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

ANDERSON PINHEIRO CAVALCANTI

Análise Automática de Feedback em Ambientes Virtuais de Aprendizagem

Recife

2022

ANDERSON PINHEIRO CAVALCANTI

Análise Automática de Feedback em Ambientes Virtuais de Aprendizagem

Trabalho apresentado ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para obtenção do grau de Doutor em Ciência da Computação.

Área de Concentração: *Inteligência Computacional*

Orientador (a): Frederico Luiz Gonçalves de Freitas

Coorientador (a): Rafael Ferreira Leite de Mello

Recife

2022

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

C377a Cavalcanti, Anderson Pinheiro
Análise automática de feedback em ambientes virtuais de aprendizagem / Anderson Pinheiro Cavalcanti. – 2022.
189 f.: fig., tab.

Orientador: Frederico Luiz Gonçalves de Freitas.
Tese (Doutorado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2022.
Inclui referências e apêndice.

1. Inteligência computacional. 2. Ambientes virtuais de aprendizagem. I. Freitas, Frederico Luiz Gonçalves de (orientador). II. Título.

006.31 CDD (23. ed.) UFPE - CCEN 2023-39

Anderson Pinheiro Cavalcanti

“Análise Automática de Feedback em Ambientes Virtuais de Aprendizagem”

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.

Aprovado em: 14/12/2022.

_____

Orientador: Prof. Dr. Frederico Luiz Gonçalves de Freitas

BANCA EXAMINADORA

_____

Prof. Dr. Luciano de Andrade Barbosa
Centro de Informática / UFPE

_____

Prof. Dr. Evandro de Barros Costa
Institutp de Computação / UFAL

_____

Profa. Dra. Elyda Laisa Soares Xavier Freitas
Universidade de Pernambuco – Campus Caruaru

_____

Profa. Dra. Isabel Dillmann Nunes
Instituto Metrópole Digital / UFRN

_____

Prof. Dr. Hilário Tomaz Alves de Oliveira
Instituto Federal do Espírito Santo – Unidade Serra

Eu dedico esta dissertação aos meus familiares, amigos e professores que me deram o suporte necessário para chegar até aqui.

AGRADECIMENTOS

Agradeço aos meus pais que sempre me incentivaram a seguir o caminho dos estudos. Em especial a minha esposa que sempre esteve do meu lado me apoiando em todos os momentos de pesquisa e estudo. A todos os colegas do doutorado que de uma forma ou de outra contribuíram com meu crescimento acadêmico. Agradeço imensamente ao meu orientador Fred Freitas e meu coorientador Rafael Ferreira por toda ajuda, dedicação e paciência ao longo desses 5 anos de doutorado. Agradeço também a CAPES pelo apoio financeiro a este projeto de pesquisa.

RESUMO

O feedback é um componente muito importante no processo de ensino-aprendizagem, pois ajuda o aluno a identificar as lacunas e avaliar o seu progresso no aprendizado. Em cursos a distância o feedback se torna ainda mais importante, pois é um dos recursos mais utilizados na interação entre professor e aluno, já que ambos estão separados fisicamente. No entanto, devido ao crescimento significativo da quantidade de alunos em cursos a distância, é difícil para os instrutores fornecer um feedback de alta qualidade. Nesse contexto, este trabalho tem como objetivo propor uma abordagem para analisar automaticamente os feedbacks fornecidos por professores em cursos online. Para isso, foram utilizados recursos linguísticos para extrair as características dos textos e algoritmos de aprendizagem de máquina para classificação. O objetivo é que os modelos de aprendizagem de máquina aprendam os padrões de bons feedbacks textuais que foram apresentados durante o treinamento. Foram realizados experimentos com base em duas teorias educacionais de feedback propostas na literatura utilizando diferentes classificadores, entre eles, o *Random Forest*, *AdaBoost*, *XGBoost* e o BERT (*Bidirectional Encoder Representations from Transformers*). Além de verificar qual algoritmo obtém os melhores resultados, também propomos a análise de quais as características são mais relevantes para cada classificador usando a medida *Mean Decrease Gini* (MDG) e o método *eXplainable Artificial Intelligence* (XAI). Os experimentos seguiram uma sequência de análise de cada classificador, começando pelo *Random Forest* e com base nos resultados obtidos, outros classificadores eram analisados com o objetivo de melhorar a acurácia do modelo. Os resultados obtidos demostram uma boa acurácia e os modelos gerados podem ser utilizados integrados em sistemas educacionais para ajudar o professor a fornecer um bom feedback ao aluno, onde a ferramenta apresentará quais os níveis de feedback e boas práticas de feedback o texto do professor possui, e com base nessa informação ele irá ajustar o seu texto para conseguir atingir o maior número de boas práticas.

Palavras-chaves: feedback educacional; mineração de texto; ambientes virtuais de aprendizagem.

ABSTRACT

Feedback is a very important component in the teaching-learning process, as it helps students to identify gaps and assess their learning progress. In distance courses, feedback becomes even more important, as it is one of the most used resources in the interaction between teacher and student, since both are physically separated. However, due to the significant growth in the number of students in distance courses, it is difficult for instructors to provide high quality feedback. In this context, this work aims to propose an approach to automatically analyze the feedback provided by teachers in online courses. For this, linguistic resources were used to extract the characteristics of the texts and machine learning algorithms for classification. The goal is for the machine learning models to learn the good textual feedback patterns that were presented during training. Experiments were carried out based on two educational feedback theories proposed in the literature using different classifiers, including Random Forest, AdaBoost, XGBoost and BERT (Bidirectional Encoder Representations from Transformers). In addition to verifying which algorithm obtains the best results, we also propose the analysis of which features are most relevant for each classifier using the measure Mean Decrease Gini (MDG) and the method eXplainable Artificial Intelligence (XAI). The experiments followed a sequence of analysis of each classifier, starting with Random Forest and based on the results obtained, other classifiers were analyzed with the aim of improving the accuracy of the model. The results obtained demonstrate good accuracy and the generated models can be used integrated in educational systems to help the teacher to provide good feedback to the student, where the tool will present what levels of feedback and good feedback practices the teacher's text has, and based on that information, he will adjust his text to achieve the highest number of good practices.

Keywords: educational feedback; text mining; online learning.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizagem de Máquina
AVA	Ambiente Virtual de Aprendizagem
BERT	*Bidirectional Encoder Representations from Transformers*
EAD	Educação a Distância
ENA	*Epistemic Network Analysis*
LIWC	*Linguistic Inquiry and Word Count*
LMS	*Learning Management System*
MDG	*Mean Decrease Gini*
PLN	Processamento de Linguagem Natural
RNAs	Redes Neurais Artificiais
XAI	*eXplainable Artificial Intelligence*

LISTA DE SÍMBOLOS

γ	Letra grega Gama
$\in$	Pertence
δ	Delta
θ	Teta
σ	Sigma
μ	Mi

SUMÁRIO

1 INTRODUÇÃO

O feedback educacional é um processo no qual o desempenho de um estudante é avaliado e informações são fornecidas para ajudá-lo a melhorar seu desempenho. O feedback é um fator essencial no processo de aprendizagem, pois permite aos alunos identificar lacunas no aprendizado, pode ajudar os alunos a identificar áreas de melhoria em seus conhecimentos ou habilidades, e refletir em suas estratégias de aprendizagem (SADLER, 1989).

Existem vários tipos de feedback educacional, incluindo feedback formativo, que é fornecido durante o processo de aprendizagem para ajudar os estudantes a ajustar sua compreensão, e feedback sumativo, que é fornecido no final de um período de aprendizagem para avaliar o desempenho geral. O feedback formativo, que é o foco desse trabalho, tem sido mostrado como eficaz na melhoria do desempenho dos estudantes. Isso se deve ao fato de que ele permite que os estudantes identifiquem rapidamente seus erros e corrijam-nos antes que eles se tornem hábitos difíceis de mudar. Além disso, o feedback formativo também pode aumentar a motivação dos estudantes, pois eles podem ver o progresso que estão fazendo e se sentir mais envolvidos no processo de aprendizagem (HENDERSON et al., 2019a).

O feedback foi identificado como um dos dez principais aspectos da aprendizagem para melhorar o desempenho do aluno (HATTIE; GAN, 2011). De acordo com Sadler (1989), o feedback precisa fornecer informações relevantes relacionadas a uma tarefa ou processo de aprendizagem e evitar discrepâncias entre o conhecimento adquirido pelo aluno e o que a disciplina deveria ensinar. Além disto, Laurillard (1993) afirma que o ensino sem feedback é completamente improdutivo para o aluno. Os princípios de boas práticas de feedback reconhecem que o feedback não é um produto, mas um processo complexo que deve conscientizar os alunos sobre como o comportamento, as emoções e a cognição do estudo real influenciam seus resultados (BOUD; FALCHIKOV, 2007; HENDERSON et al., 2019a).

Apesar do reconhecimento generalizado da importância do feedback para a aprendizagem, grande parte da literatura recente indica a difusão do feedback de baixa qualidade no ensino superior (HATTIE; GAN, 2011; HENDERSON et al., 2019a). A qualidade do feedback é consistentemente avaliada como uma das maiores causas de insatisfação para estudantes do ensino superior (FERGUSON, 2011). Weaver (2006a) relata que, embora os acadêmicos reconheçam o valor do feedback para facilitar a aprendizagem, eles consideram os comentários dos instrutores incompreensíveis e ineficazes. Ferguson (2011) identifica a falta do fornecimento de feedback

oportuno, expectativas pouco claras e pouca utilidade como as principais preocupações entre os alunos.

A Educação a Distância (EAD) teve um grande crescimento nos últimos anos, pois tornou-se uma alternativa em relação à educação tradicional. Um dos motivos desse crescimento é o fato de que a EAD se torna mais econômica e conveniente do que os ambientes educacionais tradicionais, já que muitas pessoas não têm tempo disponível para estudar presencialmente (SIMONSON; ZVACEK; SMALDINO, 2019). A EAD possui plataformas, por exemplo o Ambiente Virtual de Aprendizagem (AVA), que foram desenvolvidas para dar suporte aos professores e alunos. O uso do AVA tem aumentado nos últimos anos devido ao uso das tecnologias de informação e comunicação como ferramenta de apoio educacional (CORREIA; SANTOS, 2013). Esses ambientes possuem ferramentas que permitem uma grande interação entre professores e alunos, por exemplo: chat, fórum, wiki, entre outras. Entre elas, há a ferramenta de envio de atividades em que os alunos podem enviar suas respostas às atividades em andamento. Em geral, esse recurso de avaliação é o principal espaço onde os instrutores enviam feedback (COATES; JAMES; BALDWIN, 2005).

1.1 PROBLEMA DE PESQUISA

Segundo Ypsilandis (2002), o feedback é um fator crucial para o sucesso ou o fracasso de um curso a distância, pois esses cursos geralmente têm uma alta taxa de desistência, e também pelo fato de alunos, professores e tutores estarem separados fisicamente. Assim, interações e feedback informativos e oportunos se tornam ainda mais críticos para a construção do conhecimento e o sucesso acadêmico (JOULANI; GYORGY; SZEPESVÁRI, 2013). No entanto, devido ao crescimento significativo da quantidade de alunos nos AVAs que precisam de um envolvimento contínuo, é difícil para os instrutores fornecerem um feedback personalizado e de alta qualidade (ESPASA; MENESES, 2010).

Para melhorar a qualidade do feedback fornecido aos alunos, alguns trabalhos na literatura propuseram modelos ou princípios que ajudam a aumentar o impacto do feedback na aprendizagem do aluno. Por exemplo, no trabalho de Hattie e Timperley (2007), um modelo é proposto para a construção de feedback efetivo. Este modelo identifica três perguntas principais que o feedback eficaz deve responder: *"Para onde vou?"*, *"Como estou indo?"*, *"Para onde ir depois?"*. Cada pergunta de feedback opera em quatro níveis: feedback da tarefa (FT), feedback sobre o processamento da tarefa (FP), feedback sobre a auto-regulação (FR) e fe-

edback sobre a pessoa (FS). Evans (2013) realiza uma extensa revisão da literatura e propõe um cenário de feedback que identifica áreas específicas para futuras pesquisas sobre feedback de avaliação no ensino superior.

Nicol e Macfarlane-Dick (2006) propuseram um modelo conceitual de auto-regulação com base em uma revisão da literatura de pesquisa sobre avaliação formativa e feedback. A ideia principal do trabalho é identificar como os processos formativos de avaliação e feedback podem ajudar a promover a auto-regulação. Com base no modelo conceitual, foram definidos sete princípios de boas práticas de feedback que o professor pode usar para refletir sobre o projeto, e avaliar seus próprios procedimentos de feedback.

À medida que as instituições de ensino superior adotam a tecnologia, há um portfólio crescente de abordagens que utilizam a coleta de dados para melhorar os processos de aprendizagem. De acordo com a revisão sistemática realizada por Cavalcanti et al. (2021), vários trabalhos estão explorando ativamente soluções de feedback automatizado que podem permitir que os instrutores identifiquem e empreguem boas práticas de feedback de maneira eficiente e aumentem a velocidade de entrega de feedback aos alunos. Nesse sentido, alguns estudos examinaram o uso de métodos de mineração de dados para gerar feedback textual automatizado (LIU et al., 2017; MA et al., 2017; VILLALóN et al., 2008). Essas análises são frequentemente limitadas a áreas específicas de domínio, como programação de computadores ou redação, ou não seguem uma teoria educacional como base.

Diante deste contexto, esse trabalho tem como objetivo propor uma abordagem que combina diferentes recursos linguísticos para classificar mensagens de feedback usando algoritmos de Aprendizagem de Máquina (AM) que serão treinados com base em conceituados modelos de feedback da literatura (HATTIE; TIMPERLEY, 2007; NICOL; MACFARLANE-DICK, 2006). Com os algoritmos de classificação treinados, será possível integrá-los em algum AVA para ajudar o professor a identificar as boas práticas de feedback no seu texto e assim fornecer um feedback de qualidade ao aluno.

1.2 PERGUNTAS DE PESQUISA

A literatura sobre feedback educacional enfatiza que é crucial fornecer métodos automáticos para auxiliar na elaboração de feedback de boa qualidade (HENDERSON et al., 2019a). Embora vários estudos descrevam o que o feedback deve conter (HATTIE; TIMPERLEY, 2007; NICOL; MACFARLANE-DICK, 2006), a literatura não relata uma ampla gama de trabalhos que realizam

análise de conteúdo em feedback textual (CAVALCANTI et al., 2022a). Dessa forma, como o principal objetivo desse trabalho é analisar algoritmos de aprendizagem de máquina na classificação de mensagens de feedbacks com base em teorias de feedback educacional, nossa primeira questão de pesquisa é:

> **PERGUNTA DE PESQUISA 1:** *Qual algoritmo de aprendizagem de máquina obtém o melhor desempenho na classificação de mensagens textuais de feedback obtidas de um Ambiente Virtual de Aprendizagem de acordo as teorias educacionais de bom feedback de Hattie e Timperley (2007) e Nicol e Macfarlane-Dick (2006)?*

Para responder essa pergunta de pesquisa foram analisados os algoritmos: *Random Forest, Adaboost, XGBoost* e *Bidirectional Encoder Representations from Transformers* (BERT). Além de verificar qual algoritmo obtém os melhores resultados, também propomos a análise de quais as características são mais relevantes para cada classificador, levando à nossa segunda pergunta de pesquisa:

> **PERGUNTA DE PESQUISA 2:** *Quais as características mais importantes na classificação dos feedback?*

Para analisar as características mais importantes nos algoritmos *Random Forest, Adaboost e XGBoost* foi utilizada a medida popular *Mean Decrease Gini* (MDG), que é baseada na redução na medida de impureza de Gini [5]. Neste trabalho, adotamos o MDG para avaliar a relevância de diferentes características para o resultado dos classificadores baseados em árvores de decisão. Tradicionalmente, os métodos de aprendizado profundo são categorizados como algoritmos de caixa preta. O método *eXplainable Artificial Intelligence* (XAI) (MILLER, 2019) ajuda a entender o modelo de aprendizado profundo com representações visuais da saída. Utilizamos o XAI para entender o funcionamento do classificador BERT. As perguntas de pesquisa 1 e 2 são respondidas pelos Capítulos 5, 6, 7, 8 e 9.

Existem duas teorias de feedback que são amplamente usadas para o contexto específico de feedback textual (HATTIE; TIMPERLEY, 2007; NICOL; MACFARLANE-DICK, 2006). Por um lado, é recomendado que um texto de feedback contenha assuntos relacionados não só as atividades em si, mas também textos motivacionais e que incentivem a autorregulação (NICOL; MACFARLANE-DICK, 2006). Por outro lado, também é importante considerar os níveis de conteúdo que o feedback deve tratar (HATTIE; TIMPERLEY, 2007). Diante deste contexto,

esse trabalho buscou analisar a relação entre os indicadores de bons feedback propostos por Nicol e Macfarlane-Dick (2006) e Hattie e Timperley (2007). Mais especificamente, buscamos estudar o quão relacionados são os indicadores de avaliação de feedback propostos e com isso direcionar quais boas práticas podem potencializar o impacto do feedback nos alunos. Dessa forma, nossa terceira pergunta de pesquisa é:

> **PERGUNTA DE PESQUISA 3:** *Qual a relação existente entre as teorias educacionais de feedback propostas por Nicol e Macfarlane-Dick (2006) e Hattie e Timperley (2007) com base em mensagens textuais de feedback obtidas de um Ambiente Virtual de Aprendizagem utilizando redes epistêmicas?*

Para responder essa pergunta foi principalmente utilizado o método computacional de análise de redes epistêmicas (SHAFFER; COLLIER; RUIS, 2016). Este método propõe análises estatísticas e visuais para relacionar diferentes categorias no contexto educacional e tem sido largamente utilizado na literatura educacional (FERREIRA-MELLO et al., 2019). A pergunta de pesquisa 3 é respondida pelo Capítulo 10.

1.3 OBJETIVOS

1.3.1 Objetivo Geral

O objetivo principal desta tese é analisar algoritmos de aprendizagem de máquina na classificação de mensagens de feedbacks com base em teorias de feedback educacional.

1.3.2 Objetivos Específicos

Para atingir o objetivo geral foram definidos os seguintes objetivos específicos:

- Realizar uma revisão sistemática da literatura sobre feedback automático;
- Criar uma base de dados com feedbacks escritos de cursos a distância;
- Anotar a base de dados com base nas boas práticas de Nicol e Macfarlane-Dick (2006) e níveis de Hattie e Timperley (2007);
- Treinar algoritmos de aprendizagem de máquina para classificação dos feedbacks e obter os modelos com melhor acurácia;

1.4 ORGANIZAÇÃO DA TESE

Esta tese está dividida em dez capítulos e foi escrita com base em sete artigos publicados ou submetidos. O Capítulo 1 introduz o problema e os objetivos do trabalho proposto. Capítulo 2 detalha alguns conceitos necessários para o entendimento deste trabalho. Capítulo 4 apresenta um estudo com uma revisão sistemática da literatura sobre feedback automático em ambientes virtuais de aprendizagem. Capítulo 5 apresenta os resultados de uma análise da qualidade do feedback fornecido por instrutores em um curso online seguindo as boas práticas de Nicol e Macfarlane-Dick (2006) e também utiliza um algoritmo de aprendizagem de máquina (*Random Forest*) para identificar a presença das boas práticas em mensagens de feedback. Nesse trabalho o algoritmo de aprendizagem de máquina verifica apenas se o texto possui ou não pelo menos uma das boas práticas de feedback. Capítulo 6 apresenta um estudo semelhante, mas utilizando os algoritmos de aprendizagem de máquina AdaBoost e XGBoost.

Capítulo 7 mostra os resultados da classificação automática de feedbacks de um curso online utilizando o algoritmo *Random Forest* seguindo os níveis propostos por Hattie e Timperley (2007), ou seja, nesse trabalho são criados classificadores binários para cada nível de feedback. Capítulo 8 apresenta uma melhoria nos resultados dos classificadores de níveis de feedback do trabalho anterior (Capítulo 7) utilizando os algoritmos AdaBoost e XGBoost em conjunto com um novo recurso de extração de características e também apresenta os resultados dos classificadores para as boas práticas de feedback (NICOL; MACFARLANE-DICK, 2006). Diferentemente do Capítulo 5 que cria apenas um classificador para verificar se o texto tem ou não alguma boa prática, o Capítulo 8 apresenta diversos classificadores binários, um para cada boa prática.

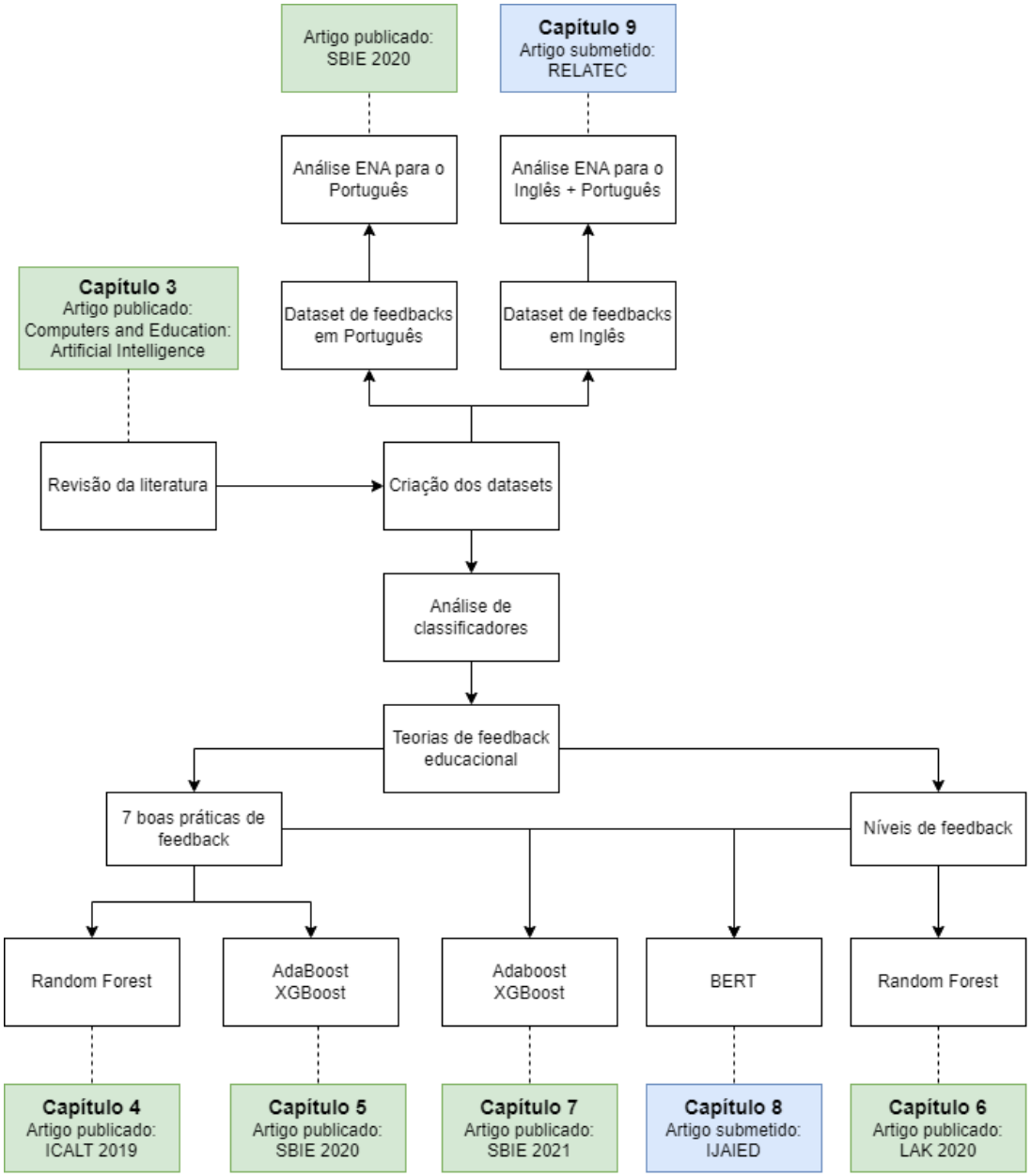
Capítulo 9 apresenta os resultados de classificação utilizando uma abordagem de *deep learning* com o BERT. Nesse trabalho são treinados classificadores binários para cada boa prática e nível de feedback. Além disso, é utilizada uma abordagem de XAI para explicar as classificações realizadas pelo modelo BERT.

Capítulo 10 descreve um estudo com a relação entre as teorias de feedbacks (NICOL; MACFARLANE-DICK, 2006; HATTIE; TIMPERLEY, 2007) utilizando datasets de feedback em Inglês e Português em conjunto com uma técnica de análise de dados chamada *Epistemic Network Analysis* (ENA). Por fim, o capítulo 11 apresenta as considerações finais e as contribuições da pesquisa.

A Figura 1 mostra um diagrama com a sequência e divisão dos artigos que foram publicados

e submetidos nesta tese.

Figura 1 – Diagrama dos artigos publicados.



Fonte: Elaborada pelo autor (2022)

1.5 PUBLICAÇÕES RELACIONADAS A TESE

Conforme mencionado anteriormente, a tese está dividida em um conjunto de sete artigos. A lista a seguir apresenta onde o artigo principal de cada tópico foi aceito ou submetido.

Capítulo 4: Artigo publicado na revista *Computers and Education: Artificial Intelligence* (CAVALCANTI et al., 2021).

Capítulo 5: Artigo publicado na *IEEE 19th International Conference on Advanced Learning Technologies* (ICALT) (CAVALCANTI et al., 2019).

Capítulo 6: Artigo publicado no Simpósio Brasileiro de Informática na Educação (SBIE) (CAVALCANTI et al., 2020a).

Capítulo 7: Artigo publicado na *10th International Learning Analytics and Knowledge Conference* (LAK) (CAVALCANTI et al., 2020).

Capítulo 8: Artigo publicado no Simpósio Brasileiro de Informática na Educação (SBIE) (CAVALCANTI et al., 2021).

Capítulo 9: Artigo submetido para revista *International Journal of Artificial Intelligence in Education* (CAVALCANTI et al., 2022a).

Capítulo 10: Artigo submetido para Revista Latinoamericana de Tecnología Educativa-RELATEC (CAVALCANTI et al., 2022b), tendo um estudo inicial previamente publicado no Simpósio Brasileiro de Informática na Educação (SBIE) (CAVALCANTI et al., 2020)

2 FUNDAMENTAÇÃO

Este capítulo apresenta alguns conceitos fundamentais para o entendimento deste trabalho que não estão descritos no conteúdo dos artigos publicados.

2.1 PROCESSAMENTO DE LINGUAGEM NATURAL

Processamento de Linguagem Natural é uma área de Ciência da Computação que estuda o desenvolvimento de programas de computador que analisam, reconhecem e/ou geram textos em linguagens humanas ou linguagens naturais. O objetivo do Processamento de Linguagem Natural (PLN) é fornecer aos computadores a capacidade de entender e compor textos. "Entender" um texto significa reconhecer o contexto, fazer análise sintática, semântica, léxica e morfológica, criar resumos, extrair informação, interpretar os sentidos, analisar sentimentos e até aprender conceitos com os textos processados (VIEIRA; LOPES, 2010).

Segundo Perna, Delgado e Finatto (2010), o PLN não é uma tarefa trivial devido à rica ambiguidade da linguagem natural. Essa ambiguidade torna o PLN diferente do processamento das linguagens de programação de computador, as quais são formalmente definidas evitando, justamente, a ambiguidade. O PLN visa promover um nível mais alto de compreensão da linguagem natural através do uso de recursos computacionais, com o emprego de técnicas para o rápido processamento de texto (MACHADO et al., 2010).

Para Loula (2011) embora a abordagem computacional tradicional analise a linguagem natural como uma mera sequência sintática de *tokens*, a linguagem natural pode e deve ser vista sob suas diversas faces, envolvendo não só aspectos sintáticos, mas também semânticos, pragmáticos, sociais, cognitivos, biológicos, semióticos, dentre outros.

O PLN é uma área de estudo que teve início no campo da Ciência da Computação e em seguida se estendeu para outros domínios como a Inteligência Artificial, Linguística e Ciência da Informação. O PLN tem como objetivo pesquisar os problemas relacionados à geração e à compreensão de linguagens naturais como as faladas por nós, seres humanos (JURAFSKY; MARTIN, 2000).

Segundo Jurafsky e Martin (2000) o objetivo do PLN é assegurar que computadores executem tarefas relevantes envolvendo a linguagem humana, tarefas como permitir a comunicação entre homem e máquina, melhorar a comunicação entre humanos ou simplesmente fazer pro-

cessamento relevante de texto ou fala.

2.1.1 Níveis de Conhecimento

Para que um sistema de PLN seja robusto, é preciso contemplar algumas tarefas base ou níveis de conhecimento. Cada uma dessas tarefas concentra-se em resolver parte de um problema maior. Entre elas estão as análises fonética, morfológica, sintática, semântica, pragmática e de discurso. Por exemplo, a análise fonética pretende analisar as palavras levando em consideração a maneira como estas são pronunciadas. A análise morfológica lida com a composição das palavras e sua natureza, divididas em morfemas. Já a análise sintática, faz o estudo das palavras tendo em conta a relação entre elas numa frase. A análise semântica pretende dar significado às palavras de forma a que seja possível perceber o que realmente o texto pretende transmitir. Por fim, a análise pragmática e de discurso preocupa-se essencialmente com o contexto do texto. A seguir, cada um desses níveis será detalhado.

2.1.1.1 Análise fonética

A análise fonética consiste no estudo dos sons de uma língua, ou seja, como as palavras são faladas e ouvidas. Segundo Jurafsky e Martin (2008), a ambiguidade está presente neste nível através das palavras homófonas, ou seja, palavras onde a pronúncia é semelhante, como por exemplo, as palavras *houve* e *ouve*. Esse nível envolve, principalmente, a transformação de áudio para texto ou até o reconhecimento de voz.

2.1.1.2 Análise morfológica

A análise morfológica lida com a composição das palavras e sua natureza, divididas em morfemas. Estes, por sua vez, são fragmentos mínimos que contêm significado, mas ainda não são palavras. A análise morfológica separa o texto em átomos ou *tokens*, fazendo um estudo a cada um desses átomos isoladamente. Aos átomos que formam palavras é feita a identificação da sua classe gramatical, o seu lema e o seu radical.

2.1.1.3 Análise sintática

O objetivo da análise sintática é estabelecer as relações sintáticas entre as palavras da linguagem natural. Essa análise permite, por exemplo, reconhecer que um determinado adjetivo está classificando um determinado nome em uma frase. A análise sintática é uma tarefa muito importante no PLN devido ao seu papel na determinação da estrutura de expressões linguísticas de forma que se possam extrair os seus significados. Descobrir que uma determinada palavra pertence a uma classe gramatical ajuda a determinar que tipo de palavras são as mais prováveis de serem suas vizinhas (JURAFSKY; MARTIN, 2008).

2.1.1.4 Análise semântica

Na área da Linguística, a semântica busca estudar o significado das palavras, de acordo com o contexto inserido. Esse nível é fundamental (e um pouco mais complexo) em PLN, já que envolve o entendimento real da linguagem. Dessa forma, o objetivo dessa análise é descobrir o significado das palavras em um determinado texto. Após essa análise é possível obter um texto em linguagem formal, passível de ser compreendido por um computador, ou seja, um texto sem ambiguidade.

2.1.1.5 Análise pragmática e de Discurso

A análise pragmática é o campo da linguística em que se estuda a língua em uso, ou seja, a linguagem em seu contexto comunicacional, seja escrita, falada ou sinalizada. Ela consiste em determinar a relação entre a linguagem e o contexto. O contexto inclui como a língua se refere a pessoas e objetos, como o discurso está estruturado e como este é interpretado pelo ouvinte (JURAFSKY; MARTIN, 2008).

2.2 APRENDIZAGEM DE MÁQUINA

A área de AM é uma área especializada no estudo e construção de sistemas que sejam capazes de aprender de forma automatizada a partir de dados (BRINK; RICHARDS; FETHEROLF, 2016). De acordo com Simon (1983) aprendizado é qualquer mudança em um sistema que melhore o seu desempenho na segunda vez que ele repetir a mesma tarefa, ou outra tarefa

da mesma população. Após efetuado o aprendizado, também denominado treinamento, um sistema pode ser utilizado para classificar ou estimar saídas para instâncias desconhecidas.

Os métodos para aprendizagem de máquina são divididos em duas categorias: supervisionado e não-supervisionado. A abordagem de aprendizado supervisionado consiste em utilizar uma série de exemplos (chamados de instâncias), já classificados, para induzir um modelo que seja capaz de classificar novas instâncias de forma precisa, com base no aprendizado obtido na fase de treinamento. Nessa abordagem tem-se a figura de um "professor externo", o qual apresenta um conhecimento do ambiente representado por conjuntos de exemplos na forma entrada- saída. Neste caso, o algoritmo de AM é treinado a partir de conjuntos de exemplos rotulados com o objetivo de aprender uma função desejada.

Já na abordagem de aprendizado não-supervisionado, o conjunto de dados utilizado não possui classificação, ou seja, a saída do conjunto de dados de treinamento não possui uma saída pré-definida para cada uma de suas instâncias. Esta é a abordagem indicada quando o objetivo do sistema não é construir um modelo de predição, e sim um modelo cuja função seja encontrar regularidades nos dados que possam vir a ser úteis (THEODORIDIS; KOUTROUMBAS, 2001).

2.3 REDES NEURAIS PROFUNDAS

As Redes Neurais Artificiais (RNAs) são modelos computacionais que surgiram de uma inspiração biológica, na tentativa de criar um modelo matemático que fosse capaz de replicar o funcionamento de neurônios biológicos (MCCULLOCH; PITTS, 1943). As redes neurais permitem que modelos computacionais compostos de várias camadas de processamento aprendam representações de dados com vários níveis de abstração. Esses métodos melhoraram drasticamente o estado da arte em reconhecimento de fala, reconhecimento visual de objetos, detecção de objetos e muitos outros domínios (NIELSEN, 2015). Devido ao aumento de dados e poder computacional, essas redes foram ficando cada vez maiores, com várias camadas de processamento, o qual denominou-se de Aprendizado Profundo (do inglês *Deep Learning*). O Aprendizado Profundo descobre estruturas complexas em grandes conjuntos de dados usando o algoritmo de retropropagação (do inglês *BackPropagation*) para indicar como uma máquina deve alterar seus parâmetros internos que são usados para calcular a representação em cada camada a partir da representação na camada anterior.

O aprendizado profundo tem desempenhado um papel cada vez maior nas tarefas de PLN

relacionadas à modelagem/classificação de tópicos (YOUNG et al., 2018). Arquiteturas de redes neurais profundas cada vez mais sofisticadas precisavam de mais dados de treinamento, devido à quantidade de parâmetros nos modelos que aumentaram substancialmente. Para lidar com os requisitos excessivos, foram desenvolvidos métodos de aprendizagem por transferência onde os pesquisadores fazem uso de grandes modelos de aprendizagem profunda previamente treinados em corpora linguísticos massivos, como a Wikipedia. Um modelo de linguagem pré-treinado pode ser definido como uma caixa preta que possui conhecimento prévio sobre linguagem natural e pode ser aplicada e ajustada para resolver vários problemas de PLN. O processo de pré-treinamento usa dados não rotulados para aprender os parâmetros iniciais de um modelo de rede neural. Um exemplo de tal modelo é o BERT (*Bidirectional Encoder Representations from Transformers*), que é um modelo de linguagem profundamente bidirecional treinado em conjuntos de dados muito grandes (ou seja, *Books corpus* e Wikipedia) com base em representações contextuais (DEVLIN et al., 2018). O modelo BERT pode ser ajustado usando uma camada de rede neural densa para diferentes tarefas de classificação.

2.4 CLASSIFICADORES

Na literatura, há vários algoritmos para realizar a classificação automática de padrões. Neste trabalho foram utilizados os seguintes classificadores baseados em árvores de decisão: *Random Forest*, AdaBoost e XGBoost.

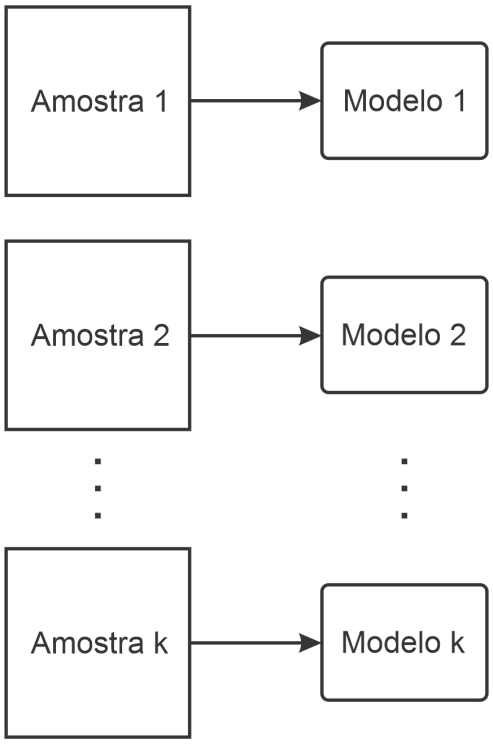
2.4.1 Ensemble

Um classificador ensemble consiste em um conjunto de classificadores treinados individualmente cujas decisões são de alguma forma combinadas. O uso de vários classificadores é uma estratégia bastante utilizada para aumentar o desempenho de sistemas de reconhecimento de padrões (MARQUÉS; GARCÍA; SÁNCHEZ, 2012). O objetivo principal é que os erros sejam minimizados através do uso de múltiplos classificadores ao invés de um único classificador. Existem duas técnicas principais para combinação de classificadores: o *bagging* e o *boosting*, descritos a seguir.

2.4.1.1 Bagging

O *Bagging* (*Bootstrap Aggregating*) é um método proposto por Breiman (1996) que tem como objetivo combinar modelos gerados por um mesmo algoritmo base para reduzir a variância de funções preditivas. O *Bagging* utiliza uma técnica de amostragem chamada *bootstrap* que gera conjuntos sucessivos e independentes de amostras dos dados originais. No Bagging os classificadores são treinados de forma independente a partir das amostras de *bootstraps* (Figura 2). Ao final do treinamento, deve-se combinar os classificadores através de um método de combinação apropriada, tal como a maioria de votos (BREIMAN, 1996).

