [4].

3.2 Cluster Analysis

The current study adopted a cluster analysis to identify students' emerging roles during online discussions (RQ1). Clustering algorithms are unsupervised machine learning techniques applied to identify relevant subgroups with similar patterns within an extensive data collection [63]. Cluster analysis has been widely used to analyse online discussions under different perspectives [15, 29], including analysis of student roles [13, 28, 33, 60]. Among the possible clustering algorithms, several previous studies in educational settings have adopted the k-means algorithm, with the use of the euclidean distance, to uncover patterns in online discussions [13, 15, 33]. K-means usually reaches a good performance with a small processing time [63]. However, k-means requires the number of clusters (k) as an input parameter. The literature recommends the use of the silhouette coefficient to determine an optimal number of clusters [24, 55]. The silhouette coefficient is generally evaluated with a number from 2 to n (depends on the dataset), and the higher value indicates the ideal number of clusters for the k-means algorithm.

As presented in Section 3.1, our data was collected from 82 students, who could assume the expert and practicing researcher roles in different discussion threads. It is important to note that not every student assumed both roles during

Table 2. Social Presence Indicators

Category	Indicator	Count	Percent Agreement
Affective	Expression of emotions	288 (16.5%)	84.4
	Use of humor	44 (2.52%)	93.1
	Self-disclosure	322 (18.4%)	84.1
Interactive	Continuing a Thread	1664 (95.2%)	98.9
	Quoting from others messages	65 (3.72%)	95.4
	Referring explicitly to otherâ€™s messages	91 (5.21%)	92.7
	Asking questions	800 (45.21%)	89.4
	Complementing, expressing appreciation	1391 (79.6%)	90.7
	Expressing agreement	243 (13.9%)	96.6
Cohesive	Vocatives	1433 (82%)	91.8
	Addresses or refers to the group using inclusive pronouns	144 (8.24%)	88.8
	Phatics, salutions	1281 (73.3%)	96.1

the course (i.e., some students simply decided not to perform their scripted role). To answer the first research questions, the unit of analysis of the clustering algorithm was a pair of students in a specific scripted role. Therefore, we did not evaluate only 82 instances (i.e., number of students), but a total of 163 instances, which is equivalent to all pairs of (student,role). We adopted social presence indicators and the phases of cognitive presence as features for the clustering algorithm. The final feature vector of each instance contained the number of indicators of each message posted in the online discussion by the pairs of (student,role).

3.3 Epistemic Network Analysis

To answer the second and the third research questions we adopted epistemic network analysis (ENA) [54]. ENA is a technique based on network analysis employed to explore rich relationships between a set of concepts. Initially applied to examine the relationships between concepts in discourse transcripts [54], ENA has been applied to study different educational phenomena such as the assessment of students, time management, learning strategies, the analysis of verbal data in CSCL, the evaluation of the development of different students groups in an online discussion, and measurement of the performance of an educational intervention [3, 14, 38, 39]. ENA can model complex domains as networks especially adopted for problems with a small set of concepts characterized by highly dynamic and dense interactions [52]. It provides a mechanism to compare the differences between different groups of analysis units – such as between the emerging and scripted roles in the present study.

ENA models connections among different *codes* (e.g., indicators for social presence) for each *unit of analysis* (e.g., student) based on co-occurrences of codes in data subsets called *stanzas* or *conversations* (e.g., co-occurrence of codes for indicators of social presence in every post of a online discussion). Based on code co-occurrences of all units of analysis, ENA builds a high-dimensional network representation called *analytic space* . Then, the units of analysis are plotted in a *projection space*, which is derived from the analytic space through singular value decomposition (SVD), a dimension reduction technique similar to principal component analysis. In the end, the output of ENA is a series of two-dimensional graph models which capture the relationships between the different codes and which are based on two SVD dimensions which typically explain the high proportion of variance in data[53].

Typical diagrams used in ENA include: (i) projection graphs, which plot centroids[2] of each unit of analysis as dots in the diagram and with squares representing the mean values of centroids of different groups; ii) the network graphs, which represent connections among the codes of a unit of analysis or a group of units of analysis; (iii) the subtraction network graph, which displays the differences in the connections of the codes related to two groups. Furthermore, it is also possible to use ENA to build a trajectory graph, which explains how networks of different groups changed over time. That is, trajectory graphs plot connections between centriods of the same unit of analysis.

To answer research question 2, each discussion message was coded with several binary codes capturing the presence/absence of the social presence indicators and the cognitive presence phases. Both units of analysis and stanzas were students (i.e., all student messages) within different roles (emerging and scripted). The use of students as units of analysis here enabled us to see, for each student, their connections between the phases of cognitive presence and the indicators of social presence. Specifically, we answered research question 2 through the projection and subtraction graphs. Moreover, the differences between the roles on both svd1 and svd2 values were then compared by using a series of the Mann-Whitney tests [48] with α=0.05.

As research question 3 aimed to analyse the temporal development in the relationship between cognitive and social presences for emerging and scripted roles, we used a different configuration in the application of ENA. We adopted week and day as the unit of analysis and stanza, respectively. This analysis produced a different network graph (i.e., the values of the svd dimensions differed from those produced by ENA applied to research questions 2). The goal is to use the trajectory graph to assess the change of each scripted and emerging role over the weeks of the course discussion.

4 RESULTS

4.1 Research Question 1

The analysis of different values for k (2..30) identified 3 as the optimal number of clusters with the value of 0.3644 for the silhouette coefficient. Then, we employed the k-means algorithm to identify the three groups (cluster 0, cluster 1, and cluster 2), which are considered the *emerging roles* in this analysis. Table 3 shows the percentage of indicators for each cluster identified and the scripted roles (expert and practicing research) and the number of students and messages for each role.

Clusters 0 and 1 presented similar values for cognitive presence and affective category of social presence. However, cluster 0 showed a strong focus on interactivity in the discussions (i.e., interactive category of social presence) and creating an environment where the students are deeply involved in a conversation (Cohesive category). Moreover, it is possible to see that students in cluster 1 were less active during the discussion (with an average of 5.55 posts by students). In short, students in both clusters (0 and 1) can be described as cognitively engaged, but cluster 0 also had relevant social interactions. Therefore, we named clusters 0 and 1 as the ***socio-cognitively engaged*** and ***cognitively engaged*** emerging roles, respectively.

Cluster 2 had a similar trend with that of the practicing-researcher scripted role according to the indicators of social presence. This trend could be explained by the small number of students who performed the role of experts in this cluster (only one). However, this cluster also contained most of its messages in the triggering event (which could also be associated with asking questions, an indicator of social presence) and exploration phases, both of which represent low levels of cognitive presence. As the main focus of students in cluster 2 was the social presence, we categorise this cluster as the **socially-focused** emerging role.

[2] A centroid is the mean value of edge weights for a given epistemic network [53]

Table 3. Percentage of indicators for emerging (Clusters 0–2) and scripted (Expert and Practicing Researcher) roles

	Emerging roles			**Scripted roles**	
	Cluster 0: Socio-cognitively engaged	Cluster 1: Cognitively engaged	Cluster 2: Socially-focused	Expert	Practicing
Social presence					
Affective					
Emotions	16.43%	17.20%	15.84%	22.12%	11.15%
Humor	02.28%	02.42%	00.42%	02.94%	02.12%
Self discosure	17.00%	17.74%	23.91%	17.06%	19.73%
Interactive					
Continuing Thread	95.73%	90.32%	99.38%	90.24%	100.00%
Quoting Message	04.75%	03.23%	00.93%	06.47%	01.11%
Referring Message	05.32%	05.91%	04.04%	05.65%	04.79%
Asking Question	40.36%	37.37%	73.29%	14.47%	75.47%
Complimenting	81.67%	64.25%	90.68%	70.94%	87.85%
Agreement	14.81%	16.67%	07.76%	17.18%	10.81%
Cohesive					
Vocatives	84.14%	66.94%	92.55%	77.41%	86.40%
Group	08.55%	08.87%	06.52%	09.41%	07.13%
Salutations	75.78%	60.48%	80.12%	72.00%	74.58%
Cognitive presence					
Triggering Event	16.14%	16.13%	24.22%	17.65%	17.61%
Exploration	38.46%	37.10%	43.79%	39.76%	38.57%
Integration	31.53%	29.30%	20.81%	29.18%	28.99%
Resolution	06.46%	06.99%	04.04%	06.12%	06.13%
Others	07.41%	10.48%	07.14%	07.29%	08.70%
No. of students	82	67	14	82	81
No. of students - Expert	45	36	1	82	0
No. of students - Practicing	37	31	13	0	81
No. of posts	1053	372	322	850	897
No. of posts per student	12.84	5.55	23	10.36	11.07

4.2 Research Question 2

Figure 2 presents the projection of the average networks of each emerging and scripted roles. The rectangles represent group-average networks (95% CI are also outlined) for students in the expert (red), practicing researcher (blue), socio-cognitively engaged (green), cognitively engaged (orange), and socially-focused (purple) roles. The graph shows that svd1 and svd2 dimensions accounted for 18.8 and 12.2 percent of variance in data, respectively. The differences between the students in the expert and practicing researcher roles can be observed visually, with the key difference along the svd1 (U=283.50, p=0.00, r=0.91). In terms of emerging roles, the difference between socio-cognitively engaged students and cognitively engaged students was along the svd2(U=2206.00, p=0.04, r=0.20), while the students when played the socially-focused role differentiated from the cognitively engaged role along both svd1(U=190.00, p=0.00, r=0.59) and svd2(U=672.00, p=0.01, r=0.43), and from the socio-cognitively engaged role on the svd1 (U=254.00, p=0.00, r=0.56). Finally, Figure 2 also shows that socially-focused role overlapped with the practicing researcher scripted role, and the socio-cognitively engaged and cognitively engaged roles were relatively close to the expert role.

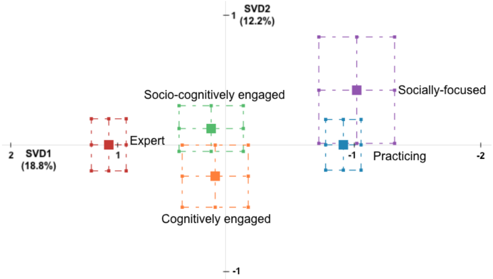


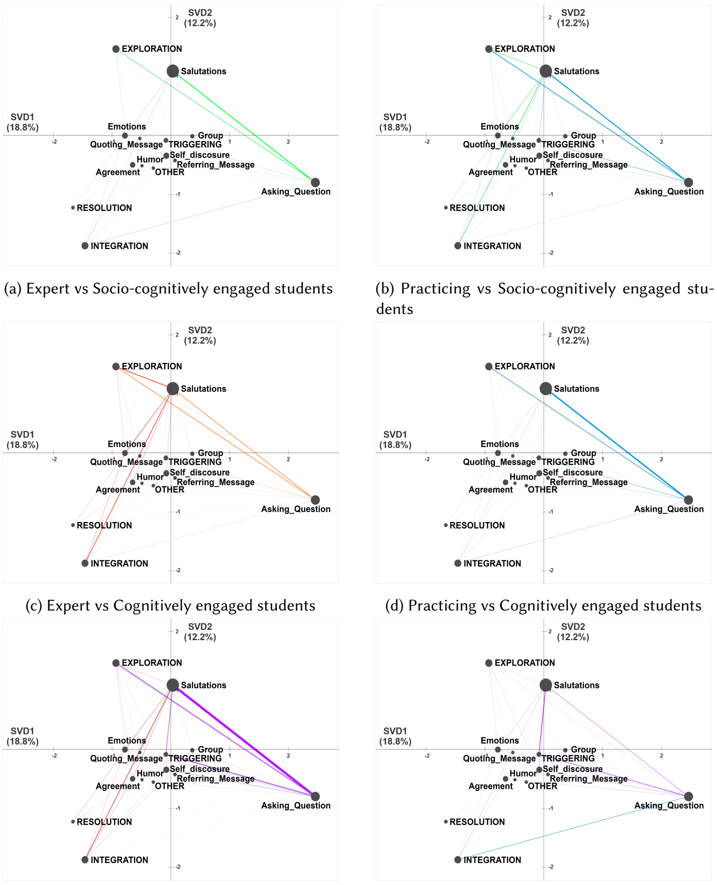
Fig. 2. ENA projections of the means values of group networks of the students who assumed different scripted and emerging roles. Each colored square (95% CI are outlined around the group means) represent a mean network of one role (emerging or scripted).

Figure 3 shows the subtraction networks comparing the scripted and emerging roles. It reveals that the socio-cognitively engaged emerging role had a similar level of participation compared to that of the expert script role; however, the students in the socio-cognitively engaged emerging role made more connections with the indicators of social presence (Asking_Question, Salutation, Referring_Message, Group) than the expert emerging role did (Figure 3a). Following a similar trend, the cognitively engaged emerging role also possessed connections with the aforementioned indicators of social presence (Figure 3c), but the expert scripted role had stronger connections of the exploration and integration phases with the indicators of social presence plotted on the left side of the svd1-axis. Finally, Figure 3f confirms the similarity between socially-focused and the practicing researcher role, while Figure 3e indicates a weak connection of socially-focused role and the cognitive presence phases when compared to the expert role.

4.3 Research Question 3

The ENA trajectory graph was employed to answer research question 3, providing details on how networks of the students in different roles changed in terms of connections between social and cognitive presences over the four weeks of the discussion. Figure 4a presents the final network built using the entire dataset. Again, the visualisation was done using svd1 and svd2, which accounted for 63.8 and 10.1 percent of variability, respectively. The network in Figure 4a shows that svd1 and svd2 tended to explain the indicators of social presence and the cognitive presence phases, respectively. In terms of social presence, the right-hand side of svd1 contains indicators related to basic interactions (asking questions and referring messages) and affective indicators (emotion and humor), while the left-hand side encompasses more indicators related to group cohesion (group, agreement, self_disclosure). Another interesting observation is that the integration and exploration phases of cognitive presence were plotted in distinct parts of the graph, leaving the resolution and triggering event categories in the middle.

The trajectories of the role over the four weeks (Figure 4b) were plotted in the same high-dimensional space as it shown in Figure 4a. It means that we could analyse the trajectory of each group in relation to the position of social and cognitive presence in Figure 4a. As mentioned before, the trajectory graph shows the location of the central activity performed by students from each role in each week of the discussion connected by a line that represents how the groups evolved from one week to the other. This analysis revealed a similar trend between students in the socially-focused emerging role and practicing researcher scripted role, where they were more related to indicators such as asking



(a) Expert vs Socio-cognitively engaged students

(b) Practicing vs Socio-cognitively engaged students

(c) Expert vs Cognitively engaged students

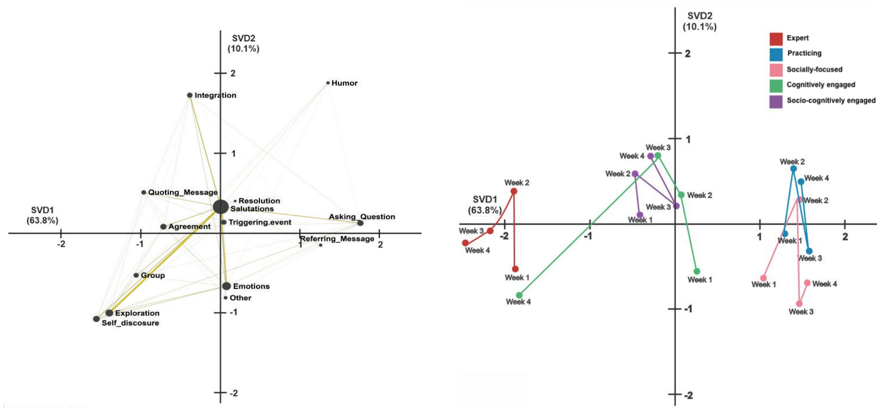
(d) Practicing vs Cognitively engaged students

(e) Expert vs Socially-focused students

(f) Practicing vs Socially-focused students

Fig. 3. ENA subtraction of student networks related to the scripted (Expert and Practicing) and emerging roles (clusters 0, 1 and 2)

questions, referring to previous messages, and triggering events. The socio-cognitively engaged and cognitively engaged emerging roles had some similarities, as they were plotted on the central part of the graph. The socio-cognitively engaged emerging role was closer to the loop between the triggering event-resolution phases of cognitive presence, while the cognitively engaged emerging role initiated week 1 closer to the exploration phase of social presence. Finally, the trajectory graph revealed that the students in the expert scripted role favoured the exploration and group cohesion indicators.



(a) Group-average networks graph of the relationship between social and cognitive presences for the development of all students over the four weeks of discussions.

(b) Mean trajectory graph for the students in the different groups including expert (red), practicing (blue), cluster 0 (salmon), cluster 1 (green), and cluster 2 (purple).

Fig. 4. Students' trajectory analysis over four weeks of discussions.

5 DISCUSSION

5.1 Interpretation of the Results

The results related to research question 1 showed that the indicators of social presence and phases of cognitive presence could be suitable to identify emerging roles (defined as clusters) as confirmed with the silhouette value of 0.3644 [55]. The results presented in Table 3 detail the main features of different emerging roles, which were categorised as socio-cognitively engaged, cognitively engaged and socially-focused roles. It is note that the emerging roles demonstrated some similarities and differences with scripted roles.

The emerging roles identified with the use of the NASC approach had some parallels with the existing research on scripted and emerging roles. First, the literature shows that when an instructor assign the students to scripted roles, the general trend is that students participate according to their scripted roles, especially when the participation in group work counts towards final marks [36, 56]. However, when the same student needs to act differently in distinct online discussions, a previous role the student was asked to play could influence the attitudes of the students in the future role assignments [42]. Our results confirm this trend as two out of the three emerging roles (except the socially-focused emerging role) contained students from both scripted roles. Second, it is important to highlight that the scripted roles investigated in this study were related to simple instructions (experts – initiating and maintain a discussion and practicing researchers – asking questions and post comments). However, the CSCL community usually suggests the adoption of more cognitively-related (e.g., argumentative roles) scripted roles [16, 56]. Therefore, the emerging roles identified in this study show a higher similarity with previous CSCL literature papers and provide additional information on the students' behaviour during the discussion.

Research question 2 aimed to unpack the association between social and cognitive presences for scripted and emerging roles in asynchronous online discussions. The results confirmed that each emerging role tended to be related to one of the scripted roles. While socio-cognitively engaged and cognitively engaged emerging roles were closer to the expert scripted role, students in the socially-focused emerging role had a participation level similar to that of the practicing researcher scripted role. Despite the similarities of socio-cognitively engaged and cognitively engaged emerging roles with and the expert scripted role, the results showed that the expert role leaned towards higher cognitive presence levels. This result is sensible as the expert role was designed to be in charge of discussions [1]. Moreover, the participation of students in socio-cognitively engaged and cognitively engaged emerging roles was slightly different. On the one hand, socio-cognitively engaged role predominantly showed social presence in terms of interactive indicators; this can explain the strong relation of the socio-cognitively engaged role with the exploration and integration phases, which are reported in the literature to be more related to the the interactive category of social presence [3, 46]. In contrast, students in the cognitively engaged emerging role focused on affective messages and asked questions representative of the triggering event phase [3, 46].

As the socially-focused emerging role encompassed a subgroup of the students who assumed the roles of practicing researchers, its relationship with the expert role is similar to the relationship between the practicing and expert scripted roles as reported in the literature [3]. In short, the socially-focused emerging role mostly remained on low levels of cognitive presence and was related to the salutation and asking question indicators of social presence. The direct comparison between the socially-focused emerging role and the practicing researcher script role showed that students in the socially-focused emerging role emerging role did not engage deeply in the discussion as their messages were more related to the triggering event phase and the salutation and asking questions indicators of social presence. Thus, students in the socially-focused role could potentially be in the lower performance group in the discussion [13].

The last research question aimed to analyse changes in the participation of each emerging role over time through the use of trajectory analysis provided by ENA. This analysis reemphasised the need for analysing the emerging roles even if the course design previously introduced scripted roles. It is clear that the students acting as experts focused on answering questions (i.e., exploration phase) and encourage group cohesion (i.e., group and self-disclosure). However, they did not entirely act as expected as the goal of this role was the initiation and conclusion of the discussion that could be related to the triggering and resolution phases, respectively [1, 5]. On the other hand, both socio-cognitively engaged and cognitively engaged roles were closely related to this goal.

5.2 Implications

The findings of the study showed that the use of the NASC approach is promising for the analysis of emerging roles in online discussions. Not only could the approach automatically identify theoretically meaningful emerging roles, but it could compare them to the script roles and track progression over time in terms of well-established higher-order constructs of social and cognitive presence. This can offer useful insights to the teachers to inform their decisions and improve overall teaching presence in a community of inquiry [17]. NASC can provide teachers with insights in the efficacy of their pedagogical intentions (i.e., script roles), so that they can take relevant action such as changes to the scripts of roles. It can also help them track student progress in terms of social and cognitive presences and create a foundation for enhancement of educational experience by giving timely feedback and recognising students that are not cognitively engaged in the discussion [3]. Moreover, NASC also holds a potential to be used in formative assessment of collaboration skills [13] by emphasising the developmental nature of assessment [35] through the use of trajectory graphs of ENA .

While the NASC approach showed premising results in the analysis of student roles even bigger potential lies in its integration with the existing body of research in learning and collaboration analytics [51]. The approach can effectively be used together with existing approaches for automatic classification of online discussion messages for cognitive and social presence [2, 8, 37]. This in turn eases potential adoption of the proposed analytic approach without a need for the users to manually code messages, as it was done in the current study. It would however be a promising research direction to investigate how the NASC approach can be complemented with existing work in collaboration analytics that makes use of social network analysis to further probe the social dimension of collaboration [23, 41] and automatic methods for automatic analysis of other constructs of the cognitive dimension of collaboration [27].

5.3 Limitations and future work

While this proposed approach shows a promise in addressing significant issues in the research on the association of emerging and scripted roles with social and cognitive presence in developing communities of inquiry, there are several limitations of the present study that should be acknowledged. First, this study was based on the data from six offerings of the same course at a single institution, which can negatively affect the broader implications of the presented analytic approach and the generalisability of the results obtained. Second, the findings of the present study findings may be somewhat limited, given the specifics of the adopted course design and the scripted roles. Therefore, it is important to apply the same analytic approach to other datasets collected in different course settings and with online discussions in a different language other than English to address those problems. Finally, as common for cluster analysis, it involves making many methodological decisions, such as deciding on the number of clusters and the features used, it might be that the findings in the present study would be different if different algorithms and parameters were used. Thus, it would be prudent to evaluate different clustering algorithms and experiment with other features in cluster analysis as suggested in the related literature [12, 13, 42].

REFERENCES

[1] ××× . 2015. Blinded for peer review. (2015).

[2] ××× . 2016. Blinded for peer review. (2016).

[3] ××× . 2019. Blinded for peer review. (2019).

[4] ××× . 2019. Blinded for peer review. (2019).

[5] ××× . 2019. Blinded for peer review.

[6] ××× . 2020. Blinded for peer review. (2020).