Figura 2 – *Bagging* - classificadores em paralelo.



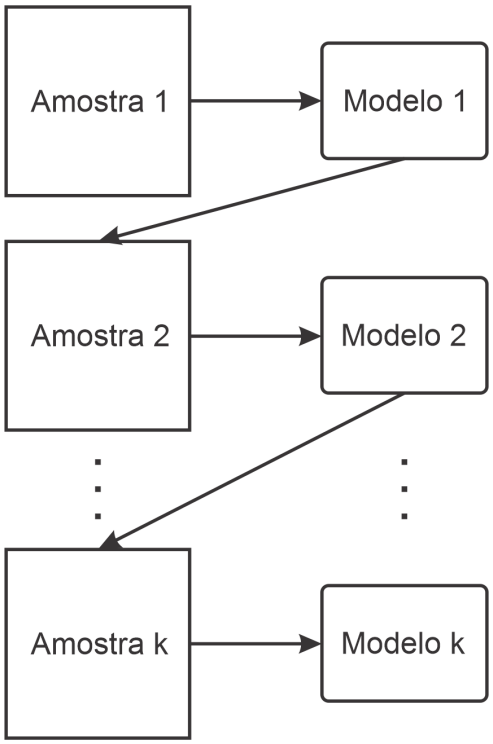
Fonte: Elaborada pelo autor (2022)

2.4.1.2 Boosting

Boosting é uma técnica que combina os classificadores gerados por um mesmo algoritmo de aprendizado e com isso consegue formar um classificador "forte" com base em classificadores mais simples, ditos "fracos". No *Boosting*, de forma semelhante ao *Bagging*, cada classificador é treinado usando um conjunto de treinamento diferente (FREUND; SCHAPIRE et al., 1996). A principal diferença em relação ao *bagging* é que o *boosting* gera conjuntos de treino e

classificadores de forma sequencial com base nos resultados da iteração anterior (Figura 3), enquanto que o *bagging* gera os conjuntos de treino de forma aleatória e pode gerar os classificadores paralelamente (FREUND; SCHAPIRE et al., 1996).

Figura 3 – *Boosting* - classificadores em sequência.



Fonte: Elaborada pelo autor (2022)

2.4.1.3 AdaBoost

Adaboost (*Adaptative Boosting*) é um algoritmo de classificação que utiliza a técnica *Boosting* para classificar os exemplos. O AdaBoost mantém um conjunto de pesos sobre os exemplos de treinamento. Em cada iteração, o algoritmo de aprendizado é chamado para minimizar o erro ponderado no conjunto de treinamento. Esse erro é utilizado para ponderar os pesos nos exemplos de treinamento. Dessa forma, o algoritmo coloca mais peso nos exemplos classificados incorretamente e menos peso nos exemplos classificados corretamente (DIETTERICH, 2000).

2.4.1.4 XGBoost

O algoritmo XGBoost é um modelo de aprendizado de máquina escalonável para o aprimoramento de árvore baseado em árvores de Decisão de Intensidade de Gradiente. O XGBoost fornece um aumento de árvore paralelo (também conhecido como GBDT - *Gradient Boosted Decision Tree*) que resolve muitos problemas de ciência de dados de maneira rápida e precisa. Diferente dos métodos tradicionais de reforço que pesam amostras positivas e negativas, o GBDT faz convergência global do algoritmo seguindo a direção do gradiente negativo (CHEN; GUESTRIN, 2016a).

2.4.1.5 Random Forest

O classificador *Random Forest* ou Florestas Aleatórias consiste em um conjunto de árvores de decisão geradas dentro de um mesmo objeto. Esse método foi proposto por (BREIMAN, 2001) e consiste em um conjunto de árvores de decisão construídas no momento de treinamento do método. As árvores são construídas selecionando aleatoriamente alguns dos atributos contidos dentro do vetor de características. Cada objeto (conjunto de árvores) passa por um mecanismo de votação (*bagging*), que elege a classificação mais votada. A classificação encontra-se nos nós terminais das mesmas. A saída do classificador é dada pela classe que foi retornada como resposta pela maioria das árvores pertencentes à floresta. É um método estatístico, de aprendizagem supervisionada, podendo ser utilizado em problemas de classificação e na realização de previsões. As árvores de Decisão são um dos modelos mais práticos e mais usados em inferência indutiva.

Muitas vezes as Árvores de Decisão são aplicadas a problemas de classificação binária. Uma das vantagens é permitir que se perceba quais as decisões que o sistema tomou, ou seja, facilmente se percebe quais foram as características mais importantes para o sistema. Entretanto, as árvores de decisão com baixo viés e alta variância tendem a superajustar os dados. Portanto, a técnica de *bagging* torna-se uma solução muito boa para diminuir a variância em uma árvore de decisão. Em vez de usar um modelo *bagging* com um modelo subjacente como árvore de decisão, também podemos usar a floresta aleatória, que é mais conveniente e bem otimizada para árvores de decisão. O principal problema com o *bagging* é que não há muita independência entre os conjuntos de dados amostrados, ou seja, há correlação. A vantagem das florestas aleatórias sobre os modelos de *bagging* é que as florestas aleatórias fazem um

ajuste no algoritmo de trabalho do modelo de *bagging* para diminuir a correlação nas árvores. A ideia é introduzir mais aleatoriedade ao criar árvores que ajudarão a reduzir a correlação.

A floresta aleatória é um algoritmo de aprendizado supervisionado usado tanto para classificação quanto para regressão. No entanto, é usado principalmente para problemas de classificação. Como sabemos, uma floresta é feita de árvores e mais árvores significa uma floresta mais robusta. Da mesma forma, um algoritmo de floresta aleatória cria árvores de decisão em amostras de dados e, em seguida, obtém a previsão de cada uma delas e, finalmente, seleciona a melhor solução.

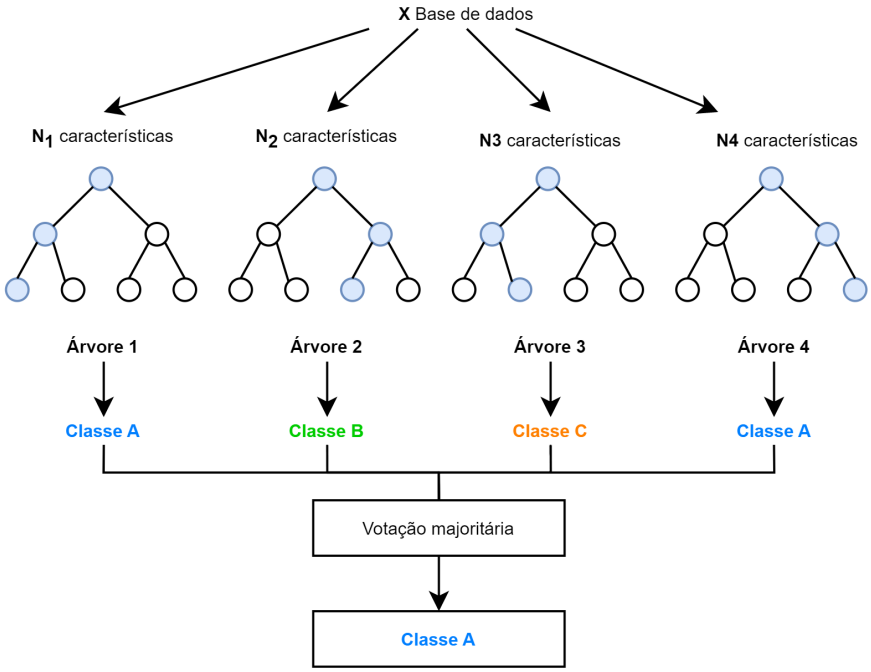
O algoritmo de florestas aleatórias funciona da seguinte forma:

1. Selecione aleatoriamente "k" características do total de "m" características. (Onde k « m)
2. Entre as características "k", calcule o nó "d" usando o melhor ponto de divisão.
3. Divida o nó em nós filhos usando a melhor divisão.
4. Repita 1 a 3 etapas até que o número "l" de nós seja alcançado.
5. Construa a floresta repetindo as etapas 1 a 4 por "n" números de vezes para criar "n" números de árvores.

O início do algoritmo de floresta aleatória começa com a seleção aleatória de "k" características do total de "m" características. Na próxima etapa, estamos usando as "k" características selecionadas aleatoriamente para encontrar o nó raiz usando a melhor abordagem de divisão. Na próxima etapa, calcularemos os nós filhos usando a mesma abordagem de melhor divisão. Serão as 3 primeiras etapas até formarmos a árvore com um nó raiz e tendo como alvo o nó folha. Por fim, repetimos de 1 a 4 etapas para criar "n" árvores criadas aleatoriamente. Essas árvores criadas aleatoriamente formam a floresta aleatória.

Para executar a previsão usando o algoritmo de floresta aleatória treinado, é necessário obter as características do teste e usar as regras de cada árvore de decisão criada aleatoriamente para prever o resultado e armazenar a classe prevista. Em seguida calcular os votos para cada classe prevista. Então, consideramos a classe prevista mais votada como a previsão final do algoritmo de floresta aleatória. Este conceito de votação é conhecido como votação por maioria (do inglês: *majority voting*). A figura 4 mostra um exemplo de como é feita a predição de uma classe usando a votação por maioria.

Figura 4 – Exemplo de classificação com florestas aleatórias.



Fonte: Elaborada pelo autor (2022)

2.5 AVALIAÇÃO DO DESEMPENHO DO CLASSIFICADOR

Validação cruzada (do inglês *Cross Validation*) é uma técnica estatística que particiona uma amostra dos dados em subconjuntos de tal modo que a análise é inicialmente executada em um único subconjunto, enquanto os outros subconjuntos são mantidos para treino (KOHAVI et al., 1995).

O K-*Fold Cross Validation* é um método de avaliação. Os documentos são aleatoriamente divididos em K partições mutuamente exclusivas ("*folds*") de tamanho aproximadamente igual a n/K, onde n é o tamanho do conjunto de documentos. Então, são realizados K experimentos, onde, em cada experimento, uma partição diferente é escolhida para o teste e as K-1 partições restantes são escolhidas para o treinamento. A medida de eficiência é a média das medidas de eficiência calculadas para cada uma das partições. Neste trabalho foi utilizada a taxa de acerto como medida de eficiência. A grande vantagem dessa técnica é que todos os documentos são usados tanto para treinamento quanto para teste. Assim, a forma como a divisão do corpus foi feita influi menos no resultado final, qualquer documento será usado exatamente uma vez para teste e K-1 vezes para treinamento.

2.6 EPISTEMIC NETWORK ANALYSIS

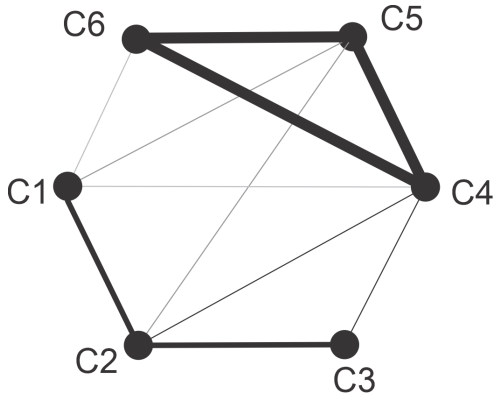
ENA é um conjunto de técnicas que identifica e mede conexões entre elementos em dados codificados e os representa em uma rede de relações dinâmicas. Esses modelos ilustram a estrutura das conexões e medem a força da associação entre os elementos de uma rede, e quantificam as mudanças na composição e força das conexões ao longo do tempo. Dentro da ENA, as conexões entre os códigos são derivadas para cada unidade de análise com base nas coocorrências de código nos subconjuntos de dados chamados estrofes (por exemplo, frase, parágrafo e documento). ENA é um método versátil que pode ser usado para modelar padrões de associação em qualquer sistema caracterizado por uma rede complexa de relacionamentos dinâmicos entre um conjunto fixo de elementos relativamente pequeno.

A partir de coocorrências de código, ENA primeiro cria uma representação de alta dimensão, chamada de espaço analítico, de todas as unidades de análise. As unidades de análise são então projetadas em um espaço de representação inferior, denominado espaço de projeção, que é derivado do espaço analítico por meio da decomposição de valores singulares (SVD). No final, a saída de ENA é uma série de modelos gráficos que capturam as relações entre os diferentes códigos (SHAFFER; COLLIER; RUIS, 2016).

Em contextos educacionais, a ENA é normalmente usada para investigar as associações entre códigos de um esquema de codificação (por exemplo, fases de presença cognitiva ou indicadores de presença social) em que o esquema de codificação é aplicado para analisar transcrições de fóruns educacionais (FERREIRA et al., 2018; ROLIM et al., 2019). Ao contrário de outras técnicas de análise de redes, a ENA foi projetada principalmente para problemas com um conjunto relativamente pequeno de conceitos, caracterizado por interações densas e altamente dinâmicas. Também pode ser usado para comparar as diferenças entre os diferentes grupos das unidades de análise.

Quanto maior a espessura da conexão entre 2 conceitos maior é a relação entre eles. Por exemplo, na Figura 5 os conceitos C4, C5 e C6 têm uma forte relação, pois a espessura da aresta de conexão entre os conceitos é a maior entre os outros conceitos. Da mesma forma, quanto menor for a espessura entre 2 conceitos, menor é a relação (por exemplo as relações entre C1 e C6 ou entre C1 e C5). Outro ponto a analisar na rede ENA é a proximidade entre os conceitos, quanto mais próximo estão os conceitos no gráfico mais similares eles são.

Figura 5 – Exemplo de relações com ENA.



Fonte: Elaborada pelo autor (2022)

2.7 FEEDBACK EDUCACIONAL

O feedback é um fator essencial no processo de aprendizagem, pois permite aos alunos identificar lacunas no aprendizado, pode ajudar os alunos a identificar áreas de melhoria em seus conhecimentos ou habilidades, e refletir em suas estratégias de aprendizagem (SADLER, 1989). Além disso, os professores usam o feedback para identificar as necessidades dos alunos, para que possam adaptar seus métodos e conteúdo com base nessas necessidades (LANGER, 2011). A literatura demonstra os benefícios e o impacto positivo que um bom feedback causa no aprendizado do aluno (NICOL; MACFARLANE-DICK, 2006), mas também ressaltam que um feedback ruim pode desmotivar e até levar a evasão do aluno (HATTIE; TIMPERLEY, 2007). Por exemplo, no estudo de Burke (2009) foram levantados alguns fatores de insatisfação do aluno com relação ao feedback recebido. Entre esses fatores, estão o comprimento (o feedback é muito breve), a polaridade (o feedback é muito negativo) e a complexidade (é muito difícil decifrar ou entender o feedback).

Vários estudos foram conduzidos nas últimas décadas na tentativa de identificar como os alunos devem receber e agir sobre o feedback fornecido (MUTCH, 2003). Por exemplo, Weaver (2006b) investigou, qualitativa e quantitativamente, se os alunos valorizam o feedback que recebem e relatam dois problemas: (1) o feedback muitas vezes não contém conteúdo suficiente para orientar ou motivar os alunos; e (2) os alunos não têm compreensão suficiente do discurso acadêmico para interpretar o feedback do instrutor com precisão. Além disso, Weaver (2006b) mostrou que mais de 50% dos estudantes universitários nunca receberam nenhuma orientação sobre *como entender e usar o feedback*, enquanto 75% dos alunos não receberam nenhum conselho sobre como entender e usar feedback antes de seus estudos

universitários. O aumento das matrículas de alunos de muitos ambientes educacionais reforça a importância e a dificuldade de fornecer feedback suficientemente detalhado e personalizado (KULKARNI et al., 2013).

2.7.1 Boas Práticas de Feedback

A literatura também oferece recomendações de boas práticas de feedback. Nicol e Macfarlane-Dick (2006) propôs um modelo conceitual de autorregularão com base em uma revisão da literatura de pesquisa sobre avaliação formativa e feedback. A ideia principal do trabalho é identificar como os processos formativos de avaliação e feedback podem ajudar a promover a autorregularão. Com base no modelo conceitual, foram definidos sete princípios de boas práticas de feedback que o professor pode usar para refletir sobre o projeto e avaliar seus próprios procedimentos de feedback. Abaixo estão as sete boas práticas[1] de feedback (NICOL; MACFARLANE-DICK, 2006). De acordo com os autores, boas práticas de feedback são amplamente definidas como qualquer coisa que possa fortalecer a capacidade dos alunos de autorregularem seu desempenho.

- **BP 1**: Ajuda a esclarecer o que é um bom desempenho (metas, critérios, padrões esperados);
- **BP 2**: Facilita o desenvolvimento da autoavaliação (reflexão) na aprendizagem;
- **BP 3**: Fornece informações de alta qualidade aos alunos sobre seu aprendizado;
- **BP 4**: Incentiva o diálogo entre professores e colegas sobre o aprendizado;
- **BP 5**: Incentiva crenças motivacionais positivas e autoestima;
- **BP 6**: Oferece oportunidades para fechar a lacuna entre o desempenho atual e o desejado;
- **BP 7**: Fornece informações aos professores que podem ser usadas para ajudar a moldar o ensino;

Cada princípio de feedback pode ser conectado a diferentes estratégias que o instrutor pode implementar em sala de aula. De acordo com Nicol e Macfarlane-Dick (2006), boas práticas de

[1] Foi utilizado o acrônimo BP para Boas Práticas

feedback são amplamente definidas como qualquer estratégia ou conteúdo que pode aumentar a capacidade dos alunos de autorregular seu desempenho de aprendizagem.

2.7.2 Níveis de Feedback

Hattie e Timperley (2007) analisaram várias condições que poderiam maximizar os efeitos positivos do feedback na aprendizagem, incluindo o aumento da consciência do aluno sobre uma meta geral de aprendizagem, o progresso em direção à meta e as metas subsequentes necessárias para atingir o objetivo principal. Assim, Hattie e Timperley (2007) propuseram quatro perspectivas, propostas como níveis, que o feedback deve abordar a fim de melhorar sua eficácia. Eles postularam que seu modelo é mais adequado para examinar o feedback textual porque o modelo é focado em aspectos relacionados a tarefas de aprendizagem, processo de aprendizagem e autorregulação do aluno. A Tabela 1 mostra os níveis propostos por (HATTIE; TIMPERLEY, 2007).

Tabela 1 – Níveis de feedback (HATTIE; TIMPERLEY, 2007)

#	Nível	Descrição
FT	Feedback sobre a tarefa	O feedback pode ser sobre uma tarefa, como se o trabalho está correto ou incorreto, pode incluir instruções para mais informações ou informações diferentes.
FP	Feedback sobre o processamento da tarefa	O feedback pode ser direcionado ao processo usado para criar um produto ou concluir uma tarefa, é mais direcionado ao processamento de informações ou processos de aprendizagem que requerem compreensão ou conclusão da tarefa.
FR	Feedback sobre autorregulação	O feedback para os alunos pode ser focado no nível de autorregulação, incluindo maior autoavaliação ou habilidades de confiança, que podem ter grandes influências na autoeficácia, na proficiência autorregulatória e nas crenças pessoais dos alunos.
FS	Feedback pessoal.	O feedback pode ser pessoal no sentido de que é direcionado a si mesmo. Frequentemente, não está relacionado ao desempenho da tarefa.

Fonte: Elaborada pelo autor (2022)

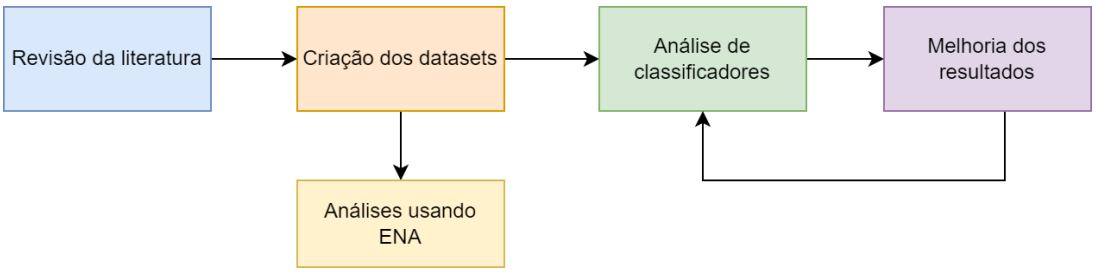
O nível a que pertence o feedback tem influência na sua eficácia, pelo que um feedback que incide sobre as qualidades do trabalho realizado e sobre o processo ou estratégias utilizadas dá maior ajuda ao aluno. O feedback que orienta o aluno para o desenvolvimento de estratégias autorregulatórias também tende a ser mais eficaz. Por exemplo, a meta-análise realizada por Wisniewski, Zierer e Hattie (2020) revelou que o feedback tem um impacto maior nos resultados das habilidades cognitivas e motoras do que nos resultados motivacionais

e comportamentais. Os comentários focados nas características pessoais dos alunos são muito vagos e não levam o aluno a focar em seu aprendizado (BROOKHART, 2017).

3 METODOLOGIA

Como foi citado anteriormente, esse trabalho foi escrito baseado nos artigos que foram publicados ou que foram submetidos e estão em processo de revisão. A imagem 6 mostra a sequência de passos que foram desenvolvidas durante esta tese.

Figura 6 – Sequência de atividades que foram desenvolvidas durante a tese.



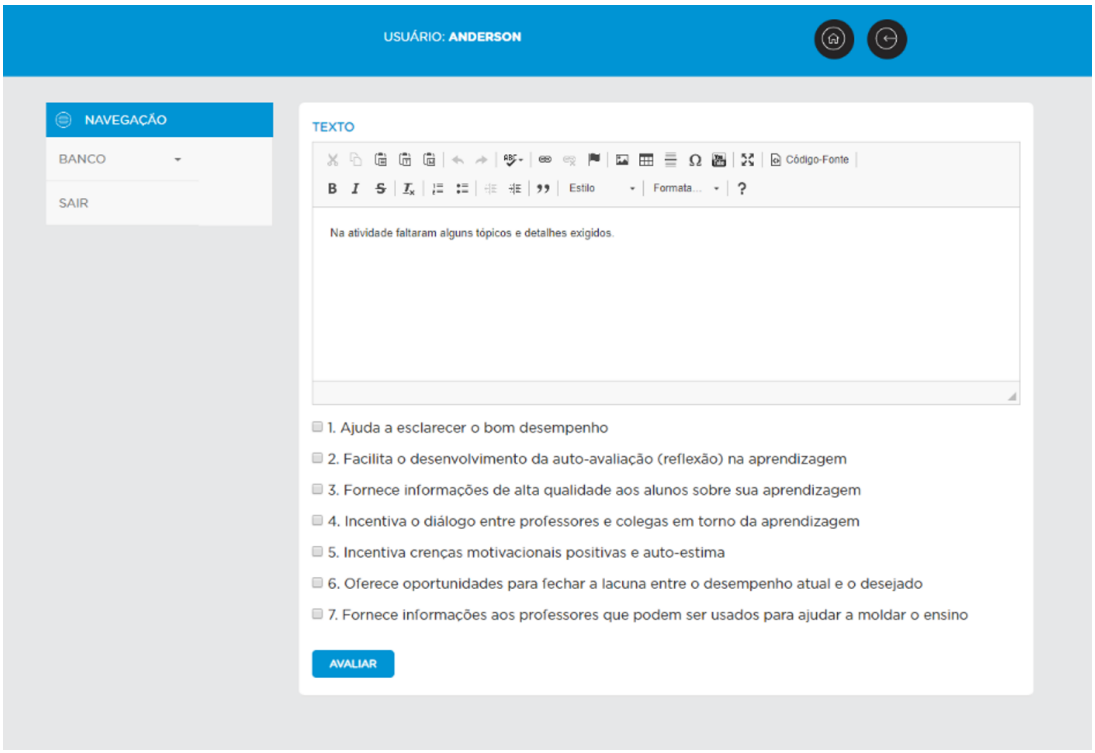
Fonte: Elaborada pelo autor (2022)

Inicialmente foi realizada uma revisão sistemática da literatura entre os anos de 2009 à 2018 sobre feedback automático em ambientes virtuais de aprendizagem. A ideia desse trabalho seria verificar o que está sendo desenvolvido na literatura sobre feedback automático e com base nesse resultado estruturar os próximos passos da pesquisa. O resultado dessa revisão foi publicado na revista *Computers and Education: Artificial Intelligence* e encontra-se no Capítulo 4.

Muitos estudos incluídos nesta revisão visam fornecer feedback automático. No entanto, poucos trabalhos tentaram analisar a qualidade do feedback fornecido por meio de formulários aplicados a alunos e/ou instrutores. Dessa forma, com base nessa lacuna da pesquisa sobre feedback educacional, foram analisados artigos que tinham como objetivo melhorar a qualidade do feedback fornecido, sendo os principais trabalhos de Nicol e Macfarlane-Dick (2006) e Hattie e Timperley (2007).

Em seguida, um dataset com textos de feedback de um ambiente virtual de aprendizagem foi anotado com base nessas duas teorias sobre feedback educacional ((NICOL; MACFARLANE-DICK, 2006) e (HATTIE; TIMPERLEY, 2007)). Inicialmente foi criada uma ferramenta online para anotar o dataset baseado nas boas práticas de Nicol e Macfarlane-Dick (2006). Essa ferramenta apresentava o texto do feedback e caixas de seleção com as 7 boas práticas, para que o anotador escolhesse quais das 7 boas práticas o texto de feedback estava seguindo, conforme mostra a Figura 7.

Figura 7 – Ferramenta de anotação para as boas práticas.



Fonte: Elaborada pelo autor (2022)

Os anotadores receberam um arquivo PDF de suporte com as informações referentes às boas práticas de feedback (para mais detalhes verificar a Figura 28 no Apêndice A). Esse processo também foi realizado da mesma forma para os níveis de Hattie e Timperley (2007). A Tabela 2 mostra alguns exemplos de feedback presentes na base de dados e as categorias pertencentes ao texto. Cada exemplo de feedback possui 11 rótulos diferentes (7 boas práticas e 4 níveis).

Com a base de dados criada com as categorias das boas práticas de Nicol e Macfarlane-Dick (2006), inicialmente foi analisado o algoritmo de aprendizagem de máquina (*Random Forest*) para identificar a presença das boas práticas nas mensagens de feedback. Nesse trabalho o algoritmo de aprendizagem de máquina verifica apenas se o texto possui ou não pelo menos uma das boas práticas de feedback, ou seja, foi treinado apenas um classificador binário que quando recebe um texto de feedback o classifica em 0 (não possui nenhuma das boas práticas) ou 1 (possui pelo menos uma das boas práticas). A utilização do *Random Forest* foi baseada em trabalhos semelhantes da área educacional (KOVANOVIĆ et al., 2016; FERREIRA et al., 2018; FERREIRA-MELLO et al., 2019). Esse trabalho é apresentado no Capítulo 5.

Em seguida, foram analisados classificadores mais recentes, que também eram baseados

Tabela 2 – Exemplos de feedback presentes na base de dados.

#	Texto	Categorias
1	Atividade muito bem elaborada. Parabéns. Faltou apenas o cabeçalho.	GP5, GP6, FT, FP ,FS
2	Texto super coerente com a proposta do título, linguagem direta e concisa.	GP1, GP5, FT, FS
3	Prezada, Parabéns! O trabalho atende os pontos pertinentes na sistemática proposta. Logo, observa-se que faltou uma última leitura ao finalizar o texto, por não ter sido feito a correção de algumas palavras que faltaram os acentos. Nesse sentido, aconselho que ao finalizar um trabalho releia para evitar essa ocorrência. Um abraço.	GP1, GP3, GP6, FT, FP
4	Bom trabalho. A utilização de jogos transposta os conceitos do conteúdo para o concreto da aprendizagem... Como ficaria então a questão da avaliação? Bons estudos!	GP1, GP2, GP3, GP5, FT, FR

Fonte: Elaborada pelo autor (2022)

em árvores de decisão, na base de dados de boas práticas binária, onde a classe 0 não possui nenhuma das boas práticas e a classe 1 possui pelo menos uma das boas práticas. Os algoritmos utilizados foram o AdaBoost e o XGBoost, por serem considerados mais robustos e eficientes que o *Random Forest*. Esse trabalho encontra-se no Capítulo 6.

Na sequência, a base de dados baseada nos níveis de Hattie e Timperley (2007) foi analisada usando o algoritmo *Random Forest* em conjunto com uma análise das características mais importantes retornadas pelo classificador. Nesse trabalho são treinados 3 classificadores binários para cada nível de feedback (FT, FP e FS). O resultado desse trabalho é apresentado no Capítulo 7.

Em seguida, os algoritmos AdaBoost e XGBoost foram utilizados para analisar tanto os níveis de feedback quanto as boas práticas em conjunto com um novo recurso de extração de características. Diferentemente do estudo no Capítulo 5 que cria apenas um classificador para verificar se o texto tem ou não alguma boa prática, o Capítulo 8 apresenta diversos classificadores binários, um para cada boa prática. O resultado desse estudo é apresentado no Capítulo 8.

Finalmente, seguindo as análises de classificadores, foi realizada uma análise utilizando uma abordagem de *deep learning* com o BERT. Nesse trabalho são treinados classificadores binários para cada boa prática e nível de feedback. Além disso, é utilizada uma abordagem de XAI para explicar as classificações realizadas pelo modelo BERT. O resultado desse estudo

encontra-se no Capítulo 9.

Um último estudo foi conduzido para analisar a relação entre as duas teorias de feedbacks (NICOL; MACFARLANE-DICK, 2006; HATTIE; TIMPERLEY, 2007) utilizando os 2 datasets de feedback criados nesse trabalho para a língua portuguesa em conjunto com 2 datasets de feedback da língua inglesa que tiveram o mesmo processo de anotação mencionado nesse trabalho de tese. Para isso, foi utilizada uma técnica de análise de dados chamada ENA que cria grafos com as relações entre determinadas categorias.

4 AUTOMATIC FEEDBACK IN ONLINE LEARNING ENVIRONMENTS: A SYSTEMATIC REVIEW OF THE LITERATURE

Authors: Anderson Pinheiro Cavalcanti, Arthur Diego, Ruan Carvalho, Rafael Ferreira Mello, Fred Freitas, Yi-Shan Tsai and Dragan Gašević

4.1 INTRODUCTION

Online learning has grown tremendously in recent years as an alternative or complementary option to traditional education which is primarily based on face-to-face teaching. According to Sung e Mayer (2012), online learning has grown because it is more flexible than traditional educational environments. For the purpose of facilitating online learning, various platforms such as learning management systems (LMSs) have emerged in the past decades. The use of LMSs has increased in recent years due to the use of information and communication technologies as an educational support tool (REISER; DEMPSEY, 2012). These environments have several resources (e.g., chat, forum, and wiki) that allow numerous interaction between instructors, students, and content (JOKSIMOVIĆ et al., 2015). Despite the advantages of online learning, there are some challenges for instructors. Among these, it is particularly notable that instructors struggle to follow the progress and activities of a cohort that is potentially unlimited in size (HERNÁNDEZ-GARCÍA et al., 2015).

Feedback is an essential component in the teaching-learning process as it allows students to identify gaps and assess their learning progress (BUTLER; WINNE, 1995a). According to Sadler (1989), feedback needs to provide specific information related to a learning task or process that fills a gap between the desired and the real understanding of the content or the development of abilities. Through feedback, students seek to hone some inadequate or poor knowledge or skills that hinder their learning progress. Several studies have shown that useful feedback brings benefits to learning (HATTIE; TIMPERLEY, 2007; NICOL; MACFARLANE-DICK, 2006; PARIKH; MCREELIS; HODGES, 2001). For instance, Black e Wiliam (1998) analyzed more than 250 feedback studies, and concluded that feedback produced significant gains in student learning and satisfaction. Recently, the study of Henderson et al. (2019b) analyzed seven case studies, through multiple stages of thematic analysis, case comparison and reliability verification, and proposed 12 main conditions that support effective feedback. These conditions highlight the importance of carefully designing feedback processes and have been organized

into three categories: capacity, projects, and culture.

In online learning contexts, feedback plays a crucial role due to the lack of face-to-face interaction among the participants of the course (YPSILANDIS, 2002). As instructors and students are separated in space and/or time in online contexts, the instructor must provide high-quality feedback to assist students in their learning and motivation (NICOL; MACFARLANE-DICK, 2006). However, the large size of student cohort in online learning environments can make it challenging for the instructor to provide useful and sufficient feedback to students. In light of this, several automatic tools have been proposed to enhance feedback practice (MARIN et al., 2017; GULWANI; RADIČEK; ZULEGER, 2014; BELCADHI, 2016).

There has been a lacuna in studies that systematically analyze automatic feedback systems in online environments. One exception is a technical report on studies about automatic feedback generation for programming exercises (KEUNING; JEURING; HEEREN, 2016). One key finding of this study is that existing tools often do not give feedback on how to solve problems and take next steps. This has also made it difficult for instructors to quickly adapt tools and resources to their own needs. The difference of the study presented in (KEUNING; JEURING; HEEREN, 2016) from the systematic literature review presented in this paper is that we do not limit automatic feedback to programming exercise tools only. Instead, we include all the automatic feedback generation systems in online learning environments.

In this context, this paper presents a *systematic literature review focusing on tools and resources that enable automatic feedback in learning management systems*. It allows the identification, evaluation, and interpretation of all available research relevant to a research question, subject, or event of interest (KITCHENHAM, 2004). Moreover, a literature review should conduct a critical evaluation of research studies that address a specific issue and must have a well-defined structure so that the results are not biased. Finally, the rigour of a systematic literature review needs to be strengthened by reducing random effects and ensuring reproductivity Becheikh, Landry e Amara (2006).

The current systematic literature study followed the guidelines and model of systematic review protocol proposed by Keele et al. (2007), which included three main steps:

1. The *planning step* identified the goals of the systematic literature review and defined the review protocol;

2. The *execution step*, which was the main stage of the review, and included the following activities: (i) formulated focused research questions, (ii) searched for and selected the

primary studies, (iii) defined the papers needed to answer the research questions, and (iv) extracted the data and synthesized the results.

3. The *reporting step* presented the summarized results with interpretation and discussion.

All the steps of the systematic review were performed using a systematic review management tool called StArt (State of the Art through Systematic Review)[1]. StArt assists the researcher in the development of an systematic literature review (LAPES, 2014), i.e., the steps presented previously. In summary, we selected 63 studies based on relevant keywords related to `feedback` and `online learning environments`, published between 2009 and 2018. In order to present a concise analysis, we have extract 19 features from papers selected. There features were grouped into categories such as basic information (e.g., year and title), goals, results, and methods to provide feedback of the papers selected. The results and their implications are further discussed in this paper.

4.2 METHOD

4.2.1 Research Questions

Automatic feedback emerges as a solution to an instructor's heavy workload due to the need to support a large number of students enrolled in online courses. However, it is necessary to analyze whether the studies that propose an automatic feedback approach help the instructor and/or the student. To do so, we defined the overarching research question:

RESEARCH QUESTION: What are the approaches used for generating automatic feedback in online learning environments?

Based on this overarching research question, we divided our work into four sub-questions:

> **RESEARCH QUESTION 1 (RQ1):** *Does automated feedback in online learning environments improve student performance in activities?*

This question aims to identify whether the papers selected support the expectation that automated feedback approaches improve student performance in activities compared to conditions without automatic feedback tools.

[1] <http://lapes.dc.ufscar.br/tools/start_tool>

RESEARCH QUESTION 2 (RQ2): *What are the main goals in using automatic feedback generation techniques in online learning environments?*

In particular, this research question aimed to explore whether the objective of the studies was to: (i) help students with specific content, (ii) support students to improve self-regulation, and (iii) assist instructors in the creation of feedback.

RESEARCH QUESTION 3 (RQ3): *Is there any evidence suggesting that automatic feedback can help instructors?*

This question examines whether approaches proposed by existing studies provide evidence that the use of automatic feedback tools/resources enhances the capability of instructors to develop better feedback.

RESEARCH QUESTION 4 (RQ4): *What techniques are commonly used to generate automatic feedback?*

This question aimed to investigate which techniques and resources had been used to generate automatic feedback. The techniques could be, for example, machine learning, natural language processing, and ontologies.

4.2.2 Search Strategy

According to Kitchenham (2004), in a systematic literature review, it is necessary to determine and follow a search strategy. The first stage is to define the keywords and their possible combinations. In this step, we followed the same approach used by Tenório et al. (2016). The following keywords (and their synonyms) were used:

- feedback;
- online learning environment (virtual learning environments, massive open online courses, MOOC, intelligent tutoring system, e-learning, online courses, distance education, educational environment, learning management system);
- student (learner);
- instructor (tutor, teacher).

After defining the keywords and their synonyms, we built a search string using the logical operators (OR) and (AND). The operators (OR) and (AND) were used between the synonyms and keywords, respectively. Therefore, the following search strings were generated:

1. “feedback”;
2. “online learning environment” OR “virtual learning environments” OR “educational environment”;
3. “massive open online courses” OR “MOOC”;
4. “intelligent tutoring system”;
5. “e-learning” OR “online courses” OR “distance education” OR “educational environment” OR “learning management system”;
6. “student” OR “learner”;
7. “teacher"OR “tutor” OR “instructor”.

The ultimate combination of the search string used was:
((1) AND (2 OR 3 OR 4 OR 5) AND (6 OR 7))

We employed the proposed search string in the following databases that are prominent in publishing research in the field of educational technology Tenório et al. (2016):

- ACM – (<https://dl.acm.org/>)
- IEEExplorer – (<https://ieeexplore.ieee.org/>)
- Engineering Village (<https://www.engineeringvillage.com/>)
- Science Direct – (<https://www.sciencedirect.com>)
- Scopus – (<http://www.scopus.com>)
- SpringerLink – (<https://link.springer.com/>)

4.2.3 Selection Criteria

In this step, the studies have to meet the selection criteria (inclusion and exclusion) to be included in the systematic review (KEELE et al., 2007). Table 3 summarizes the step-by-step of our selection criteria.

Tabela 3 – Selection criteria.

Number	Type	Description
1	Inclusion	Primary study
2	Inclusion	Study that proposes an automatic feedback approach in Online Learning Environments
3	Inclusion	Study published from January 2009 to December 2018 (10 years)
4	Exclusion	Secondary and tertiary studies
5	Exclusion	Short papers (<5 pages)
6	Exclusion	Duplicated studies
7	Exclusion	Non-English written papers
8	Exclusion	Grey literature
9	Exclusion	Incomplete Studies

Fonte: CAVALCANTI et al. (2021)

4.2.4 Selection Process

In step one, the reviewers only read the title and abstract and decide to include or exclude the study based on inclusion and exclusion criteria. If the reviewers did not have enough information to exclude, the study went to the next step, where the reviewers read the introduction and final considerations in order to define the relevance of the paper to the review.