[7] Terry Anderson, Liam Rourke, D. Randy Garrison, and Walter Archer. 2001. Assessing Teaching Presence in a Computer Conferencing Context. *Journal of Asynchronous Learning Networks* 5 (2001), 1–17.

[8] Arthur Barbosa, Máverick Ferreira, Rafael Ferreira Mello, Rafael Dueire Lins, and Dragan Gasevic. 2021. The impact of automatic text translation on classification of online discussions for social and cognitive presences. In *LAK21: 11th International Learning Analytics and Knowledge Conference*. 77–87.

[9] Gian Barbosa, Raissa Camelo, Anderson Pinheiro Cavalcanti, Péricles Miranda, Rafael Ferreira Mello, Vitomir Kovanović, and Dragan Gašević. 2020. Towards automatic cross-language classification of cognitive presence in online discussions. In *Proceedings of the tenth international conference on learning analytics & knowledge*. 605–614.

[10] Maarten De Laat and Vic Lally. 2004. It's not so easy: Researching the complexity of emergent participant roles and awareness in asynchronous networked learning discussions. *Journal of Computer Assisted Learning* 20, 3 (2004), 165–171.

[11] Bram De Wever, Hilde Van Keer, Tammy Schellens, and Martin Valcke. 2007. Applying multilevel modelling to content analysis data: Methodological issues in the study of role assignment in asynchronous discussion groups. *Learning and instruction* 17, 4 (2007), 436–447.

[12] Bram De Wever, Hilde Van Keer, Tammy Schellens, and Martin Valcke. 2010. Roles as a structuring tool in online discussion groups: The differential impact of different roles on social knowledge construction. *Computers in Human Behavior* 26, 4 (2010), 516–523.

[13] Nia MM Dowell and Oleksandra Poquet. 2021. SCIP: Combining Group Communication and Interpersonal Positioning to Identify Emergent Roles in Scaled Digital Environments. *Computers in Human Behavior* (2021), 106709.

[14] Rafael Ferreira, Vitomir Kovanović, Dragan Gašević, and Vitor Rolim. 2018. Towards combined network and text analytics of student discourse in online discussions. In *International conference on artificial intelligence in education*. Springer, 111–126.

[15] Rafael Ferreira-Mello, Máverick André, Anderson Pinheiro, Evandro Costa, and Cristobal Romero. 2019. Text mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9, 6 (2019), e1332.

[16] Frank Fischer, Ingo Kollar, Karsten Stegmann, and Christof Wecker. 2013. Toward a script theory of guidance in computer-supported collaborative learning. *Educational psychologist* 48, 1 (2013), 56–66.

[17] D Randy Garrison. 2015. *Thinking collaboratively: Learning in a community of inquiry.* Routledge.

[18] D Randy Garrison, Terry Anderson, and Walter Archer. 2000. Critical inquiry in a text-based environment: Computer conferencing in higher education. *The internet and higher education* 2, 2-3 (2000), 87–105. https://doi.org/10.1016/S1096-7516(00)00016-6

[19] D. Randy Garrison, Terry Anderson, and Walter Archer. 2001. Critical thinking, cognitive presence, and computer conferencing in distance education. *American Journal of Distance Education* 15, 1 (2001), 7–23. https://doi.org/10.1080/08923640109527071

[20] D. Randy Garrison, Terry Anderson, and Walter Archer. 2010. The first decade of the community of inquiry framework: A retrospective. *The Internet and Higher Education* 13, 1-2 (2010), 5–9.

[21] D Randy Garrison and Martha Cleveland-Innes. 2005. Facilitating cognitive presence in online learning: Interaction is not enough. *The American journal of distance education* 19, 3 (2005), 133–148.

[22] D Randy Garrison, Martha Cleveland-Innes, and Tak Shing Fung. 2010. Exploring causal relationships among teaching, cognitive and social presence: Student perceptions of the community of inquiry framework. *The internet and higher education* 13, 1-2 (2010), 31–36.

[23] Dragan Gašević, Srećko Joksimović, Brendan R Eagan, and David Williamson Shaffer. 2019. SENS: Network analytics to combine social and cognitive perspectives of collaborative learning. *Computers in Human Behavior* 92 (2019), 562–577.

[24] Greg Hamerly and Charles Elkan. 2004. Learning the k in k-means. *Advances in neural information processing systems* 16 (2004), 281–288.

[25] Friedrich Hesse, Esther Care, Juergen Buder, Kai Sassenberg, and Patrick Griffin. 2015. Assessment and Teaching of 21st Century Skills. *The Internet and Higher Education* (2015), 37–56.

[26] Srećko Joksimović, Dragan Gašević, Vitomir Kovanović, Bernhard E Riecke, and Marek Hatala. 2015. Social presence in online discussions as a process predictor of academic performance. *Journal of Computer Assisted Learning* 31, 6 (2015), 638–654.

[27] Srećko Joksimović, Jelena Jovanović, Vitomir Kovanović, Dragan Gašević, Nikola Milikić, Amal Zouaq, and Jan Paul Van Staalduinen. 2020. Comprehensive analysis of discussion forum participation: from speech acts to discussion dynamics and course outcomes. *IEEE Transactions on Learning Technologies* 13, 1 (2020), 38–51.

[28] Min Kyu Kim and Tuba Ketenci. 2019. Learner participation profiles in an asynchronous online collaboration context. *The Internet and Higher Education* 41 (2019), 62–76.

[29] Vitomir Kovanović, Dragan Gašević, Srećko Joksimović, Marek Hatala, and Olusola Adesope. 2015. Analytics of communities of inquiry: Effects of learning technology use on cognitive presence in asynchronous online discussions. *The Internet and Higher Education* 27 (2015), 74–89.

[30] Vitomir Kovanović, Srecko Joksimović, Dragan Gašević, and Marek Hatala. 2014. What is the source of social capital? The association between social network position and social presence in communities of inquiry. In *Workshop on Graph-Based Educational Data Mining.* 1.

[31] Vitomir Kovanović, Srećko Joksimović, Oleksandra Poquet, Thieme Hennis, Iva Čukić, Pieter De Vries, Marek Hatala, Shane Dawson, George Siemens, and Dragan Gašević. 2018. Exploring communities of inquiry in massive open online courses. *Computers & Education* 119 (2018), 44–58.

[32] Vitomir Kovanović, Srećko Joksimović, Oleksandra Poquet, Thieme Hennis, Pieter de Vries, Marek Hatala, Shane Dawson, George Siemens, and Dragan Gašević. 2019. Examining communities of inquiry in Massive Open Online Courses: The role of study strategies. *The Internet and Higher Education* 40 (2019), 20–43.

[33] Nale Lehmann-Willenbrock, Stephenson J Beck, and Simone Kauffeld. 2016. Emergent team roles in organizational meetings: Identifying communication patterns via cluster analysis. *Communication Studies* 67, 1 (2016), 37–57.

[34] José-Antonio Marcos-García, Alejandra Martínez-Monés, and Yannis Dimitriadis. 2015. DESPRO: A method based on roles to provide collaboration analysis support adapted to the participants in CSCL situations. *Computers & Education* 82 (2015), 335–353.

[35] Roberto Martinez-Maldonado, Dragan Gaševic, Vanessa Echeverria, Gloria Fernandez Nieto, Zachari Swiecki, and Simon Buckingham Shum. 2021. What Do You Mean by Collaboration Analytics? A Conceptual Model. *Journal of Learning Analytics* 8, 1 (2021), 126–153.

[36] Rosa M Mayordomo and Javier Onrubia. 2015. Work coordination and collaborative knowledge construction in a small group collaborative virtual task. *The Internet and Higher Education* 25 (2015), 96–104.

[37] Valter Neto, Vitor Rolim, Anderson Pinheiro Cavalcanti, Rafael Dueire Lins, Dragan Gasevic, and Rafael Ferreiramello. 2021. Automatic Content Analysis of Online Discussions for Cognitive Presence: A Study of the Generalizability across Educational Contexts. *IEEE Transactions on Learning Technologies* (2021).

[38] Nataša Pantić, Sarah Galey, Lani Florian, Srećko Joksimović, Gil Viry, Dragan Gašević, Helén Knutes Nyqvist, and Krystallia Kyritsi. 2021. Making sense of teacher agency for change with social and epistemic network analysis. *Journal of Educational Change* (2021), 1–33.

[39] Luc Paquette, Theodore Grant, Yingbin Zhang, Gautam Biswas, and Ryan Baker. 2021. Using Epistemic Networks to Analyze Self-regulated Learning in an Open-Ended Problem-Solving Environment. In *Proceedings of the 2nd International Conference on Quantitative Ethnography*, Andrew R. Ruis and Seung B. Lee (Eds.). Springer International Publishing, Cham, 185–201.

[40] Abelardo Pardo, Kathryn Bartimote, Simon Buckingham Shum, Shane Dawson, Jing Gao, Dragan Gašević, Steve Leichtweis, Danny Liu, Roberto Martínez-Maldonado, Negin Mirriahi, et al. 2018. OnTask: Delivering data-informed, personalized learning support actions. *Journal of Learning Analytics* 5, 3 (2018), 235–249.

[41] Oleksandra Poquet and Jelena Jovanovic. 2020. Intergroup and interpersonal forum positioning in shared-thread and post-reply networks. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge.* 187–196.

[42] Francesca Pozzi. 2011. The impact of scripted roles on online collaborative learning processes. *International Journal of Computer-Supported Collaborative Learning* 6, 3 (2011), 471–484.

[43] Anika Radkowitsch, Freydis Vogel, and Frank Fischer. 2020. Good for learning, bad for motivation? A meta-analysis on the effects of computer-supported collaboration scripts. *International Journal of Computer-Supported Collaborative Learning* 15, 1 (2020), 5–47.

[44] Jennifer Richardson, Karen Swan, Patrick Lowenthal, and Phil Ice. 2016. Social presence in online learning: Past, present, and future. In *Global Learn*. Association for the Advancement of Computing in Education (AACE), 477–483.

[45] Jennifer C Richardson, Yukiko Maeda, Jing Lv, and Secil Caskurlu. 2017. Social presence in relation to students' satisfaction and learning in the online environment: A meta-analysis. *Computers in Human Behavior* 71 (2017), 402–417.

[46] Liam Rourke, Terry Anderson, D. Randy Garrison, and Walter Archer. 1999. Assessing Social Presence In Asynchronous Text-based Computer Conferencing. *The Journal of Distance Education* 14, 2 (1999), 50–71. https://www.learntechlib.org/p/92000/

[47] Liam Rourke, Terry Anderson, D. Randy Garrison, and Walter Archer. 2007. Assessing Social Presence In Asynchronous Text-based Computer Conferencing. *The Journal of Distance Education / Revue de l'Éducation à Distance* 14, 2 (2007), 50–71. http://eric.ed.gov/?id=EJ616753

[48] Graeme D Ruxton. 2006. The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test. *Behavioral Ecology* 17, 4 (2006), 688–690.

[49] Mohammed Saqr and Olga Viberg. 2020. Using Diffusion Network Analytics to Examine and Support Knowledge Construction in CSCL Settings. In *In Proceedings of the 15th European Conference on Technology Enhanced Learning*. Springer, Cham, Switzerland, 158–172.

[50] Tammy Schellens, Hilde Van Keer, Bram De Wever, and Martin Valcke. 2007. Scripting by assigning roles: Does it improve knowledge construction in asynchronous discussion groups? *International Journal of Computer-Supported Collaborative Learning* 2, 2 (2007), 225–246.

[51] Bertrand Schneider, Nia Dowell, and Kate Thompson. 2021. Collaboration Analytics—Current State and Potential Futures. *Journal of Learning Analytics* 8, 1 (2021), 1–12.

[52] David Williamson Shaffer. 2004. Epistemic Frames and Islands of Expertise: Learning from Infusion Experiences. In *Proceedings of the 6th International Conference on Learning Sciences (ICLS '04)*. International Society of the Learning Sciences, Santa Monica, California, 473–480.

[53] David Williamson Shaffer, Wesley Collier, and Andrew R Ruis. 2016. A tutorial on epistemic network analysis: Analyzing the structure of connections in cognitive, social, and interaction data. *Journal of Learning Analytics* 3, 3 (2016), 9–45.

[54] David Williamson Shaffer, David Hatfield, Gina Navoa Svarovsky, Padraig Nash, Aran Nulty, Elizabeth Bagley, Ken Frank, André A Rupp, and Robert Mislevy. 2009. Epistemic network analysis: A prototype for 21st-century assessment of learning. *International Journal of Learning and Media* 1, 2 (2009).

[55] Artur Starczewski and Adam Krzyżak. 2015. Performance evaluation of the silhouette index. In *International Conference on Artificial Intelligence and Soft Computing*. Springer, 49–58.

[56] Jan-Willem Strijbos and Armin Weinberger. 2010. Emerging and scripted roles in computer-supported collaborative learning. *Computers in Human Behavior* 26, 4 (2010), 491–494.

[57] Jenny Thomas. 2013. Exploring the use of asynchronous online discussion in health care education: A literature review. *Computers & Education* 69 (2013), 199–215.

[58] Howard T Welser, Eric Gleave, Danyel Fisher, and Marc Smith. 2007. Visualizing the signatures of social roles in online discussion groups. *Journal of social structure* 8, 2 (2007), 1–32.

[59] Alyssa Friend Wise and Ming Ming Chiu. 2011. Analyzing temporal patterns of knowledge construction in a role-based online discussion. *International Journal of Computer-Supported Collaborative Learning* 6, 3 (2011), 445–470.

[60] Alyssa Friend Wise and Ming Ming Chiu. 2014. The impact of rotating summarizing roles in online discussions: Effects on learners' listening behaviors during and subsequent to role assignment. *Computers in human behavior* 38 (2014), 261–271.

[61] Alyssa Friend Wise, Marzieh Saghafian, and Poornima Padmanabhan. 2012. Towards more precise design guidance: Specifying and testing the functions of assigned student roles in online discussions. *Educational Technology Research and Development* 60, 1 (2012), 55–82.

[62] Alyssa Friend Wise and Baruch B Schwarz. 2017. Visions of CSCL: Eight provocations for the future of the field. *International Journal of Computer-Supported Collaborative Learning* 12, 4 (2017), 423–467.

[63] Rui Xu and Don Wunsch. 2008. *Clustering*. Vol. 10. John Wiley & Sons.

APÊNDICE G – ADOPTING LEARNING ANALYTICS TO PROMOTE COLLABORATION IN ONLINE DISCUSSIONS WRITTEN IN PORTUGUESE

Adopting Learning Analytics to Promote Collaboration in Online Discussions Written in Portuguese

**[α]Máverick André, [α]Anderson Pinheiro, [α]Vitor Rolim, [β]Rafael Dueire Lins,
[β]Taciana Pontual Falcão, [β]Rafael Ferreira Mello**

[α]Universidade Federal de Pernambuco, Centro de Informática
[β]Universidade Federal Rural de Pernambuco, Departamento de Computação

Corresponding author: rafael.mello@ufrpe.br

Abstract. Learning Management Systems (LSM) have been adopted to provide interaction between instructors and students in online learning courses. In these environments, large volumes of messages are exchanged, making it difficult for the participants to follow all the discussions.
When messages are ignored, students' opportunities to effectively construct knowledge through online discussions are reduced. In this context, this paper proposes a method for mining the content of messages written in Brazilian Portuguese to enhance students and instructors' interactions in online discussion forums. Four different automatic methods were developed, one for each of the following four dimensions of the originality and collaboration processes: (i) Semantic textual similarity to measure originality; (ii) Expressing appreciation towards other participants; (iii) Recognizing group presence; (iv) Sharing information and resources; (v) Soliciting feedback/Answering questions. The use of the methods for analyzing online discussions forums was validated in a case study conducted in a real course. A discussion forum tool was developed to provide feedback for students and instructors about the originality and collaboration level during online discussions. Results showed the potential of the proposal in promoting students' collaboration, enhancing interaction and reducing plagiarism in discussion forums.

André, M., Pinheiro, A., Rolim, V., Lins, R.D., Falcão, T.P., Mello, R.F. (2021) Adopting Learning Analytics to Promote Collaboration in Online Discussions Written in Portuguese. Computer-Based Learning in Context, 3 (2), 1-15.

Keywords: Learning Analytics, Online Discussion, Natural Language Processing, Collaboration, Originality Detection

1 Introduction

Learning Analytics (LA) can offer practical solutions to analyze the massive amount of data generated in Learning Management Systems (LMS), for example by applying automatic methods to extract information from the LMS generated data (Lang et al., 2017). It can be used to improve student engagement and performance (Coffrin et al., 2014), measuring the time elapsed by students in performing tasks (Kovanović et al., 2017a), and to model students' "behaviour" and participation (KickmeierRust et al., 2016). In its early days, much research in LA focused on extracting information from log data about the interaction with technology. However, recent research has focused on analyzing the textual contents of educational resources, such as analyzing essays, the production of academic texts, answers to open-ended questions, and online discussion forums (Lárusson and White, 2012; Simsek et al., 2015; Dascalu et al., 2015; Kovanović et al., 2015).

Among the resources available in a LMS, educational forums are widely used to encourage student participation, to answer questions, and to share resources (Hew and Cheung, 2008), being an asynchronous tool that can enhance collaboration in online learning courses (Kovanović et al., 2017b). Although the adoption of discussion forums can bring many benefits, a large volume of posts makes it difficult for the instructor and students to keep up with all interactions and leverage opportunities for social knowledge construction (Marbouti and Wise, 2016). Thus, a relevant issue arises: "*Can large volumes of online discussion data be automatically analyzed to provide meaningful insights for the students and instructors?*" There are different approaches to deal with this problem. One line of research proposes methods that analyze interactions within online educational discussion using shallow features, which basically uses statistics about the interactions and simple text analysis. For instance, methods based on direct statistics (e.g., number of posts and replies) and word cloud (Moreno-Marcos et al., 2018; Hu et al., 2018). In this case, the methods do not consider deeper text analysis such as text semantics.

Another possibility is to adopt natural language processing methods, such as sentiment analysis (Wen et al., 2014), text classification (Kovanović et al., 2016), and text summarization (Bhatia et al., 2014). Although the works cited produced significant results, they were primarily designed for texts in English only. The mapping of such methods to another language, such as Brazilian Portuguese, is not always easy or possible due to the absence of text mining tools and resources for languages other than English, e.g. LIWC (Linguistic Inquiry and Word Count) (Tausczik and Pennebaker, 2010) and Coh-Metrix (Graesser et al., 2004; Barbosa et al., 2021).

The present work aimed to provide methods for analyzing large volumes of collaborative discourse among students in educational forums written in Brazilian Portuguese, thus contributing towards a broader support for learning analytics in different languages. Particularly, this work focused on the analysis of collaborative indicators within an online discussion. Moreover, we also analyzed the messages posted by the students in terms of the originality of the contributions. In order to accomplish this, we proposed:

i. Using machine learning and rule-based approaches to automatically identify the four analytics for the detection of collaborative indicators in discussion transcripts in Brazilian Portuguese namely, expressing appreciation towards other participants, recognizing group presence, sharing information and resources and asking for feedback/answering questions.

ii. Adapting a measure for semantic textual similarity originally developed in English to Brazilian Portuguese. Such measure allowed assessing the level of novelty of a post by comparing it with the preceding messages in the forum.

2 Background

2.1 **Discussion forums.** An educational discussion forum is an asynchronous communication resource often used in LMS to promote discussions and interactions amongst participants about some educational topic (Pendry and Salvatore, 2015). It is a communication channel where students and instructors can share questions, opinions, comments and answers. Educational forums are usually mediated by instructors, who assist and supervise students in their learning process (De Wever et al., 2006; Garrison et al., 2010; Ferreira-Mello et al., 2019). Instructors can adopt the forum for:

- Encouraging the creation of bonds between students (Joksimovic et al., 2014)
- Developing the capacity for critical discussion about a theme or subject (Yoo and Kim, 2014)
- Answering questions and comments (Lin et al., 2009); and
- Assessing students (Rubio and Villalon, 2016).

Despite the advantages brought by adopting forums as communication resources in educational environments, it may also lead to the problem of information overload (Wulf et al., 2014), hindering students from effectively constructing knowledge (Marbouti and Wise, 2016), and instructors from being able to facilitate the online discussions (Garrison et al., 2010). Following up a large number of posts may become a heavy workload for the instructor, which, if not properly handled, may lower the motivation level of the students, leading to a decrease in the quantity and quality of their posts. These problems may be addressed by adopting the methods proposed in this paper to extract learning analytics related to collaboration and originality.

2.2 **Collaboration in educational forums.** Several researchers, over the last few years, have been working on strengthening the collaborative aspects of discussion forums in online learning. Murphy (2004) lists general aspects of collaboration that are relevant to educational forums, including: social presence; articulating individual perspectives; accommodating or reflecting the perspectives of others; co-constructing shared perspectives and meanings; building shared goals and purposes; and producing shared artefacts. Usually, the works to improve collaboration in educational forums address some of these aspects.

Garrison et al. (1999) highlight the importance of social interaction in online discussions with the concept of social presence, which is "the ability of participants in a community of inquiry to project themselves socially and emotionally, as 'real' people (i.e., their full personality), through the medium of communication being used" (p. 94). Social presence includes three categories with several indicators, which form a road-map to interpret the concept of interactions.

All indicators proposed by Murphy (2004) and Garrison et al. (1999) are relevant to analyze social interactions. However, four of them appear in both works and are listed as important psychological factors to improve social participation in online discussions (Amichai-Hamburger et al., 2016; Castellanos-Reyes, 2020):

- Expressing appreciation towards other participants: This binary indicator relates to students' expression of positive sentiments towards previous posts or forum participants.
- Recognizing group presence: This indicator relates to students' acknowledgement of the presence of peers.
- Sharing information and resources: This indicator relates to the identification of links or resources shared in the educational forum.
- Asking for feedback/Answering questions: This indicator is used to identify if students are answering the questions proposed by the instructors or raised by peers, or asking for help to understand the questions.

Therefore, the method we proposed aimed to automatically analyze those four different collaboration features to provide useful analytics to students and instructors. Moreover, we provided a learning analytics visualization tool to promote self-regulation among students.

2.3 **Approaches to online discussion analysis.** The literature offers several approaches to using learning analytics in the context of online discussions (Joshi and Rosé, 2007; Yen, 2013; Kim et al., 2016; Wise et al., 2014; Ferreira et al., 2020), with some emerging work on methods for automated analysis of online discussions (Mu et al., 2012; Kovanović et al., 2016). For example, text classification can be used for assessing online participation of students and monitoring of the learning progress through collaboration, among others (Lui et al., 2007). Moreover, Rosé et al. (2017) summarized several applications and the importance of collaboration within online discussion in terms of student assessment and to support instructors' decision-making.

Profiling students' interactions in threaded discussions is one of the most relevant applications of text analysis in learning analytics. Ravi and Kim (2007) proposed the classification of students' posts individually as "discourse acts" which include the following categories: (i) complement (complement of a previous message); (ii) information (information, command or announcement); (iii) correction (correction or objection to an earlier message); (iv) elaboration (elaboration of an earlier message or description, including the elaboration of questions and answers); (v) question (a question about a problem, including questions about previous messages); and (vi) answer (response to an earlier question or suggestion). They used a set of n-gram (n varying from 1 to 4) and a Support Vector Machine (SVM) classifier to categorize the posts. In the best case, the method reached an accuracy of 92.70%.