4.2.5 Extraction Process

In step three, the reviewers read the full text of the articles to extract data relevant to answering our research questions. Table 4 shows all the fields that were extracted from the articles.

Tabela 4 – Data extraction form fields.

#	Field	Description
1	ID	Unique identifier for the study
2	Title	Title of the paper.
3	Authors	Authors of the paper.
4	Year	Year which the paper was published.
5	Country	Country of the first author of the paper.
6	Type	Conference, journal, and workshop.
7	Educational Tool	Does it propose a new tool?
8	Is the tool available?	If yes, what is the URL?
9	Database	If the study uses or proposed a corpus for analyzing a feedback system.
10	Tools	Tools used in the study.
11	Type of evaluation	Experiment, case study, application in the real environment, and questionnaires, among others.
12	Subject area	Subject area of the course in which the system was applied.
13	Main results	What are the main results of the paper?
14	Educational Level	Higher education, secondary education, primary education, N/A (i.e., no enough data to conclude).
15	Impact on student performance (RQ1)	Evidence of positive or negative impact
16	Main Goal (RQ2)	What are the main goals of the paper?
17	Impact on teaching? (RQ3)	Evidence of positive or negative impact
18	Methods (RQ4)	What techniques were used to generate automatic feedback?

Fonte: CAVALCANTI et al. (2021)

4.3 EXECUTION OF THE SYSTEMATIC LITERATURE REVIEW

This section describes the execution of the systematic review of the literature. The first step was to use the search string in the digital libraries and download all returned articles in the .bibTex format. This step was performed manually for each digital library. Table 5 shows the number of articles obtained in each of the digital libraries.

The next step was to import the files of each digital library into the StArt tool. This step was divided into three phases: (1) Automatic removal of duplicate articles using the StArt tool; (2) The reviewers read the title and abstract of the article and applied the inclusion and exclusion criteria; (3) The reviewers read the introduction and conclusion sections of the article and applied the inclusion and exclusion criteria. Figure 8 shows the number of articles

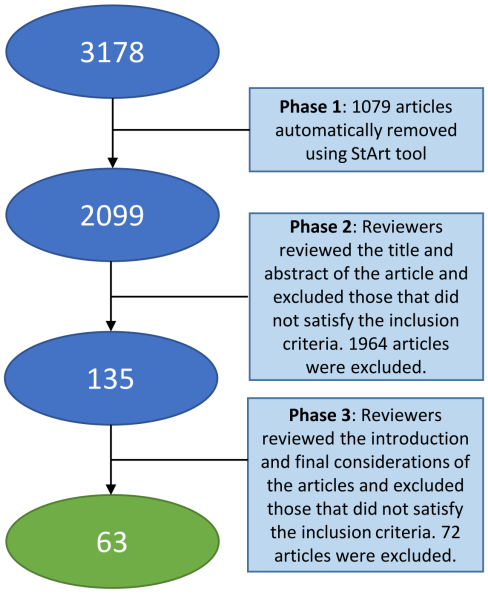
Tabela 5 – Number of articles returned for the search string in each digital library.

Digital Library	Number of articles
ACM	354
IEEExplorer	361
Engineering Village	25
Science Direc	667
Scopus	1371
Springer Link	400
Total	3178

Fonte: CAVALCANTI et al. (2021)

selected in each phase.

Figura 8 – Selection phases of articles.



Fonte: CAVALCANTI et al. (2021)

In Phase 1, duplicate articles were automatically removed using the StArt tool. The tool can detect same articles comparing texts between the articles. In this phase, 1079 articles were removed. In Phase 2, the reviewers excluded 1964 articles that did not satisfy the inclusion criteria. About 92% of articles were excluded because they were out of scope, 3% grey literature, 3% short papers, 1% secondary and tertiary studies, 0.9% duplicated studies, and 0.1% incomplete studies. It is important to note that an article may have been removed by more than one exclusion criteria.

Some information, such as the number of pages of an article or keywords, sometimes did not appear in the StArt tool. Therefore, the reviewers did not have enough information to determine whether the article would be included or excluded in phase 2. As a result, researchers reviewed these articles manually in phase 3. In this phase, some articles such as short papers and articles not written in English were discovered and excluded. In summary, in Phase 3, the reviewers excluded 72 articles that did not satisfy the inclusion criteria. About 74% of articles were excluded because they were out of scope, 20% short papers, 2% incomplete studies, 2% duplicated studies, and 2% non-English written papers. The final number of included studies in this systematic literature review was 63.

4.4 RESULTS

This section summarises findings of the systematic literature review based on 63 selected studies. The attributes extracted from each study are shown in Table 4.

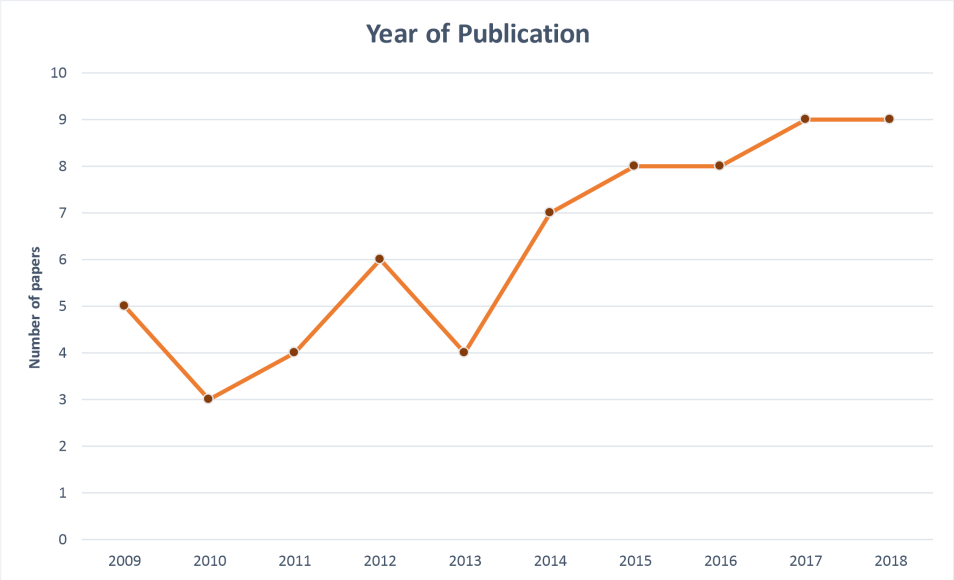
4.4.1 Year of Publication

The first attribute to be analyzed is the year of publication. Figure 9 shows the division of studies by year of publication. The figure shows an increase in publications in recent years on automatic feedback. The last four years (2015 to 2018) had more than 50% of the articles in comparison with the early years (2009 to 2014). During the period analyzed, the year with the lowest publications was 2010 (n = 3) and the years with the highest publications were 2017 and 2018 (n = 9).

4.4.2 Type of Publication

The second attribute is the type of publication, that is, whether the source of the publication was conference proceedings, journal, workshop proceedings, or others. Table 6 shows the division of types of publications by digital library. Most of the publications were conference papers (71%), followed by journal articles (27%), and workshop papers (2%). Examples of the most commonly publication venues are the IEEE Transactions on Learning Technologies, Assessment & Evaluation in Higher Education, Computers in Human Behavior, International Conference on Advanced Learning Technologies, and International Conference on Computing

Figura 9 – Year of publication of the selected studies.



Fonte: CAVALCANTI et al. (2021)

Education Research.

Tabela 6 – Type of publication by Digital Library of the selected studies.

Digital Library	Conference	Journal	Workshop
ACM	25 (39%)	1 (2%)	0 (0%)
IEEE	15 (24%)	7 (11%)	0 (0%)
Scopus	4 (6%)	1 (2%)	0 (0%)
Science Direct	0 (0%)	3 (4%)	0 (0%)
Springer	1 (2%)	5 (8%)	1 (2%)
TOTAL	**45 (71%)**	**17 (27%)**	**1 (2%)**

Fonte: CAVALCANTI et al. (2021)

4.4.3 Publication Country

Table 7 shows the number of articles by country, which is derived from the address of the first author of the articles included into the review. The country with the most publications was the USA (n = 10), followed by the United Kingdom (UK) (n = 6), and The Netherlands and China (n = 5 each).

Tabela 7 – Number of articles by country.

Country	Number of articles
USA	10
UK	6
Netherlands, China	5
Spain	4
Finland, Japan	3
Belgium, Tunisia, India, Germany, Taiwan	2
Australia, Indonesia, Bahrain, Turkey, Korea, Romania, Chipre, Singapore, Serbia, Malaysia, New Zealand, Brazil, Croatia, Ireland, South Korea, Colombia and Canada	1
Total	**63**

Fonte: CAVALCANTI et al. (2021)

4.4.4 Subject area

As the articles aimed to propose an automatic feedback system in educational environments, we analyzed subject areas in which automatic feedback was applied and categorized them as shown in Table 8.

Tabela 8 – Course that the system was applied.

Course	Number of articles
Programming	19
Computer science	7
Game Exercises	4
Mathematics	3
Circuits	3
Engineering course	2
Different courses	2
Foreign Language Learning	2
Software Engineering	2
None	5
Other subject area	17

Fonte: CAVALCANTI et al. (2021)

The most popular subject area (n = 19) was computer programming, that is, courses

that teach some programming language to computer science students. We did not take into account which programming languages were covered in those courses. The second most common subject area (n = 7) was related to different aspects of computer science. It is important to mention that an article may have had more than one course involved. Some more specific subject areas were categorized as "Other subject areas". Among them are courses such as Computer Hardware, Chemistry, Data Networking, Biotechnology, Biochemistry, Data Structures, SQL Programming, Handwriting, and Electrical Engineering. Some articles did not present any course (n = 5).

4.4.5 Research Questions

In addition to the above attributes, we extracted information relevant to the five research questions, as described in Table 4 in section 4.2.5.

4.4.5.1 Research Question 1

The first research question "**Does automated feedback in online learning environments improve students performance in activities?**" investigated if automatic feedback helped student performance. We coded the articles based on: (i) the result, if the feedback had a positive or negative influence, and (ii) the evaluation, if the study presented an empirical evaluation of an automatic feedback system or not. We also coded those papers with "no evidence" in cases where an feedback practice/tool was described without a consistent evaluation. Table 9 shows the results.

Tabela 9 – Statistics about the papers related to student performance.

Evidence	**Number of articles (%)**
No evidence	22 (34.92%)
Positive with empirical evaluation	32 (50.79%)
Positive without empirical evaluation	9 (14.28%)
Negative with empirical evaluation	0 (0%)
Negative without empirical evaluation	0 (0%)
Total	**63 (100%)**

Fonte: CAVALCANTI et al. (2021)

As already reported in other studies, manual instructor feedback helps student performance (HATTIE; TIMPERLEY, 2007; PARIKH; MCREELIS; HODGES, 2001; NICOL; MACFARLANE-DICK, 2006). This has also been reflected in automatic feedback, with 65.07% of positive results (50.79% with and 14.28% without empirical evaluation), proving that feedback is an important factor in the teaching/learning process, whether it is manual or automatic. For instance, Krusche e Seitz (2018) administered a survey in a programming language course to the students of a computer science major program to analyze the impact of feedback. Krusche e Seitz (2018) concluded that automatic feedback increased the students' participation in the exercises and submission of solutions. They also reported that more than 60% of the students successfully completed the course tasks.

Some papers showed positive results with empirical evaluation used a methodology where they first proposed an activity without the use of an automatic feedback tool and then another activity with the aid of the automatic feedback tool. Thus, they showed the change in the students' behavior and their increase in performance, since feedback informed student learning by showing mistakes and successes. For instance, in the study by Krusche e Seitz (2018) students stated that the test results and feedback helped solve the exercises, in an introductory programming course, and they enjoyed working with the ArTEMiS tool as it provided instant feedback. In an online questionnaire, the authors found that over 90% of students found interactive instructions useful to improve exercise performance. Kebodeaux, Field e Hammond (2011) presented a sketch recognition based tutoring system (Mechanix) that provides immediate feedback on problems in statics for engineering. The system was evaluated in an introduction to engineering course for 2 semesters on 2 different tasks. The results showed that students who used the tool to answer the task had a significantly higher score (p-value <0.001), with an average difference of 2.5 out of 10 points that those who took the course before Mechanix is introduced. These results were attributed to the fact that the tool provides immediate feedback to students before sending the final answer.

Several other studies compared between conditions with and without the use of automatic feedback. Al-Hamad e Mohieldin (2013) proposed an E-assessment tool that supports the design of the assessment of a chemistry course. The tool showed the learning outcomes to the student where the learning outcomes were augmented with qualitative feedback written by the instructor,. Their results showed that the quality of teaching and learning could be improved by using technology to increase faculty efficiency and provide students with real-time feedback mechanisms to help them develop a culture of self-monitoring and self-assessment. In the

study by Wong, Taylor e Beaumont (2012) a tool was proposed for facilitating an efficient and transparent coursework assessment and feedback process. The tool was evaluated in a computer science degree program. Most students who participated in the survey reported that they preferred to receive feedback through the proposed system because the feedback was easy to read and it highlighted mistakes that students made. Results from a comparative study between an experimental group and a control group, of students from a programming course, showed that using the proposed tool (an on-line multiple choice questions system integrated with a neural network) improved the learning outcomes (ALEMÁN; PALMER-BROWN; DRAGANOVA, 2010).

Studies that did not perform empirical evaluation but indicated positive results (14.28%) generally focused on assessing student satisfaction with tools. For example, the results of the study by Wang, Singh e Su (2018), which proposed a data-driven program repair framework to automate feedback generation for introductory programming exercises. The study showed that the system was effective and could generate concise and useful feedback for 89.7% of the incorrect student submissions, in just two seconds on average. The study by Keuning, Heeren e Jeuring (2014) presents a prototype of a programming tutor to help students with feedback and hints. The authors found that they could recognize between 33% and 75% of the exercises solutions collected during two programming courses. Zhou et al. (2018) analyzed the design of existing online judge systems (WASIK et al., 2018) and their advantages and problems in applying to programming education. The authors state that after applying the system in a course on the C programming language, the students' performance and satisfied grades increase. However, the article does not show details about this assessment.

Papers that did not show any evidence (34.92%) are more descriptive, showing details about the tools and how they work. It means that they did not present evidence on how feedback potentially enhanced student performance. For example, the study by Lan et al. (2015) presents the development of a framework for mathematical language processing (MLP). As a result, the authors stated that the structure could substantially reduce the human effort required for classification in large-scale courses and also allows instructors to visualize solution groups to help them identify groups of students with the same misconceptions. The study by Lodder, Heeren e Jeuring (2017) describes a system that is part of a set of tools that help students study logic by providing automatic feedback. The authors state that the performance of the system for resolution logic proofs reached a quality comparable to that of a group of experts. The work proposed by Ying e Hong (2011) presents a SQL (Structured Query

Language) teaching system with an automatic feedback mechanism. The system helps the student in the construction of SQL queries. This system provides tips to assist the students in the understanding of a specific concept of SQL more quickly and then verify the effectiveness of the student solution to the exercise.

4.4.5.2 Research Question 2

The answer for the second research question "**What are the main goals in using automatic feedback generation techniques in online learning environments?**" is shown in Table 10. The articles had very specific objectives. We grouped the articles included into this systematic literature review into four categories based on the objectives of the feedback approaches.

Tabela 10 – Main goals for using automatic feedback generation.

Goal	**Number of articles (%)**
Use feedback to help students on a specific content/-course	33 (52.38%)
Use feedback to support self-regulation	26 (41.26%)
Use feedback to help instructors	3 (4.76%)
Use feedback to reduce plagiarism behavior	1 (1.60%)
Total	**63 (100%)**

Fonte: CAVALCANTI et al. (2021)

The articles that help students in a particular content/course (52.38%) are systems developed to assist programming courses (KARAVIRTA; HELMINEN; IHANTOLA, 2012; ARENDS et al., 2017), teaching of circuit analysis (BANERES et al., 2014; WEYTEN et al., 2010), and teaching of a foreign language (ONO; ISHIHARA; YAMASHIRO, 2013; MURAD et al., 2018), among others. These systems often provide feedback to show students what went wrong (showing where the error is and giving tips on how to get the answer right) or well (showing a congratulations message) (KRUSCHE; SEITZ, 2018; MARIN et al., 2017). In the article by Weyten, Rombouts e Maeyer (2008), a new web-based system for training students in electrical and electronic circuit theory is presented. The system can be used to gain valuable information from students and thereby bring improvements in instructor teaching. The article by Bryfczynski et al. (2013) describes an intelligent tutoring system called beSocratic, which assists students who

study data structures; the students can be evaluated and the results of their task completions are analyzed automatically to help instructors refine their activities and improve future performance. The study by Ono, Ishihara e Yamashiro (2013) proposed a new type of feedback system based on a text mining method. The system encourages students to reflect on their own presentation and has shown positive results in the use of foreign language teaching in Japan. The tool proposed by D'antoni et al. (2015) aims to provide feedback for the construction of a deterministic finite automaton that accepts chains corresponding to a described pattern. The system provides automatic feedback with counterexamples or tips so that students can complete the activity.

In contrast to the first main goal, the second goal (41.26%) of the included studies was to provide more general feedback to promote self-regulated learning. These articles generally are focused on providing personalized feedback (DEMAIDI; GABER; FILER, 2018), gamification (UTOMO; SANTOSO, 2015; YING; YANG; DENG, 2012), or dashboards (DAVIS et al., 2017; YU, 2016) in an online environment to motivate students, detect poor performance and reduce dropouts (KHAN; PARDO, 2016). The paper by Jin (2017) presents a visualization tool to motivate students to participate in collaborative online learning communities actively. The work of Smithies et al. (2010) presents a tool called CONSPECT, which aims to provide formative feedback and monitor students' conceptual development. It uses an NLP method, based on latent semantic analysis, to compare student answers to generated reference models. The article by Alencar e Netto (2014) introduces TUCUMÃ, an intelligent 3D virtual agent integrated with Moodle for virtual learning. The tool automatically simulates a distance course tutor, monitors student activities, and answering student questions through dialogue.

As Table 10 shows, only 3 studies aimed to assist instructors. Several studies that introduce approaches to help students in online learning operate under the assumption that automatic feedback can also benefit instructors in terms of teaching efficiency (XIE; LI, 2018; MARTIN et al., 2009). We hypothesise that if the student can learn from automatically generated feedback, these systems have great potential to reduce the effort of the instructors in answering questions or giving feedback to the students. Our third research question explored this assumption.

The study by Akçapınar (2015) is the only article which presented an automatic feedback system with the goal of reducing the plagiarism behaviour of students. This study aimed to reduce students' plagiarism in written tasks by providing automated feedback based on text mining analysis.

4.4.5.3 Research Question 3

The third research question "**Are there indications that automatic feedback helps instructors?**" aimed to understand if the approaches proposed in the literature provided insights and assisted instructors during preparation and teaching phases. Table 11 shows the information found in the literature as response to this question.

Tabela 11 – Numbers of papers that show the support of automatic feedback system to instructors.

Evidence	**Number of articles (%)**
No evidence	29 (46.03%)
Positive with empirical evaluation	6 (9.52%)
Positive without empirical evaluation	28 (44.44%)
Negative with empirical evaluation	0 (0%)
Negative without empirical evaluation	0 (0%)
Total	**63 (100%)**

Fonte: CAVALCANTI et al. (2021)

Most articles (46.03%) have not shown any evidence in their findings whether automatic feedback helps the instructor. This result crosspond to the objectives of existing studies (Section 4.4.5.2), where 93.64% of studies (articles that use feedback to help students on specific content or in specific disciplines and articles that use feedback to support learning) aimed to assist student learning using automated feedback. Only 3 studies (4.76%) aimed to help instructors (see Table 10). Among these studies, Martin et al. (2009) were able to support instructors' needs when they tried to integrate various learning systems to improve students' learning process. The authors proposed a system called MAGADI that helps instructors with visualizations of relevant information about students. Trausan-Matu, Dascalu e Rebedea (2014) proposed the PolyCAFe system, which provides tools that support a polyphonic analysis of chat conversations and discussions of small student groups on online forums. The system uses NLP to identify topics, semantic similarities, and links between utterances. A statement chart is created with the detected links, which is the central element for polyphonic analysis and for providing automatic feedback and support for instructors and students. The study by Xie e Li (2018) proposes a system that combines a recommendation model based on big data content and a clustering model to personalize exercises and feedback in online education.

The majority of the papers presented positive results, with and without empirical evaluation, (53.96 %). The main goal of the studies selected in this review, as shown in Table 10, was to assist online learning in specific disciplines. Furthermore, these studies also indicate success in reducing the instructor's workload, since the amount of questions and problems from students is reduced using the automatic feedback system (KRUSCHE; SEITZ, 2018). Most articles (44.44%) claim that automatic feedback helps instructors but offer no empirical evaluation to support such claims. These studies demonstrate in their results the satisfaction reported by instructors regarding the reduction of students doubts (WONG; TAYLOR; BEAUMONT, 2012; KRUSCHE; SEITZ, 2018) or the reduction of instructors' workload (MARIN et al., 2017; FAST et al., 2013). There were no negative results in this analysis.

4.4.5.4 Research Question 4

The fourth research question aimed to analyze which methods and techniques are used in the automatic generation of feedback. Table 12 shows the main methods and techniques found in the articles.

Tabela 12 – Main methods and techniques used to generate automatic feedback.

Method	**Number of articles**
Comparison with desired solution	15
No details	14
Dashboard	7
Natural Language Processing	7
Ontology	4
Graphs	3
Neural Network	2

Fonte: CAVALCANTI et al. (2021)

The main technique used was the comparison between student answers and the desired solution (15 articles). Among these articles are those that aim to propose feedback to help students solve specific exercises in subjects, such as programming, circuit analysis, automaton, among others. In this way, the proposed systems provided instant feedback comparing a student's response with a possible response already registered in the system. The research by (LODDER; HEEREN; JEURING, 2017) used this method to determine the quality of LOGAX (a

tutoring tool that helps students to build an axiomatic proof), comparing the proofs generated by experts and student solutions.

Many articles (n=14) have not detailed the methods or techniques used to generate automatic feedback. Most of these articles propose prototype systems that are still in the development phase (RIOFRÍO-LUZCANDO; RAMIREZ; BERROCAL-LOBO, 2017; EFSTATHIOU et al., 2018; JEREMIĆ; JOVANOVIĆ; GAŠEVIĆ, 2012). Other articles do not describe how the proposed systems were developed, but only describe how they were applied in a real environment and results of the implementation (AL-HAMAD; MOHIELDIN, 2013; WONG; TAYLOR; BEAUMONT, 2012; YING; YANG; DENG, 2012; KEBODEAUX; FIELD; HAMMOND, 2011).

The second most used technique was dashboard (n=7). These studies used graphical elements to motivate the students or the class to carry out activities in the online environment. For example, the article by (KHAN; PARDO, 2016) presents a study that categorised students based on how they interact with the dashboard, taking into account time, number and timing of hits.

The third most used technique was Natural Language Processing (NLP). NLP is a field of computer science applied to manipulate text or speech in natural language that can process and analyze large amounts of data in natural language using algorithms for semantic and syntactic analysis. For example, Trausan-Matu, Dascalu e Rebedea (2014) proposed a system which provides tools that support the polyphonic analysis of chat conversations and online discussion forums, and NLP is used in order to identify topics, semantic similarities and links between utterances. In the study by Ono, Ishihara e Yamashiro (2013) text mining technology was used to produce instant feedback in a foreign language presentation course.

Other specific methods found in the articles are: longest common subsequences (LCS), feature extraction with clustering, cybernetic principles, linguistic analysis engine, tree edit distance, abstract syntax trees (ASTs), knowledge databases, predictive analytics, mobile sensors, and data mining. It is important to note that 1 article can have more than 1 method.

4.5 DISCUSSION

Based on the insights obtained from this systematic literature review, we highlight three factors that should be considered when researching and developing systems to provide feedback. These factors include methods and goals, relevance for instructors, and techniques adopted. Based on our results, these factors were considered critical in the process of sending feedback.

Each of the three factors is discussed in the remainder of the section.

4.5.1 Feedback impact and educational goals

The first research question aimed to assess the impact of the automatically provided feedback on students' performance. In this case, performance could be related to a specific activity or the final marks. Unsurprisingly, the majority of the papers retrieved in this review, about 65% (Table 9), concluded that the feedback had a positive impact on students performance (HATTIE; GAN, 2011). However, the papers do not provide enough information to determine if the positive impact was caused by the use of the tool or the feedback final product (dashboard/message). For instance, several papers proposed tools to evaluate programming activities automatically (GULWANI; RADIČEK; ZULEGER, 2014; D'ANTONI et al., 2015; BIRCH; FISCHER; POPPLETON, 2016). In this case, the authors do not analyse if the improvement in the students' abilities was due to the usage of the entire system or just because of the feedback. Price et al. (2010b) suggests that the perceived value of feedback and the students' final performance should be analysed separately.

Additionally, a few papers reported an increase in the students' performance, but some degrees of dissatisfaction with the feedback message. In this direction, Burke (2009) presented several factors that led to poor evaluation of the feedback, even with the improvement of final marks. The students listed the feedback length (brief), polarity (always negative), and complexity (difficult to interpret) as the main drawbacks (BURKE, 2009). Possible reasons for this can be the lack of training related to good feedback practices. Weaver (2006b) showed that more than 50 % of university students never received any guidance on "how to understand and use feedback", and three-quarters of students received no advice on how to understand and use feedback before university, and Mutch (2003) highlighted the need for more research on how students "receive and respond" to feedback.This is inline with what Carless e Boud (2018) refers to as the importance of the student feedback literacy to enhance the feedback impact. Carless e Boud (2018) also states that the instructors have a key role to enable students to appreciating, making judgments, managing affect, and taking action on the feedback messages. Moreover, the current literature also offers recommendations for good feedback practices. Nicol e Macfarlane-Dick (2006) proposed a conceptual model of self-regulation based on a review of the research literature on formative assessment and feedback. The main idea of the work is to identify how the training processes of evaluation and feedback can help promote self-

regulation. Based on the conceptual model, seven principles of good feedback practices to enhance teaching feedback were proposed.

In our review, we also analysed the educational goals of the feedback systems. Table 10 revealed that more than half of the systems (52.38%) aimed to provide feedback about a specific content/course. More specifically, the majority of these papers were applied to student performance (Table 9) and more procedural and specific activities, such as Programming and circuit analysis (Table 8). This result could explain the possible reasons for the weaknesses such as length (brief), polarity (always negative), and complexity (difficult to interpret).

A total of 41.26% of the articles in this review used the feedback as a method to support self-regulation. This goal is more aligned with the literature on good practises of feedback (NICOL; MACFARLANE-DICK, 2006; HATTIE; TIMPERLEY, 2007). However, these works did not present an analysis to support the effectiveness of the feedback in terms of improving the students performance and self-regulation processes (41.27% listed as no evidence in Table 9). Moreover, the majority of these papers proposed the adoption of a dashboard to support students. However, the literature shows that this kind of visualisation does not guarantee effective feedback and does not offer sufficient support for self-regulated learning (MATCHA et al., 2019a).

The papers in this review have not considered several factors that are well-established in the literature to enhance the feedback process. They do not align the proposed feedback systems with educational research on the provision of feedback, which could be extremely helpful in order to improve the final result of the feedback process, in terms of learning outcomes, learning processes, and students' satisfaction.

There are several popular frameworks for good feedback practices that are proposed in educational research. For instance, Nicol e Macfarlane-Dick (2006) suggested incorporating more than just simple instructions in feedback messages. According to Nicol e Macfarlane-Dick (2006), good feedback practice is broadly defined as anything that might strengthen students' capacity to self-regulate their performance. Although Nicol e Macfarlane-Dick (2006) suggested several good practises to enhance feedback, the papers included in this literature review just focused o provide specific information related to the activities. It is a limitation that could had influenced the students' satisfaction level reported in the studies.

Educational research has documented factors that should be considered when creating feedback. Hattie e Timperley (2007) investigate several conditions that could maximise the positive effects of feedback on student learning, including the increase in student awareness

about the overall learning goal, the progress towards the goal, and the subsequent goals required to achieve the overall goal. Hattie e Timperley (2007) also propose a model that encapsulates four levels of information to be considered in feedback messages: (i) task level such as whether the activity is correct or incorrect, can include instructions for more or different information; (ii) process level includes suggestions about study methods to the student to create a product or complete a task and is more directed to information processing, or learning processes that require understanding or completing the task; (iii) self-regulation level which includes greater self-assessment or confidence skills, can have major influences on self-efficacy, self-regulatory proficiency, and students' personal beliefs as learners; (iv) self level, feedback can be personal in the sense that it is directed to the self; self-level is often unrelated to task performance. Hattie e Timperley (2007) research showed that the most potent feedback is on process and self-regulation levels, while self-level is usually ineffective for learning. Task level is typically ineffective, unless it is combined with either process or self-regulation levels. Feedback in the systems proposed in the papers included in this review is focused on the task level only, which reduces the potential of feedback to positively impact in students motivations and participation in class (ROBISON; MCQUIGGAN; LESTER, 2009).

Finally, the papers in this review do not consider feedback as a dialogic process. Pardo (2018), Pardo et al. (2019) proposed that feedback should be a process where students and instructors have a conversation about the course, assisting not only the students to understand the course content better, but raising the capability of the instructor to personalise the content and improve the course design and orchestration (DILLENBOURG et al., 2013; PRIETO et al., 2016). Furthermore, Pardo et al. (2018) also advised that timely feedback increases the chances to help students to reach the learning goals and improve their final performance. Among other things, this concept could reduce student dropout rates (LEE; CHOI, 2011).

4.5.2 Feedback relevance for instructors

The results of this study suggest that the existing feedback systems do not take in consideration the instructors needs. Table 10 shows that only 4.76% of the papers initially intended to support instructors. However, none of the systems proposed a platform to assist instructors/-teachers to better write feedback (MULLINER; TUCKER, 2017; HARVEY, 2003). The arguments for the importance of students in the feedback process are undeniable (BROOKHART, 2017). However, recent literature also advises that the instructor role is crucial in the adoption of

automatic tools for the provision of automated feedback (PARDO et al., 2018; LIM et al., 2019). More broadly, the importance of the instructor in the adoption of education technology has already been made by several researchers (GAsEVIć et al., 2019; GAšEVIć et al., 2017; ALI et al., 2013; ROGERS, 2000; ZHAO; CZIKO, 2001).

The results also point out that several feedback systems (53.96%) have impacted the instructor experience positively, as shown in Table 11. Possible reasons for this can be found in the capability of automation and personalization provided by the feedback systems, which can potentially decrease the instructors' workload (HENTEA; SHEA; PENNINGTON, 2003; MANOHARAN, 2016; SHERIDAN, 2006). The majority of the works retrieved in this review that reduce the activities performed by instructors are based on student dashboards showing simple statistics (KEBODEAUX; FIELD; HAMMOND, 2011; YU, 2016), systems comparing the students' results with a pre-defined desired solution (LAN et al., 2015; BANERES et al., 2014), or feedback in a particular domain, e.g., programming language problems (KARAVIRTA; HELMINEN; IHANTOLA, 2012; ARENDS et al., 2017; ALEMÁN; PALMER-BROWN; DRAGANOVA, 2010; HELMINEN; MALMI, 2010; KEUNING; HEEREN; JEURING, 2014) and essays evaluation (WHITELOCK et al., 2015; TOSHNIWAL et al., 2015; USENER, 2015). In a nutshell, this result reveals a preponderance of papers related to intelligent tutoring systems in the provision of automatic feedback which explains the decrease of instructors' workload (POLSON; RICHARDSON, 2013). Nevertheless, this approach fails to provide analytics to inform the instructor to support the feedback process and inform their teaching alongside an automatic feedback system.

Current educational research indicates that supplying instructors with relevant information about the students and the learning environment could enhance the capability of instructors to provide more informative feedback at scale and adjust the course content/methodology to reach better educational results (PARDO et al., 2019; DILLENBOURG et al., 2013; PRIETO et al., 2016). Learning dashboards focusing on the use of visualizations to support instructors are potentially a powerful instrument to understand student behaviour supporting the provision of feedback (CHARLEER; KLERKX; DUVAL, 2014; VERBERT et al., 2014). However, learning dashboards have to be carefully designed to support instructional decision-making. Wise e Jung (2019) concluded that a learning dashboard for instructors should contain informative content regarding the students activities and learning context; otherwise, the instructors will not engage or take action based on the visualization.

Many authors define feedback as a dialogical process whereby learners obtain information on their performance and instructors better understand students needs (BOUD; MOLLOY, 2013;

PARDO, 2018). In other words, feedback should not be unidirectional from instructors to learners, but it has to incorporate information for both actors. From the instructor's point of view, the dialogue enabled by feedback could aid the process of refining course design and the orchestration of activities (DILLENBOURG et al., 2013; WISE, 2014; PRIETO et al., 2016). Therefore, an essential improvement in the current feedback systems is to provide support for both instructors and students and the entire process of feedback instead of just support specific tasks in a course design such as programming tasks.

Finally, Dawson et al. (2019) advise including only content related information for students is not enough to provide a good feedback message, it is also important to include affective aspects in feedback that encourage positive motivational beliefs and provide information that can be used to help shape teaching (NICOL; MACFARLANE-DICK, 2006; HATTIE; TIMPERLEY, 2007). Thus, future research on systems that aim to assist instructors with feedback provision is the analysis of the message content with the aim to suggest improvement of quality of the overall feedback, including non-content aspects (CAVALCANTI et al., 2019; CAVALCANTI et al., 2020).

4.5.3 Techniques adopted to provide feedback

The last research question of this study aimed to identify which methods, tools and techniques were applied to provide the feedback and to discuss their alignment with the educational goals and student performance. Specifically, we analysed how researchers develop approaches to create and send feedback messages in online environments. Our analysis suggests that commonly adopted methods are direct comparison of students answers with the desired solutions (pre-defined by instructors) (DUTCHUK; MUHAMMADI; LIN, 2009; MITROVIC et al., 2011; USENER, 2015; BIRCH; FISCHER; POPPLETON, 2016; LODDER; HEEREN; JEURING, 2017), dashboards/graph visualizations (YU, 2016; KHAN; PARDO, 2016; JIN, 2017; BODILY et al., 2018), and natural language processing/machine learning (ONO; ISHIHARA; YAMASHIRO, 2013; JUGO; KOVAČIĆ; SLAVUJ, 2014; TRAUSAN-MATU; DASCALU; REBEDEA, 2014; CORRIGAN et al., 2015).

The majority of the papers that focused on comparing students answers with the desired solutions were reported in programming or automation courses where the main goal is to evaluate programming activities automatically, providing information on the students' performance and possible improvements to enhance software programs (as shown in Table 8). The literature confirms the importance of answer comparison to provide feedback to students (IHANTOLA et

al., 2010; KEUNING; JEURING; HEEREN, 2016), but it has two main limitations: (i) it is necessary for the instructor to register the answer in the system beforehand; and (ii) the student must respond exactly the same as the answer given by the instructor. Unsurprisingly, it provides minimal information to students and is not connected with the good practices of feedback found in the educational research literature (NICOL; MACFARLANE-DICK, 2006). For instance, Nicol e Macfarlane-Dick (2006) suggests that for feedback to be effective, it is necessary to provide more valuable information such as helping to clarify good performance or encouraging students with positive motivational beliefs. More importantly, it should offer guidance to the students in terms of learning strategy (HATTIE; TIMPERLEY, 2007) they can adopt to learn the concept they missed to answer correctly. To achieve this, automatic feedback should not only consider students responses on assessments, but it should also include data about how students' learning proces. Therefore, the recent literature recommendation for this kind of information is to inform instructors on students progress systematically, so they could write effective feedback messages on student activity and performance (BLIKSTEIN et al., 2014; PARDO et al., 2019).

Dashboards and visualisations have also been widely used to provide student feedback on their learning process and progress (YU, 2016; SCHWENDIMANN et al., 2016; KHAN; PARDO, 2016; JIN, 2017; BODILY et al., 2018). Few studies however demonstrated that these visualisations were effective in improving students performance (UTOMO; SANTOSO, 2015; DAVIS et al., 2017). However, a systematic review on learning analytics dashboards by Matcha et al. (2019a) reveals negative effects of dashboards on students and the need for improvement in dashboards to address the recommendations for effective feedback. None of the studies included in this review presented enough evidence of effectiveness in the provision of feedback for students by dashboards.

Some authors used natural language processing and machine learning techniques to provide or assist in the feedback process (ONO; ISHIHARA; YAMASHIRO, 2013; JUGO; KOVAČIĆ; SLAVUJ, 2014; TRAUSAN-MATU; DASCALU; REBEDEA, 2014; CORRIGAN et al., 2015). The development of the fields of learning analytics and educational data mining could explain the increase in the adoption of these methods that should become a trend in the field of (semi-)automatic feedback systems (ER; DIMITRIADIS; GAŠEVIĆ, 2020; TEMPELAAR; NGUYEN; RIENTIES, 2020; TSIAKMAKI et al., 2020; CAVALCANTI et al., 2020). However, this kind of application requires a substantial amount of data to build a consistent model that works for different contexts (BARBOSA et al., 2020; ROMERO; VENTURA, 2020). In sum, this line of work has a considerable potential to provide useful information, but problems such as data contamination should be

carefully avoided (FARROW; MOORE; GAŠEVIĆ, 2019).

Finally, machine learning approaches could solve several shortcomings of the feedback systems by proposing methods for:

- **Analysis of feedback quality**: Many studies included in this review aimed to provide automatic feedback. However, few papers have attempted to analyse the quality of feedback provided through forms applied to students and/or instructors. The recent paper by (CAVALCANTI et al., 2019) focused on the analysis of the feedback quality extracted from evaluations collected in an online course offered at a Brazilian higher education institution. It shows the potential of using machine learning to achieve this goal.

- **Automatic feedback generation**: Almost all work reviewed in this study aimed to provide feedback for a specific context, for instance, introduction to programming, circuit analysis, and foreign language essays evaluation. Yet, the papers reviewed did not present any evidence of the generalizability of their approaches. One possible solution is to use Natural Language Generation (NLG) techniques to produce automatic feedback (PERERA; NAND, 2017). NLG is defined as a systematic approach to produce human-understandable natural language texts based on analytics or representations of meaning.