Recognition of the kind of message (genre) in online discussion posts is another important topic that has been addressed in the literature. Lin et al. (2009) proposed a system that classifies the genre of a post using a bag of words (using all words, and only verbs, adverb, or contextual words) as features and the decision tree algorithm for classification. The genres that this system sought to identify were: (i) announcements (clarifies doubts); (ii) questions (proposes a question); (iii) interpretation (interpretation based on facts and ideas exposed); (iv) conflict (conflicting opinion); (v) affirmation (maintaining and defending the idea disagreed by others); and (vi) others (varied messages that are difficult to categorize). Better results were achieved using the combination of verbs with contextual words and adverbs with contextual words.

In addition to the classification of posts, the extraction of indicators from forums that may aid instructors in following up on the discussion is another topic considered in the literature. McLaren et al. (2007) proposed the use of Awareness Indicators to provide an interface that allowed the instructor to supervise all of the online discussion. This indicator classified the posts as positive or negative, which meant whether the student needed help or not. Although they used machine learning, the features (e.g., text length, shape type, number of in-links, number of out-links, and number of indirect links) were provided manually. Different decision tree algorithms were evaluated, reaching an accuracy of 85.00%.

Most of the existing research has focused on the analysis of online discussions in English. Recently, Neto et al. (2018) proposed the adoption of LIWC and Coh-Metrix, which are text mining tools to measure text readability and coherence, to classify higher-level classes as the cognitive presence, in Portuguese. Moreover, Barbosa et al. (2020) and Barbosa et al. (2021) used the same tool to proposed approaches to evaluate texts in English and Portuguese according to cognitive and social presences. While Barbosa et al. (2020) presented Multilanguage classifier, Barbosa et al. (2021) applied text translation to create a model for both languages. Although both works provided interesting results, they offer classifiers for high-level categories and not for specific indicators. Moreover, there has been little attention devoted to developing systems that can transform the information extracted by text analysis into insights for the instructors and students.

This paper aimed to address those gaps by providing both: (i) a method for the automated analysis of student-generated content in online discussions in Brazilian Portuguese in detailed categories; and (ii) student and instructor dashboards with insights obtained from the automated analysis of the texts.

3 Methodology

This section presents the indicators extracted from online discussions to create the proposed dashboards. The unit of analysis for both levels was the message posted in an online discussion. Each post can contain more than one indicator of the analyzed dimensions. The following sections present further details on each of the proposed text analysis methods.

3.1 **Expressing appreciation towards other participants.** Sentiment Analysis (SA) techniques were used to analyze the degree of explicit friendliness in the students' messages in the forum. The SA algorithm proposed by (Cambria, 2016) was used here to identify if a text message was either positive, negative or neutral compared with the previous messages posted by other students in the forum. A traditional approach to SA adopts semantic resources (e.g., lexical dictionaries) and machine learning methods. Following this idea, a Brazilian Lexical Dictionary called SentiLex was used (Silva et al., 2012). SentiLex encompasses more than 7,000 words divided into different grammatical functions (verb, noun, adverb, adjective) and polarities (positive, negative, neutral) as one feature for a machine learning algorithm.

As the problem was to identify the degree of appreciation between the participants, this work focused on extracting the sentiment related to other students in the discussion. Thus, besides the SentiLex, other four features were used:

- ***Proper name***: As the goal was to identify the appreciation to the post of another participant, it was important to identify the proper names listed in the post. Thus, this feature stored the number of proper names found. To accomplish this, a Named Entity Recognizer (NER) system was used (Sarmento, 2006);
- ***Distance Name/Complement***: This feature was used to determine if an instance of appreciation is related to a proper name or not, by measuring the distance between the proper name and the positive word extract from SentiLex.
- ***Information Sharing***: This feature used the information sharing analytics (presented in "Sharing information and resources" section) to eliminate the appreciation related to a resource, by looking at compliments made by a participant to another;
- ***Post Size***: This feature captured the size of a post, in terms of the the number of characters. The idea was that longer posts tend to have more information, thus increasing the chance of being allusive to a previous post or another participant in the discussion.

The above features were applied to an SVM classifier to categorize posts according to their appreciation towards other participants. Such an SVM classifier was chosen because it reached good results in a wide range of applications (Fernández-Delgado et al., 2014).

3.2 **Recognizing group presence.** We proposed a rule-based approach to recognize Group Presence (Murphy, 2004). If a post adhered to any of the four rules below, the system classified it as group presence:

i. The first rule identified the existence of words tagged as an interjection in the analyzed post's first three positions. This strategy was adopted because interjections such as "*hi*" and "*hello*" customarily appear at the beginning of a conversation, indicating group presence. The implementation of this rule made use of the part-of-speech class algorithm available at Spacy[1].

ii. The second rule followed a very similar idea, but it identified the greetings at the beginning of a post instead of the interjection. For instance, expressions such as "good morning", "good afternoon" and "good evening".

iii. The third rule searched for plural personal pronouns (e.g., us and you) and treatment pronouns (e.g., Mr., Mrs., Prof.) in the message.

iv. Finally, the last rule extracted terms representing collectives in educational settings, such as colleagues, friends, and companions, that are classified as nouns. If one or more expressions were found, then the analyzed post was said to have the group presence indicator.

3.2 **Sharing information and resources.** Students often share information and resources by posting websites in online discussions. Thus, this feature was based on a link identification approach. Moreover, an automatic approach was used to check if the links retrieved were related to the forum's topics.

[1] https://spacy.io/usage/linguistic-features

Initially, we used a regular expression to verify if the post contained any URL in its body. Then, the link was validated to check if it was still active. Next, based on the content of the web page, a TFISF (Term Frequency-Inverse Sentence Frequency) was performed to find the main concepts from the retrieved content (Neto et al., 2000). The computation of the TFISF followed the same idea of TFIDF presented in Section Originality, but here instead of calculating the TFIDF of a document, we used a sentence. The same TFISF analysis was made with the nouns in the previous posts in the forum. If the intersection of both outputs was non-empty, we considered that the information shared was related to the proposed content for discussion in the forum. Otherwise, the instructor received an alert to check the content of the posts. It is important to mention that this feature only covered resources written in Portuguese.

3.3 **Asking for feedback/Answering questions.** This module aimed to classify the posts in discussion forums in: asking for feedback (questions), answering questions (answer), and neutral comment. To accomplish that, different features were used to train an SVM classifier (Joachims, 2002). A Genetic Algorithm (GA) (Mitchell, 1998) was used in finding the best set of parameters for the SVM classifier.

In the feature extraction step, the information required for the classification process was retrieved from the posts. The feature extraction used CoGrOO (Silva and Finger, 2013) to perform a part of speech tagging in each post. Next, word vectors of each class (questions, answer and neutral) were created. To avoid vectors with high dimension, only verbs, nouns, pronouns, adjectives and adverbs were considered.

The word vectors were ranked decreasingly concerning the representativeness (measured by TFIDF), for each grammatical class, at the beginning of the vectors. Here, TF was the number of times a word appeared performing the same grammatical function within the universe of postings of a specific class (questions, neutral or answer); and IDF measured the relationship between the number of times a word appeared in a specific class divided by the total number of posts that had this word. Finally, for each post, a vector was generated using 15 features, encompassing the pair $< class, grammatical\ class >$, where the classes were question, answer and neutral; and the grammatical classes were verb, noun, pronoun, adjective and adverb. Each position of the vector received the number of words with the same grammatical function, which appeared in the post and in one of the word vector classes.

After the feature extraction, the postings' classification was performed through the application of an SVM classifier. To improve the results of the classifiers, the parameters were optimized using a GA.

The SVM received the feature vectors and performed the training using each set of parameters generated by the GA until a termination condition was met. The GA had its initial population represented by solutions (individuals) whose genetic material consisted of eight genes referred to the SVM parameters: $SVMType$, $probability$, $kernelType$, $gamma$, nu, $cacheSize$, $cost$, and $epsilon$.

The execution cycle of the GA was performed with the following steps:

i. ***Initial population generation***: initialization of the algorithm with individuals (solutions) with random parameters (genes)
ii. ***Evaluation of individuals***: running an SVM algorithm on the set of parameters of each individual of the population.

Each GA solution's objective was to parameterize the SVM to classify the posts. For each execution, the F-measure was calculated and used as the function to be optimized; in GA, the F-measure was the “fitness" of each individual.

To create the new generations of individuals, the following methods were used (Melanie, 1999):

i. ***Crossover***: aimed to mix the genetic material of the selected individuals (parents) with the purpose of generating new solutions (children) capable of exploring new spaces in the search field, using uniform crossing;
ii. ***Mutation***: inserted diversity into the population using the change (mutation) of some genes, given a probability defined a priori.

The last step was to select the surviving individuals. In this step, the steady-state method (Melanie, 1999) was used, which consists of selecting the best solutions within a universe constituted by the previous population and the population acts.

The final output of this module used an optimized SVM to classify the posts into question, answer and neutral. It was used to measure the degree of collaboration in asking for feedback and answers to questions.

3.4 **Originality.** The semantic textual similarity measure for Brazilian Portuguese proposed in (Cavalcanti et al., 2017) was adapted to educational settings to detect the level of originality of the post in a forum. It measures the degree of similarity between posts in the entire forum. To accomplish that, four features were extracted: TFIDF similarity, Word2Vec similarity, binary similarity, and the size of sentences. Details about each of the features are presented in the following subsections.

- **TFIDF.** The TFIDF (Term Frequency-Inverse Document Frequency) is a statistical measure that stands for the importance of a word in a set of documents (i.e., a set of forum messages in this case) (Ricardo, 1999), widely used in natural language processing. Two text processing steps were used: (i) each word was expanded using two synonyms from TeP (Thesaurus for Brazilian Portuguese) (Aluísio et al., 2008). As there is typically a number of words in a forum message, the values of TFIDF are typically too small. (ii) application of a stemming algorithm to reduce the sparsity of the data (Hartmann, 2016). The final similarity between two posts was calculated by applying the Cosine distance between their TFIDF vectors.
- **Word2vec.** The Word2vec model makes use of a neural network to build “classes of equivalence" of a given word. Through training, Word2vec translates texts into a numerical Kdimensional vector space. Each word in a text is represented as a vector, allowing to measure the degree of similarity between words as the distance between two vectors (Mikolov et al., 2013). The Word2vec model used in this paper was built using the original implementation of Word2vec[2] based on Wikipedia[3] and news texts obtained from the G1[4] portal from September 15, 2016, to December 5, 2016. The default parameters for word2vec training were adopted in the current study. At this step, the stopwords were removed from each of the posts, which underwent lemmatization before the similarity calculation. Then, the matrix similarity method presented in (Ferreira et al., 2016) was applied.
- **The Binary Similarity Matrix.** The method detailed in (Ferreira et al., 2016) of the similarity matrix between sentences was used here, but the similarity values between words were set to 1 if the words were equal and to 0, otherwise.
- **Size of Posts.** Based on the idea that two posts with a different number of words potentially convey different information, an additional feature representing the size of the post was added (Zhao et al., 2014). To obtain a value that represented the size of the post (Post Size), the number of words of the shortest post (shortPost) was divided by the number of words of the longest post (longPost), as shown in Equation 1. Before applying this method, the stopwords were removed.

$$PostSize = \frac{shortPost}{longPost} \tag{1}$$

- **Linear Regression.** A multiple linear regression model was applied to the four similarity measures outlined in previous sections to obtain the final similarity value between the posts. The linear regression step aimed to verify a functional relationship between a dependent variable and one or more independent variables (Seber and Lee, 2012).

In the educational scenario proposed, the final similarity measure was used to identify the degree of similarity among posts from an online discussion forum. In such a case, the originality of a message is inversely proportional to the degree of similarity between a message and all the other posts in a forum (Sánchez-Martí et al., 2018). Thus, a higher degree of similarity means a lower degree of originality.

4 Evaluation

4.1 **Context and Data.** The proposed methods were evaluated on a database containing 600 posts from a “*Programming Language*" module of a Computer Science undergraduate online course held at a Federal University in Brazil in 2017. The forum was designed to account for 20% of the students’ final mark. A total of 35 students posted messages over four weeks of discussions. The course was specifically redesigned to promote online discussions with an emphasis on the originality of discussions and collaboration. The pedagogical goals of the forum were to be a Question & Answering (Q&A) virtual space, and also to promote discussions among the students. In this context, the

[2] http://code.google.com/p/word2vec

[3] https://dumps.wikimedia.org/ptwiki/20160920/

[4] http://g1.com.br/

student had to post at least two messages a week (one question and one reply) in order to increase the interactivity in the discussion. In the end, only 10% of the messages were posted by the instructor.

Two expert coders manually analyzed each post from the database, and a third person acted as a referee to solve occasional divergences. The posts were categorized according to the presence or absence of each indicator considered: Expressing appreciation towards other participants; Recognizing group presence; and Sharing information and resources. Table 1 shows the distribution of the category of the posts and the Cohen's Kappa coefficient (κ) (Cohen, 1960) of agreement between the evaluators.

Table 1. Post distribution in the evaluation dataset

Analytic	Presence	Absence	κ
Appreciation	78	522	0.86
Group presence	19	581	0.75
Sharing information	17	583	0.93

In addition to the experiment, a case study was performed in which the learning analytics extracted was applied in a real-time application to assess students' participation in the online discussion. An educational forum, called iFórum, was developed to present the extracted learning analytics for instructors and students. This case study was applied in a class of "*Advanced topics in artificial intelligence*" with 12 students in the last year of a Computer Science course at the Federal University in Brazil, during the first semester of 2017. During this course, the students discussed scientific papers sent to the forum by the instructor. The course guidelines recommended that the students presented their views on the paper. The students had one week to make posts for each paper assigned. The online discussion accounted for 20% of the final mark.

4.2 **Experiments.** The following evaluation measures were used to assess the proposal: (i) precision: the number of correctly predicted classes divided by the number of all predicted classes; (ii) recall: the number of correctly predicted classes divided by the number of all instances that should have been classified; and (iii) F-measure: the uniform harmonic mean of the precision and the recall (Forman, 2003).

As presented in the Method section, the proposed approach used an SVM classifier to identify the appreciation and a rulebased method to extract the group presence and the sharing of information. Table 2 shows the results obtained with the proposed approach.

Table 2. Proposed Methods Results

Analytic	Precision	Recall	F-Measure
Appreciation	94.50	94.70	94.60
Group presence	80.70	76.60	78.50
Sharing information	84.60	81.10	82.80

All features obtained a performance above 75% of F-measure, reaching up to 94.60% for appreciation, 78.50% for Group Presence and 82.80% for sharing information. It is important to note that the database was unbalanced due to the nature of the forum, in which most of the posts were questions or answers. Despite that, the proposed approach reached good results.

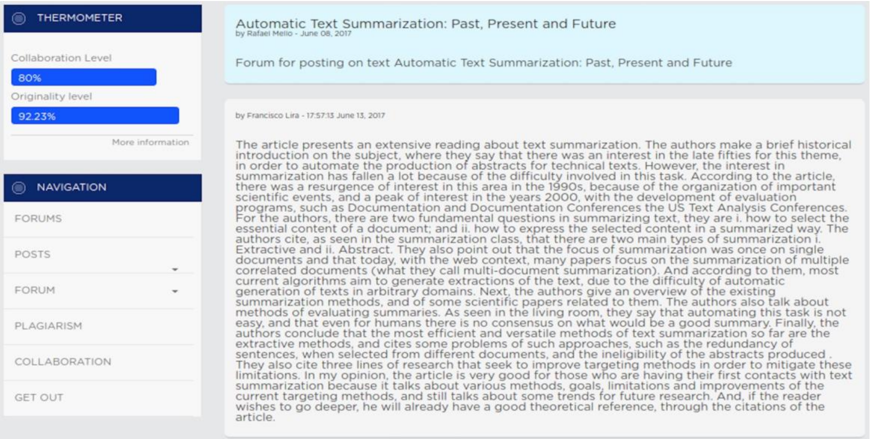
Moreover, as presented in the Method section, the problem of identifying "request for feedback" and "answering to questions" was translated into a classification problem regarding questions, answers and neutral comments. The same dataset was used to evaluate such an indicator. The 600 posts were divided into 60 posts for questions, 96 neutral comments and 444 answers. The coding was performed using the same methodology previously described, and the agreement reached $\kappa = 0.89$. Two scenarios were developed to evaluate the proposed classification: Scenario 1 (S1) - classification of posts using the features proposed; Scenario 2 (S2) - classification of posts using a TDIDF vector as features. For each scenario, the results were evaluated using a simple SVM classifier with default parameters and adopting the Genetic Algorithm (GA) to optimize the SVM parameters. The evaluation was performed using a 10-fold cross-validation approach.

Table 3 presents the results of the F-measure for each scenario. As the classes “question" and “answer" were equally important, the table shows the F-measure for each class.

Table 3. Question/Answer Evaluation: F-Measure

Method	Questions	Neutral	Answers
S1 (SVM)	87.40	97.90	98.70
S1 (SVM+GA)	97.80	99.70	98.80
S2 (SVM)	28.30	99.00	95.20
S2 (SVM+GA)	54.90	99.40	96.40

The results show that the adoption of the GA enhanced the F-measure for all the classes. We highlight the improvement in the class “question", raised from 87.40% to 97.80% and from 28.30% to 54.90% for S1 and S2, respectively. Besides, S1 outperformed S2 in all cases. For example, S1 improved the result of the identification of “question" in 77.04%, showing the proposed method’s efficacy. The combination S1 (SVM+GA) reached better results than the others.



Automatic Text Summarization: Past, Present and Future
by Rafael Mello - June 08, 2017

Forum for posting on text Automatic Text Summarization: Past, Present and Future

by Francisco Lira - 17:57:13 June 13, 2017

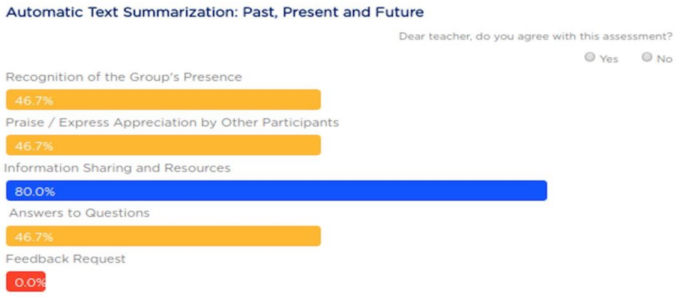
The article presents an extensive reading about text summarization. The authors make a brief historical introduction on the subject, where they say that there was an interest in the late fifties for this theme, in order to automate the production of abstracts for technical texts. However, the interest in summarization has fallen a lot because of the difficulty involved in this task. According to the article, there was a resurgence of interest in this area in the 1990s, because of the organization of important scientific events, and a peak of interest in the years 2000, with the development of evaluation programs, such as Documentation and Documentation Conferences the US Text Analysis Conferences. For the authors, there are two fundamental questions in summarizing text, they are i. how to select the essential content of a document; and ii. how to express the selected content in a summarized way. The authors cite, as seen in the summarization class, that there are two main types of summarization i. Extractive and ii. Abstract. They also point out that the focus of summarization was once on single documents and that today, with the web context, many papers focus on the summarization of multiple correlated documents (what they call multi-document summarization). And according to them, most current algorithms aim to generate extractions of the text, due to the difficulty of automatic generation of texts in arbitrary domains. Next, the authors give an overview of the existing summarization methods, and of some scientific papers related to them. The authors also talk about methods of evaluating summaries. As seen in the living room, they say that automating this task is not easy, and that even for humans there is no consensus on what would be a good summary. Finally, the authors conclude that the most efficient and versatile methods of text summarization so far are the extractive methods, and cites some problems of such approaches, such as the redundancy of sentences, when selected from different documents, and the ineligibility of the abstracts produced . They also cite three lines of research that seek to improve targeting methods in order to mitigate these limitations. In my opinion, the article is very good for those who are having their first contacts with text summarization because it talks about various methods, goals, limitations and improvements of the current targeting methods, and still talks about some trends for future research. And, if the reader wishes to go deeper, he will already have a good theoretical reference, through the citations of the article.

Figure 1. Proposed educational forum.

4.2 **Case Study.** This section presents the tool created to support the collaboration in online discussion and the results of the proposed case study. Figure 1 presents an overview of the proposed tool. It shows the forum subject, the first post in the forum, a navigation menu and the “thermometer", which allowed to visualize the results of the learning analytics of the extracted indicators. This interface indicates the potential practical value of the developed algorithms in a real course. In terms of visualization for students, the novelty proposed in this forum was the inclusion of the two different “thermometer" bars (in the top-left part of figure 1) showing:

i. the student’s collaboration level, which encompassed the four collaboration features proposed. The final value of this level was measured by dividing the number of posts containing each of the collaboration features by the total number of posts;
ii. the level of originality, which is inversely proportional to the level of similarity between the students’ posts.

Those “thermometers” bars were visible to all participants (students and instructors) and intended to stimulate self-regulation among the students, especially for enhancing collaborative discussion and the quality of posts, avoiding similar messages. Moreover, an increase in the number of messages was also expected.



Automatic Text Summarization: Past, Present and Future

Dear teacher, do you agree with this assessment?

○ Yes ○ No

Recognition of the Group's Presence

46.7%

Praise / Express Appreciation by Other Participants

46.7%

Information Sharing and Resources

80.0%

Answers to Questions

46.7%

Feedback Request

0.0%

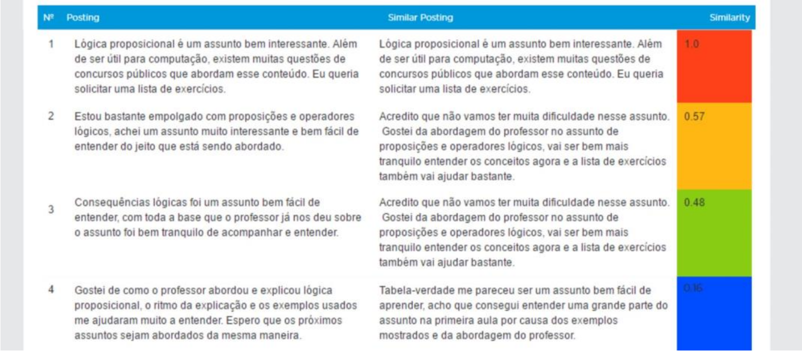
Figure 2. Collaboration per forum.

The proposed tool also contained learning analytics that could be accessed only by instructors, with details on the collaboration analytics per forum and per student and the similarity among posts. Figure 2 presents the level of collaboration analytics from a specific forum. It was also possible to visualize the same information for each student. Such information could aid the formative evaluation and to measure student participation in the online discussion.

Figure 3 presents the analytics about the similarity in the forum, which could be accessed only by instructors. For each post, it listed the most similar post and the similarity value. The similarities were split into 4 groups: blue (0-0.3), green (0.310.5), yellow (0.51-0.7) and red (above 0.7). Those groups were defined by the instructor responsible for this case study, but they can be configured according to the course's design. The colors blue and green meant that the posts did not bear much resemblance to the previous posts in the forum, i.e. they were potentially original posts. The yellow and red colors meant that the posts had /higher levels of similarity to previous posts.