4.6 LIMITATIONS

The primary limitation of this study is related to the search process in which we focused on papers that only contain feedback provided in online environments. This could potentially exclude papers that describe feedback systems, but that were not evaluated in a virtual context.

Second, a few papers had limited information about the methods and techniques used, which led to several categories such as “no details"and “no evidence"in the result tables. We decided to keep these papers nevertheless because they contain information relevant to at least one research question.

4.7 CONCLUSIONS

This article presents an overview of existing studies on automatic feedback in online learning environments from 2009 to 2018. It analyzed the benefits that automatic feedback generation

can bring in relation to instructors and students. The systematic literature review showed the main techniques used and the main objectives in applying automatic feedback in online learning environments. The research questions were answered by analyzing the results of the articles and verifying whether an empirical or non-empirical evaluation was performed with positive or negative results. Research questions examined whether automatic feedback helps student performance, whether it helps the instructor, and whether it can override and be more efficient than manual feedback.

We concluded that there is evidence that automatic feedback increases student performance in activities (50.79% of articles). The main purpose of using automatic feedback systems was to help students on a specific content/discipline. Moreover, the majority of these articles have the same type of assessment: comparing students' scores in a discipline before using the system and after using the system. In this case, the studies did not show an analysis of other factors besides the feedback that could influence these results.

This study also assessed if automatic feedback can also help the instructor. As described in many of the articles included in this review, the objective of the automatic feedback systems is precisely to decrease the instructor's effort in correcting various student exercises. Our results confirm this statement showing that there is evidence that automatic feedback also helps reduce instructor effort (53.96% of articles). Finally, we found that the main methods and techniques used to generate automatic feedback were: comparison with desired solution, dashboards and natural language processing.

This systematic literature review highlighted that the main shortcoming in the research literature about the automatic provision of feedback are: (i) the insufficient use of educational research on feedback to inform development tools for automatic feedback; and (ii) the exclusive focus on students which neglect the role of teachers in feedback practice. Providing tools for instructors would inform their teaching practice and even involve them in the improvement of feedback for students.

5 AN ANALYSIS OF THE USE OF GOOD FEEDBACK PRACTICES IN ONLINE LEARNING COURSES

Authors: Anderson Pinheiro Cavalcanti, Vitor Rolim, Máverick André, Fred Freitas, Rafael Ferreira, Dragan Gašević

5.1 INTRODUCTION

Online learning is an education modality where instructors and students are physically separated in space and/or time. Online learning has widely been adopted in several contexts like primary and secondary education, higher education, MOOCs, and workplace. According to the 2017/2018 census of the Brazilian Association of Distance Education (ABED)[1], in 2017, 1,320,025 students were enrolled in fully regulated distance courses, an increase of 135.01% over the year 2016.

The growth of online learning led to the increased adoption of *Learning Management System* (LMS) (CAVALCANTI et al., 2017). A LMS provides several tools that allow interactions between instructors and students. Among them, there is an assessment resource where students can submit their responses to activities in a course. In general, this assessment resource is the main space where instructors send feedback (COATES; JAMES; BALDWIN, 2005). However, due to the significant growth of the student body and need continuous engagement, it is difficult for the instructors to provide the high quality of feedback. This can cause problems in student learning and even lead to the decline of motivation(HATTIE; TIMPERLEY, 2007).

Feedback is essential as it allows students to identify gaps in their learning and improve their learning strategies. Besides, feedback allows instructors to adapt their methods and content to students' learning needs (AL-YAHYA; GEORGE; ALFARIES, 2015). Several literature reviews have shown that useful feedback brings benefits to learning (BLACK; WILIAM, 1998; HATTIE; TIMPERLEY, 2007; NICOL; MACFARLANE-DICK, 2006). For example, The Black e Wiliam (1998) study analyzed more than 250 feedback studies conducted between 1988 and 1998, covering all education levels. This study revealed that feedback produced significant benefits in student learning and satisfaction. However, poor feedback can cause problems for student learning and motivation. Thus, it is necessary for the instructor to follow some of the principles of good feedback practices (NICOL; MACFARLANE-DICK, 2006).

[1] http://www.abed.org.br/site/pt/midiateca/censo_ead/

The paper reports on a study that aimed to analyze the quality of feedback extracted from the assessment resource collected from an online course offered by a Brazilian higher education institution. The analysis was performed by following the good feedback practices proposed by Nicol e Macfarlane-Dick (2006). The study also aimed to automate the classification of textual feedback messages to students by distinguishing between messages that had at least one good feedback practices and messages with none of them. The automation was done by developing a random forest classifier.

5.2 BACKGROUND

5.2.1 Feedback

Feedback is a topic that has been addressed in much literature published over the past two decades (HATTIE; TIMPERLEY, 2007), (NICOL; MACFARLANE-DICK, 2006), (JERIA; VILLALON, 2017), (MOREL, 2016) because it is a fundamental part of the learner's learning process. Some publications cited that the factors of student dissatisfaction with the feedback they receive are (BURKE, 2009):

- **The length**: feedback is very brief;
- **The polarity:** feedback is very negative;
- **The complexity**: feedback is very difficult to decipher or understand.

Several other issues in feedback are also emphasized in the literature. Mutch (2003) highlighted the need for more research on how students "receive and respond" to feedback. Weaver (2006b) showed that over 50% of college students had never received any guidance on 'how to understand and use feedback', and three-quarters of students had not received any advice on how to understand and use feedback before university. The increase in student enrollments of many educational environments reinforces the importance and difficulty of providing sufficiently detailed and personalized feedback (KULKARNI et al., 2013).

The current literature also offers recommendations of good feedback practice. Nicol e Macfarlane-Dick (2006) defined seven principles of good feedback practices in order to assist teaching staff, see Table 13 for details. According to Nicol e Macfarlane-Dick (2006), good

Tabela 13 – Good practices of feedback according to Nicol e Macfarlane-Dick (2006)

Good Practice	Description	Label
1	Helps clarify what good performance is (goals, criteria, expected standards)	*GP 1*
2	Facilitates the development of self-assessment (reflection) in learning	*GP 2*
3	Delivers high-quality information to students about their learning;	*GP 3*
4	Encourages teacher and peer dialogue around learning	*GP 4*
5	Encourages positive motivational beliefs and self-esteem	*GP 5*
6	Provides opportunities to close the gap between current and desired performance	*GP 6*
7	Provides information to teachers that can be used to help shape teaching	*GP 7*

Fonte: CAVALCANTI et al. (2019)

feedback practice is broadly defined as anything that might strengthen students' capacity to self-regulate their performance.

5.2.2 Automatic Text Analysis and Feedback

There has been much less literature that automatically analyzes the application of good feedback practice in feedback messages students receive from their instructors. Maitra et al. (2018) proposed an automatic classifier to investigate feedback by using the Naïve Bayes classifier. It classified the feedback provided for each student into two classes valid or invalid categories based on a proposed feedback measure. The classifier takes into account the independent contribution of different student features (i.e., effort, academic background, course outcomes achieved), in the feedback provided by the instructor. The approach was evaluated using 1000 feedback extract from an Indian Higher Education Institution. The main drawback of this approach was the lack in the details about the influence of the features used and the further implications of their approach.

Text mining methods have also been used to provide automatic feedback to students. Akçapınar (2015) aimed to reduce students' plagiarism in online tasks by providing automated feedback based on text mining analysis. Before this automatic feedback, 81.4% of the students committed plagiarism. After feedback, 83% of the students reduced post-plagiarism attempts. The number of non-plagiarists increased by 42.37% after feedback was provided.

5.2.3 Contributions

The study presented in this paper offers three main contributions: (1) a quantitative content analysis of feedback according to a well-established model (NICOL; MACFARLANE-DICK, 2006) was performed; (2) co-occurrence analysis of the components of good feedback practices was carried out; (3) an automatic classifier of textual feedback messages written in Portuguese was developed and evaluated.

5.3 RESEARCH QUESTIONS

As mentioned before, this study is based on the seven good practices for feedback as proposed by Nicol e Macfarlane-Dick (2006). Thus, we intended to analyze the occurrence of the use of and co-occurrences in use among the seven feedback practices in a study conducted in a Brazilian higher education institution. Hence, our first research question was:

What is the presence of the seven practices for good feedback in actual use and the extent to which the seven feedback practices co-occurs within the same feedback messages?

After this initial analysis, the study developed a supervised feedback message classifier to detect the presence of good practices automatically. That is, our second research question was:

Is it possible to use text mining methods to extract the indicators of good practices of feedback from texts written in Portuguese?

5.4 METHOD

5.4.1 Dataset

The dataset used in this work was obtained from a LMS used in online courses offered by a Brazilian public university. The dataset was composed of textual messages the instructor sent as feedback to the activities of their student submitted by the LMS. We collected a total of 1,000 feedbacks from the biology and literature courses divided into 41 messages for biology and 959 for literature.

5.4.2 Analysis - Research Question 1

Expert coders classifier these feedback messages according to the seven good feedback practices (NICOL; MACFARLANE-DICK, 2006), see Table 13. The inter-rater agreement was moderate with percent agreement of 76.7% and Cohen's $\kappa = 0.43$. Cohen's kappa coefficient (κ) is a statistic which measures inter-rater agreement for qualitative items. The closer to 1, the greater is the indication that there is agreement between the judges and the closer to 0, the greater is the indication that the agreement is purely random (LANDIS; KOCH, 1977a). A third evaluator resolved the divergences that occurred in the first stage of the coding (23.3% messages in total).

To understand the extent to which the seven practices co-occurred in the feedback messages, we performed the Phi correlation. In other words, we compared pairs of good practices studied to evaluate at which extent they are related to each other.

5.4.3 Analysis - Research Question 2

The dataset was divided into two classes: **class 0**: if the feedback does not belong to any of the good practices; **class 1**: if the feedback belongs to at least one of the seven good practice. Table 14 shows the distribution of the feedback messages by the two classes.

Tabela 14 – Distribution of feedback messages by the two class – (i) messages with one or more good feedback practices; and (ii) no occurrence of good feedback practice

Course	Class Number	Class Description	Quantity
Literature	0	Zero Good Practice	207
Literature	1	One or more Good Practice	752
Biology	0	Zero Good Practice	15
Biology	1	One or more Good Practice	26
		Total	**1000**

Fonte: CAVALCANTI et al. (2019)

5.4.3.1 Feature Extraction

This study followed a similar approach as applied in previous studies that made use of linguistic features in text classification rather than traditional (e.g., n-grams) text classification features (NETO et al., 2018; KOVANOVIĆ et al., 2016). One of the reasons is that the classic features are very "dataset dependent", as data itself defines the classification space (NETO et al., 2018). Therefore, the set of features was based mainly on the LIWC (TAUSCZIK; PENNEBAKER, 2010) and Coh-Metrix frameworks (MCNAMARA et al., 2014).

The *Linguistic Inquiry and Word Count* (LIWC) (TAUSCZIK; PENNEBAKER, 2010) extracts a large number of word counts that are indicative of different psychological processes, such as affective, cognitive, social and perceptual. The LIWC has 127,149 entries, where each entry can be assigned to one or more categories. The core of this tool is a lexical resource, better known as the LIWC dictionary, which was also made available in the Portuguese language (FILHO; PARDO; ALUÍSIO, 2013). A total of 64 LIWC features were extracted in the study.

Coh-Metrix is a computational tool that analyzes many different words, sentences, and multi-sentence texts. Coh-Metrix provides several measures of text coherence, linguistic complexity, text readability, and lexical category use (MCNAMARA et al., 2014). The Portuguese version of Coh-Metrix has 48 different measures (SCARTON; GASPERIN; ALUISIO, 2010).

In addition to the LIWC and Coh-Metrix features, we extracted including the number of named entities, sentiment polarity (i.e., positive, negative, and neutral), greeting presence ("Bom dia", "Boa tarde", "Como vai"), and positive expressions ("Muito bem", "Excelente", "Bom trabalho"). These features were also proposed in works of text classification (NETO et al., 2018; KOVANOVIĆ et al., 2016). The rationale for the extraction of these additional features is as follows. **The number of named entities** is hypothesized to reflect on the quality of feedback, since a large number of entities may mean that the teacher is giving effective examples of a given subject. The presence of **sentiment, greeting, and positive expression** is directly related to good practices 3 and 5, where the teacher needs to provide praise alongside constructive criticism and encourage positive motivational beliefs.

SentiLex-PT was used for the sentiment analysis, which is a lexicon explicitly created for the sentiment and opinion analysis about human entities in texts of the Portuguese language. Each input of the SentiLex-PT contains the polarity of a word, which can be positive, neutral or negative (CARVALHO; SILVA, 2015). The spaCy library[2] was used to extract the number of

[2] https://spacy.io

named entities. The final number of features was 116, with LIWC (64), Coh-Metrix (48), and four additional features (4).

5.4.3.2 Data Preprocessing

In machine learning it is necessary to divide the dataset into training and testing sets to avoid overestimating the performance of the model (FRIEDMAN; HASTIE; TIBSHIRANI, 2001). The dataset was split into training (80%) and test (20%) sets. Table 15 shows the distribution of the dataset by class.

Tabela 15 – Dataset division.

Course	Class	Training	Test	Total
Literature	Class 0	166 (20.75%)	41 (20.50%)	207 (20.70%)
Literature	Class 1	602 (75.25%)	150 (75.00%)	752 (75.20%)
Biology	Class 0	11 (01.37%)	4 (02.00%)	15 (01.50%)
Biology	Class 1	21 (02.62%)	5 (02.50%)	26 (02.60%)
	Total	**800 (100%)**	**200 (100%)**	**1000 (100%)**

Fonte: CAVALCANTI et al. (2019)

Table 15 shows that the dataset was unbalanced, i.e., the classes had different amounts of feedback messages in the training set, where 22.12% is for class 0 (177 feedbacks) and 77.87% is for class 1 (623 feedbacks). The class imbalance can have very negative effects on the results of the classification analyses (TAN et al., 2007). Thus, the SMOTE algorithm was used to address the dataset imbalance. This algorithm creates additional synthetic data points as a linear combination of the existing data points (CHAWLA et al., 2002). After applying the SMOTE algorithm class 0 increased from 177 to 619 and class 1 continued with 623, totaling 1242 instances.

5.4.3.3 Classification

According to (FERNÁNDEZ-DELGADO et al., 2014) random forests and Gaussian kernel SVMs were the top performing algorithms in a sizeable comparative analysis of 179 general-purpose classification algorithms over 121 different datasets. For this reason and for the fact that it is

a white-box algorithm, this study adopted a random forest.

The Random Forest classifier consists of a set of decision trees generated within the same object. This method was proposed by (BREIMAN, 2001) and consists of a set of decision trees constructed at the time of training. Trees are created by randomly selecting some of the attributes contained within the features vector. Each object (set of trees) goes through a voting mechanism (bagging), which elects the most voted classification. The classification is at the terminal nodes thereof. The output of the classifier is given by the class that was returned in response by most trees belonging to the forest.

The random forest classifier has two parameters that need to be configured: *ntree*: the number of trees generated by the algorithm; and (ii) *mtry*: the number of random features selected by each tree. To obtain the optimal random forest parameters, 10-fold cross-validation was used on the training data and then reported the classification accuracy of the best performing model on the testing data.

5.4.4 Implementations

All feature extraction was performed using the Python language. The classifier was coded using R. The packages and libraries used were:

- spaCy[3] and NLTK[4] package, for natural language processing;
- scikit-learn[5], for stratified sampling of the test and train datasets;
- the randomForest R package, for classifier development, and;
- The caret R package, for model training, selection, and validation.

5.5 RESULTS

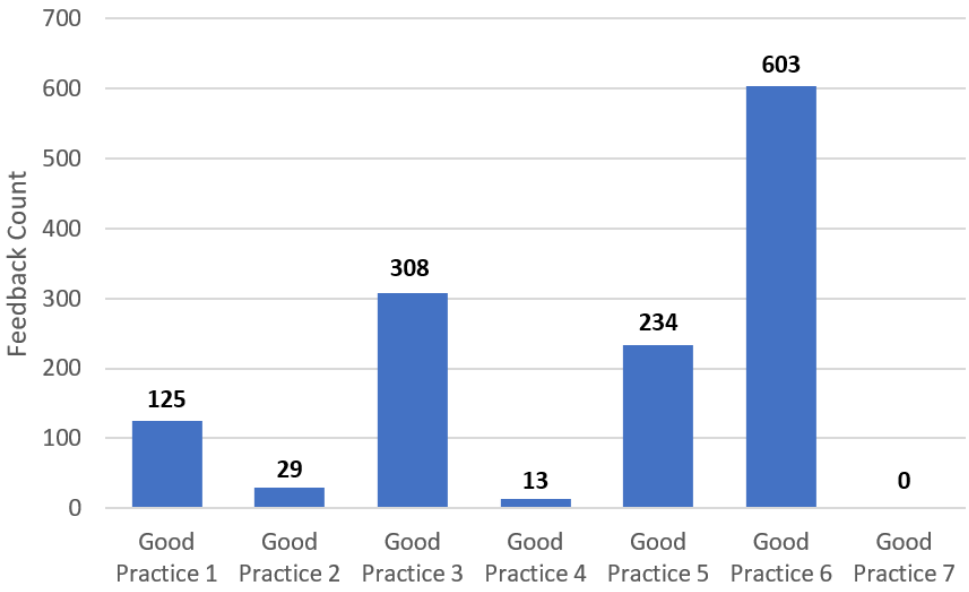
5.5.1 Research question 1

Initially, we intended to analyze the number related to each group studied, as shown in Figure 10. The good practice 6 has almost two times more messages then the second largest

[3] https://spacy.io
[4] http://www.nltk.org/
[5] https://scikit-learn.org/

category. Besides, GP3, GP5, and GP1 also had a reasonable number of feedback texts. On the other side, we did not find any message related to GP7.

Figura 10 – Distribution of feedback messages by seven practices proposed by Nicol e Macfarlane-Dick (2006) and as outlined in Table 13



Fonte: CAVALCANTI et al. (2019)

The Phi correlations between the occurrences of the seven principles of good feedback practices are shown in Table 16. This table shows that the occurrence of good practice 3 had the highest correlation with the presence of other feedback practices. Only the relationship between good practice 1 and 5 also reached a comparable value.

Tabela 16 – Phi's correlations in the occurrence of the seven feedback practices

	GP 1	GP 2	GP 3	GP 4	GP 5	GP 6	GP 7
GP 1	1	**0.13****	**0.29****	0.09*	**0.18****	0.07	0
GP 2	-	1	**0.14****	0.13**	-0.03	0.01	0
GP 3	-	-	1	0.01	**0.25****	**0.24****	0
GP 4	-	-	-	1	-0.06	-0.03	0
GP 5	-	-	-	-	1	**-0.14****	0
GP 6	-	-	-	-	-	1	0
GP 7	-	-	-	-	-	-	1

*$p < 0.05$; **$p < 0.001$

Fonte: CAVALCANTI et al. (2019)

5.5.2 Research question 2

The classifier was initially applied on the training dataset with and without the use of the SMOTE algorithm. The average accuracy with unbalanced data was 0.8725 and with balanced data was 0.9256. Table 17 shows the minimum and maximum values of *mtry* parameter for balanced data. In the best case (*mtry* = 19), the classifier achieved the accuracy of 0.9256 with Standard Deviation (SD) of 0.02 and Cohen's κ of 0.85 (SD = 0.04).

Tabela 17 – The results of parameter tuning for the random forest classifier

	mtry	Accuracy	Kappa
Min	103	0.9136 (0.03)	0.8274 (0.05)
Max	19	0.9256 (0.02)	0.8516 (0.04)
	Difference	0.012	0.024

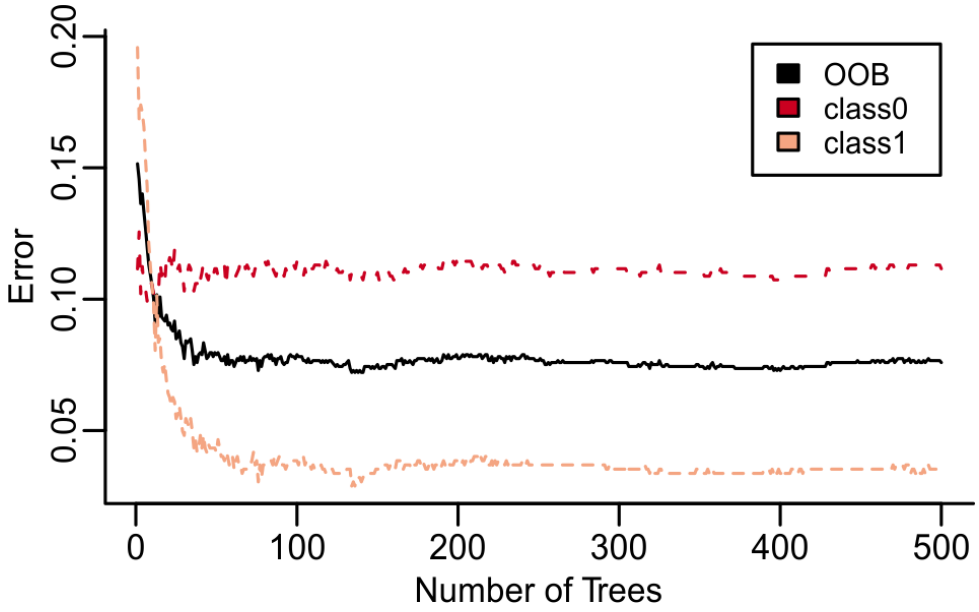
Fonte: CAVALCANTI et al. (2019)

The difference between the best- and worst-performing configurations was 0.012 and 0.024 for accuracy and Cohen's κ respectively, which shows that the parameter optimization played an essential role in the final classifier performance.

Figure 11 indicates that 500 trees were enough to guarantee a good classifier performance (the error stabilized before 100 trees). After 50 trees the average out-of-bag (OOB) error rate remained reasonably stable under 0.1, which suggests that less than 10% of the data points were misclassified. The highest error rate was observed for class 0 and this result was expected due to the low number of instances for this class (Table 15).

For the testing data (20%), the classifier obtained 0.715 of accuracy (95% CI : [0.6471, 0.7764]) with Cohen's κ of 0.1204 in the best model without SMOTE and 0.75 of accuracy (95% CI : [0.684, 0.8084]) and Cohen's κ of 0.2082 in the best model with SMOTE. Again the results show that unbalanced data is an important factor in the classification process. The difference between the accuracy obtained by the best model with and without SMOTE was 0.035.

Figura 11 – Best random forest configuration performance.



Fonte: CAVALCANTI et al. (2019)

5.6 CONCLUSIONS

This study aimed to evaluate the use of the seven good feedback practices in textual messages written in Portuguese (NICOL; MACFARLANE-DICK, 2006). Within this context, the study offered two main contributions. First, we investigated the statistics about different categories of good feedback practices found in the literature. The study showed that the most used aspect of feedback was trying to assist a student that have not reached a good performance while the features that did not appear much was related to the behavior of teachers. This outcome is well aligned with the literature, as traditionally instructors tend to send more elaborated messages to the students at risk and, in general, they do not use feedback to reshape their courses (NICOL; MACFARLANE-DICK, 2006; HATTIE; TIMPERLEY, 2007; MUTCH, 2003).

Second, we proposed a new classifier to code teachers' feedback written in Portuguese. It used 116 features, mainly extracted using LIWC and Coh-Metrix. The classifier achieved good accuracy, on both the training and testing sets, for the binary classification task that analyzed if the feedback messages contained or not at least one good practice of feedback. Systems of this type can help instructors provide good student feedback and can also automate feedback in some LMS.

The main limitations of the approach presented here are related to the dataset. The dataset was from the task tool of biology and literature courses, where students submitted their

responses to the course assignments and the teachers evaluated and provided feedback to the students. Thus, the study may not be fully representative of the different feedback types that can be used in online learning courses. Also, the size of the data set and the unbalanced categories can affect the performance of the classifier.

As future work, we plan to assess the feedback in another education context (i.e., blended vs. fully online) and in others e-learning resources (e.g., forum and chats).

6 ANÁLISE AUTOMÁTICA DE FEEDBACK EM AMBIENTES DE APRENDIZAGEM ONLINE

Authors: Anderson Pinheiro Cavalcanti, Rafael Ferreira Mello, Péricles Miranda, Fred Freitas

6.1 INTRODUÇÃO

O feedback é um fator essencial no processo de aprendizagem, pois permite aos alunos identificar lacunas no aprendizado, pode ajudar os alunos a identificar áreas de melhoria em seus conhecimentos ou habilidades, e refletir em suas estratégias de aprendizagem (SADLER, 1989). Além disso, os professores usam o feedback para identificar as necessidades dos alunos, para que possam adaptar seus métodos e conteúdo com base nessas necessidades (LANGER, 2011). A literatura demonstra os benefícios e o impacto positivo que um bom feedback causa no aprendizado do aluno (NICOL; MACFARLANE-DICK, 2006), mas também ressaltam que um feedback ruim pode desmotivar e até levar a evasão do aluno (HATTIE; TIMPERLEY, 2007).

Outro fator relevante neste contexto é que o uso de Ambientes Virtuais de Aprendizagem (AVAs) tem aumentado nos últimos anos devido ao uso das tecnologias de informação e comunicação como ferramenta de apoio educacional (CORREIA; SANTOS, 2013). Esses ambientes possuem ferramentas que permitem uma grande interação entre professores e alunos, como por exemplo: chat, fórum, wiki, entre outras. Entre elas, há a ferramenta de envio de atividades em que os alunos podem enviar suas respostas às atividades em andamento. Em geral, esse recurso de avaliação é o principal espaço onde os instrutores enviam feedback (COATES; JAMES; BALDWIN, 2005). No entanto, devido ao crescimento significativo da quantidade de alunos nos AVAs que precisam de um envolvimento contínuo, é difícil para os instrutores fornecerem um feedback personalizado e de alta qualidade (ESPASA; MENESES, 2010).

Com o objetivo de reduzir o esforço do instrutor e melhorar a qualidade do feedback, pesquisas recentes trazem algumas abordagens para automatizar o feedback para os alunos (MARIN et al., 2017; BELCADHI, 2016). Entretanto, esses trabalhos abordam o feedback em um contexto específico, como por exemplo cursos de programação introdutórios.

Nesse contexto, este artigo propõe uma abordagem para analisar automaticamente os feedbacks enviados pelos instrutores com base em alguns princípios de boas práticas de feedback (NICOL; MACFARLANE-DICK, 2006). Para isso, foram analisados os algoritmos XGBoost e Ada-

Boost para extrair indicadores que podem levar à melhoria do feedback enviado aos alunos. A abordagem proposta alcançou resultados até quatro vezes melhores quando comparado aos resultados de trabalhos anteriores.

6.2 TRABALHOS RELACIONADOS

O feedback é um componente essencial da aprendizagem, uma vez que direciona os alunos para o tipo apropriado de estudo ou prática. Esse tópico é uma parte fundamental do processo de aprendizado do aluno e vem sendo abordado em diversas pesquisas nas últimas duas décadas (NICOL; MACFARLANE-DICK, 2006; JERIA; VILLALON, 2017). Embora muitas pesquisas relatem efeitos positivos do feedback, nem todo feedback é igualmente eficaz. Por exemplo, no estudo de Burke (2009) foram levantados alguns fatores de insatisfação do aluno com relação ao feedback recebido. Entre esses fatores, estão o comprimento (o feedback é muito breve), a polaridade (o feedback é muito negativo) e a complexidade (é muito difícil decifrar ou entender o feedback).

A literatura também oferece recomendações de boas práticas de feedback. Nicol e Macfarlane-Dick (2006) propuseram um modelo conceitual de autorregularão com base em uma revisão da literatura de pesquisa sobre avaliação formativa e feedback. A ideia principal do trabalho é identificar como os processos formativos de avaliação e feedback podem ajudar a promover a autorregularão. Com base no modelo conceitual, foram definidos sete princípios de boas práticas de feedback que o professor pode usar para refletir sobre o projeto e avaliar seus próprios procedimentos de feedback. Abaixo estão as sete boas práticas[1] de feedback (NICOL; MACFARLANE-DICK, 2006). De acordo com os autores, boas práticas de feedback são amplamente definidas como qualquer coisa que possa fortalecer a capacidade dos alunos de autorregularem seu desempenho.

- **BP 1**: Ajuda a esclarecer o que é um bom desempenho (metas, critérios, padrões esperados);
- **BP 2**: Facilita o desenvolvimento da autoavaliação (reflexão) na aprendizagem;
- **BP 3**: Fornece informações de alta qualidade aos alunos sobre seu aprendizado;
- **BP 4**: Incentiva o diálogo entre professores e colegas sobre o aprendizado;

[1] Foi utilizado o acrônimo BP para Boas Práticas

- **BP 5**: Incentiva crenças motivacionais positivas e autoestima;
- **BP 6**: Oferece oportunidades para fechar a lacuna entre o desempenho atual e o desejado;
- **BP 7**: Fornece informações aos professores que podem ser usadas para ajudar a moldar o ensino;

Na literatura também é possível encontrar alguns trabalhos que analisam automaticamente o feedback. (MAITRA et al., 2018) propôs um classificador automático para investigar o feedback usando o classificador Naïve Bayes. O autor classificou o feedback fornecido para cada aluno em duas classes, válidas ou inválidas, com base em uma medida de feedback proposta. O classificador leva em consideração a contribuição independente de diferentes recursos do aluno (ou seja, esforço, formação acadêmica, resultados do curso alcançados), no feedback fornecido pelo instrutor. A abordagem foi avaliada usando 1.000 extratos de feedback de uma instituição de ensino superior indiana. A principal desvantagem dessa abordagem foi a falta de detalhes sobre a influência dos recursos utilizados e as implicações adicionais de sua abordagem.

Os métodos de mineração de texto também foram usados para fornecer feedback automático aos alunos. Akçapınar (2015) visava reduzir o plágio dos alunos nas tarefas *online*, fornecendo feedback automatizado com base na análise de mineração de texto. Antes do feedback automático, 81,4% dos estudantes cometeram plágio. Após o feedback, 83% dos estudantes reduziram as tentativas pós-plágio. O número de não plagiadores aumentou 42,37% após o fornecimento de feedback.

Por fim, o estudo de Cavalcanti et al. (2019) teve como objetivo analisar a qualidade do feedback extraído das avaliações coletadas em um curso *online* oferecido em uma instituição de ensino superior brasileira. Um classificador *Random Forest* foi usado para classificar as mensagens de feedback, verificando se os textos seguiam as boas práticas propostas por Nicol e Macfarlane-Dick (2006). Os autores avaliaram o sistema usando um conjunto de dados que continha 1.000 exemplos de mensagens de feedback. O classificador alcançou 0,75 e 0,20 para as medidas de acurácia e kappa, respectivamente.

Dessa forma, este trabalho se difere dos trabalhos mencionados acima em duas principais contribuições: (1) Utilizar algoritmos de aprendizagem de máquina mais recentes para classificar os feedbacks; e (2) Analisar as características mais importantes para cada classificador.

6.3 PERGUNTAS DE PESQUISA

Esse trabalho tem como objetivo automatizar a classificação de feedback de cursos a distância seguindo as boas práticas de feedback propostas por Nicol e Macfarlane-Dick (2006). Dessa forma, nossa primeira questão de pesquisa é:

> **PERGUNTA DE PESQUISA 1:** *Qual classificador obtém o melhor desempenho na classificação de feedback de acordo com os critérios de boas práticas?*

Além de verificar qual classificador obtém melhores resultados, também propomos a análise de quais as características são mais relevantes para esses classificadores.

> **PERGUNTA DE PESQUISA 2:** *Quais as características mais importantes na classificação dos feedback?*

6.4 MÉTODO

6.4.1 Dados

O conjunto de dados utilizado neste trabalho é o mesmo utilizado por Cavalcanti et al. (2019), Cavalcanti et al. (2020) e foi obtido de um AVA usado em cursos *online* oferecidos por uma universidade pública brasileira. O conjunto de dados é composto por mensagens de texto que o instrutor enviou como feedback para as atividades de seus alunos enviadas pelo AVA. Foram coletados um total de 1.000 feedbacks dos cursos de biologia e literatura divididos em 41 mensagens para biologia e 959 para literatura. Todo o conjunto de dados foi anotado por especialistas que analisaram se o feedback seguia as boas práticas propostas por Nicol e Macfarlane-Dick (2006). O conjunto de dados está dividido em duas classes: classe 0: se o feedback não pertencer a nenhuma das boas práticas; classe 1: se o feedback pertencer a pelo menos uma das sete boas práticas. A Tabela 18 mostra a quantidade de feedback por classe.

6.4.2 Extração de Características

A extração de características é o processo que transforma o texto do feedback em dados numéricos. Nesse trabalho foram utilizados dois sistemas que extraem características linguís-

Tabela 18 – Distribuição do banco de dados por classe.

Classe	Descrição	Quantidade
0	Não possui boa prática	222
1	Possui 1 ou mais boas práticas	778
	Total	**1000**

Fonte: CAVALCANTI et al. (2020a)

ticas do texto em conjunto com 4 características adicionais utilizando mineração de texto. Todos esses recursos são detalhados a seguir.

O LIWC (*Linguistic Inquiry and Word Count*) extrai um grande número de contagens de palavras que são indicativas de diferentes processos psicológicos, como afetivos, cognitivos, sociais e perceptivos (TAUSCZIK; PENNEBAKER, 2010). O LIWC possui 127.149 entradas, nas quais cada entrada pode ser atribuída a uma ou mais categorias. O núcleo desta ferramenta é um recurso lexical, mais conhecido como dicionário LIWC, que também foi disponibilizado no idioma português (FILHO; PARDO; ALUÍSIO, 2013). Um total de 64 recursos LIWC foram extraídos no estudo.

Coh-Metrix é uma ferramenta computacional que analisa muitas palavras, frases e textos com várias frases diferentes. Coh-Metrix fornece várias medidas de coerência do texto, complexidade linguística, legibilidade do texto e uso da categoria lexical (MCNAMARA et al., 2014). A versão em português do Coh-Metrix possui 48 medidas diferentes.

Além dos recursos LIWC e Coh-Metrix, foram extraídos do texto: o número de entidades nomeadas, polaridade de sentimento (positiva, negativa e neutra), presença de saudação ("Bom dia", "Boa tarde", "Como vai") e expressões positivas ("Muito bem", "Excelente", "Bom trabalho"). Esses recursos também foram propostos em trabalhos de classificação de texto educacional (FERREIRA-MELLO et al., 2019). Para isso, foi utilizado o SentiLex-PT, que é um léxico criado explicitamente para a análise de sentimentos e opiniões sobre entidades humanas em textos da língua portuguesa. Cada entrada do SentiLex-PT contém a polaridade de uma palavra, que pode ser positiva, neutra ou negativa (CARVALHO; SILVA, 2015).

Por fim, a biblioteca spaCy [2] foi usada para extrair o número de entidades nomeadas. O número final de características foram 116, com 64 do LIWC, 48 do Coh-Metrix e 4 características adicionais.

[2] https://spacy.io

6.4.3 Processamento de Dados e Classificação

No aprendizado de máquina, para evitar superestimar o desempenho do modelo, é necessário dividir o conjunto de dados em conjuntos de treinamento e teste (FRIEDMAN; HASTIE; TIBSHIRANI, 2001). O conjunto de dados foi dividido em conjuntos de treinamento (80%) e teste (20%). Além disso, os dados estão desbalanceados, como mostra a Tabela 18, a classe 1 tem quase 4 vezes mais exemplos que a classe 0. O desbalanceamento de classe pode ter efeitos muito negativos nos resultados das análises de classificação. Assim, o algoritmo SMOTE foi usado para resolver o desequilíbrio do conjunto de dados. Esse algoritmo cria pontos de dados sintéticos adicionais como uma combinação linear dos pontos de dados existentes (CHAWLA et al., 2002).

Para classificar os feedbacks foram utilizados 2 algoritmos de aprendizagem de máquina baseados em árvores de decisão, o AdaBoost[3] e o XGBoost[4]. Esses classificadores possuem alguns parâmetros de treinamento que devem ser definidos. Neste trabalho foi analisado o parâmetro *ntree* que significa o número de árvores geradas pelo algoritmo. Para obter o parâmetro ideal, foi utilizada a técnica validação cruzada de 10 vezes nos dados de treinamento e depois aplicado o modelo com melhor acurácia nos dados de teste. Os 2 classificadores também fornecem métodos para analisar a importância das características na classificação, que foram utilizados para abordar a questão de pesquisa 2 no presente estudo. Uma medida popular para calcular a importância do recurso é o *Mean Decrease Gini (MDG)*, que se baseia na redução na medida de impureza de Gini (BREIMAN, 2001). Neste artigo, foi utilizado o MDG para avaliar a relevância de diferentes características para o resultado dos classificadores.

6.4.3.1 AdaBoost

Adaboost (*Adaptative Boosting*) é um algoritmo de classificação que utiliza a técnica *Boosting* para classificar os exemplos. *Boosting* é uma técnica que combina os classificadores gerados por um mesmo algoritmo de aprendizado e com isso consegue formar um classificador "forte" com base em classificadores mais simples, ditos "fracos". O AdaBoost mantém um conjunto de pesos sobre os exemplos de treinamento. Em cada iteração, o algoritmo de aprendizado é chamado para minimizar o erro ponderado no conjunto de treinamento. Esse

[3] https://scikit-learn.org/
[4] https://xgboost.readthedocs.io/

erro é utilizado para ponderar os pesos nos exemplos de treinamento. Dessa forma, o algoritmo coloca mais peso nos exemplos classificados incorretamente e menos peso nos exemplos classificados corretamente (DIETTERICH, 2000).

6.4.3.2 XGBoost

O algoritmo XGBoost é um modelo de aprendizado de máquina escalonável para o aprimoramento de árvore baseado em árvores de Decisão de Intensidade de Gradiente. O XGBoost fornece um aumento de árvore paralelo (também conhecido como GBDT - *Gradient Boosted Decision Tree*) que resolve muitos problemas de ciência de dados de maneira rápida e precisa. Diferente dos métodos tradicionais de reforço que pesam amostras positivas e negativas, o GBDT faz convergência global do algoritmo seguindo a direção do gradiente negativo (CHEN; GUESTRIN, 2016a).