In the first week of discussions (week1), a traditional forum using the Moodle platform was adopted to discuss the articles. At the end of week1, only ten posts were recorded. Two students did not participate in the discussions, and all others posted only once. It is important to mention that the posts entered were simply a summary of the paper proposed for discussion.

The second paper assignment was sent in the proposed tool using the learning analytics extracted. Again, the students had one week (week2) to interact in the forum. As a result, the number of posts on week2 was 30. Besides that, all the students enrolled in the course participated in the forum, and there was a more significant number of interactions between the students, counting not only posts that summarized the article but also texts that included all four categories of collaboration identified by the proposed approach. Each student posted at least twice.



Nº	Posting	Similar Posting	Similarity
1	Lógica proposicional é um assunto bem interessante. Além de ser útil para computação, existem muitas questões de concursos públicos que abordam esse conteúdo. Eu queria solicitar uma lista de exercícios.	Lógica proposicional é um assunto bem interessante. Além de ser útil para computação, existem muitas questões de concursos públicos que abordam esse conteúdo. Eu queria solicitar uma lista de exercícios.	1.0
2	Estou bastante empolgado com proposições e operadores lógicos, achei um assunto muito interessante e bem fácil de entender do jeito que está sendo abordado.	Acredito que não vamos ter muita dificuldade nesse assunto. Gostei da abordagem do professor no assunto de proposições e operadores lógicos, vai ser bem mais tranquilo entender os conceitos agora e a lista de exercícios também vai ajudar bastante.	0.57
3	Consequências lógicas foi um assunto bem fácil de entender, com toda a base que o professor já nos deu sobre o assunto foi bem tranquilo de acompanhar e entender.	Acredito que não vamos ter muita dificuldade nesse assunto. Gostei da abordagem do professor no assunto de proposições e operadores lógicos, vai ser bem mais tranquilo entender os conceitos agora e a lista de exercícios também vai ajudar bastante.	0.48
4	Gostei de como o professor abordou e explicou lógica proposicional, o ritmo da explicação e os exemplos usados me ajudaram muito a entender. Espero que os próximos assuntos sejam abordados da mesma maneira.	Tabela-verdade me pareceu ser um assunto bem fácil de aprender, acho que consegui entender uma grande parte do assunto na primeira aula por causa dos exemplos mostrados e da abordagem do professor.	0.16

Figure 3. Similarity among posts from forum.

Table 4. Question/Answer Evaluation: F-Measure

	Analytic	Number of Posts	Percentage of Posts
Week1	Appreciation	00	00.00%
	Group presence	03	30.00%
	Sharing information	00	00.00%
Week 2	Appreciation	06	20.00%
	Group presence	13	43.33%
	Sharing information	13	43.33%

Table 4 presents the increase in the number and percentage of collaborative posts from week1 to week2. It is important to observe that week1 and week2 had 10 and 30 posts, respectively. Moreover, Figure 4 shows students' individual behavior on week1 and week2 regarding the collaborative indicators. It is possible to see a change of behavior in the students' interactions after adopting the iForum tool. Every student enhanced the quality of their posts, in relation to the features analyses, in week2 (except student 2). This shows that the proposed visualization helped students not only to write more messages but also to incorporate the social skills analyzed.

Regarding the degree of originality of posts in the first week, the average similarity between posts was 0.28, with a standard deviation of 0.13; while in the second week (forum tool with the proposed measure), the mean was 0.27 with standard deviation 0.09. Although there was a reduction in the degree of originality, it did not represent a statistic difference. However, it is important to highlight that the level of plagiarism was lower with the use of the proposed tool, even though it had three times as many posts and almost ten times as many comparisons in the forum (the first and second week had 45 and 435 similarity comparisons, respectively). Moreover, in the second week, the posts had a higher degree of similarity with the instructor's first message.

As an automatic approach to improve originality, the tool sent a private message to the student who had written a post classified in the red group before its publication. It suggested that the student could re-write their post, referring to the previous messages (with a high degree of similarity).

Finally, the instructor used the learning analytics proposed to indicate the assignment of course grades of the students participating in the forum. The participation in this forum represented 10% of the final marks of the course. This 10% was determined using the automatically generated analytics and not the content of the post itself.

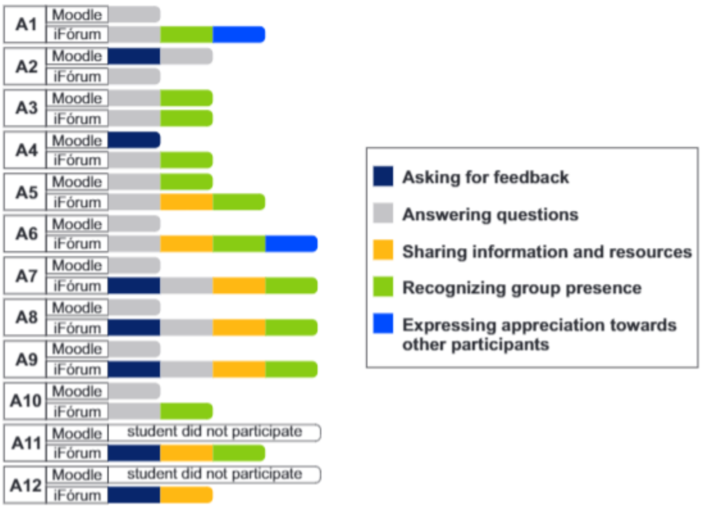


Figure 4. Behavior of students in the first week (Moodle) and second week (iFórum).

5 Discussion

It is important to notice that there is no other work in the literature that adopts automatic methods to identify: (i) expressing appreciation towards other participants, (ii) recognizing group presence and (iii) sharing information and resources for texts written in Portuguese. There are studies that proposed manual approaches to reach similar goals. For instance, a study that sought to identify collaboration in forums and chats was reported by (de Melo Ferreira et al., 2013), where two pairs of linguistics experts in Portuguese

were responsible for creating the rules based on a syntactic analysis. One of the pairs of experts analyzed the discussion forum and then proposed the rules. The second pair of linguists proposed rules without viewing the forum, creating more general rules. The results showed that the rules explicitly created for the forum obtained a better result with 93.95% accuracy, while the generic rules resulted only in 58.90% accuracy. The generic rules approach's low performance when creating the rules independently of the forum is the limitation of (de Melo Ferreira et al., 2013). The rule-based system was created to automatically generate learning analytics that allows the students to evaluate their performance and the instructors to measure the participants' performance in the discussion forums.

Gomes (2012) advocates a teaching-learning process centred on student interactions. He presents a system based on natural language processing techniques to analyze interactions in discussion forums automatically. Using the TFIDF approach, the postings were classified into the following classes: greeting, discussion, motivation, social, information, confirmation, negation, task, clarification, inquiry, and thanks. The highest accuracy obtained was 40% with the Bayesian classifier.

Recently, Ferreira et al. (2020) proposed a classifier based on structural features to develop a classifier for social presence in English. Although the authors reported an accuracy value up to 0.88, it would not be possible to reproduce the results for Portuguese directly due to the lack of resources are available to support text mining applications in Portuguese.

As an attempt to address the problem of limited resources for automatically analyze Portuguese Language, Barbosa et al. (2021) proposed the use of text translation techniques in combination with classification methods. The authors found that it was possible to increase the classification of online discussion messages in Portuguese by up to 55%, translating it to English before extracting the features.

There are two main differences between the approach proposed in the current paper and the previous works Barbosa et al. (2021); Ferreira et al. (2020): (i) both of the presented work relies somehow on English resources and data, which is not always available; (ii) the previous studies focused on high-level social presence indicators, while we proposed the analysis of low-level aspects.

Additionally, several works focus on the automatic cognitive presence for Portuguese (Barbosa et al., 2020; Neto et al., 2018). However, none of them had been reproduced to the analysis of low-level indicators of social presence.
Finally, the case study presented here showed that adopting the proposed learning analytics was associated with an increase in collaboration in educational forums in this scenario. However, further analysis is required for a generalizable conclusion. The number of collaborative posts increased whenever the students interacted with the indicators of originality and collaboration proposed. Moreover, the case study also presented details about the change of behavior for each student, as they started to write more messages considering the social elements evaluated in this study.

6 Conclusion

This work has two main contributions. First, text mining techniques were used to analyze the posts in a forum written in Brazilian Portuguese extracting four different features: (i) Expressing appreciation towards other participants; (ii) Recognizing group presence; (iii) Sharing information and resources; (iv) Asking for feedback/Answering questions. To the best of our knowledge, this paper is the first to address these problems in texts written in Portuguese. Previous works in the literature (Barbosa et al., 2020; Neto et al., 2018) presented attempts to classify the cognitive and social presences in Portuguese; however, they focused on higher-level classes rather than on specific indicators as proposed here.

Second, a case study in a real classroom was performed. In this case study, the instructor proposed a discussion activity about a scientific paper related to the course topic. In the first week, they used the traditional forum, and in the second week, a forum that shows the originality and collaboration learning analytics to the students. In terms of content, the difference between the two weeks was the paper used as the theme of discussion.

The number of posts increased from 10 (in the first week) to 30 in the second week, and the number of posts with collaborative indicators also rose. Besides the positive impact of the proposed analytics, other aspects may have influenced the number of posts in the second week, such as the number of activities of the other modules, the level of interest on the paper focus of the discussion, and the motivation of the students. Besides, the learning analytics information was used by the instructor to assess the interactions among students.

This work's main limitation is the small size of the dataset for the quantitative experiment and the cohort of students (12) in the case study, and the case study's duration (two weeks). The following points may be pursued as lines for further work:

- Replicating the case study in other courses and with a larger number of students;
- Using data from other online discussions to increase the number of posts used in the training step, as some machine learning algorithms (such as deep learning algorithms) perform better on larger corpora. This would also improve the reliability of the results in the quantitative evaluation;
- Adapting the proposed methods to other languages, such as English and Spanish. As the methods for natural language processing are usually language-dependent, the proposed methods cannot be directly adopted;
- Increasing the number of collaboration indicators (Murphy, 2004) such as: (i) Summarizing or reporting on content, (ii) Directly disagreeing with/challenging statements made by another participant, and (iii) Introducing new perspectives;
- Conducting a focus group to understand the students' perception of the proposed tool;
- Performing an empirical study using a control group to evaluate the efficacy of the proposed method.

Acknowledgements

This research was partially supported by the fellowship 310060/2018-4 of the National Council for Scientific and Technological Development – CNPq.

References

Aluísio, S.M., Specia, L., Pardo, T.A., Maziero, E.G., Fortes, R.P., 2008. Towards brazilian portuguese automatic text simplification systems, in: *Proceedings of the Eighth ACM Symposium on Document Engineering*, ACM, New York, NY, USA. pp. 240–248. doi:10.1145/1410140.1410191.

Amichai-Hamburger, Y., Gazit, T., Bar-Ilan, J., Perez, O., Aharony, N., Bronstein, J., Dyne, T.S., 2016. Psychological factors behind the lack of participation in online discussions. *Computers in Human Behavior* 55, 268–277.

Barbosa, A., Ferreira, M., Mello, R.F., Lins, R.D., Gaševic, D., 2021. The impact of automatic text translation on´ classification of online discussions for social and cognitive presences, in: *Proceedings of the International Conference on Learning Analytics & Knowledge*, pp. 605–614. Doi: 10.1145/3448139.3448147.

Barbosa, G., Camelo, R., Cavalcanti, A.P., Miranda, P., Mello, R.F., Kovanović, V., Gaševi´ c, D., 2020. Towards automatic´ cross-language classification of cognitive presence in online discussions, in: *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, pp. 605–614.