6.5 RESULTADOS

6.5.1 Análise dos classificadores

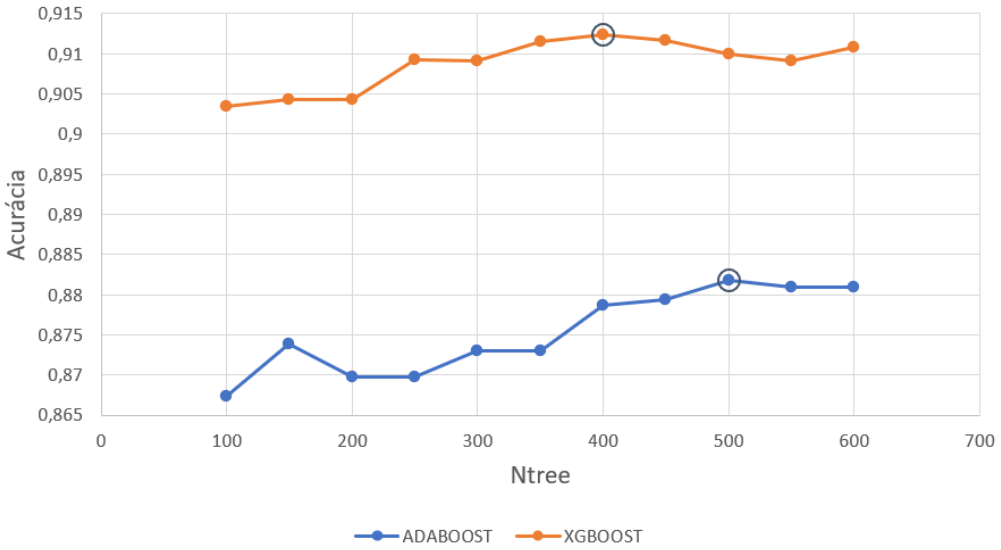
Inicialmente, os classificadores foram aplicados no conjunto de treinamento utilizando a técnica de validação cruzada 10 vezes para diferentes valores de *ntree*. A Figura 12 mostra os resultados obtidos para os dois classificadores.

A melhor acurácia do AdaBoost foi de 0,88 (Kappa = 0,76) obtida com *ntree* = 500. Para o XGBoost a melhor acurácia foi de 0,91 (Kappa = 0,82) com *ntree* = 400. Pela Figura 12 é possível perceber que o XGBoost obteve a maior acurácia em todos os casos quando comparado com o AdaBoost. A diferença entre a pior e a melhor acurácia foi de 0,0145 e 0,0089 para AdaBoost e XGBoost, respectivamente. Isso mostra que a otimização do parâmetro *ntree* é essencial no desempenho final do classificador.

O melhor modelo obtido no conjunto de treinamento foi aplicado no conjunto de teste. O AdaBoost obteve 0,89 de acurácia (Kappa = 0,78) e o XGBoost obteve 0,91 de acurácia (Kappa = 0,82). A Tabela 19 mostra os resultados obtidos e compara com o resultado obtido por Cavalcanti et al. (2019).

Da mesma forma que no conjunto de treinamento, o XGBoost obteve o melhor resultado comparado ao AdaBoost e também obteve um aumento substancial na acurácia e principal-

Figura 12 – Resultados dos classificadores no conjunto de treinamento.



Fonte: CAVALCANTI et al. (2020a)

Tabela 19 – Comparação resultados na base de teste.

	Acurácia	Kappa
AdaBoost	0,89	0,78
XGBoost	**0,91**	**0,82**
(CAVALCANTI et al., 2019)	0,75	0,2

Fonte: CAVALCANTI et al. (2020a)

mente no Kappa quando comparado ao resultado obtido por Cavalcanti et al. (2019).

6.5.2 Análise top-10 características

Assim como trabalhos recentes, (SIROTHEAU et al., 2019), esse estudo também analisou as contribuições das diferentes características no desempenho final do classificador. Como mencionado anteriormente, foi utilizada a medida *Mean Decrease Gini (MDG)* para definir o grau de relevância de uma característica (BREIMAN, 2001). As Tabelas 20 e 21 mostram as 10 características mais importantes para os classificadores AdaBoost e XGBoost, respectivamente.

A característica mais importante do AdaBoost foi a frequência de palavras de conteúdo (substantivos, verbos, adjetivos e advérbios). Essa característica pode estar diretamente relacionada ao tamanho do texto fornecendo exemplos para o aluno (usando substantivos) ou elogiando a atividade do aluno (usando adjetivos) seguindo as boas práticas 3 e 5 (NICOL;

Tabela 20 – Top-10 de características mais importantes do AdaBoost.

#	Característica	Descrição	MDG
1	cw_frequencies:cw_freq,	Frequência de palavras de conteúdo	4,20
2	liwc.we	Contagem de pronomes na 1ª pessoa do plural	3,60
3	liwc.posemo	Contagem de palavras positivas	3,60
4	liwc.number	Contagem de números	2,80
5	conn:tmp_neg_conn_inc	Incidência de conectivos de negação	2,60
6	liwc.certain	Contagem de palavras de certeza	2,60
7	liwc.friend	Contagem de palavras de amizade	2,60
8	liwc.excl	Contagem de pontos de exclamação	2,40
9	liwc.affect	Contagem de palavras de afeto	2,40
10	liwc.auxverb	Contagem de verbos auxiliares	2,20

Fonte: CAVALCANTI et al. (2020a)

MACFARLANE-DICK, 2006), respectivamente. Através da Tabela 20 podemos verificar também que as características *liwc.posemo*, *liwc.friend* e *liwc.affect* podem estar relacionadas a boa prática 5, cujo objetivo é fornecer crenças motivacionais e positivas para o aluno.

A característica mais importante do XGBoost foi a proporção de sentenças adjacentes que compartilham palavras de conteúdo (substantivos, verbos, adjetivos e advérbios). Da mesma forma que o AdaBoost, essa característica em conjunto com *basic_counts:adverbs*, *liwc.auxverb*, *bc:words_per_sentence*, *tokens.ttr* e *cw_frequencies:cw_freq* estão relacionadas ao tamanho do texto fornecido pelo professor. Isso significa que o professor forneceu informações de alta qualidade aos alunos usando exemplos ou dicas para melhorar as próximas atividades e dessa forma seguiu as boas práticas 3, 5 e 6 (NICOL; MACFARLANE-DICK, 2006).

Tabela 21 – Top-10 de características mais importantes do XGBoost.

#	Característica	Descrição	MDG
1	coreference:adj_cw_ovl	Palavras de conteúdo sobrepostas em frases adjacentes	7,86
2	basic_counts:adverbs	Contagem básica de advérbios	5,43
3	liwc.auxverb	Contagem de verbos auxiliares	4,46
4	constituents:np_incidence	Incidência de Sintagmas	4,23
5	bc:words_per_sentence	Contagem básica de palavras por sentença	4,15
6	anaphoras:anaphoric_refs	Referência anafóricas	3,02
7	cw_frequencies:cw_freq	Frequência de palavras de conteúdo	2,76
8	liwc.we	Contagem de pronomes na 1ª pessoa do plural	2,27
9	liwc.conj	Contagem de conjunções	1,97
10	tokens:ttr	Número de palavras divididas pelo número de tokens	1,92

Fonte: CAVALCANTI et al. (2020a)

6.6 DISCUSSÃO

Este trabalho analisou automaticamente se os textos de feedback fornecidos por professores em cursos a distância seguiam as boas práticas de feedback propostas por Nicol e Macfarlane-Dick (2006). A 1ª pergunta de pesquisa teve como objetivo analisar qual classificador obtém a melhor acurácia para classificar os feedbacks. Foram analisados os classificadores AdaBoost e XGBoost (ambos utilizam árvores de decisão) e comparados com um estudo anterior que utilizou o *Random Forest*. O classificador XGBoost obteve o melhor resultado em ambos os conjuntos de treinamento e teste comparado ao AdaBoost. Os resultados também mostraram que tanto o AdaBoost quanto o XGBoost conseguiram obter uma melhor acurácia e Kappa comparado ao trabalho de Cavalcanti et al. (2019) com uma melhoria de 18,66% e 21,33% com relação a acurácia e 290% e 310% com relação ao Kappa, respectivamente.

Além de analisar a acurácia dos classificadores, a questão de pesquisa 2 tinha como objetivo analisar as características mais relevantes no processo de classificação dos feedbacks. Como os classificadores são baseados em árvores de decisão, é possível obter as características que são mais relevantes para cada classificador. Através dos resultados obtidos foi possível verificar relações entre as características mais relevantes e as boas práticas propostas por Nicol e

Macfarlane-Dick (2006). A abordagem proposta, baseada principalmente nos sistemas Coh-Metrix e LIWC, mostrou que é possível fornecer um sistema totalmente automatizado para auxiliar na verificação se o texto do professor segue as boas práticas.

6.7 LIMITAÇÕES E TRABALHOS FUTUROS

Uma das limitações desse estudo é relacionado ao conjunto de dados. Os textos obtidos são de apenas dois cursos a distância diferentes (Literatura e Biologia) que foram fornecidos através da ferramenta de envio de atividades. Dessa forma, os textos podem não ser totalmente representativos com relação ao tipo de feedback fornecido em cursos a distância. Além disso, o tamanho do conjunto de dados e o desbalanceamento das classes podem afetar o desempenho do classificador. Como trabalhos futuros pretendemos aumentar o conjunto de dados com textos de outros cursos e também criar uma ferramenta online que analise textos de feedback.

7 HOW GOOD IS MY FEEDBACK? A CONTENT ANALYSIS OF WRITTEN FEEDBACK

Authors: Anderson Pinheiro Cavalcanti, Arthur Diego, Rafael Ferreira Mello, Katerina Mangaroska, André Nascimento, Fred Freitas, Dragan Gašević

7.1 INTRODUCTION

Feedback is an essential part of learning that helps students to identify gaps, self-assess, and act upon the provided insights (BUTLER; WINNE, 1995a; SADLER, 1989). Moreover, it is through feedback that instructors guide students to improve their performance (LANGER, 2011; MATCHA et al., 2019b). As Laurillard (1993) stressed "action without feedback is completely unproductive for the learner" (LAURILLARD, 1993, p.61). Thus, contemporary principles of good feedback practice recognize that feedback is not a product, but a complex process that should make students aware of how actual study behavior, emotions, and cognition influence their outcomes (BOUD; FALCHIKOV, 2007; HENDERSON et al., 2019a). Such an understanding could help instructors gauge when and how to communicate the right information.

According to Sadler (1989) feedback needs to provide relevant information related to a learning task or process, and avoid discrepancy between what a student understands and what a student should understand. Past research have shown that timely feedback brings benefits to learning (HATTIE; TIMPERLEY, 2007; NICOL; MACFARLANE-DICK, 2006; PARIKH; MCREELIS; HODGES, 2001). However, although feedback holds a central part in the student development (HATTIE; TIMPERLEY, 2007), there is a significant body of research demonstrating that the current feedback practices are underdeveloped, thereby limiting student progress (BOUD; FALCHIKOV, 2007; HENDERSON et al., 2019a).

In the context of online learning, where instructors and students are physically separated in space and/or time, the scenario is even worse because the student relies significantly on the material that the instructor provides within the educational platform (ESPASA; MENESES, 2010). Thus, informative and timely interactions and feedback become even more critical to knowledge construction and academic success (JOULANI; GYORGY; SZEPESVÁRI, 2013; IRONS, 2007).

Among the resources provided by online educational environments, the activity submission tool becomes a critical communication element, as that is the way students receive comments

on their activities and progress (COATES; JAMES; BALDWIN, 2005). However, the increasing numbers of students enrolled in online learning magnify the challenges for instructors to provide informative and high-quality feedback, which can assist students in self-regulation. According to Hattie e Timperley (2007), poor feedback can cause problems in student learning and even lead to dropout. Aiming to reduce instructor effort and improve the feedback quality, many researchers have proposed approaches to automating feedback for students (MARIN et al., 2017; GULWANI; RADIČEK; ZULEGER, 2014; BELCADHI, 2016; FERREIRA-MELLO et al., 2019). However, their approaches address feedback within a particular context like introductory programming courses.

Consequently, this paper proposes a tool that can automatically analyze comments shared by instructors based on well-known practices of good feedback. More specifically, this work introduces a set of binary classifiers of text, build on supervised machine learning, to extract indicators that can lead to the improvement of feedback sent to students via the activity submission tool. The proposed approach reached results up to 87% and 0.39 for accuracy and Cohen's kappa, respectively. The results and their implications for theory and practice are further discussed in the paper.

7.2 BACKGROUND

7.2.1 Feedback

Feedback is an essential factor in the learning process that triggers reflection among students and display changes between current and expected performance (KLEIJ; FESKENS; EGGEN, 2015). Moreover, feedback can help students to identify areas for improvement in their knowledge or skills, and reflection on their learning strategies (PARIKH; MCREELIS; HODGES, 2001). Instructors use feedback to identify students' needs, so that they can tailor their methods and content based on those needs (ORRELL, 2006). Past research reported benefits and impact on learning from using feedback practices (NICOL; MACFARLANE-DICK, 2006; HATTIE; TIMPERLEY, 2007; BLACK; WILIAM, 1998; PARIKH; MCREELIS; HODGES, 2001). However, poor feedback can be harmful to students, creating distrust in the feedback process and the teacher, which in turn might be detrimental to students' self-efficacy and motivation (BOUD; FALCHIKOV, 2007). Hence, generating effective feedback is a challenging task (MATCHA et al., 2019b).

Several studies have been conducted over the past decades in an attempt to identify how

students should receive and act upon provided feedback (MUTCH, 2003). For instance, Weaver (2006b) investigated, qualitatively and quantitatively, whether students value the feedback they receive, and they report two problems: (1) feedback often does not contain enough content to guide or motivate students; and (2) students do not have sufficient understanding of academic discourse to interpret the instructor's feedback accurately. Moreover, Weaver (2006b) showed that over 50% of college students had never received any guidance on *how to understand and use feedback*, while 75% of students had not received any advice on how to understand and use feedback before their university studies.

Burke (2009) analyzed proposed guidelines for feedback use, from 350 students coming from disciplines related to humanities. Their responses revealed many dissatisfactions with feedback they receive, particularly:

- **The length**: feedback is very brief;
- **The polarity:** feedback is very negative;
- **The complexity**: feedback is very difficult to decipher or understand.

Consequently, the educational literature proposed several methods to increase the impact of feedback on the students' learning.

7.2.2 Practices of Good Feedback

According to Nicol e Macfarlane-Dick (2006), good feedback practice is broadly defined as any strategy or content that could enhance students' capacity to self-regulate their learning performance. This study proposed seven general principles of good feedback practices in order to assist teaching staff: (i) Helps clarify what good performance is (goals, criteria, expected standards), (ii) Facilitates the development of self-assessment (reflection) in learning, (iii) Delivers high-quality information to students about their learning, (iv) Encourages teacher and peer dialogue around learning, (v) Encourages positive motivational beliefs and self-esteem, (vi) Provides opportunities to close the gap between current and desired performance, (vii) Provides information to teachers that can be used to help shape teaching.

Although the work by Nicol e Macfarlane-Dick (2006) demonstrates an initial set of indicators to evaluate the quality of feedback, these indicators are too general and in some cases difficult to apply for written feedback. In this direction, Hattie e Timperley (2007) analyzed

several conditions that could maximize positive effects of feedback on learning, including the increase in a student's awareness about an overall learning goal, the progress towards the goal, and the subsequent goals required to achieve the main goal. Thus, Hattie e Timperley (2007) proposed four perspectives, proposed as levels, that feedback should approach in order to improve its effectiveness. They posited that their model is more suitable for examining textual feedback because the model is focused on aspects related to learning tasks, learning process, and student self-regulation. HATTIE; TIMPERLEY also proposed several text examples on how the feedback for different levels should be written/identified. Table 32 shows a description and examples for each of the proposed levels. The Hattie e Timperley (2007) study also suggests that the most potent feedback is on the process (FP) and self-regulation (FR) levels. The task feedback level (FT) is only valuable if combined with either the process (FP) or self-regulation (FR) levels. The value of feedback on the self-level (FS) is negligible and suggested that this level of feedback should generally be avoided.

Tabela 22 – Levels of feedback according to Hattie e Timperley (2007)

Label	Level	Description	Example
FT	Feedback About the Task	Feedback can be about a task, such as whether the job is correct or incorrect, can include instructions for more or different information.	"You need to include more about the Treaty of Versailles."
FP	Feedback About the Processing of the Task	Feedback can be directed to the process used to create a product or complete a task, is more directed to information processing, or learning processes that require understanding or completing the task.	"You need to edit this piece of writing by attending to the descriptors you have used so the reader is able to understand the nuances of your meaning," ;"This page may make more sense if you use the strategies we talked about earlier."
FR	Feedback About Self-Regulation	Feedback for students can be focused on the level of self-regulation, including greater self-assessment or confidence skills, which can have major influences on self-efficacy, self-regulatory proficiency, and students' personal beliefs as learners.	"You already know the key features of the opening of an argument. Check to see whether you have incorporated them in your first paragraph."
FS	Feedback About the Self as a Person.	Feedback can be personal in the sense that it is directed to the self. It is often unrelated to task performance.	"Good girl"; "Great effort";"You're really great because you have diligently completed this task by applying this concept";"You are a great student";"That's an intelligent response, well done."

Fonte: CAVALCANTI et al. (2020)

7.2.3 Written Feedback Analysis

In recent years, several researchers aimed to examine textual feedback using different approaches of content analysis. For example, Lewkow et al. (2016) proposed a platform that allows the ingestion of heterogeneous educational data from multi-source systems. This platform serves as basis for an end-to-end automated writing feedback system. The system allows students to view written feedback in near real-time, make edits based on the provided feedback, and follow the progress over time. The feedback provided by this system is composed of seventeen writing competencies, including traditional writing metrics, such as spelling and grammatical accuracy, as well as more advanced metrics that capture sentiment and writing flow.

The work of Lee e Lim (2016) used text mining to analyze feedback examples to highlight students' main concerns based on key terms in the feedback they received. In their findings, Lee e Lim (2016) reported that it is possible to process and understand a large amount of unstructured data generated from concept maps.

Moreover, Maitra et al. (2018) proposed an automatic classifier to investigate feedback using a Naïve Bayes classifier. The feedback provided to each student was classified into two classes, valid or invalid. A total of 1,000 higher education student feedback extracts were evaluated. The method attempted to address inaccuracy to overcome the limitations of the traditional model. The main pitfall of this approach is that the authors do not provide information on what they consider as valid feedback, while insufficient methodological details are provided about the analysis of the results.

Finally, the study of Cavalcanti et al. (2019) focused on analyzing the quality of feedback extracted from evaluations collected in an online course offered at a Brazilian higher education institution. A random forest classifier was used to classify feedback messages. The authors coded the dataset following the seven principles proposed by Nicol e Macfarlane-Dick (2006). They evaluated the system using a data set that contained 1,000 examples of feedback messages, divided in two classes: (1) feedback with at least one good practice; and (2) feedback with no good practice. The classifier reached 0.75 and 0.20 for accuracy and Cohen's κ, respectively.

The previous research draws limited conclusions connecting the results with the educational literature on good practices of feedback. Moreover, to the best of our knowledge, no work performed a content analysis under the perspective of the four levels proposed by Hattie e

Timperley (2007).

7.3 RESEARCH QUESTIONS

As stated in the previous sections, it is crucial to provide automatic methods to assist the elaboration of good quality feedback. Although several studies describe what feedback should contain (HATTIE; TIMPERLEY, 2007; NICOL; MACFARLANE-DICK, 2006), the literature does not report a wide range of work that performs content analysis in textual feedback. For example, Cavalcanti et al. (2019) proposed a method to analyze feedback under the perspective of (NICOL; MACFARLANE-DICK, 2006). However, the four levels proposed in (HATTIE; TIMPERLEY, 2007) are more suitable to analyze textual feedback, details in section 7.2.2. Thus, this paper proposed a content analysis using Hattie e Timperley (2007) model as input, raising our first research question as:

RQ 1: To what extent can text mining methods accurately and automatically identify different aspects of good feedback practice?

Besides the automatic identification of good feedback indicators, we also intend to provide additional insights into the features that are more relevant for different categories. To do so, we explored a method similar to the one used by Kovanović et al. (2016) and Neto et al. (2018). Hence, our second research question is:

RQ 2: Which features do best predict each indicator of good feedback?

7.4 METHOD

7.4.1 Data and course design

The dataset used in this work is the same one used by Cavalcanti et al. (2019) and was generated from a Moodle learning management systems used in online courses at a public university in Brazil. The dataset contains individual written feedback in Portuguese provided by instructors who posted the feedback messages through the activity submission tool in the LMS. The dataset is composed of 1,000 examples of feedback from 2 courses: biology (41 examples of feedback) and literature (959 examples of feedback). The average length of feedback texts is 30 words.

Each feedback example was classified by experts who followed the instructions described by

Hattie e Timperley (2007) on the 4 levels of feedback. Each feedback content was analyzed by two expert coders separately. After this step, the differences between each pair of experts were verified. Another two experts who did not participate in the first step resolved the divergent cases (27.8% of the total). The inter-rater agreement had a percentage of 72.2% and Cohen's Kappa was 0.44, enough values for content analysis experimentation (NEUENDORF, 2016).

At the end of the annotation process, we derived a dataset with binary classes: **class 0** if a feedback message does not belong to a particular level; **class 1** if the feedback message belongs to the feedback level. The dataset was divided as shown in Table 23, i.e., from 1,000 feedback examples 888 were classified as level FT, 501 level FP, 8 level FR, and 151 level FS (see Table 32 for the definitions of the feedback levels).

Tabela 23 – Dataset division after annotation.

	Level			
	FT	**FP**	**FR**	**FS**
class 0	112	499	992	849
class 1	888	501	8	151
Total	1000	1000	1000	1000

Fonte: CAVALCANTI et al. (2020)

The FR level had only eight feedback examples out of a total of 1,000, which was not enough to fit a machine learning model (FRIEDMAN; HASTIE; TIBSHIRANI, 2001). Thus, we decided to use only the other levels (FT, FP and FS). Table 24 shows the dataset for these levels divided by course.

Tabela 24 – Dataset division by course.

		Level		
Course	**Class**	**FT**	**FP**	**FS**
Literature	0	109	462	811
Literature	1	850	497	148
Biology	0	3	37	38
Biology	1	38	4	3
	Total	1000	1000	1000

Fonte: CAVALCANTI et al. (2020)

7.4.2 Feature extraction

Feature extraction was performed based on reviewing relevant studies (FARROW; MOORE; GAŠEVIĆ, 2019; NETO et al., 2018; KOVANOVIĆ et al., 2016; CAVALCANTI et al., 2019). These studies suggest using linguistic features (TAUSCZIK; PENNEBAKER, 2010; MCNAMARA et al., 2014) rather than traditional text classification features because Kovanović et al. (2016) argues, traditional features (e.g., unigrams, bigrams, trigrams, POS) are "data set dependent" and the data itself defines the classification space. Hence, Kovanović et al. (2016) suggests that using traditional text classification features leads to an unnecessary increase in the training space, that generates thousands of resources (even for small data sets), which can lead to the overfitting of the training data.

Kovanović et al. (2016) performed an automatic content analysis of discussion transcripts by extracting 205 features based mainly on LIWC (Linguistic Inquiry and Word Count) (TAUSCZIK; PENNEBAKER, 2010) and Coh-Metrix (MCNAMARA et al., 2014) for the English language. However, due to the limitation of the natural language processing resources for Portuguese, only 116 features could be used in Cavalcanti et al. (2019) to perform a similar analysis of online discussions to that done by Kovanović et al. (2016). Therefore, the current study used the LIWC and Coh-Metrix text analysis tools, and extracted four additional features, namely, the number of named entities, polarity of feeling (obtained through sentiment analysis, presence of greeting, and positive expressions, which were proven relevant for this problem (CAVALCANTI et al., 2019). In the following sections, we present all of the features we extracted in this study.

7.4.2.1 LIWC

LIWC is a lexical dictionary that groups words into categories with psychological significance, such as emotions, cognitive processes, life concerns, social words, and various categories of functional words (PENNEBAKER et al., 2015). The distribution of these categories in the text may provide information about the author's psychological state or reflect an author's personal condition (WISSEN; BOOT, 2017). The LIWC dictionary has 127,149 entries, where each entry can be assigned to one or more categories (FILHO; PARDO; ALUÍSIO, 2013) and provides a large number of word counts that are indicative of different psychological processes, such as affective, cognitive, social and perceptual (KOVANOVIĆ et al., 2016). In this study, a total of 64 LIWC features were extracted. As good feedback practices are not entirely related to

the content of the text but also the style of the writing, these features are relevant to our problem as they provide structural characteristics of text. LIWC also provides features related to emotions which can be useful to identify the FS level.

7.4.2.2 Coh-Metrix

According to Scarton e Aluísio (2010), Coh-Metrix is a computational linguistics tool that calculates indexes that assess the cohesion, coherence, and difficulty of comprehension of a text using various levels of linguistic analysis, such as lexical, syntactic, discursive, and conceptual. The free version of Coh-Metrix for English language has 60 indices ranging from simple metrics (such as word count) to more complex measurements involving anaphoric resolution algorithms. The version of Coh-Metrix for Portuguese language has 48 different measures (SCARTON; ALUÍSIO, 2010). Coh-Metrix provides a set of features about the text that are widely adopted in the educational literature to evaluate the quality of written activities and text (MCNAMARA et al., 2014).

7.4.2.3 Additional Features

In addition to LIWC and Coh-Metrix, this work also applied four additional features of the feedback text messages: number of named entities, the polarity of feeling, presence of greeting, and positive expressions. These features could influence especially the FS level as sentiment is important for it (CAVALCANTI et al., 2019). **Number of Named Entities** is the name given to the task of identifying and classifying proper names in texts, including locations such as Brazil and Australia; people, such as Arthur and Daniela; and organizations such as Apple and Coca-Cola (MCCALLUM; LI, 2003). **Polarity of Feeling** is divided into positive, negative or neutral sentiments; polarity is obtained using a lexical dictionary of sentiment analysis (CARVALHO; SILVA, 2015). **Presence of Greeting** represents the presence of phrases such as *"Good morning"*, *"Good afternoon"*, *"Good evening"*, *"How are you?"*, *"Are you ok?"*. **Positive Expressions** reflect the presence of phrases such as *"Very good"*, *"Excellent"*, *"Congratulations"*, *"Good work"*, *"Great"*, *"Good studies"*, and others. These features were also proposed in text classification studies (KOVANOVIĆ et al., 2016; NETO et al., 2018; CAVALCANTI et al., 2019). According to Cavalcanti et al. (2019), the number of named entities may reflect the quality of feedback, as the large number of entities may indicate that the teacher

is referencing key concepts and authors by giving effective examples of a particular subject. The presence of feeling, greeting and positive expression is related to compliments that the teacher can make to a student and attempts to establish social presence.

7.4.3 Data processing

Initially, the data were divided into training and testing sets (70% training and 30% testing), a necessary step taken in machine learning to avoid overestimating model performance (which can occur if model accuracy estimated on the same data as model parameters has been learned) (FRIEDMAN; HASTIE; TIBSHIRANI, 2001). Thus, the split data set included 700 and 300 instances for the training and test data sets, respectively (Table 40).

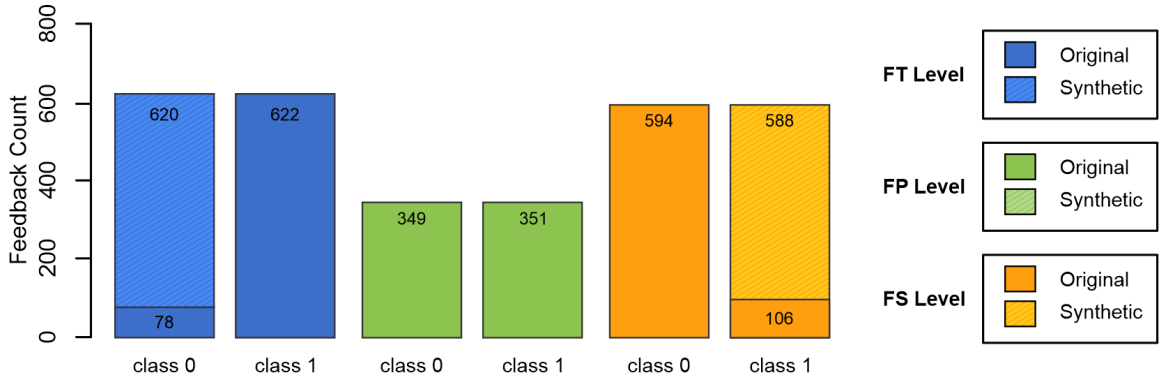
Tabela 25 – Dataset division in train and test.

		Class 0	Class 1	Total
FT	**Train**	78 (11.15%)	622 (88.85%)	700 (100%)
	Test	34 (11.33%)	266 (88.67%)	300 (100%)
FP	**Train**	349 (49.85%)	351 (50.15%)	700 (100%)
	Test	150 (50%)	150 (50%)	300 (100%)
FS	**Train**	594 (84.85%)	106 (15.15%)	700 (100%)
	Test	255 (85%)	45 (15%)	300 (100%)

Fonte: CAVALCANTI et al. (2020)

From Table 40 it is possible to see that the classes FT and FS are unbalanced. In the training set, class FT had 78 feedback examples belonging to class 0 and 622 belonging to class 1 and class FS had 594 and 106 feedback examples belonging to classes 0 and 1, respectively. According to Tan et al. (2007), class imbalance can have very negative effects on the results of classification analysis. To address the data set imbalance, we used the SMOTE algorithm, an approach also suggested by Kovanović et al. (2016), Neto et al. (2018), and Cavalcanti et al. (2019). SMOTE processes the data points in a two-dimensional resource space of a specific class selected for re-sampling and creates synthetic data points as a linear combination of existing data points (CHAWLA et al., 2002). The algorithm considers the nearest K-neighbors for each x_i instance belonging to the majority class, for a given value of K. The nearest K-neighbors are defined as the K elements of the majority class whose Euclidean distance from each other and instance x_i has the lowest value. To create a new "synthetic" instance,

Figura 13 – Distribution of the feedback messages for the training set by class for FT, FP and FS levels after the application of the SMOTE algorithm.



Fonte: CAVALCANTI et al. (2020)

SMOTE randomly selects one of the nearest K-neighbors from the x_i instance, subtracts the x_i instance from its nearest neighbor, multiplies this difference by a random number between 0 and 1, and adds to the instance value (x_i).

Figure 13 shows the new distribution of feedback messages for the training set of each level after applying the SMOTE algorithm. Class 0 increased from 78 to 620 for FT level and class 1 increased from 106 to 588 for FS level. Besides, the figure presents the final values for FP level without synthetic data because the data is already balanced.

7.4.4 Model Selection and Evaluation

Fernández-Delgado et al. (2014) showed that Gaussian kernel random forests and SVMs were the best performing algorithms in comparative analyses, taking into account 179 commonly used classification algorithms across 121 different data sets. Because of this performance advantage over other algorithms and its popularity as a technique for classification, prediction, variable importance study, variable selection and outlier detection (VERIKAS; GELZINIS; BACAUSKIENE, 2011), this study adopted the Random Forest algorithm to address research question 1.

The Random Forest algorithm consists of a set of decision trees generated on the same object. The method presented by Breiman (2001) consists of a set of decision trees built at the time of training, forming the Random Forest. Each tree is created by randomly selecting a sample of attributes contained in a resource vector. The object (set of trees) goes through a voting (bagging) mechanism which elects the most voted classification, resulting in a class

prediction (BREIMAN, 2001). The classifier output is provided by the class that is returned in response to most of the classifications at the endpoints of the trees belonging to the forest.

In order to evaluate our model we used accuracy and Cohen's kappa (). To calculate accuracy, the confusion matrix which is composed by the following calculations was used in the current study: True positive (TP) – number of positive elements classified as positive; True negative (TN) – number of positive elements classified as false; False positive (FP) – number of false elements classified as positive; and False Negative (FN) – Number of false elements classified as false. Thus, accuracy is obtained from Chapter 7.1.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{7.1}$$

We also report the Cohen's kappa () result which is largely used in the content analysis literature and address the accuracy limitations (NEUENDORF, 2016). Cohen's kappa () consists of a coefficient developed to measure the agreement between two raters (COHEN, 1960). In our case, that was the agreement between automatically and manually coded messages in our dataset.

The two parameters used in the Random Forest classifiers (BREIMAN, 2001) were set up: (i) *ntree* – the number of trees generated by the algorithm; and (ii) *mtry* – the number of random features selected by each tree. To obtain optimal random forest parameters, 10-fold cross-validation was used in the training data and then reported the accuracy of the best performing model classification in the test data.

The Random Forest algorithm provides methods to analyze the importance of the classification features, which were used to address research question 2 in the current study. One popular measure to calculate the feature importance is Mean Decrease Gini (MDG) which is based on the reduction in Gini impurity measure (BREIMAN, 2001). In this paper, we adopted MDG to evaluate the relevance of different features to the outcome of our classifiers.

7.4.5 Implementation

The feature extraction was performed using the Python language. The classifier was coded using R. The packages and libraries used were:

- the spaCy[1] and NLTK[2] packages, for natural language processing;

[1] https://spacy.io
[2] http://www.nltk.org/

- scikit-learn[3], for stratified sampling of the test and training datasets; (PEDREGOSA et al., 2011)
- the Random Forest R package, for classifier development (LIAW; WIENER et al., 2002);
- The caret R package, for model training, selection, and validation (KUHN, 2015);
- The Coh-Metrix Portuguese version (SCARTON; GASPERIN; ALUISIO, 2010), and;
- The LIWC Portuguese version [4](FILHO; PARDO; ALUÍSIO, 2013).

7.5 RESULTS

7.5.1 Model training and evaluation – RQ1

To obtain the optimal Random Forest parameters (*ntree* and *mtry*), we adopted a 10-fold cross-validation approach on the training data for each binary classifier. Then we reported the accuracy and Cohen's κ of the best performing model on the testing data, which is completely independent of the training set. In addition, we compare the results with and without the use of the SMOTE algorithm (for cases where data is unbalanced, level FT and FS).

7.5.1.1 FT level

Table 26 shows the minimum and maximum values of the *mtry* parameter for the balanced data. In the best case ($mtry = 2$), the classifier achieved 0.95 (0.02) and 0.91 (0.04) of accuracy and Cohen's κ, respectively. Figure 14 shows the performance of the Random Forest model using the optimal *mtry* value ($mtry = 2$) for different numbers of trees. The figure shows that slightly after 100 trees, the results converged because above that the average out-of-bag (OOB) error rate remains stable. Breiman (2001) suggests that it is better to pick a number higher than the one that initially shows a convergence, thus, we decided to adopt 500 for the parameter *ntree*. The OOB error rate suggests that about 5% of the data points were misclassified. Class 1 had a tiny error rate, less than 0.05. This value was expected because class 1 (contains the feature) was the majority class, having almost eight times more examples than class 0 (error rate around 0.10).

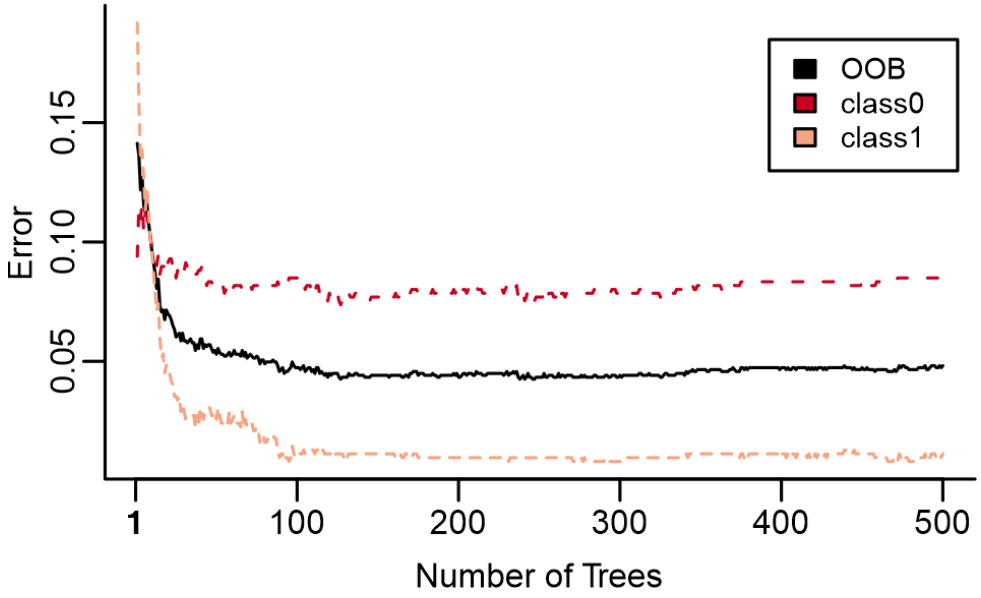
[3] https://scikit-learn.org/
[4] http://www.nilc.icmc.usp.br/portlex/index.php/en/liwc

Tabela 26 – The results of parameter tuning for the Random Forest classifier for the FT level.

	mtry	Accuracy	Kappa
Min	115	0.94 (0.01)	0.89 (0.02)
Max	2	0.95 (0.02)	0.91 (0.04)
	Difference	0.01	0.02

Fonte: CAVALCANTI et al. (2020)

Figura 14 – The best Random Forest configuration performance for level FT.



Fonte: CAVALCANTI et al. (2020)

For the testing set, the classifier obtained 0.58 of accuracy (95% CI : [0.5219, 0.6365]) with Cohen's κ of 0.2351 in the best model without SMOTE and 0.75 of accuracy (95% CI : [0.697, 0.798]) and Cohen's κ of 0.2985 in the best model with SMOTE. It is important to remark that the SMOTE algorithm was applied just to the training set. The difference in accuracy between balanced and unbalanced data was 0.17. These results show that, in this case, the unbalanced data influences the accuracy of the classifier.

7.5.1.2 FP level

Table 27 shows the minimum and maximum values of the *mtry* parameter for the FP classifier. Differently for the previous one, the best value for *mtry* was 37. Using this parameter, the classifier achieved an accuracy of 0.81 (0.04) and Cohen's κ of 0.62 (0.08). Figure 15 shows the performance of the Random Forest model using the optimal *mtry* value on the training

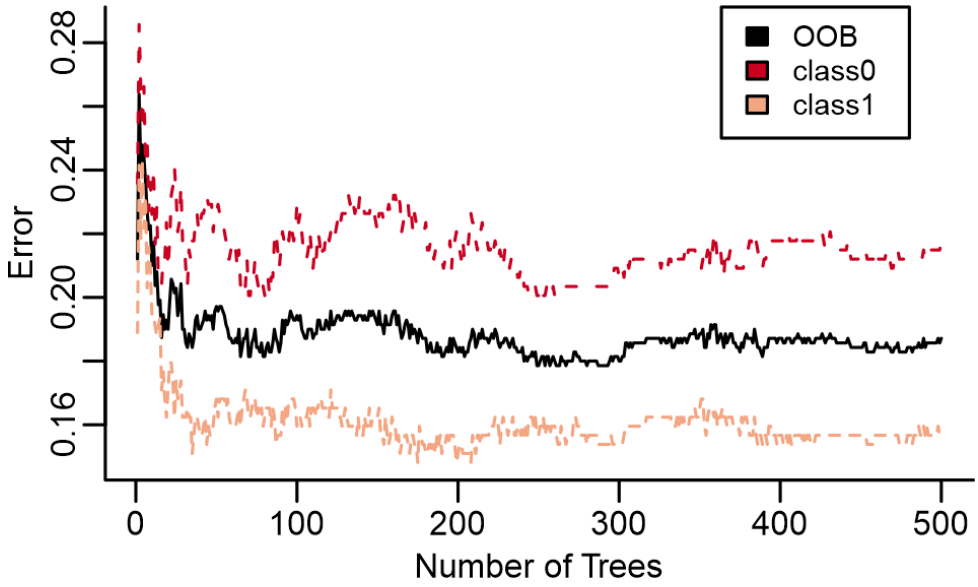
set. The OOB error rate had many variations between 0 and 400 trees, but after that, it converged. So we adopted *ntree* = 500. We can also see that the difference in error between classes 0 and 1 was small (around 6%). This slight difference probably occurred because of the data was balanced.

Tabela 27 – The results of parameter tuning for the Random Forest classifier for the FP level.

	mtry	Accuracy	Kappa
Min	103	0.79 (0.04)	0.58 (0.08)
Max	37	0.81 (0.04)	0.62 (0.08)
	Difference	0.02	0.04

Fonte: CAVALCANTI et al. (2020)

Figura 15 – The best Random Forest configuration performance for level FP.



Fonte: CAVALCANTI et al. (2020)

In the evaluation with the test set, the classifier obtained 0.64 of accuracy (95% CI : [0.5863, 0.6975]) and Cohen's κ of 0.2867 in the best model.

7.5.1.3 FS level

Finally, the last model evaluated was FS level. Table 26 shows the minimum and maximum values of the *mtry* parameter for the balanced data. The classifier reached the best outcome for *mtry* = 31. In this case, the accuracy and Cohen's κ were 0.97 (0.01) and 0.95 (0.03),

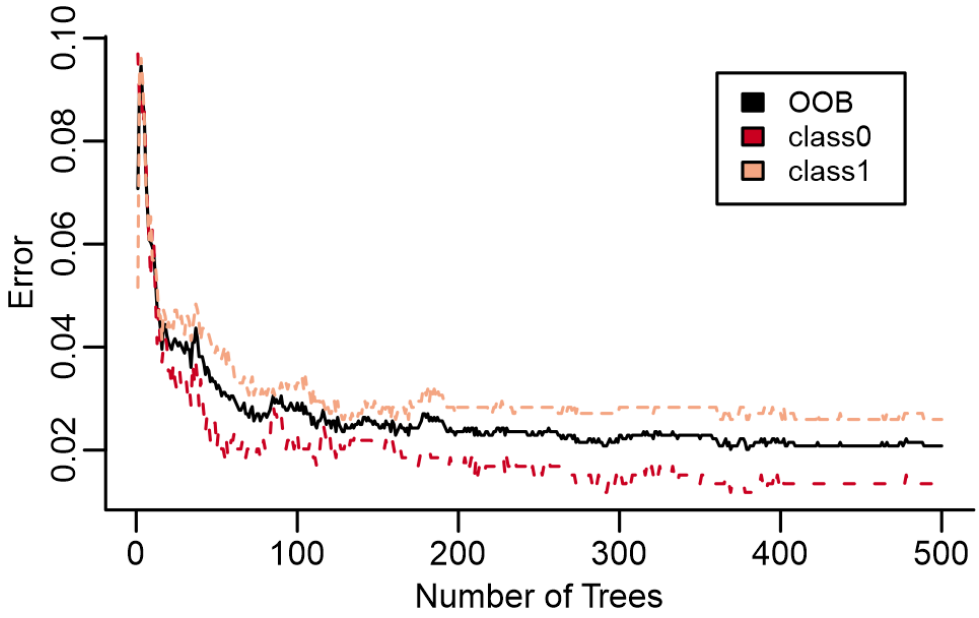
respectively. Figure 16 shows that 200 trees were sufficient to ensure the convergence of the error rate. The OOB remained stable after 200 trees, with approximately 3% misclassified data points. Again, the error difference between the classes was irrelevant, and class 0 (majority class) had an error of less than 0.02. This result was the best error rate obtained among the three levels of feedback analyzed.

Tabela 28 – The results of parameter tuning for the Random Forest classifier for the FS level.

	mtry	Accuracy	Kappa
Min	115	0.96 (0.01)	0.93 (0.03)
Max	31	0.97 (0.01)	0.95 (0.03)
	Difference	0.01	0.02

Fonte: CAVALCANTI et al. (2020)

Figura 16 – The best Random Forest configuration performance for level FS.



Fonte: CAVALCANTI et al. (2020)

For the test set, the classifier obtained 0.8533 of accuracy (95% CI : [0.8082, 0.8914]) with Cohen's κ of 0.3794 in the best model without SMOTE and 0.87 of accuracy (95% CI : [0.8266, 0.9059]) and Cohen's κ of 0.3963 in the best model with SMOTE. The results with both training and test sets showed a slight difference in the accuracy between the balanced and unbalanced data.

7.5.2 Features importance analysis – RQ2

This study also analyzed the contributions of the different features of the final performance of the classifier. As mentioned before, we adopted the Mean Decrease Gini (MDG) measure to define the degree of relevance of a feature (BREIMAN, 2001). Tables 29, 30 and 31 show the 20 most important features for the FT, FP and FS level, respectively.

The most important variable for the FT level was *bc_paragraphs*, i.e., the basic paragraph count. Since the purpose of the FT level is to provide students with corrective information (HATTIE; TIMPERLEY, 2007), this feature is directly related to the amount of corrective information provided to students. The features *cm.cw_fq:min_cw_freq* (rarer content word frequency), *cm.bc:sentences* (basic sentence count), and *cm.bc:words_per_sen* (basic word count per sentence) showed similar trends.

The second most important feature for the FT level was the pronoun count by phrases (*cm.tk:pron_per_np*). This feature was relevant due to the fact that the instructors predominately use the personal pronoun in the second person to refer to the student. For example: "*Student X, you need to include more about the Versailles Treaty.*"The *cm.tk:per_pronouns* and *liwc.ppron* features that count the number of personal pronouns also support this argument.

On the other hand, the FP level provides higher-level feedback focused on the learning process and not only on a specific task. Thus, the text of feedback messages on this level provided more details than the text of messages on the FT level. Given this, the most predictive features of this category were the adverb (liwc.adverb) and cognition mechanism (*liwc.cogmech*) counts. These features are indicative of the richness of details in the instructors' text in feedback messages. Also, feedback related to the process, in general, possesses a larger amount of content. So, features related to the text amount are also highlighted in Table 30: number of words (cm.bc:words), total function words (liwc.funct), minimum among content words frequency (cm.cw_fq:min_cw_freq), number of inhibition words (liwc.inhib), and number of words per sentence (cm.bc:words_per_sen).

Finally, we analyzed the top features for the FS level. The FS level aims to provide feedback about the student (i.e., self) and provide motivation for the student (HATTIE; TIMPERLEY, 2007). In this sense, Table 31 presents several features related to compliments: *add.compliment*, *liwc.feel*, *liwc.posemo* (positive emotion), *liwc.affect*, *cm.bc:adjectives* and *liwc.leisure*. The most important feature was the frequency of the rarest content word. Since

Tabela 29 – Top-20 features importance for FT level.

Order	Variable	Description	MDG
1	cm.bc:paragraphs	Number of paragraphs	15.69
2	cm.tk:pron_per_np	Pronouns by syntagmas	13.68
3	cm.cw_fq:min_cw_freq	Mininum among content words frequency	11.93
4	liwc.health	Number of health words	11.71
5	liwc.they	Number of 3rd person pronouns	11.44
6	cm.bc:sentences	Number of sentences	11.19
7	cm.tk:personal_pron	Incidence of personal pronouns	10.91
8	add.compliment	Number of compliments	10.60
9	cm.crf:stem_ovl	Word Radical Overlap	9.91
10	cm.con:cau_pos_inc	Connective incidence classified as positive causal	9.29
11	cm.crf:adj_stem_ovl	Adjacent Word Radical Overlap	9.17
12	cm.hyp:hyp_verbs	Hypernyms verbs	9.13
13	liwc.negemo	Number of words with negative emotion	9.04
14	cm.crf:adj_cw_ovl	Content words overlap in adjacent sentences	9.03
15	liwc.sexual	Number of sexual words	8.95
16	liwc.motion	Number of motion words	8.89
17	liwc.future	Number of words in future	7.34
18	liwc.ppron	Number of personal pronouns	7.05
19	cm.bc:wd_per_sen	Number of words per sentence	6.94
20	cm.anp:adj_refs	Adjacent anaphoric reference	6.91

Fonte: CAVALCANTI et al. (2020)

the text of feedback messages on the FS level was related to compliments, these feedback messages were usually short and used rare words (i.e., words with the lowest frequency among all content words).

Tabela 30 – Top-20 variable importance for FP level.

Order	Variable	Description	MDG
1	liwc.adverb	Number of adverbs	31.04
2	liwc.cogmech	Mechanism of cognition	22.75
3	cm.bc:words	Number of words	17.25
4	add.greeting	Number of greetings	14.16
5	liwc.quant	Number of quantifiers	13.60
6	cm.const:np_inc	Incidence of Syntagms	12.71
7	liwc.funct	Total function words	12.48
8	liwc.time	Number of time words	12.30
9	liwc.affect	Number of affective processes	10.68
10	liwc.cause	Number of causation words	8.87
11	cm.amb:verbs_amb	Verb Ambiguity	8.62
12	liwc.money	Number of money words	7.99
13	cm.cw_fq:min_cw_freq	Minimum among content words frequency	7.37
14	cm.const:mod_per_np	Mean pronouns per noun phrase	7.11
15	cm.con:add_neg_con_inc	Connective incidence classified as negative additives	6.67
16	liwc.inhib	Number of inhibition words	6.49
17	cm.bc:words_per_sen	Number of words per sentence	6.31
18	cm.bc:flesch	Flesch Readability Index	6.28
19	cm.amb:nouns_amb	Noun ambiguity	6.28
20	liwc.discrep	Number of discrepancy words	6.20

Fonte: CAVALCANTI et al. (2020)

Tabela 31 – Top-20 variable importance for FS level.

Order	Variable	Description	MDG
1	cm.cw_fq:min_cw_freq	Minimum among content words frequency	88.97
2	add.compliment	Number of compliments	74.75
3	liwc.past	Number of words in past	52.43
4	liwc.feel	Number of feeling words	52.22
5	liwc.posemo	Number of words with positive emotion	46.94
6	liwc.auxverb	Number of auxiliary verbs	22.15
7	liwc.quant	Number of quantifiers	15.42
8	liwc.future	Number of words in future	14.17
9	liwc.affect	Number of affective processes	12.69
10	add.ners	Number of Named Entities	12.08
11	cm.amb:adj_amb	Adjective ambiguity	10.57
12	cm.cw_fq:cw_freq	Content words frequency	10.51
13	liwc.certain	Number of certainty words	8.21
14	liwc.excl	Number of exclamation points	7.98
15	cm.bc:adjectives	Number of adjectives	7.91
16	liwc.humans	Number of human words	7.73
17	liwc.leisure	Number of leisure words	7.60
18	cm.lo:neg_inc	Negation incidence	7.02
19	liwc.space	Number of space words	6.98
20	liwc.body	Number of body words	6.91

Fonte: CAVALCANTI et al. (2020)

7.6 DISCUSSION

The focus of RQ1 was investigating whether natural language processing methods can be effective in automatically identifying different aspects of good feedback practice. The results of the classifiers developed showed that the combination of additional features and word counts extracted from LIWC and Coh-Metrix were effective in classifying feedback texts in three different levels, reaching an accuracy of 0.75, 0.64, and 0.87 for levels FT, FP, and FS, respectively. The values of Cohen's κ of 0.29 (FT), 0.28 (FP) and 0.39 (FS) represent a medium to substantial inter-rater agreement (LANDIS; KOCH, 1977b).

Although we did not find much research work conducting similar analyses of good feedback practice to compare to, it is important to highlight that the method presented in this paper displayed better results than the classifier method proposed by Cavalcanti et al. (2019), in which Cavalcanti et al. (2019) performed a feedback analysis using a similar approach. Furthermore, the optimization of the *mtry* (i.e., the number of attributes used in each tree of the forest) and *ntree* (i.e., the number of trees used in each iteration) parameters, as well as the adoption of SMOTE for balancing the data, improved the final results in all cases.

The focus of RQ2 was identifying what features can best predict the different feedback levels defined by Hattie e Timperley (2007). To answer the second RQ, we created tables with the 20 most important features, considering the mean decrease impurity index, for each level. Next, we try to derive practical implications from the relationship between the output in these tables and the objectives of each level.

Feedback about the task (FT) focuses specifically on how well a student, or a group of students, are performing or accomplishing a specific task. Hattie e Timperley (2007) suggests that having information where you did wrong on the task is a precondition for understanding the process of how you arrive at the solution, which in turn can help a student to build self-regulation skills. Hence, as shown in Table 29, among the most important features for FT are **pronouns** – which distinguish whether feedback is given on individual and group situations, as not every feedback is perceived as relevant by everyone equally; **the number of words in future** – also a very important feature, as it shows whether feedback includes actions (that Hattie e Timperley (2007) called feed-forward) that students need to do to accomplish the task; features related to **text complexity** – for this level of feedback, the instructors should provide very specific and detailed guidance on how to resolve a low-level problem; and **number of words with negative emotions** – having correct information about the potential negative

effects that not solving a task can have on your learning, might increase the likelihood that the student can more easily remember the error connected with negative emotion (i.e., effect on learning)(XIE; ZHANG, 2017).

Feedback about the processing of the task (FP) intends to provide corrective information on a higher level, which involves the process of knowledge construction and the cognitive process of the student during the work on a task (HATTIE; TIMPERLEY, 2007). This level of feedback should help students to shape their individual strategies on how to approach a specific task and how to pursuit goals in their learning process. Therefore, Table 30 shows the dominance of features related to the size of the feedback provided, such as **the number of words per sentence** and **total number of function words**. In other words, the instructor needs to provide information about a wider scenario, and not only focusing on a specific task. This way, students could develop the self-regulation skills necessary to create their own ideas on how to solve a particular problem. Another dominant feature was the **number of adverbs** – since the function of an adverb is to modify adjectives, verbs, or other adverbs by adding a particular precision or nuance (TAUSCZIK; PENNEBAKER, 2010), and this feature may be related to the richness of detail in the instructors' text. One more dominant feature to highlight is the **number of cognition mechanism constructors** – this feature counts the cognition mechanism in the comments (e.g., cause, know, ought). According to Pennebaker et al. (2015), these words are indicative of more complex language because they are being used to create causal explanations for organizing students' thoughts. The importance of this feature can be interpreted in the context of the FP purpose – to support the development of students' cognition.

The feedback about self as a person (FS) aims to motivate or complement the student, and, in general, contains less task-related comments. This feedback level can be effective only if it influence engagement, effort, or feelings in relation to the learning or the task that is being performed. tab:variable3 presents features related to **the number of compliments**, **number of feeling words**, and **words with positive emotions** – these are typically used by the instructor to try to change the state of mind of a student related to a task, or to compliment a good performance in an activity; **the minimum amount of content words** – is very representative of this feedback level as, in this case, the instructor should not present any deeper concept about the task; and finally the **number of human and leisure words** - which shows that the instructors have tried to be more informal with this type of feedback.

The MDG measure for the features on the FS level had much higher values than at the

other levels. For example, the most important feature of the FS level had an MDG of 88.97, a value almost three times greater than the first feature of the FP level (31.04) and almost six times greater than the highest-ranked feature of the FT level (15.69). We can draw several conclusions from this result: (i) It shows that the nature of the FS level is more restrictive than the others; (ii) The range of features used by the FS classifiers is reduced compared to the ones applied by the FT and FP ones; (iii) These values may explain why we reached better results in the classifier for FS. In general, the more high MDG features a class has, more probable it is to get higher accuracy because the features are more descriptive of this class (BREIMAN, 2001).

Moreover, the results of the most important features for the FS level additionally corroborate the suggestions by Hattie e Timperley (2007) that feedback on the self level is least valuable for learning. While feedback messages on the FS level may have offered some reassurance to the students, they hardly offer any helpful guidance to the learner on how to improve self-regulation of their performance and how to adjust the use of their learning strategies. In practice, when feedback on this level is detected with the classifier, suggestions should be offered to instructors to supplement them with relevant indicators on the process (FP) and regulation levels (FR) to increase the quality of feedback.

The low number of messages on the self-regulation level (FR) is probably revealing of generally poor feedback practice happening in real courses. Sadly, in our case, we had so few instances of feedback messages (eight) with this feedback level that we could not train a classifier. This low number might also be a potential reason why many studies generally report low perceived value of students with feedback in higher education (HENDERSON et al., 2019a). Given that the FR level together with the FP level (process) are the two most potent feedback, the potential of the automatic tools for checking the quality at the time when the instructors write feedback is high.

7.6.1 Study Limitations

One of the limitations of the study is related to the database. The data collected are feedback messages from two online courses (biology and literature). Thus, the study may not be fully representative of the different feedback in online learning courses. Also, many feedback messages had similar or repeated texts, as one had data from only two courses, the same instructor may have provided the same feedback to several students. Another important

factor related to the database is that the FR level had only eight instances from a total of 1000, which was not enough to train a classifier, and the FT and FS levels were unbalanced and this may affect the final classifier performance.

7.7 CONCLUSIONS

This study proposed two contributions. First, we developed several binary classifiers to encode different criteria of good feedback proposed by Hattie e Timperley (2007) in Portuguese written feedbacks from an online course at a public university in Brazil. For this, text features were extracted using natural language processing methods. The results presented good performance comparing with similar works that used the same data set (CAVALCANTI et al., 2019) and other content analysis methods found in the educational literature (NETO et al., 2018; KOVANOVIĆ et al., 2016; FARROW; MOORE; GAŠEVIĆ, 2019). The proposed approach, mainly based on features extracted from Coh-Metrix and LIWC, shows that it is possible to provide a fully automated system for coding the three levels of feedback.

Second, an analysis of the most important features was performed, explaining the characteristics that are more descriptive of each level of feedback. This analysis is fundamental to understand which features play a key role in the classification and it can be further used to provide insights to instructors on how to improve their feedbacks.

As future works, we intend to assess the feedback in another field (i.e., computer science) and other e-learning resources (e.g., forum and chats). Besides, we plan to use the outcomes of this work to develop a tool to assist teaching staff in the process of providing written feedback to students.

8 UTILIZAÇÃO DE RECURSOS LINGUÍSTICOS PARA CLASSIFICAÇÃO AUTOMÁTICA DE MENSAGENS DE FEEDBACK

Authors: Anderson Pinheiro Cavalcanti, Rafael Ferreira Mello, Péricles Miranda, André Nascimento, Fred Freitas

8.1 INTRODUÇÃO

Uma educação de qualidade requer atenção e apoio pessoal. Um elemento crucial para isso é o feedback, cuja definição muda dependendo da literatura de pesquisa educacional. Neste artigo, o feedback é conceituado como o fluxo de informações de um agente para outro a respeito de uma decisão do aluno (HATTIE; TIMPERLEY, 2007). Vários estudos destacam o papel significativo que o feedback desempenha no processo de aprendizagem (BATISTA; SALGADO; BARRETO, 2018; CAVALCANTI et al., 2021). Price et al. (2010a) afirmam que o feedback é o componente mais crítico das avaliações. O feedback direciona os alunos para o tipo apropriado de estudo ou prática e ajuda os indivíduos a reconhecer áreas de deficiência, que podem ser usadas para aprimorar as táticas e estratégias de aprendizagem (PARIKH; MCREELIS; HODGES, 2001; WEAVER, 2006b). Dweck (1999) relata que o feedback pode afetar a motivação do aluno, bem como os conteúdos e habilidades aprendidas. Laurillard (1993) enfatizou que a ação sem feedback é completamente improdutiva para o aluno.

Apesar do reconhecimento generalizado da importância do feedback para a aprendizagem, grande parte da literatura atual indica a difusão do feedback de baixa qualidade no ensino superior (HATTIE; GAN, 2011). A qualidade do feedback é consistentemente avaliada como uma das maiores causas de insatisfação para estudantes do ensino superior (FERGUSON, 2011). Weaver (2006b) relata que, embora os acadêmicos reconheçam o valor do feedback para facilitar a aprendizagem, eles consideram os comentários dos instrutores incompreensíveis e ineficazes. Ferguson (2011) identifica a falta do fornecimento de feedback oportuno, expectativas pouco claras e pouca utilidade como as principais preocupações entre os alunos.

À medida que as instituições de ensino superior adotam a tecnologia, há um portfólio crescente de abordagens que utilizam a coleta de dados para melhorar os processos de aprendizagem. De acordo com a revisão sistemática realizada por Cavalcanti et al. (2021) vários trabalhos estão explorando ativamente soluções de feedback automatizado que podem permitir que os instrutores identifiquem e empreguem boas práticas de feedback de maneira eficiente

e aumentem a velocidade de entrega de feedback aos alunos. Nesse sentido, alguns estudos examinaram o uso de métodos de mineração de dados para gerar feedback textual automatizado (LIU et al., 2017; MA et al., 2017; VILLALóN et al., 2008). Essas análises são frequentemente limitadas a áreas específicas de domínio, como programação de computadores ou redação, ou falta de base na teoria educacional.

Na literatura é possível encontrar trabalhos que apresentam abordagens para analisar feedback automaticamente (OSAKWE et al., 2021; CAVALCANTI et al., 2020a). Contudo, são abordagens que focam em textos de feedback da língua inglesa ou utilizam poucos recursos linguísticos para extrair informações dos textos.

Diante deste contexto, esse artigo tem como objetivo propor uma abordagem que combina diferentes recursos linguísticos para classificar mensagens de feedback usando o algoritmo XGBoost com base em conceituados modelos de feedback da literatura (HATTIE; TIMPERLEY, 2007; NICOL; MACFARLANE-DICK, 2006). Os resultados indicam um desempenho de classificação eficaz para ambos os modelos de feedback alcançando melhor desempenho do que outros trabalhos da literatura.

8.2 TRABALHOS RELACIONADOS

Alguns trabalhos teorizam que o objetivo principal do feedback é reduzir a discrepância entre o conhecimento/desempenho do exercício e uma meta (HATTIE; TIMPERLEY, 2007). Portanto, o feedback é frequentemente examinado através da autorregulação da aprendizagem (CLARK, 2012; NICOL; MACFARLANE-DICK, 2006). Butler e Winne (1995b) reconhecem que os alunos autorregulados são os alunos mais eficazes. A autorregulação é um ciclo de estabelecimento de metas de construção de conhecimento, seleção de estratégias que maximizem o progresso em direção a tais metas e monitoramento do progresso com a possibilidade de alterar estratégias dependendo do nível de progressão (CLARK, 2012).

De acordo com Nicol e Macfarlane-Dick (2006), a boa prática de feedback é amplamente definida como qualquer estratégia ou conteúdo que possa aumentar a capacidade dos alunos de autorregular seu desempenho de aprendizagem. Nicol e Macfarlane-Dick (2006) propôs sete princípios gerais de boas práticas de feedback para auxiliar o corpo docente: (i) Ajuda a esclarecer o que é o bom desempenho (objetivos, critérios, padrões esperados), (ii) Facilita o desenvolvimento de autoavaliação (reflexão) na aprendizagem , (iii) Oferece informações de alta qualidade aos alunos sobre sua aprendizagem, (iv) Incentiva o professor e o diálogo com

os colegas sobre a aprendizagem, (v) Incentiva crenças motivacionais positivas e auto-estima, (vi) Oferece oportunidades para fechar a lacuna entre o desempenho desejado, (vii) Fornece informações aos professores que podem ser usadas para ajudar a moldar o ensino.

Hattie e Timperley (2007) analisaram várias condições que poderiam maximizar os efeitos positivos do feedback na aprendizagem, incluindo o aumento da consciência do aluno sobre uma meta geral de aprendizagem, o progresso em direção à meta e as metas subsequentes necessárias para atingir o objetivo principal. Assim, Hattie e Timperley (2007) propuseram quatro perspectivas, propostas como níveis, que o feedback deve abordar a fim de melhorar sua eficácia. Eles postularam que seu modelo é mais adequado para examinar o feedback textual porque o modelo é focado em aspectos relacionados a tarefas de aprendizagem, processo de aprendizagem e autorregulação do aluno. A Tabela 32 mostra os níveis propostos por (HATTIE; TIMPERLEY, 2007).

Tabela 32 – Níveis de feedback (HATTIE; TIMPERLEY, 2007)

#	Nível	Descrição
FT	Feedback sobre a tarefa	O feedback pode ser sobre uma tarefa, como se o trabalho está correto ou incorreto, pode incluir instruções para mais informações ou informações diferentes.
FP	Feedback sobre o processamento da tarefa	O feedback pode ser direcionado ao processo usado para criar um produto ou concluir uma tarefa, é mais direcionado ao processamento de informações ou processos de aprendizagem que requerem compreensão ou conclusão da tarefa.
FR	Feedback sobre autorregulação	O feedback para os alunos pode ser focado no nível de autorregulação, incluindo maior autoavaliação ou habilidades de confiança, que podem ter grandes influências na autoeficácia, na proficiência autorregulatória e nas crenças pessoais dos alunos.
FS	Feedback pessoal.	O feedback pode ser pessoal no sentido de que é direcionado a si mesmo. Frequentemente, não está relacionado ao desempenho da tarefa.

Fonte: CAVALCANTI et al. (2021)

Cavalcanti et al. (2020a) propôs uma análise de conteúdo do texto de feedback fornecido pelos instrutores com base nas boas práticas de feedback propostas por Nicol e Macfarlane-Dick (2006). Os autores se concentraram em analisar a qualidade do feedback extraído das avaliações coletadas em um curso online oferecido em uma instituição de ensino superior brasileira. Um dos problemas desse trabalho é a divisão dos feedbacks em classes binárias (classe 0: não tem boa prática; classe 1: tem pelo menos 1 boa prática). Esse tipo de classificação pode não ser totalmente eficaz, pois o professor apenas tem a informação se o texto tem ou não boa prática. Para que o professor consiga fornecer um feedback de qualidade para os

alunos, é necessário que ele saiba quais dentre as 7 boas práticas o seu texto está seguindo ou não (NICOL; MACFARLANE-DICK, 2006). Além disso, a combinação de ambas as teorias pode ajudar o professor no fornecimento de um feedback de qualidade (CAVALCANTI et al., 2020b).

Para treinar um classificador é necessário transformar o texto de feedback em características numéricas. Existem ferramentas linguísticas que auxiliam na extração de características. O trabalho de Camelo, Justino e Mello (2020) propõe uma adaptação da ferramenta Coh-metrix para a língua Portuguesa. Essa ferramenta extrai diferentes recursos linguísticos do texto, incluindo legibilidade, coerência e coesão textual. Esses recursos podem ser utilizados para análise de diferentes tipos de textos educacionais como redações, mensagens em fóruns educacionais ou até mesmo feedback educacional. Ferreira et al. (2020) propõem uma abordagem baseada em uma combinação de recursos tradicionais de mineração de texto e contagem de palavras extraídas com o uso de estruturas linguísticas para rotulagem automática do conteúdo das mensagens em discussões online.

O trabalho de Cavalcanti et al. (2020) propõe um abordagem para analisar textos de feedback automaticamente utilizando o algoritmo *Random Forest*. Os autores seguiram o modelo que classifica o feedback em quatro níveis diferentes (HATTIE; TIMPERLEY, 2007). Contudo, os autores utilizaram um conjunto limitado de características, o que pode reduzir a generalização do classificador para os diferentes níveis de feedback.

Dessa forma, este trabalho difere dos trabalhos mencionados em 3 aspectos principais: (1) Classificar automaticamente as boas práticas propostas por Nicol e Macfarlane-Dick (2006); (2) Classificar automaticamente os níveis de feedback propostos por Hattie e Timperley (2007); (3) Utilizar o algoritmo XGBoost em conjunto com novos recursos linguísticos para análise das mensagens de feedback (CAMELO; JUSTINO; MELLO, 2020).

8.3 MÉTODO

8.3.1 Conjunto de Dados

O conjunto de dados usado nesse estudo consiste em comentários de feedback fornecidos por instrutores (CAVALCANTI et al., 2019). O conjunto de dados possui 1.000 exemplos de feedback escritos em português dos cursos de Biologia e Literatura. O comprimento médio do comentário foi de 30 palavras por comentário. Os exemplos de feedback foram codificados por especialistas usando instruções propostas nos dois modelos de feedback (NICOL;

MACFARLANE-DICK, 2006; HATTIE; TIMPERLEY, 2007). Com isso, cada texto de feedback poderia ser classificado da seguinte forma: classe 0, se o feedback não pertence ao nível ou boa prática; classe 1: se o feedback pertence ao nível ou boa prática. A Tabela 33 mostra a divisão do conjunto de dados para cada boa prática e nível de feedback.

Tabela 33 – Divisão da base de dados para os níveis e boas práticas.

	BP1	BP2	BP3	BP4	BP5	BP6	BP7	FT	FP	FR	FS
Classe 0	875	971	692	987	766	397	1000	112	499	992	849
Classe 1	125	29	308	13	234	603	0	888	501	8	151
Total	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000

Fonte: CAVALCANTI et al. (2021)

8.3.2 Extração de características

Foram extraídas 161 características dos textos usando o LIWC (*Linguistic Inquiry and Word Count*), o Coh-Metrix proposto por Camelo, Justino e Mello (2020), e 3 características adicionais (número de entidades nomeadas, presença de saudação e presença de expressões positivas). Essas características foram escolhidas por terem apresentado bom desempenho em trabalhos de análise de textos educacionais (BARBOSA et al., 2020; CAVALCANTI et al., 2020; BARBOSA et al., 2021).

O LIWC é uma ferramenta lexical que analisa palavras de grupos de texto que refletem várias dimensões psicológicas (TAUSCZIK; PENNEBAKER, 2010). As principais categorias fornecidas pelo LIWC incluem processos cognitivos, processos sociais, linguagem informal, preocupações pessoais, afeto, relatividade, orientação temporal, impulsos e processos perceptivos. A contribuição relativa dessas categorias no texto oferece um perfil descritivo de vários construtos psicológicos envolvidos na escrita. O dicionário LIWC tem mais de 120.000 palavras, onde cada palavra pode ser atribuída a uma ou mais categorias. Foram extraídas 64 características do LIWC.

O Coh-Metrix é um sistema de linguística computacional que mede um conjunto de características sobre o texto que são amplamente adotadas na pesquisa educacional para avaliar a qualidade das atividades escritas (CAMELO; JUSTINO; MELLO, 2020). Coh-Metrix pode fornecer uma visão sobre coesão, linguagem, complexidade e legibilidade de textos (MCNAMARA et al., 2014). Foram extraídas 94 características do Coh-Metrix em sua versão para a língua

portuguesa (CAMELO; JUSTINO; MELLO, 2020).

Além do LIWC e Coh-Metrix, foram extraídas 3 características adicionais: número de entidades nomeadas, presença de saudação e presença de expressões positivas. O número de entidades nomeadas pode fornecer informações sobre o nível de detalhe nas mensagens de feedback entregues aos alunos. A presença de saudação e expressão positiva está relacionada aos elogios ou crenças motivacionais que o professor pode fazer a um aluno para atender a boa prática 5 (NICOL; MACFARLANE-DICK, 2006).

8.3.3 Classificação

Nesse trabalho foi utilizado o algoritmo *XGBoost* devido ao seu desempenho em comparação a outros métodos e por ser um algoritmo caixa branca (CHEN; GUESTRIN, 2016b). Foi demonstrado em Xiao et al. (2017), Pan (2018) que o *XGBoost* supera o Random Forest em várias tarefas de classificação. O algoritmo utiliza o aumento de gradiente, que envolve modelos de combinação sequencial (neste caso, árvores de decisão) que preveem os resíduos ou erros de modelos anteriores em cada iteração para melhorar a precisão geral (CHEN; GUESTRIN, 2016b).

Para desenvolver e avaliar um sistema de classificação de texto, primeiro precisa-se definir os conjuntos de dados de treinamento e teste. Neste trabalho, os dados foram divididos em 80% para o treinamento e 20% para o teste, onde foi utilizado a validação cruzada de 10 vezes no treinamento e depois selecionado o modelo que obteve a maior acurácia para ser aplicado nos dados de teste. Como o *XGBoost* é um algoritmo baseado em árvores de decisão, é possível definir o número de árvores geradas para realização do treinamento (parâmetro *ntree*). Nesse trabalho foi utilizado diferentes números de árvores para cada classificador durante o treinamento e o melhor resultado foi aplicado na base de teste.

Como é possível perceber pela Tabela 33, algumas classes possuem um número muito pequeno de mensagens de feedback. Por exemplo, a classe 1 da boa prática 2 (BP2) possui apenas 29 mensagens. Essa quantidade é insuficiente para treinar um modelo de aprendizado de máquina (FRIEDMAN; HASTIE; TIBSHIRANI, 2001). Dessa forma, as classes BP2, BP4, BP7 e FR foram removidas do experimento, formando um novo conjunto com 4 boas práticas (BP1, BP3, BP5 e BP6) e 3 níveis (FT, FP e FS) conforme mostram as Tabelas 34 e 35, respectivamente.

Para avaliar nosso modelo, usamos as medidas de acurácia e Kappa de Cohen. A acurácia

Tabela 34 – Divisão em treino e teste para cada boa prática do dataset.

		Classe 0	Classe 1	Total
BP1	**Treino**	700 (87.25%)	100 (12.75%)	800 (80%)
	Teste	175 (87.25%)	25 (12.75%)	200 (20%)
BP3	**Treino**	554 (69.25%)	246 (30.75%)	800 (80%)
	Teste	138 (69.00%)	62 (31.00%)	200 (20%)
BP5	**Treino**	613 (76.62%)	187 (23.38%)	800 (80%)
	Teste	153 (76.50%)	47 (23.50%)	200 (20%)
BP6	**Treino**	318 (39.75%)	482 (60.25%)	800 (80%)
	Teste	79 (39.50%)	121 (60.50%)	200 (20%)

Fonte: CAVALCANTI et al. (2021)

Tabela 35 – Divisão em treino e teste para cada nível do dataset.

		Classe 0	Classe 1	Total
FT	**Treino**	90 (11.25%)	710 (88.75%)	800 (80%)
	Teste	22 (11.00%)	178 (89.00%)	200 (20%)
FP	**Treino**	399 (49.87%)	401 (51.13%)	800 (80%)
	Teste	100 (50.00%)	100 (50.00%)	200 (20%)
FS	**Treino**	679 (84.87%)	121 (15.13%)	800 (80%)
	Teste	170 (85.00%)	30 (15.00%)	200 (20%)

Fonte: CAVALCANTI et al. (2021)

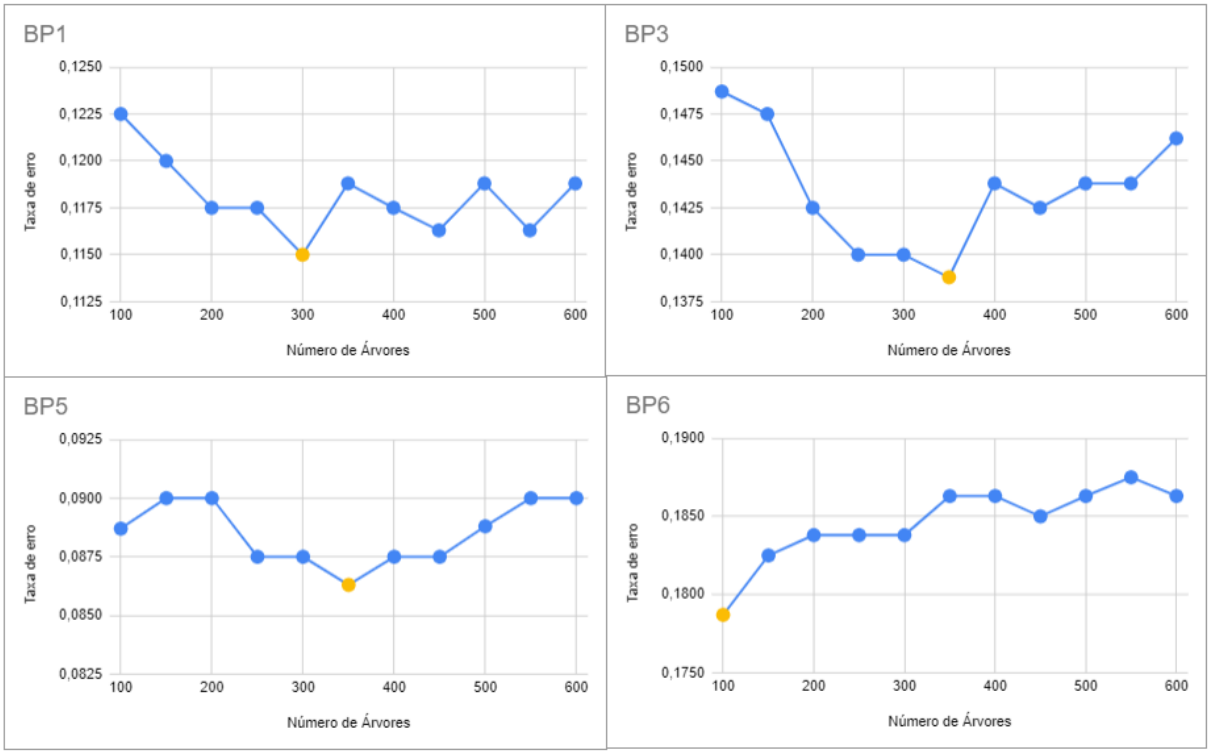
pode ser calculada usando a proporção de amostras corretamente identificadas para o número total de amostras. Além disso, o Kappa mede o acordo entre avaliadores, levando em consideração a possibilidade de acordo por acaso; assim, pode ser usado para medir o nível de concordância entre mensagens codificadas automaticamente e manualmente no conjunto de dados. O Kappa é amplamente usado na literatura de análise de conteúdo e complementa o resultado fornecido pela acurácia (NEUENDORF, 2016; FERREIRA-MELLO et al., 2019).