Bhatia, S., Biyani, P., Mitra, P., 2014. Summarizing online forum discussions-can dialog acts of individual messages help? in: *EMNLP*, pp. 2127–2131.

Cambria, E., 2016. Affective computing and sentiment analysis. *IEEE Intelligent Systems* 31, 102–107.

Castellanos-Reyes, D., 2020. 20 years of the community of inquiry framework. *TechTrends* , 1–4.

Cavalcanti, A.P., de Mello, R.F.L., Ferreira, M.A.D., Rolim, V.B., Tenório, J.V.S., 2017. Statistical and semantic features to measure sentence similarity in portuguese, in: *2017 Brazilian Conference on Intelligent Systems* (BRACIS), IEEE. pp. 342–347.

Coffrin, C., Corrin, L., de Barba, P., Kennedy, G., 2014. Visualizing patterns of student engagement and performance in moocs, in: *Proceedings of the fourth international conference on learning analytics and knowledge*, ACM. pp. 83–92.

Cohen, J., 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 37–46.

Dascalu, M., Trausan-Matu, S., Dessus, P., McNamara, D.S., 2015. Discourse cohesion: A signature of collaboration, in: *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge,* ACM. pp. 350–354.

De Wever, B., Schellens, T., Valcke, M., Van Keer, H., 2006. Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review. *Computers & Education* 46, 6–28.

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., 2014. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research* 15, 3133–3181. URL: http://jmlr.org/papers/ v15/delgado14a.html.

Ferreira, M., Rolim, V., Mello, R.F., Lins, R.D., Chen, G., Gaševic, D., 2020. Towards automatic content analysis of´ social presence in transcripts of online discussions, in: *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, pp. 141–150.

Ferreira, R., Lins, R.D., Simske, S.J., Freitas, F., Riss, M., 2016. Assessing sentence similarity through lexical, syntactic and semantic analysis. *Computer Speech & Language* 39, 1–28.

Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., Romero, C., 2019. Text mining in education. Wiley Interdisciplinary Reviews: *Data Mining and Knowledge Discovery* 9, e1332.

Forman, G., 2003. An extensive empirical study of feature selection metrics for text classification. *Journal of machine learning research* 3, 1289–1305.

Garrison, D.R., Anderson, T., Archer, W., 1999. Critical Inquiry in a Text-Based Environment: *Computer Conferencing in Higher Education. The Internet and Higher Education* 2, 87–105.

Garrison, D.R., Anderson, T., Archer, W., 2010. The first decade of the community of inquiry framework: A retrospective. *The internet and higher education* 13, 5–9.

Gomes, G.A.F., 2012. Eu-tu: O emprego da classificação automática de mensagens em fóruns eletrônicos de discussões para análise do processo de ensino e aprendizagem centrado em interações. Rio de Janeiro, *PPGI/IM/iNCE/UFRJ*.

Graesser, A.C., McNamara, D.S., Louwerse, M.M., Cai, Z., 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers* 36, 193–202. Doi: 10.3758/BF03195564.

Hartmann, N.S., 2016. Solo queue at assin: Combinando abordagens tradicionais e emergentes. *Linguamática* 8, 59–64.

Hew, K.F., Cheung, W.S., 2008. Attracting student participation in asynchronous online discussions: A case study of peer facilitation. *Computers & Education* 51, 1111–1124.

Hu, Q., Huang, Y., Deng, L., 2018. A method for analysis of online discussion forum in moodle, in: *2018 13th International Conference on Computer Science & Education (ICCSE), IEEE*. pp. 1–4.

Joachims, T., 2002. Learning to classify text using support vector machines: *Methods, theory and algorithms. Kluwer Academic Publishers.*

Joksimovic, S., Gasevic, D., Kovanović, V., Adesope, O., Hatala, M., 2014. Psychological characteristics in cognitive presence of communities of inquiry: A linguistic analysis of online discussions. *The internet and higher education* 22, 1– 10.

Joshi, M., Rosé, C.P., 2007. Using transactivity in conversation for summarization of educational dialogue, in: *Workshop on Speech and Language Technology in Education*.

Kickmeier-Rust, M.D., Bedek, M., Albert, D., 2016. Theorybased learning analytics: Using formal concept analysis for intelligent student modelling, in: *Proceedings on the International Conference on Artificial Intelligence* (ICAI), The Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp). p. 97.

Kim, D., Park, Y., Yoon, M., Jo, I.H., 2016. Toward evidencebased learning analytics: Using proxy variables to improve asynchronous online discussion environments. *The Internet and Higher Education* 30, 30–43.

Kovanović, V., Gaševi´ c, D., Hatala, M., Siemens, G., 2017a.´ A novel model of cognitive presence assessment using automated learning analytics methods, in: *Measurement in Digital Environments White Paper Series*. SRI International.

Kovanović, V., Joksimovi´ c, S., Gaševi´ c, D., Hatala, M., ´ Siemens, G., 2015. Content analytics: the definition, scope, and an overview of published research. *Handbook of Learning Analyitcs* .

Kovanović, V., Joksimovi´ c, S., Katerinopoulos, P., Michail, C., ´ Siemens, G., Gaševic, D., 2017b. Developing a mooc ex-´ perimentation platform: insights from a user study, in: *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, ACM. pp. 1–5.

Kovanović, V., Joksimovi´ c, S., Waters, Z., Gaševi´ c, D., Kitto, ´ K., Hatala, M., Siemens, G., 2016. Towards automated content analysis of discussion transcripts: A cognitive presence case, in: *Proceedings of the sixth international conference on learning analytics & knowledge, ACM*. pp. 15–24.

Lang, C., Siemens, G., Wise, A., Gaševic, D., 2017. *Handbook´ of learning analytics*.

Lárusson, J.A., White, B., 2012. Monitoring student progress through their written point of originality, in: Proceedings of the *2nd International Conference on Learning Analytics and Knowledge*, ACM. pp. 212–221.

Lin, F.R., Hsieh, L.S., Chuang, F.T., 2009. Discovering genres of online discussion threads via text mining. *Computers & Education* 52, 481–495.

Lui, A.K.F., Li, S.C., Choy, S.O., 2007. An evaluation of automatic text categorization in online discussion analysis, in: Advanced Learning Technologies, 2007. ICALT 2007. *Seventh IEEE International Conference on, IEEE*. pp. 205–209.

Marbouti, F., Wise, A.F., 2016. Starburst: a new graphical interface to support purposeful attention to others' posts in online discussions. *Educational Technology Research and Development* 64, 87–113.

McLaren, B.M., Scheuer, O., De Laat, M., Hever, R., De Groot, R., Rosé, C.P., 2007. Using machine learning techniques to analyze and support mediation of student e-discussions.*Frontiers in Artificial Intelligence and Applications* 158, 331.

Melanie, M., 1999. An introduction to genetic algorithms. Cambridge, Massachusetts London, England, Fifth printing 3, 62–75.

de Melo Ferreira, F.J., Miranda, S.K.O., de Barros Costa, E., da Costa, F.P.D., Rocha, H.J.B., 2013. Um modelo de fórum de discussão com suporte às interações entre aprendizes utilizando mapas conceituais, in: *Brazilian Symposium on Computers in Education*, p. 416.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality, in: *Advances in neural information processing systems*, pp. 3111–3119.

Mitchell, M., 1998. An introduction to genetic algorithms. *MIT press*.

Moreno-Marcos, P.M., Alario-Hoyos, C., Muñoz-Merino, P.J., Estévez-Ayres, I., Kloos, C.D., 2018. A learning analytics methodology for understanding social interactions in moocs. *IEEE Transactions on Learning Technologies* 12, 442–455.

Mu, J., Stegmann, K., Mayfield, E., Rosé, C., Fischer, F., 2012. The acodea framework: Developing segmentation and classification schemes for fully automatic analysis of online discussions. *International journal of computer-supported collaborative learning* 7, 285–305.

Murphy, E., 2004. Recognising and promoting collaboration in an online asynchronous discussion. *British Journal of Educational Technology* 35, 421–431.

Neto, J.L., Santos, A.D., Kaestner, C.A., Freitas, A.A., 2000. Generating text summaries through the relative importance of topics. *Lecture Notes in Computer Science*, 300–309.

Neto, V., Rolim, V., Ferreira, R., Kovanović, V., Gaševi´ c, D., ´ Lins, R.D., Lins, R., 2018. Automated analysis of cognitive presence in online discussions written in portuguese, in: *European conference on technology enhanced learning*, Springer. pp. 245–261.

Pendry, L.F., Salvatore, J., 2015. Individual and social benefits of online discussion forums. *Computers in Human Behavior* 50, 211–220.

Ravi, S., Kim, J., 2007. Profiling student interactions in threaded discussions with speech act classifiers. *Frontiers in Artificial Intelligence and Applications* 158, 357.

Ricardo, B.Y., 1999. Modern information retrieval. *Pearson Education India*.

Rosé, C.P., Howley, I., Wen, M., Yang, D., Ferschke, O., 2017. Assessment of discussion in learning contexts, in: *Innovative Assessment of Collaboration. Springer*, pp. 81–94.

Rubio, D., Villalon, J., 2016. A latent semantic analysis method to measure participation quality online forums, in: Advanced Learning Technologies (ICALT), *2016 IEEE 16th International Conference on, IEEE*. pp. 18–19.

Sánchez-Martí, A., Puig, M.S., Ruiz-Bueno, A., Regós, R.A., 2018. Implementation and assessment of an experiment in reflective thinking to enrich higher education students' learning through mediated narratives. *Thinking Skills and Creativity* 29, 12–22.

Sarmento, L., 2006. Siemês–a named-entity recognizer for portuguese relying on similarity rules. *Computational Processing of the Portuguese Language*, 90–99.

Seber, G.A., Lee, A.J., 2012. Linear regression analysis. volume 936. John Wiley & Sons.

Silva, M., Carvalho, P., Sarmento, L., 2012. Building a sentiment lexicon for social judgement mining. *Computational Processing of the Portuguese Language*, 218–228.

Silva, W.D.C.M., Finger, M., 2013. Improving cogroo: the brazilian portuguese grammar checker, in: Proceedings of the *9th Brazilian Symposium in Information and Human Language Technology*.

Simsek, D., Sandor, A., Shum, S.B., Ferguson, R., De Liddo, A., Whitelock, D., 2015. Correlations between automated rhetorical analysis and tutors' grades on student essays, in: *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge*, ACM. pp. 355–359.

Tausczik, Y.R., Pennebaker, J.W., 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology* 29, 24– 54. Doi: 10.1177/0261927X09351676.

Wen, M., Yang, D., Rose, C., 2014. Sentiment analysis in mooc discussion forums: What does it tell us? in: *Educational data mining* 2014.

Wise, A., Zhao, Y., Hausknecht, S., 2014. Learning analytics for online discussions: Embedded and extracted approaches. *Journal of Learning Analytics* 1, 48–71.

Wulf, J., Blohm, I., Leimeister, J.M., Brenner, W., 2014. Massive open online courses. *Business & Information Systems Engineering (BISE)* 6, 111–114.

Yen, C.H., 2013. A framework of e-learning analytics for asynchronous discussion forums, in: Advanced Learning Technologies (ICALT), *2013 IEEE 13th International Conference on, IEEE.* pp. 26–28.

Yoo, J., Kim, J., 2014. Can online discussion participation predict group project performance? investigating the roles of linguistic features and participation patterns. *International Journal of Artificial Intelligence in Education* 24, 8–32.

Zhao, J., Zhu, T., Lan, M., 2014. Ecnu: One stone two birds: Ensemble of heterogenous measures for semantic relatedness and textual entailment., in: *SemEval@ COLING*, pp. 271– 277.