8.4 RESULTADOS

Foram analisados diferentes números de árvores (parâmetro *ntree*), no intervalo entre 100 e 600, nos conjuntos de treinamento para cada boa prática e nível de feedback utilizando a técnica de validação cruzada 10 vezes. Dessa forma, o modelo com o menor erro no conjunto

de treinamento foi aplicado no conjunto de teste.

Figura 17 – Análise do número de árvores x erro para as boas práticas.



Fonte: CAVALCANTI et al. (2021)

8.4.1 Análise das boas práticas de feedback

A Figura 17 mostra os resultados obtidos no conjunto de treinamento para cada boa prática. Para a boa prática 1, o menor erro foi obtido utilizando 300 árvores com uma acurácia de 0,88 e Kappa de 0,34. Para a boa prática 3 e boa prática 5, o melhor resultado foi obtido utilizando 350 árvores com acurácia de 0,86 e 0,91 e Kappa de 0,65 e 0,75, respectivamente. Já para a boa prática 6, o melhor resultado foi obtido com 100 árvores apresentando uma acurácia de 0,82 e Kappa de 0,62.

Os melhores modelos obtidos no conjunto de treinamento foram aplicados no conjunto de teste. A Tabela 36 mostra as medidas de acurácia e Kappa para cada boa prática. As boas práticas com maior acurácia foram a BP1 e a BP5, com 0,87 e 0,88, respectivamente. Entretanto, a boa prática 1 teve um valor de Kappa menor que as outras boas práticas (semelhante ao resultado obtido no conjunto de treinamento). Esse valor de Kappa pode ser atribuído ao fato de que a boa prática 1 está desbalanceada, ou seja, tem poucos exemplos da

Tabela 36 – Resultados obtidos no conjunto de teste para as boas práticas.

	Acurácia	Kappa
BP1	0,87	0,26
BP3	0,84	0,60
BP5	0,88	0,66
BP6	0,81	0,60

Fonte: CAVALCANTI et al. (2021)

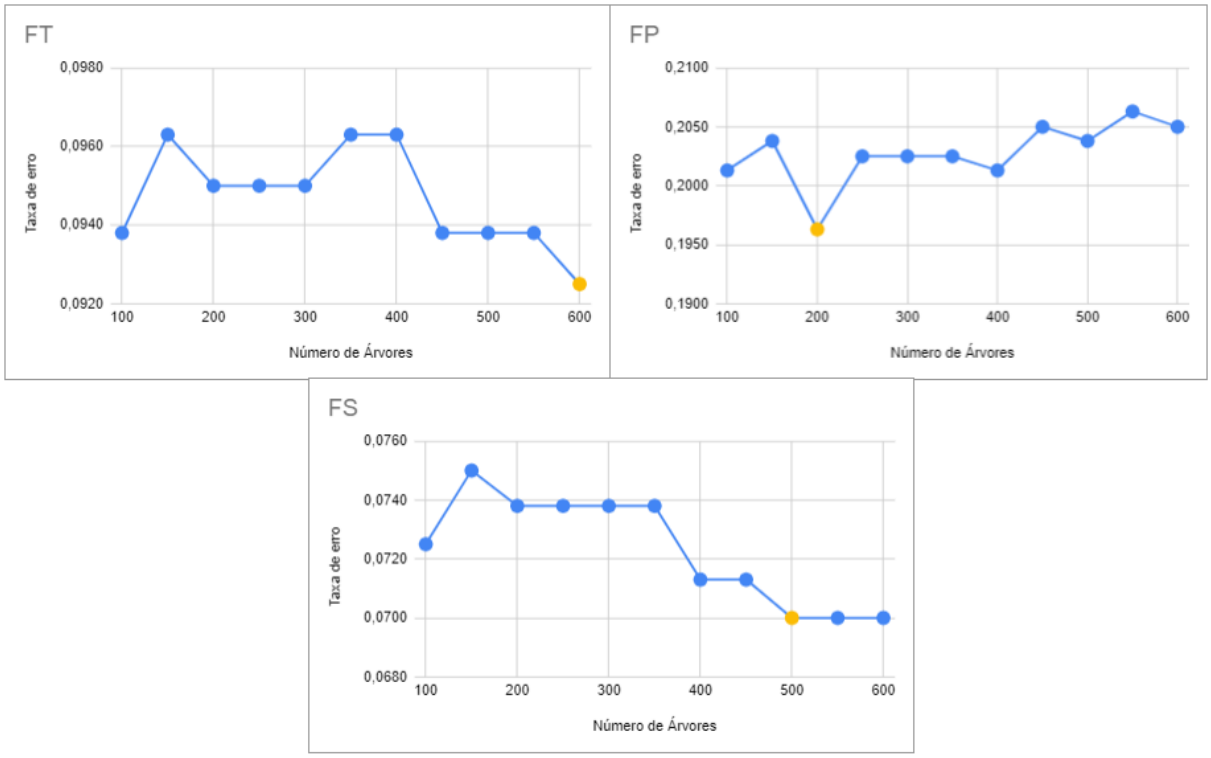
classe 1 (125 exemplos) comparando com o número de exemplos da classe 0 (875 exemplos). Dessa forma, a utilização de mais de 1 métrica de avaliação em algoritmos de aprendizagem de máquina é extremamente importante para entendermos o funcionamento do classificador (MANI; ZHANG, 2003).

8.4.2 Análise dos níveis de feedback

A Figura 18 mostra os resultados obtidos no conjunto de treinamento para cada nível de feedback. O nível de tarefa (FT) obteve o menor erro com 600 árvores, apresentando uma acurácia de 0,90 e Kappa de 0,32. Para o nível de processamento da tarefa (FP) o melhor resultado foi obtido usando 200 árvores com uma acurácia de 0,80 e Kappa de 0,60. Já para o nível pessoal (FS) foram necessárias 500 árvores para obter uma acurácia de 0,93 e Kappa de 0,70.

A Tabela 37 mostra os resultados do conjunto de teste aplicando o melhor modelo obtido no conjunto treinamento. Foram obtidos os valores de 0,89 de acurácia e 0,42 de Kappa para o nível FT, 0,77 de acurácia e 0,55 de Kappa para o nível FP e 0,89 de acurácia e 0,56 de Kappa para o nível FS. A Tabela 37 também compara o nosso resultado com um estudo anterior (CAVALCANTI et al., 2020). O trabalho de Cavalcanti et al. (2020) utilizou o classificador *Random Forest* em conjunto com a aplicação do algoritmo SMOTE para lidar com o desbalanceamento das classes e com um conjunto de 116 características extraídas. Nossa abordagem propõe a utilização do algoritmo XGBoost, que já lida com o problema de desbalanceamento de classes, e um conjunto maior de 161 características extraídas. Em todos os níveis nossa abordagem superou o resultado proposto por Cavalcanti et al. (2020) tanto na medida de acurácia quanto na de Kappa.

Figura 18 – Análise do número de árvores x erro para os níveis.



Fonte: CAVALCANTI et al. (2021)

Tabela 37 – Resultados obtidos no conjunto de teste para os níveis de feedback.

	Abordagem proposta		(CAVALCANTI et al., 2020)	
	Acurácia	Kappa	Acurácia	Kappa
FT	**0,89**	**0,42**	0,75	0,29
FP	**0,77**	**0,55**	0,64	0,28
FS	**0,89**	**0,56**	0,87	0,39

Fonte: CAVALCANTI et al. (2021)

8.5 CONCLUSÃO

Este trabalhou apresentou uma abordagem para classificação de mensagens de feedback em boas práticas e níveis propostos por Nicol e Macfarlane-Dick (2006) e Hattie e Timperley (2007), respectivamente. Foram extraídas 161 características das mensagens de feedback utilizando os recursos LIWC e Coh-Metrix em conjunto com 3 características adicionais. Foi utilizado o algoritmo XGBoost para classificação das mensagens em 4 boas práticas (NICOL; MACFARLANE-DICK, 2006) e três níveis (HATTIE; TIMPERLEY, 2007) diferentes. Foram analisados diferentes números de árvores (parâmetro do classificador XGBoost) no conjunto de

treinamento e o classificador com menor erro foi aplicado no conjunto de testes.

Os resultados dos classificadores mostraram que a combinação dos recursos LIWC e Coh-Metrix e características adicionais foram eficazes na classificação das mensagens de feedback, atingindo acurácias acima de 0,81 nos conjuntos de teste das boas práticas e acima de 0,77 nos conjuntos de testes dos níveis. Além disso, a utilização de um número maior de características extraídas (161) em conjunto com um classificador mais recente (XGBoost) trouxe resultados significativos comparado com um trabalho relacionado. Nossa abordagem teve aumentos expressivos de 18,66% para acurácia e 44,82% para o Kappa no nível FT, 20,31% para acurácia e 96,42% para o Kappa no nível FP e 2,29% para acurácia e 43,58% para o Kappa no nível FS comparados com a abordagem de Cavalcanti et al. (2020). Não foi encontrado na literatura nenhum trabalho que classificasse as boas práticas individualmente. Por esse motivo, nosso trabalho compara apenas os resultados dos níveis de feedback com o resultado obtido por Cavalcanti et al. (2020).

É importante salientar que os classificadores desenvolvidos nesse trabalho podem ser utilizados em conjunto com sistemas/ferramentas de ambientes virtuais de aprendizagem e dessa forma, pode auxiliar o professor na identificação automática de quais boas práticas o seu texto está seguindo e em qual nível de feedback o seu texto está classificado.

8.6 LIMITAÇÃO E TRABALHOS FUTUROS

Uma das limitações deste trabalho é o desbalanceamento das classes. As boas práticas BP2, BP4 e BP7 e o nível FR não tinham exemplos suficientes na classe 1 para treinar um classificador e por isso, foram removidas dos experimentos. Além disso, o conjunto de dados é formado por mensagens de feedback de apenas 2 cursos à distância de uma Universidade pública, e com isso, não generaliza os diferentes tipos de feedback em ambientes virtuais de aprendizagem. Como trabalhos futuros, pretendemos crescer o conjunto de dados com feedbacks de outros cursos e aplicar os classificadores em um ambiente real para auxiliar instrutores. Além disto, técnicas de balanceamento de dados (*over e under sampling*) devem ser analisadas para tentar melhorar a performance dos resultados finais.

9 AUTOMATIC ANALYSIS OF FEEDBACK MESSAGES USING DEEP LEARNING AND EXPLAINABLE ARTIFICIAL INTELLIGENCE

Authors: Anderson Pinheiro Cavalcanti, Rafael Ferreira Mello, Dragan Gašević, Fred Freitas

9.1 INTRODUCTION

Educational feedback is a crucial element in teaching-learning processes. It is through feedback that teachers help students to identify their strengths and weaknesses and help them on the next steps that should be taken. Several studies highlight the significant role that feedback plays in the learning process and show that the quality of this process has a strong relationship with the quality of feedback provided to students (HATTIE; TIMPERLEY, 2007). Feedback directs students to the appropriate type of study or practice and helps them recognize areas of deficiency that can be used to improve learning tactics and strategies (PARIKH; MCREELIS; HODGES, 2001; WEAVER, 2006b; PITT; CARLESS, 2021), as well as improve student performance and avoid evasion (LANGER, 2011; MATCHA et al., 2019b).

However, just the act of sending a feedback message is not enough to have an improvement in the student's performance, worse than that, sending feedback without many criteria can even be harmful (HATTIE; TIMPERLEY, 2007). According to Pitt e Carless (2021), current university feedback practices are widely criticized as not being entirely useful to students. There are several reasons that lead to dissatisfaction with the feedback received, one of them is the one-way transmission of feedback information and written comments that are usually received at the end of an evaluation sequence (WINSTONE; CARLESS, 2019). Various educational theories have been proposed over the years to ensure the positive effect of feedback. For example, Sadler (1989) suggests that feedback needs to provide accurate information related to the learning task or process that bridges a gap between what is understood and what should be understood. Furthermore, recent studies show that effective feedback should be recognized as part of a process, not just a final product (HENDERSON et al., 2019a). This kind of understanding can help teachers engage students in a dialogue beneficial to learning.

Some studies show that the quality of feedback is consistently rated as one of the biggest causes of dissatisfaction for higher education students who report severe deficiencies in the quantity and quality of feedback they receive (BOUD; MOLLOY, 2013; FERGUSON, 2011).

According to Weaver (2006b), even though students recognize the great value of feedback in facilitating learning, they find instructor comments incomprehensible and ineffective.

In online learning, feedback is a crucial factor in the success or failure, as these courses generally have a high dropout rate, and also because students, teachers and tutors are physically separated (YPSILANDIS, 2002). Thus, informative and timely interactions and feedback become even more critical for knowledge building and academic success (JOULANI; GYORGY; SZEPESVÁRI, 2013). Within online learning environments, the instructors usually propose several activities where students submit answers and receive feedback on their progress (COATES; JAMES; BALDWIN, 2005). However, it is challenging for instructors to provide high-quality, informative feedback due to the growing number of students enrolled in online learning.

In order to reduce instructor effort, many researchers have proposed systems to automate student feedback (MARIN et al., 2017; GULWANI; RADIČEK; ZULEGER, 2014; FERREIRA-MELLO et al., 2019; CAVALCANTI et al., 2021). However, their approaches address feedback within a specific context, such as introductory programming courses. An important feature that has been analyzed in recent decades is the quality of feedback (HATTIE; TIMPERLEY, 2007; NICOL; MACFARLANE-DICK, 2006). For example, Cavalcanti et al. (2020) analyzed the feedback content of a distance course based on the feedback levels proposed by Hattie e Timperley (2007).

The increase in available data and increase in computational power has led to the development of more accurate text classification techniques based on deep neural networks (LAI et al., 2015). These methods use a large amount of data, and with that, the classifier tends to achieve better results, even for different contexts. In this direction, eXplainable Artificial Intelligence (XAI) is a recent method that has emerged to explain the results of deep neural network classifiers (SAMEK; MÜLLER, 2019). The combination of these approaches has not yet been explored in the analysis of written feedback and it can bring alternative findings and analysis to the quality of feedback.

In this context, this article proposes a new approach that applies deep learning to automatically classify messages shared by instructors based on known good feedback practices. More specifically, this work presents a set of binary text classifiers using BERT to extract indicators that can lead to improved feedback sent to students. Additionally, this article uses explainable artificial intelligence to clarify the deep learning model outputs.

9.2 THEORETICAL BACKGROUND

9.2.1 Feedback on the learning process

Feedback is an essential component of learning as it directs students to the appropriate type of study or practice. Feedback has been identified as one of the top ten aspects of learning to improve student performance Hattie e Gan (2011), Wisniewski, Zierer e Hattie (2020). This topic is a fundamental part of the student learning process and has been addressed in several studies in the last two decades Nicol e Macfarlane-Dick (2006), Jeria e Villalon (2017). According to the systematic review carried out by Cavalcanti et al. (2021) several studies are actively exploring automated feedback solutions that can allow instructors to identify and employ good feedback practices efficiently and increase the speed of delivering feedback to students.

While much research reports positive effects of feedback, not all feedback is equally effective. For example, in the study by Burke (2009) some factors of student dissatisfaction regarding the feedback received were raised. Among these factors are length (feedback is very brief), polarity (feedback is very negative), and complexity (feedback is very difficult to decipher or understand). Mutch (2003) highlighted the need for more research on how students "receive and respond" to feedback. Weaver (2006b) showed that over 50 % of university students never received any guidance on "how to understand and use feedback", and three quarters of students received no advice on how to understand and use feedback before university .

To improve the quality of feedback provided to students, some works in the literature have proposed models or principles that help to increase the impact of feedback on student learning. The following section presents two well-recognized feedback models in the educational literature that help the teacher understand how feedback should be provided to the student for self-regulation of learning to occur.

9.2.2 Feedback Models

Nicol e Macfarlane-Dick (2006) proposed a conceptual model of self-regulation based on a review of the research literature on formative assessment and feedback. The key idea of the work is to identify how formative assessment and feedback processes can help to promote self-regulation. Through self-regulation, students will become more apt to use external feedback

and their own feedback and thus be able to set their own goals. Seven principles of good feedback practices were defined based on the conceptual model, where the instructor can use them to think about the project and evaluate their own feedback procedures. The seven good feedback practices are (NICOL; MACFARLANE-DICK, 2006):

- **GP 1**: Helps to clarify what good performance is (goals, criteria, expected standards);
- **GP 2**: Facilitates the development of self-assessment (reflection) in learning;
- **GP 3**: Provides high quality information to students about their learning;
- **GP 4**: Encourages dialogue between teachers and colleagues around learning;
- **GP 5**: Stimulates positive motivational beliefs and self-esteem;
- **GP 6**: Provides opportunities to close the gap between current and desired performance;
- **GP 7**: Provides information to teachers that can be used to help shape teaching;

Each feedback principle can be connected to different strategies that the instructor can implement in the classroom. According to Nicol e Macfarlane-Dick (2006), good feedback practices are broadly defined as any strategy or content that can increase students' ability to self-regulate their learning performance.

Hattie e Timperley (2007) looked at several conditions that could maximize the positive effects of feedback on learning, including increasing student awareness of an overall learning goal, progress toward the goal, and subsequent goals needed to achieve the primary goal. Thus, Hattie e Timperley (2007) proposed four perspectives, proposed as levels, that feedback should address in order to improve its effectiveness. The proposed model specifies three types of questions that feedback is designed to answer (Where do I go? How do I go? Where next?). Moreover, each feedback question operates at four levels: task feedback (FT), feedback on the task processing (FP), feedback on self-regulation (FR) and feedback on the self as a person (FS).

The level to which the feedback belongs has an influence on its effectiveness, which is why feedback that focuses on the qualities of the work carried out and on the process or strategies used gives greater help to the student. Feedback that guides the student towards the development of self-regulatory strategies also tends to be more effective. For example, the meta-analysis performed by Wisniewski, Zierer e Hattie (2020) revealed that feedback has a

greater impact on cognitive and motor skills outcomes than on motivational and behavioral outcomes. The comments focused on the students' personal characteristics are too vague and do not lead the student to focus on their learning (BROOKHART, 2017).

9.2.3 Automated Analysis of Feedback Content

Automated analysis of feedback content has been addressed in several studies recently. This type of automated analysis involves using natural language processing to extract features from textual feedback and can provide significant contributions to the quality of feedback in education. Cavalcanti et al. (2019) proposed a content analysis of the feedback text provided by instructors based on good feedback practices proposed by Nicol e Macfarlane-Dick (2006). The authors focused on analyzing the quality of feedback extracted from assessments collected in an online course offered at a Brazilian higher education institution. The authors used a random forest classifier and reached 0.75 and 0.20 for accuracy and Cohen's kappa, respectively. One of the problems of this work is the division of feedback texts into binary classes (class 0: there is no good practice; class 1: there is at least 1 good practice). This type of classification may not be fully effective, as the teacher only has information on whether or not the text has good practice. In order for the teacher to be able to provide quality feedback to students, it is necessary for them to know which of the seven good practices their text is following or not Nicol e Macfarlane-Dick (2006).

Other studies proposed automated approaches to examine feedback provided by instructors according to feedback levels proposed by Hattie e Timperley (2007). Osakwe et al. (2022) used the XGBoost (eXtreme Gradient Boosting) classifier and a dataset with feedback texts written in English. The authors extracted 166 features from the texts using LIWC (Linguistic Inquiry and Word Count) (TAUSCZIK; PENNEBAKER, 2010) and Coh-metrix (MCNAMARA et al., 2014) resources. The authors reached accuracy values of 0.87, 0.82, and 0.69 on self, task, and process levels respectively. Another objective of this article was to identify the prominent textual characteristics of the feedback components, being able to corroborate the findings of educational research on the theory of feedback. Furthermore, the transferability of features between the Portuguese and English languages was analyzed, indicating a low generalization between the two sets. Alonso et al. (2022) proposes an approach to the multi-label classification of feedback according to Hattie and Timperley's feedback levels, incorporating a hyperparameter adjustment step. The authors carry out experiments using the support vector machines, ran-

dom forest and k-nearest neighbors algorithms. The authors use feedback texts generated by a teacher to the activities sent by students in online courses on the Blackboard platform at the task, process, regulation and praise levels Hattie e Timperley (2007). The work of Cavalcanti et al. (2020) aimed to evaluate binary classifiers for the feedback levels proposed by Hattie e Timperley (2007). The authors extracted 116 features from the text using LIWC, Coh-metrix, and additional features such as the number of Named Entities, presence of praise, presence of greeting, and sentiment polarity. The authors achieved results outcomes up to 87% of accuracy and 0.39 of Cohen's kappa using a random forest classifier. The work also provides insights into the most influential feedback characteristics that predict feedback quality. However, one of the limitations of this work is related to the imbalance of the database.

Class imbalance is a typical problem in text classification tasks that can be solved by applying some popular class balancing algorithms such as SMOTE (CHAWLA et al., 2002), NearMiss (MANI; ZHANG, 2003), or Tomek Links (TOMEK, 1976). For example, to deal with the problem of class imbalance in automatic feedback classification, Osakwe et al. (2022) checks the performance of the classifier using the combination of several balancing algorithms through a genetic algorithm (approach presented by Barbosa et al. (2020)) and comparing it with the performance of the classifier only using the SMOTE algorithm. The authors found that classification performance using the SMOTE balancing algorithm had the best results and conclude that class imbalances can have a negative impact on model efficiency.

The aforementioned works use traditional machine learning algorithms, following the process of extracting various features from the text and applying them to the classifier. To the best of our knowledge, we did not find related works that performed written feedback analysis using deep learning algorithms. According to the literature, these algorithms have brought significant results in the educational area, such as: analysis of cognitive presence (HU; MELLO; GAŠEVIĆ, 2021), sentiment analysis of students' feedback (KASTRATI et al., 2021) or reflection in writing (WULFF et al., 2022; ULLMANN, 2019).

9.2.4 Deep Learning and Explainable Artificial Intelligence

Deep learning is a subfield of machine learning that uses successive layers for accurate representations or decision making. Deep learning algorithms gained notoriety when they showed excellent results in computer vision tasks (RUSSAKOVSKY et al., 2015) and also for natural language processing (NLP) tasks (YOUNG et al., 2018). These algorithms try to simulate the same

learning process as the human brain using a large number of connections generated in deep neural networks.

Deep learning has been playing an increasing role in NLP tasks related to topic modeling/classification (YOUNG et al., 2018). More and more sophisticated deep neural network architectures needed more training data, due to the amount of parameters in the models that increased substantially. To deal with the excessive requirements, transfer learning methods have been developed where researchers make use of large deep learning models previously trained in massive linguistic corpora such as the Internet dump or Wikipedia. A pre-trained language model can be defined as a black box that has prior knowledge about natural language and can be applied and tweaked to solve various NLP problems. The pre-training process uses unlabeled data to learn the initial parameters of a neural network model. An example of such a model is BERT, which is a deeply bidirectional language model trained on very large datasets (i.e., Books corpus and Wikipedia) based on contextual representations (DEVLIN et al., 2018). The BERT model can be fine-tuned using a dense neural network layer for different classification tasks.

An example is the work of André et al. (2021) that evaluated the performance of random forest-based algorithms and the BERT deep learning linguistic model for automatic detection of social presence in online discussions. The authors compare the approach with traditional text mining and linguistic features like LIWC and Coh-metrix with the approach using the fine-tuned BERT language model for social presence classification. The results demonstrate that the XGBoost and AdaBoost (Adaptive Boosting) algorithms outperformed the BERT model in online discussion messages.

Deep learning algorithms show promising results in many NLP tasks, however, one of the main barriers that artificial intelligence encounters is the inability to explain the prediction output or even the functioning of these algorithms (ARRIETA et al., 2020). The field of explainable artificial intelligence (XAI) aims to create machine learning techniques that can clarify to the human audience how it works and thus properly trust AI methods and implementations (GUNNING, 2017). XAI can complement the adoption of deep learning algorithms as it unpacks model decisions.

Khosravi et al. (2022) presents the XAI in Education (XAI-ED) framework, which is based on the areas of AI, Human-Computer Interaction and Cognitive and Learning Sciences. The XAI-ED characterizes the nature of XAI in Education in terms of questions about six main aspects: the people involved (stakeholders) and the benefits for each group; how to deliver

the explanation; the model classes widely used in education; the human-centered design of AI and interfaces to support explanation; and the potential pitfalls of providing explanations. Therefore, this article aims to use XAI to clarify deep learning algorithms in automatic feedback evaluation.

9.3 RESEARCH QUESTIONS

As emphasized in the previous sections, it is very important to provide automated methods to assist in producing good quality feedback. This work aims to analyze the use of BERT to classify feedback based on the models proposed by Nicol e Macfarlane-Dick (2006) and Hattie e Timperley (2007). It is possible to find works that perform content analysis on textual feedback (OSAKWE et al., 2022; CAVALCANTI et al., 2020), these works use traditional Machine Learning algorithms to extract the features of the text. As mentioned earlier, written feedback analysis using deep learning algorithms has not been investigated in depth. According to the literature, these algorithms have brought significant results in the educational area. For this reason, this work uses a deep learning algorithm to classify the feedback texts, raising our first research question:

RQ1: How do deep learning algorithms perform in the context of automatic classification of educational feedback?

Furthermore, the class imbalance is an important factor that can affect the performance of machine learning algorithms (CHAWLA et al., 2002; BARBOSA et al., 2020; OSAKWE et al., 2022). Thus, this work also proposes to analyze the performance of the classifier based on a data augmentation approach to balance the classes, which leads us to our second research question:

RQ2: Could class imbalance affect the performance of the deep learning algorithm in automatically classifying educational feedback?

In machine learning algorithms applied to the education area, it is important not only to provide an accurate classification of the feedback messages, but also to explain why the messages were classified in one of the feedback models (FERREIRA-MELLO et al., 2019). Traditionally, deep learning methods are categorized as black box algorithms. The eXplainable Artificial Intelligence (XAI) method (SAMEK; MÜLLER, 2019; MILLER, 2019) helps to understand the deep learning model with visual representations of the output. This is a new approach never adopted in this context. Thus, our last research question is:

What are the important features revealed of deep learning with XAI to classify feedback messages?

9.4 METHOD

9.4.1 Data and Course Design

The dataset used in this work contains individual feedback provided by instructors through the activity submission tool in the Learning Management System (LMS) for two courses: biology and literature (CAVALCANTI et al., 2020). This dataset was generated from an LMS used in online courses at a public university in Brazil and consists of 1,000 feedback messages.

Each feedback message was classified by experts who analyzed the text of the feedback messages and determined whether it belonged to the seven good practices proposed by Nicol e Macfarlane-Dick (2006) or the four levels proposed by Hattie e Timperley (2007). In summary, for each good practice and feedback level, the text was evaluated and received **label 0** if the text did not belong to that good practice or that level, or **label 1** if the text belonged to that good practice or level. The experts followed a document with information and examples of each feedback model. The inter-rater agreement was moderate for the seven good practices, reaching 76.7% and Cohen's kappa of 0.43. For the four levels, the inter-rater agreement had a percentage of 72.2%, and Cohen's kappa was 0.44. Another two experts who did not participate in the first coding stage resolved the divergent cases, 23.3% for the seven good practices and 27.8% for the four levels. Tables 19 and 20 shows the division of the dataset, in terms of the number of good practices (GP) or levels per class, respectively.

Tabela 38 – Dataset division for the 7 good practices.

	Good Practice (GP)						
	GP1	**GP2**	**GP3**	**GP4**	**GP5**	**GP6**	**GP7**
class 0	875	971	692	987	766	397	1000
class 1	125	29	308	13	234	603	0
Total	1000	1000	1000	1000	1000	1000	1000

Fonte: Elaborada pelo autor (2022)

Tables 19 and 20 show that the GP2 and GP4 good practices and the FR level had a very small amount of examples for class 1 out of a total of 1,000 feedback messages, an amount

Tabela 39 – Dataset division for the 4 levels.

	Level			
	FT	**FP**	**FR**	**FS**
class 0	112	499	992	849
class 1	888	501	8	151
Total	1000	1000	1000	1000

Fonte: Elaborada pelo autor (2022)

that is not enough to fit a machine learning model (FRIEDMAN; HASTIE; TIBSHIRANI, 2001). Similarly, good practice GP7 had no example of feedback for class 1, as this good practice aims to provide feedback from the student to the teacher to shape teaching, however, the dataset only has texts in one direction: teacher feedback to the student. Thus, this study used the following indicators in the analysis: good practices GP1, GP3, GP5, and GP6 and levels FT, FP, and FS.

9.4.2 Model Selection and Evaluation

Some recent work has used traditional machine learning algorithms to automate feedback classification (OSAKWE et al., 2022; CAVALCANTI et al., 2020). However, these approaches require frameworks or methods to extract features from the text and then apply the machine learning algorithm. In this work, we aim to evaluate the use of an approach that uses Deep Learning to classify feedback. For this, the BERT model was used, which brought significant results in several natural language processing tasks (DEVLIN et al., 2018).

BERT is a contextual language model consisting of a deep neural network with bidirectional processing. BERT generates embeddings that vary according to the textual context of each occurrence of a lexicon, which allows capturing variations of meaning (DEVLIN et al., 2018). BERT is pre-trained for Next Sentence Prediction (NSP) task. Therefore, BERT is designed to pre-train deep bi-directional representations of unlabeled text. BERT allows fine-tuning with the addition of just one output layer, in order to optimize its performance in various NLP tasks. This work uses *BERTimbau* (SOUZA; NOGUEIRA; LOTUFO, 2020), a pre-trained BERT model for the Portuguese language, and runs a fine-tuning using the feedback dataset.

To evaluate our model, we used the metrics Cohen's Kappa and accuracy. Accuracy can be

calculated using the proportion of correctly identified samples to the total number of samples. In addition, Kappa measures agreement between raters, taking into account the possibility of agreement by chance; thus, it can be used to measure the level of agreement between automatically and manually encoded messages in the dataset. Kappa is widely used in the content analysis literature and complements the result provided by the accuracy (NEUENDORF, 2016; FERREIRA-MELLO et al., 2019).

BERT fine-tuning in classification experiments utilizes the learning rate scheduler without warm-up, followed by linear decay of the learning rate across the training steps. For this, we use the AdamW optimizer from the Hugginface[1] implementation of BERT with the default values of $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The fine-tuning ran for 10 epochs, with validations every 100 steps and a learning rate of $5 * 10^{-}5$. The best model was saved and used to evaluate the test set.

Therefore, our answer to RQ1 we is the comparison of BERT with traditional machine learning algorithms. To answer the RQ2 we used the described approach to balance the dataset and compared with the previous results using the unbalanced data.

9.4.3 Explainable Artificial Intelligence

The XAI collectively refers to methods that can explore the interpretability of a given decision-making process, such as traditional and modern ML models. Enormous potential is shown from this branch of AI study that can unpack the modern “black box"ML model. LIME is a powerful tool because it provides accessibility and simplicity (RIBEIRO; SINGH; GUESTRIN, 2016). LIME inherits the basic idea of model agnosticism to explain any supervised learning model, treating it as a separate “black box". LIME provides a local explanation by weighting adjacent observations. Local explanations mean that LIME provides locally faithful explanations within the surrounding observations of the sample being explained. Due to computational limitations the LIME explainer was trained using 10 features with 50 samples. LIME produced data resource importance values for each feedback message.

The XAI was used to answer the RQ3. We adopted the XAI algorithm provided by the Python library called Lime to better understand how the deep learning algorithm classifies the text and what are the important features revealed in the classification.

[1] https://huggingface.co/

9.4.4 Data processing

To develop and evaluate a text classification system, the first step is the definition of training and testing sets. In this work, the data were divided into 70% for training, 10% for validation and 20% for testing, a necessary step taken in machine learning to avoid overestimating model performance (which can occur if model accuracy is estimated on the same data as model parameters has been learned) (FRIEDMAN; HASTIE; TIBSHIRANI, 2001). Thus, the split data set included 700, 100 and 200 instances for the training, validation and test sets, respectively (Table 40).

Tabela 40 – Dataset division in train, validation and test.

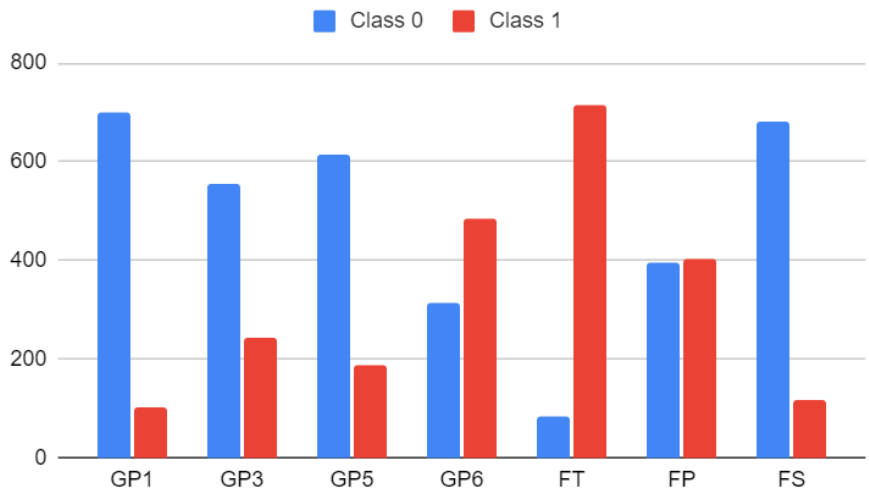
		Class 0	**Class 1**	**Total**
GP1	**Train**	613 (87.57%)	87 (12.43%)	700 (100%)
	Validation	86 (86.00%)	14 (14.00%)	100 (100%)
	Test	176 (88.00%)	24 (12.00%)	200 (100%)
GP3	**Train**	491 (70.14%)	209 (29.865%)	700 (100%)
	Validation	66 (66.00%)	34 (34.00%)	100 (100%)
	Test	135 (67.50%)	65 (32.50%)	200 (100%)
GP5	**Train**	536 (76.57%)	164 (23.43%)	700 (100%)
	Validation	77 (77.00%)	23 (23.00%)	100 (100%)
	Test	153 (76.50%)	47 (23.50%)	200 (100%)
GP6	**Train**	273 (39.00%)	427 (61.00%)	700 (100%)
	Validation	41 (41.00%)	59 (59.00%)	100 (100%)
	Test	83 (41.50%)	117 (58.50%)	200 (100%)
FT	**Train**	72 (10.29%)	628 (89.71%)	700 (100%)
	Validation	12 (12.00%)	88 (88.00%)	100 (100%)
	Test	28 (14.00%)	172 (86.00%)	200 (100%)
FP	**Train**	340 (48.57%)	360 (51.43%)	700 (100%)
	Validation	57 (57.00%)	43 (43.00%)	100 (100%)
	Test	102 (51.00%)	98 (49.00%)	200 (100%)
FS	**Train**	597 (85.29%)	103 (14.71%)	700 (100%)
	Validation	85 (85.00%)	15 (15.00%)	100 (100%)
	Test	167 (83.50%)	33 (16.50%)	200 (100%)

Fonte: Elaborada pelo autor (2022)

As Table 40 shows many classes had an imbalance number of instances between the

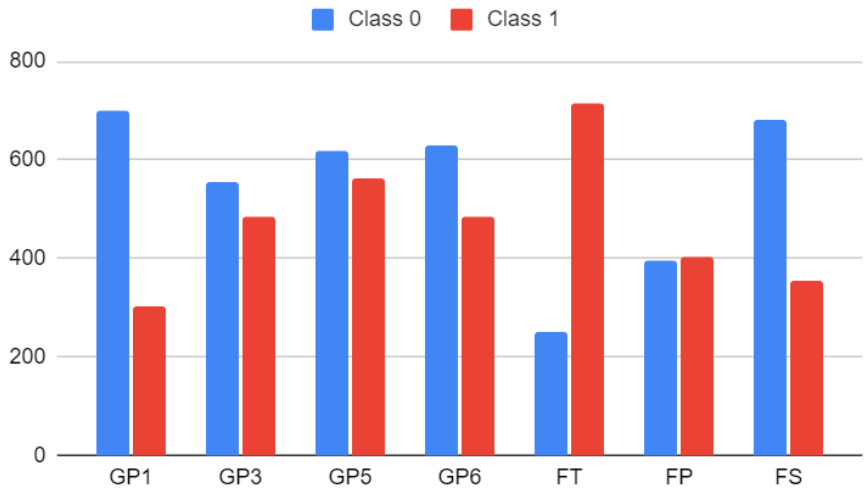
number of examples of one class in relation to the other. To deal with the imbalance problem in this work, we created a simple data augmentation approach, based on the work of Wei e Zou (2019), using the *nlpaug* Python library (MA, 2019) and the *BERTimbau* model for Portuguese (SOUZA; NOGUEIRA; LOTUFO, 2020). This data balancing approach differs from those used by related works (OSAKWE et al., 2022; BARBOSA et al., 2020) because it deals with pure text in natural language and not with extracted features. In this method, the difference between the two classes is calculated and, based on this value, new sentences were created in two different ways: (1) replacing words in the sentence with synonyms; (2) inserting new words into the text. Thus, a minority class could be increased by up to 3 times its initial number. Figures 19 and 20 presents the dataset division before and after apply the data augmentation in the training data. It is important to note that the data augmentation technique was applied to the training set only, because we need to use the same test set to compare the performance of the model before and after applying the data augmentation and make sure to avoid data contamination Farrow, Moore e Gašević (2019). It is also important to note that the FP level is balanced and it was not necessary to increase the data for this level.

Figura 19 – Division of the training set without data augmentation.



Fonte: Elaborada pelo autor (2022)

Figura 20 – Division of the training set with data augmentation



Fonte: Elaborada pelo autor (2023)

9.5 RESULTS

9.5.1 Performance of deep learning algorithms – RQ1

Table 41 compares our results obtained using *BERTimbau* with the result obtained by Cavalcanti et al. (2021) which used the LIWC and Coh-metrix frameworks to extract text features and the XGBoost algorithm to classify feedback messages in the Portuguese language.

Tabela 41 – Comparison of our approach with a related work approach.

	Our Approach		(CAVALCANTI et al., 2021)	
	Accuracy	**Kappa**	**Accuracy**	**Kappa**
GP1	**0.88** (+1.14%)	0.25	0.87	**0.26** (+4.00%)
GP3	**0.87** (+3.5%)	**0.70** (+16.66%)	0.84	0.60
GP5	**0.90** (+2.27%)	**0.73** (+10.60%)	0.88	0.66
GP6	**0.83** (+2.46%)	**0.65** (+8.33%)	0.81	0.60
FT	**0.90** (+1.12%)	**0.57** (+35.71%)	0.89	0.42
FP	**0.77** (=)	0.54	**0.77** (=)	**0.55** (+1.85%)
FS	**0.93** (+4.5%)	**0.76** (+35.71%)	0.89	0.56

Fonte: Elaborada pelo autor (2022)

Table 41 shows that the approach using deep learning achieved the best results for all levels

and best practices, with the exception of the good practice GP1 and the FP level which had a difference of 4.00% and 1.85% in the kappa measure, respectively. The accuracy measure had slight increases, especially for the GP3 model with 3.5% and for the FS level with 4.5%. As for the kappa measure, the deep learning approach brought very expressive results, highlighting the FT and FS levels, both with 35.71%. The FP level had a result very close to Cavalcanti et al. (2021), with the same accuracy value (0.77) and a difference of 0.01 kappa between the two approaches.

9.5.2 Evaluation of data augmentation approach – RQ2

As mentioned in Section 9.4.2, a fine-tuning was performed on the Portuguese BERT model using the feedback dataset. The fine-tuning ran for 10 epochs, with validations every 100 steps for each good feedback practice or feedback level. The best result obtained during fine-tuning is saved and the model is evaluated in the test set. Table 42 shows the accuracy and kappa values in the training and the number of steps that obtained the best result.

Tabela 42 – Results of the model on the training set.

	Original data			**Augmented data**		
	Step	**Accuracy**	**Kappa**	**Step**	**Accuracy**	**Kappa**
GP1	1,200	0.92	0.59	2,100	**0.94**	**0.86**
GP3	2,800	**0.89**	0.75	2,900	0.88	**0.77**
GP5	700	0.97	0.91	1,100	**0.98**	**0.96**
GP6	1,700	0.90	0.79	3,400	0.90	0.79
FT	1,300	0.92	0.57	1,400	**0.95**	**0.88**
FP	700	**0.86**	**0.71**	-	-	-
FS	400	0.96	0.85	2,500	**0.98**	**0.95**

Fonte: Elaborada pelo autor (2022)

As can be seen in Table 42, in many cases the model was able to obtain the highest kappa and accuracy values in the augmented dataset, with the exception of good practice GP3 which had a 0.01 difference in accuracy between normal and augmented data, and also the good practice GP6 which obtained the same result in both data sets. It is also possible to notice that the kappa measure had significant increases, reaching an increase of 45.76% in the GP1

good practice and 54.38% in the FT level. As emphasized in the previous section, the FP level is balanced and therefore it was not necessary to apply data augmentation.

The models obtained in the training set were applied to the test set, and the measurements of accuracy and kappa were calculated. Table 43 shows the results of models trained with and without data augmentation. Similar to the results in the training set, models trained with augmented data had better accuracy and kappa compared to models trained with normal data, with the exception of good practice GP3 which had a 6.06% increase in kappa and 1.1% in accuracy and the FT level with a 5.5% increase in kappa.

Tabela 43 – Results of the model on the testing set.

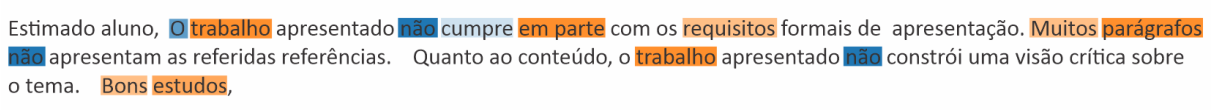
	Original data		**Augmented data**	
	Accuracy	**Kappa**	**Accuracy**	**Kappa**
GP1	0.88	0.25	**0.89**	**0.49**
GP3	**0.87**	**0.70**	0.86	0.66
GP5	0.90	0.73	**0.91**	**0.76**
GP6	0.83	0.65	**0.86**	**0.70**
FT	**0.90**	**0.57**	0.90	0.54
FP	**0.77**	**0.54**	-	-
FS	0.93	0.76	**0.93**	**0.77**

Fonte: Elaborada pelo autor (2022)

9.5.3 Important features revealed from the explainable AI — RQ3

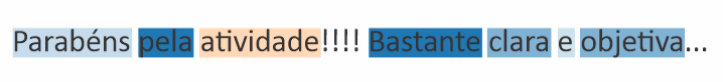
The results of XAI heatmaps were generated from the experiments of the best-performing classifiers (models GP3 and FS). These two classifiers can demonstrate purer comparative results of the word-level predictive indicators of feedback levels and good feedback practices. We provide four messages classified using the GP3 model (Figures 21 and 22) and the FS model (Figures 23 and 24) taken from the test set, as shown in the XAI views. The images have two colors: (1) blue – means that the word is more likely to belong to class 0; (2) orange – means the word is more likely to belong to class 1. As we worked with binary classifiers: orange belongs to feedback level/good practice, and blue does not belong to feedback level/good practice.

Figura 21 – The heatmap of messages in the classification model identifying GP3 practice - Message belongs to the GP3 model.



Fonte: Elaborada pelo autor (2022)

Figura 22 – The heatmap of messages in the classification model identifying GP3 practice - Message does not belong to the GP3 model.



Fonte: Elaborada pelo autor (2022)

The GP3 level aims to provide students with high-quality information. As the example text in Figure 21 shows, there are more orange words than blue ones, meaning that the model classified this text as belonging to the GP3 good practice (class 1). In this example, the teacher critiqued the students' work, highlighting what they did wrong and what they should have done. In this way, it is possible to see important words that made the model classify this text as belonging to the GP3 good practice: "work", "requirements", "partially", "paragraphs", "many", "studies", and "good". These words are strongly related to the goal of GP3 good practice which is to provide high-quality information using constructive criticism or corrective advice and limiting the text so that it is truly understood by the student.

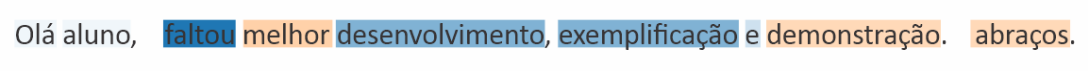
The text shown in Figure 22 has more blue words indicating that the model classified the text as not belonging to the GP3 good practice (class 0). In this example, the teacher only praised the student's work and did not bring additional information or did not show the details that made the student achieve the objective of the activity. The only word likely to belong to class 1 is the word "activity" which could be related to the teacher always using this word to detail and provide corrective advice about student activity. The other words have a strong probability of not belonging to the GP3 class, as they are just words of praise for the activity.

Figura 23 – The heatmap of messages in the classification model for self-level of feedback (FS) - Message belongs to the FS model.

Fonte: Elaborada pelo autor (2022)

Figures 23 and 24 shows two examples for the FS level. This level is intended to provide

Figura 24 – The heatmap of messages in the classification model for self-level of feedback (FS) - Message does not belong to the FS model.



Fonte: Elaborada pelo autor (2022)

praise to the student. In the example of the Figure 23 we have a text from the teacher praising the good work done by the student. Almost all the words in the text are orange (class 1) and there are no blue words (class 0). This text has a very strong relationship with the objective of the FS level and shows words of praise such as: "excellent", "fabulous", "congratulations", and "good". In Figure 24, we have an example of a text that was classified as not belonging to the FS level. The words "needed", "development" and "exemplification" have a high probability of belonging to class 0. In this example, the teacher informed the student what it would take for the student to reach the activity objective and did not provide any praise to the student.

9.6 DISCUSSION

Our first research question aimed to verify the performance of deep learning algorithms for classifying feedback texts based on Hattie e Timperley (2007) and Nicol e Macfarlane-Dick (2006) models. Table 41 compares our best results obtained in the test set with the results obtained by Cavalcanti et al. (2021) which uses the XGBoost algorithm to classify the texts. Our approach had significantly better results in 6 out of 7 trained models. The work of Cavalcanti et al. (2021) extracts features from texts using LIWC and Coh-metrix frameworks, obtaining a total of 161 features. This feature extraction process becomes computationally expensive and many of the features may have little or no importance for the trained classifier, making it necessary to analyze the most important features for this type of classifier (BARBOSA et al., 2020; CAVALCANTI et al., 2020; OSAKWE et al., 2022). On the other hand, deep learning algorithms use the concept of word embeddings which are vectors learned during training and are able to capture semantic relationships between texts based on a defined context window (MIKOLOV et al., 2013). The results obtained with our deep learning model proved to be promising for classifying feedback texts.

Our second research question aimed to verify whether class imbalance could affect the performance of the classifier. The results revealed that the data augmentation approach affected the performance of the classifiers in most models, with emphasis on the GP1 and GP6 models

which had improvements of 96% and 7.7% in the Kappa measure, respectively. However, it was not able to generalize to all models, where the GP3 and FT models had decreases in the kappa measure of 5.7% and 5.3%, respectively. These results are aligned with some findings in related work. For example, the work by Osakwe et al. (2022) analyzed different approaches to balancing data using SMOTE and a genetic algorithm that combines balancing algorithms to classify feedback levels Hattie e Timperley (2007) in English language texts. The authors show that FS level obtained the best result without applying data balancing, the FT level obtained the best result with SMOTE and the FR and FP levels obtained the best result by applying the genetic algorithm. In the same direction, Cavalcanti et al. (2020) used SMOTE to balance the FT and FS levels and had improvements of 29.31% and 2.35% in accuracy, respectively. To the best of our knowledge, no previous work has done data augmentation using BERT for the task of analyzing feedback.

Our third research question aimed to analyze the important features revealed of explainable AI. We have provided two examples for the GP3 and FS models. Through the heatmaps, it was possible to establish relationships between the most important words and the level/good practice of feedback. This type of analysis can help us better understand how deep learning algorithms rank. This approach shows us the importance of the feature at the word level. Some related works sought to find expressions corresponding to definitions based on Coh-Metrix and LIWC tools for each message (CAVALCANTI et al., 2020; BARBOSA et al., 2020). This approach is still time-consuming and not transparent. In comparison, XAI's word-level explanations are noticeable and understandable, even for short texts. However, these two explainable methods (feature importance and XAI's word-level explanations) have some similarities. For example, Cavalcanti et al. (2020) performed an analysis of the most important features of Coh-metrix and LIWC for FS level classification. In this analysis, the author presents characteristics related to the number of praises, the number of words of feeling, and words with positive emotions, which are often used by the instructor to try to change a student's state of mind in relation to a task, or basically to praise perform well in an activity. Likewise, our approach with XAI shows some words of praise or positive emotions like: "excellent", "fabulous", "congratulations", and "good". Another finding of Cavalcanti et al. (2020) is that the FS level had higher Mean Decrease Gini (MDG) values than the other levels and this leads us to a classifier capable of differentiating well between a sentence belonging to the level or not. This was also verified in our example of Figure 23, where almost all the words were colored orange, that is, the use of the words of praise caused the model to classify the text with a high probability of belonging

to the level FS.

9.7 FINAL REMARKS

This study introduced a deep learning model to automatically classify online feedback messages and also adopted an explainable artificial intelligence visualization method to provide insights into indicators for different feedback levels and good feedback practices. In future work, the trained AI models will be used in a tool that analyzes which good practices and levels the feedback text belongs to, thus helping the teacher to provide quality feedback to the student.

While previous work has focused on providing text-based structural information (lexical diversity and text readability), the proposed XAI view yields insights into relevant content for providing feedback. However, the combination of both approaches can bring a richer analysis to instructors. Despite the promising results, the present study has some limitations that can be identified, such as the small number of examples of messages used, which may have limited the potential of the deep learning models used. Then the data came from only two courses (undergraduate biology and literature undergraduate courses), so the results may have been affected by low content diversification and this analysis is not completely representative of the types of feedback in online environments.

10 A COMPARATIVE ANALYSIS BETWEEN GOOD FEEDBACK DESCRIPTORS ON ONLINE COURSES

Authors: Anderson Pinheiro Cavalcanti, Vitor Rolim, Rafael Ferreira Mello, Fred Freitas

10.1 INTRODUCTION

Feedback is a mechanism that plays a critical role in the instructor-student relationship. This mechanism allows students to improve their skills and competencies and allows instructors to tailor their methods and content based on students' learning needs. Moreover, instructors use the feedback messages to suggest insights to improve students' performance by enhancing interaction, assessments and answering questions (LANGER, 2011; MATCHA et al., 2019b). According to Sadler (1989), feedback needs to provide precise information about the learning task or process that fills a gap between what is understood and what should be understood. Several studies have shown that useful feedback brings benefits to learning (HATTIE; TIMPERLEY, 2007; NICOL; MACFARLANE-DICK, 2006; PARIKH; MCREELIS; HODGES, 2001).

Hattie e Gan (2011) highlighted that feedback is one of the top ten aspects of learning to enhance student performance and learning experience. According to the study by Sadler (2010), feedback only works when applied to fill the gap between the current and desired performance. If the information shared by feedback is not or cannot be processed by the student to produce improvements, it will not affect learning. Furthermore, inadequate feedback can negatively impact students, creating distrust in the feedback process and in the teaching, which can impair students' self-efficacy and motivation and even result in dropout (BOUD; FALCHIKOV, 2007).

Some works in the literature present feedback descriptors for the instructor to provide good feedback. Nicol e Macfarlane-Dick (2006) propose a self-regulation model and identify seven principles of good feedback practices. Each principle has specific strategies the teacher can use to facilitate student self-regulation. On the other hand, Hattie e Timperley (2007) proposed a feedback model to identify the circumstances in which feedback has the most significant impact. This model has four levels where feedback can be provided (task, processing task, self-regulation, and personal). The key idea of this work is to ensure that feedback is provided at the appropriate level, making it effective feedback.

Recently, several studies have proposed machine learning techniques designed to identify the

good practices automatically (NICOL; MACFARLANE-DICK, 2006) and levels (HATTIE; TIMPERLEY, 2007) from feedback shared through educational technology. In general, these techniques adopted random forest classifiers in combination with features related to text structure to categorize feedback messages according to the good practices and levels of feedback (CAVALCANTI et al., 2021; OSAKWE et al., 2021). Although these studies indicate a promising direction to create an automatic tool to support instructors in creating better quality feedback messages, they do not analyze the relationship between the good practices (NICOL; MACFARLANE-DICK, 2006) and levels (HATTIE; TIMPERLEY, 2007). Analyzing relationships between these two models is essential as it can inform the assessment of feedback quality; that is, whether it is sufficient to use only one of these two well-known feedback models or their combined use is preferred.

Therefore, this paper proposes a study that adopts epistemic network analysis (ENA) to perform a comparative analysis between the feedback descriptors proposed by Nicol e Macfarlane-Dick (2006) and Hattie e Timperley (2007). ENA provides techniques for quantitative analysis of qualitative aspects of learning and teaching (SHAFFER, 2017), as it has been shown through the growing adoption of ENA in learning sciences, learning analytics, and educational research (SHAFFER et al., 2009; ROLIM et al., 2019). With this analysis, we intend to produce new theoretical insights into providing good quality feedback in technology-enhanced learning.

10.2 BACKGROUND

Previous works have proposed models or principles that help increase the impact of feedback on student learning. For example, Hattie e Timperley (2007) proposed a model to assist the construction of effective feedback. This model identifies three main questions that effective feedback must answer: "*Where am I going?*", "*How am I going?*", "*Where to go next?*". Evans (2013) performs an extensive literature review and proposes a feedback scenario that identifies specific areas for future research on evaluation feedback in higher education. Nicol e Macfarlane-Dick (2006) presents seven principles of good feedback practices to assist teaching staff. The works of Hattie e Timperley (2007) and Nicol e Macfarlane-Dick (2006) have well-formulated models, are widely used in the literature, and are detailed below.

10.2.1 Seven Good Feedback Practices

Nicol e Macfarlane-Dick (2006) proposed a conceptual model of self-regulation based on a review of the research literature on formative assessment and feedback. The key idea of the work is to identify how formative assessment and feedback processes can help promote self-regulation. Based on the conceptual model, seven principles of good feedback practices were defined that the instructor can use to think about the project and evaluate his feedback procedures. The seven good feedback practices are (NICOL; MACFARLANE-DICK, 2006):

- **GP 1**: Helps clarify what good performance is (goals, criteria, expected standards);
- **GP 2**: Facilitates the development of self-assessment (reflection) in learning;
- **GP 3**: Delivers high-quality information to students about their learning;
- **GP 4**: Encourages teacher and peer dialogue around learning;
- **GP 5**: Encourages positive motivational beliefs and self-esteem;
- **GP 6**: Provides opportunities to close the gap between current and desired performance;
- **GP 7**: Provides information to teachers that can be used to help shape teaching;

Each feedback principle could be connected to different strategies that the instructor can implement in the classroom. According to Nicol e Macfarlane-Dick (2006), good feedback practices are broadly defined as any strategy or content that can increase students' ability to self-regulate their learning performance.

10.2.2 Four Feedback Levels

Hattie e Timperley (2007) proposed a feedback model that identifies the circumstances where feedback has a higher impact. The authors define the purpose of feedback as reducing discrepancies between current performance and the desired objective. The authors looked at several conditions that could maximize the positive effects of feedback on learning, including increasing student awareness of an overall learning goal, progress towards the goal, and the subsequent goals necessary to achieve the primary goal. The proposed model specifies three types of questions that feedback is designed to answer (Where do I go? How do I go? Where

next?). Moreover, each feedback question operates at four levels: task feedback (FT), feedback on the task processing (FP), feedback on self-regulation (FR) and feedback on the self as a person (FS). The goal of each level is shown below.

- **FT Level**: Feedback can be about a task, such as whether the job is correct or incorrect, can include instructions for more or different information.
- **FP Level**: Feedback can be directed to the process used to create a product or complete a task, is more directed to information processing or learning processes that require understanding or completing the task.
- **FR Level**: Feedback for students can be focused on the level of self-regulation, including greater self-assessment or confidence skills, which can have major influences on self-efficacy, self-regulatory proficiency, and students' personal beliefs as learners.
- **FS Level**: Feedback can be personal in the sense that it is directed to the self. It is often unrelated to task performance.

The authors state that feedback on the FS level is the least effective. The FR and FP levels are powerful in terms of deep processing and task domain, while FT is powerful when feedback is used to improve strategy processing or improve self-regulation (HATTIE; TIMPERLEY, 2007), that is FT is powerful when used in combination with either FP and FR.

10.2.3 Related works

Several previous papers in the literature focused on the effectiveness of feedback in educational environments. According to Gibbs e Simpson (2005), formative assessment combined with high-quality feedback can have a powerful impact on student learning. The following are some related works that guided this study.

The work of Dunworth e Sanchez (2016) presents the results of an embedded multiple case study that investigated teaching staff and students' views on written feedback at a United Kingdom university. The study results indicated a general consistency between the teaching staff and students regarding the nature of quality feedback. Therefore, Dunworth and Sanchez state that quality feedback can be described as a process in which teaching's contribution and support are actively engaged, appropriated, and used productively by students to improve their

educational experience from several different dimensions (affective or interpersonal, orientational and transformational). The results suggest that feedback needs to incorporate each of these dimensions to be perceived as good quality.

Epistemic Network Analysis (ENA) has been used in the literature to analyze and visualize the relationship between data (SHAFFER; COLLIER; RUIS, 2016). This technique provides tools to analyze data qualitatively and has been used in the educational field; for example, Rolim et al. (2019) presents an approach that uses ENA to understand the relationships between cognitive and social presence. The paper demonstrates how epistemic network analysis (ENA) can provide new qualitative and quantitative insights into the development of students' social and critical thinking skills in communities of inquiry. In addition, the authors explore how the relationship between social and cognitive presences changed over time during a course.

Mello e Gašević (2019) investigated different configurations of ENA parameters to analyze student interaction behavior in online discussions from the perspective of the research community framework. In their results, the authors state that the main implication was that removing dominant codes can lead to a better understanding of one of the two key constructs in the Community of Inquiry model (social presence), but reduce the perception of the other (cognitive presence).

Ferreira et al. (2018) analyzes students' cognitive development in asynchronous discussions in online learning environments. The authors combine natural language processing techniques and graph-based analysis with ENA to provide qualitative insight into developing students' critical and deep thinking skills. Gašević et al. (2019) propose an approach called SENS (epistemic social network signature). This approach aims to combine social network analysis (SNA) with epistemic network analysis (ENA) to model the different dimensions of collaborative learning, using social ties and content analysis of student messages.

In this context, this paper proposes a network analysis method to uncover the relationship between different constraints of good quality feedback. In this paper, we evaluate the relationship between the concepts proposed by Hattie e Timperley (2007) and Nicol e Macfarlane-Dick (2006), but the same method could be generalized to any other theory.

10.3 RESEARCH QUESTION

The previous sections demonstrate that feedback is critical to help self-regulate student learning. The feedback models proposed in Hattie e Timperley (2007), Nicol e Macfarlane-

Dick (2006) have been widely used in the literature. However, the relationship between the constraints proposed by both models was not explored. Therefore, our research question that guides this study is:

> **RESEARCH QUESTION (RQ):** *What are the relationships between indicators of the feedback models proposed by Hattie e Timperley (2007) and Nicol e Macfarlane-Dick (2006)?*

With this research question, we posit that the analysis of feedback quality can be far more insightful if feedback quality is assessed from two different and complementary perspectives.

10.4 METHOD

10.4.1 Datasets

This work used two datasets, with feedback in Portuguese and English. Each one of them is detailed below.

10.4.1.1 Portuguese Dataset

The Portuguese dataset contains individual written feedback provided by instructors who posted the feedback messages through the activity submission tool in the LMS. The dataset comprises 1,000 examples of feedback from 2 courses: biology (41 examples of feedback) and literature (959 examples of feedback). The average length of feedback texts is 30 words. Each feedback was classified by expert coders according to the 11 feedback descriptors (7 good practices (NICOL; MACFARLANE-DICK, 2006) and four levels (HATTIE; TIMPERLEY, 2007)). The feedback message received labeled 0 if it did not belong to the descriptor or label 1 if it belonged to the descriptor. This annotation process was done separately between the seven good practices and the four levels.

The measures used to analyze the quality of the annotation process were the percentage of agreement and Cohen's kappa coefficient (κ). According to Landis e Koch (1977a), Cohen's kappa (κ) is a statistic that measures inter-rater agreement for qualitative items, where values close to 1 indicate significant agreement between the raters or coders (in this case, expert classifiers) and values close to 0 indicate that the agreement between the raters or coders is

purely random. The inter-rater agreement was moderate for the seven good practices, reaching 76.7% and Cohen's $\kappa = 0.43$. For the four levels, the inter-rater agreement had a percentage of 72.2%, and Cohen's κ was 0.44. Another two experts who did not participate in the first coding stage resolved the divergent cases, 23.3% for the seven good practices and 27.8% for the four levels. Table 44 shows the Portuguese dataset division for the seven good practices and the four levels.

Tabela 44 – Portuguese dataset division for the seven good practices and the four levels.

	Good Practice (GP)							**Level**			
	GP1	**GP2**	**GP3**	**GP4**	**GP5**	**GP6**	**GP7**	**FT**	**FP**	**FR**	**FS**
class 0	875	971	692	987	766	397	1000	112	499	992	849
class 1	125	29	308	13	234	603	0	888	501	8	151
Total	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000

Fonte: Elaborada pelo autor (2022)

10.4.1.2 English Dataset

The English dataset has 272 feedback texts from instructors of two courses: Software Engineering and Learning Analytics. The average length of feedback texts is 437 words. The annotation process followed the same process as the Portuguese dataset. For the seven good practices, the inter-rater agreement had a percentage of 76.6%, and Cohen's κ was 0.53. For the four levels, the inter-rater agreement had a percentage of 63.8% and Cohen's κ was 0.38. Different coders who did not participate in the first annotation step resolved the divergent cases. Table 45 shows the English dataset division for the seven good practices and the four levels.

10.4.2 Network Analysis

To answer our research question, we use Epistemic Network Analysis (ENA) (SHAFFER et al., 2009) to analyze the relationship between feedback levels and good practices of feedback. ENA is a graph-based analysis technique to model patterns of association by a network of dynamic relationships of different concepts (SHAFFER; COLLIER; RUIS, 2016). Within ENA, the

Tabela 45 – English dataset division for the seven good practices and the four levels.

	Good Practice (GP)							Level			
	GP1	**GP2**	**GP3**	**GP4**	**GP5**	**GP6**	**GP7**	**FT**	**FP**	**FR**	**FS**
class 0	193	206	38	166	80	202	253	94	68	174	139
class 1	79	66	234	106	192	70	19	178	204	98	133
Total	272	272	272	272	272	272	272	272	272	272	272

Fonte: Elaborada pelo autor (2022)

network of relationships among different *codes* is created for each *unit of analysis*. Two codes are considered related if they appear in the same chunk of text, called *stanza* (or *conversation*). We can see ENA relationships between the two feedback models, i.e., if the good practices units are near and strongly connected with the feedback levels units.

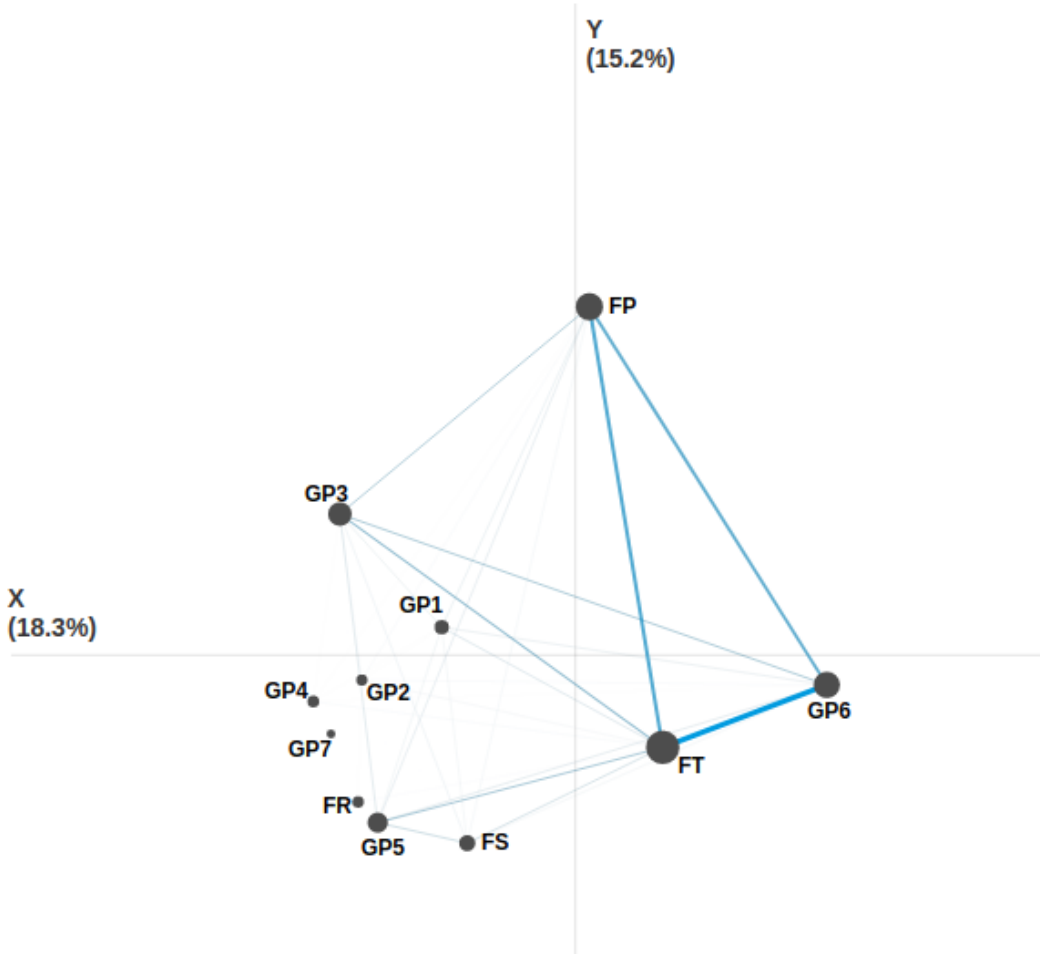
In this work, the *units of analysis* are feedback messages, the *codes* were the levels of feedback and the good practices of feedback, whereas the *stanza* we considered individual feedback messages. The network data provided by ENA was visualized as undirected graphs, and the object of interest was the interactions between the elements (SHAFFER; COLLIER; RUIS, 2016). In this network, there were three elements to observe: the *size* of the nodes represents their frequency; the *nearness* of the node represents the similarity among them; the *strength* of the code relationship represents the frequency of their co-occurrence.

10.5 RESULTS

To answer our research question, we analyzed the ENA network using: Portuguese dataset (Figure 25), English dataset (Figure 26) and both datasets (Figure 27), respectively.

Figure 25 presents the ENA graph for the Portuguese language. It shows that the most substantial relationship in the graph is between the FT and GP6 classes. The GP6 good practice aims to provide opportunities to close the gap between current and desired performance, where the instructor should indicate to the students their mistakes to help them avoid the same mistake in future tasks. Similarly, the FT level aims to provide comments that indicate where the student got it right/wrong on a specific task. In other words, feedback of these classes has a common goal: to show the student the points where he hit or miss the activity and, consequently, help the student close the gap between current and desired performance.

Figura 25 – ENA network of the relationship between the feedback descriptors for Portuguese.



Fonte: Elaborada pelo autor (2022)

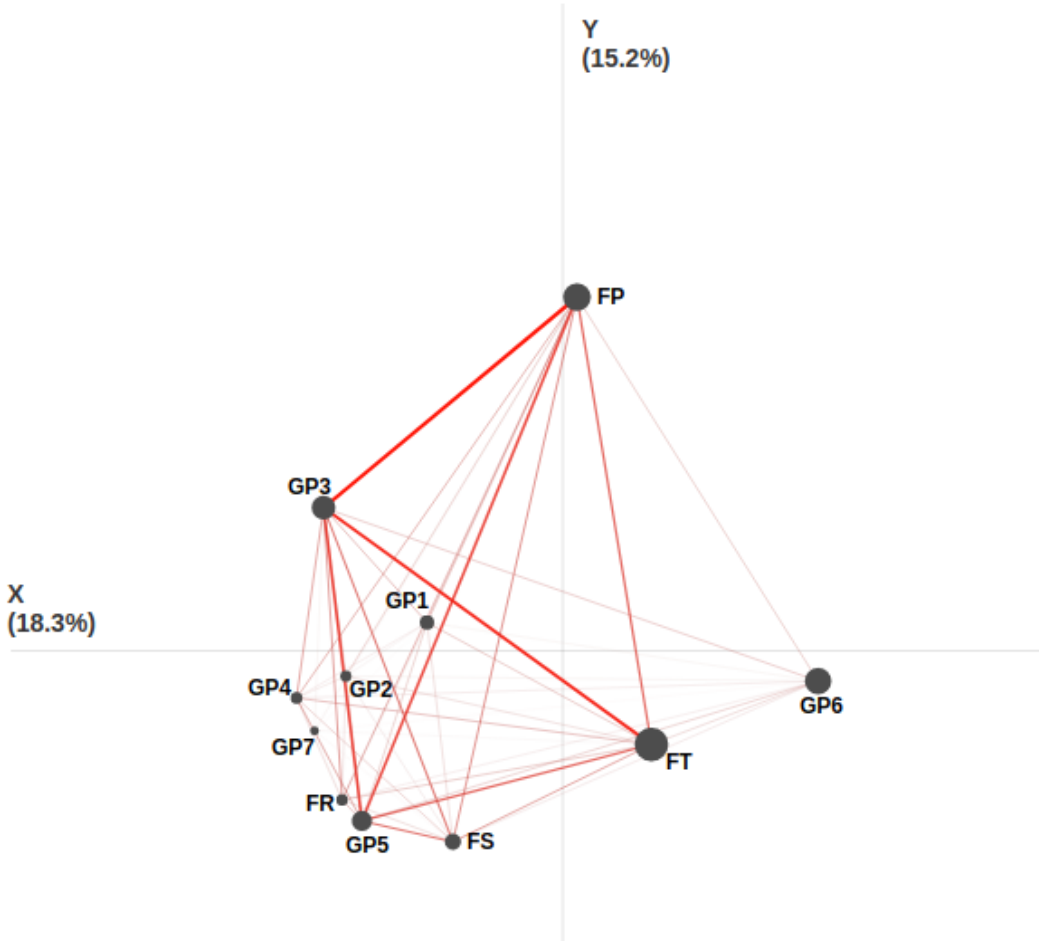
Another strong relationship in the graph is between FP and FT and between FP and GP6. An interpretation is that the FT and FP levels are related to a specific activity. The difference between these levels is that the FT focuses only on the task, checking whether it is correct or incorrect, and the FP focuses on the processing of the task (i.e., the student must reflect on his answer). Similar to our justification of the connection between FT and GP6, the connection between FP and GP6 shows that feedback about the task makes students reflect and also helps close the gap between current and desired performance.

The GP3 good practice that aims to provide students with high-quality information had a relationship with the FT and FP levels. As the FT and FP levels provide feedback on the tasks and processing, this means that on these levels, high-quality information is provided, and it helps the student to identify not only where the student got it wrong or right but why they got it wrong/right and what are the next steps for learning.

It is also possible to see in the graph, through proximity and connection, the relationship

between GP5 and FS. The goal of GP5 good practice is to encourage positive motivational beliefs and self-esteem. This relationship explains the relationship with the FS level because, on this level, feedback must provide personal praise for the student to remain motivated in his learning.

Figura 26 – ENA network of the relationship between the feedback descriptors for English.



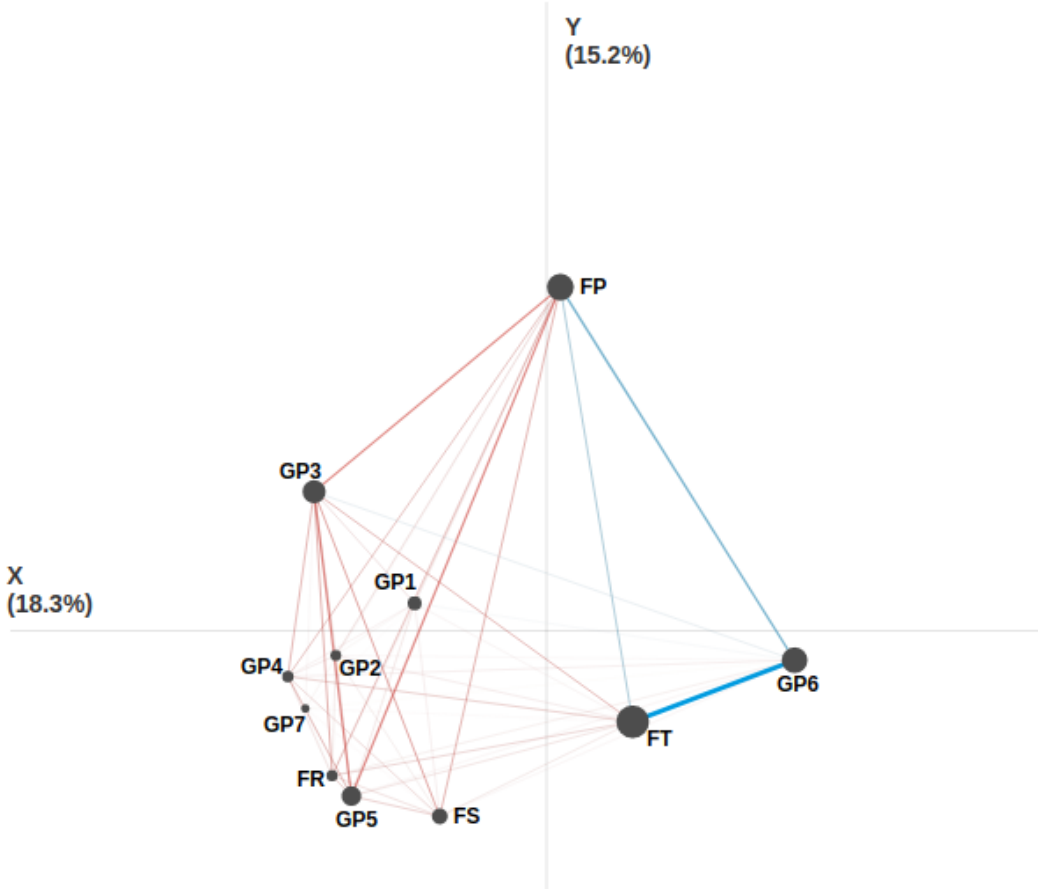
Fonte: Elaborada pelo autor (2022)

Figure 26 shows the ENA graph for feedback messages written in English. It is possible to see that the connections are well distributed, differently from the Portuguese data. The strongest relationships on the graphic are between GP3 and FP and between GP3 and FT. These relationships also happened in feedback messages written in Portuguese but with a weaker connection. We attribute this result to the fact that the English language feedback is much larger than that of the Portuguese. In other words, the English feedback provides high-quality information and makes the student reflect on his response. It is also possible to notice a good relationship between GP3 and GP5 and between GP3 and FS, which means that in addition to providing high-quality information, English language feedback also provides the

student with praise to keep them motivated.

The most significant difference between the two graphs is the relationship with the GP6. In English, fewer feedback messages had this indicator. Thus, the instructor did not provide specific information for students to close the gap between current and desired performance. In contrast, most feedback messages had good practice GP3; consequently, high-quality information helped students understand where they went wrong and what the next steps in learning should be.

Figura 27 – Subtraction between the Portuguese and English ENA network.



Fonte: Elaborada pelo autor (2022)

Figure 27 shows the subtraction of the two networks (EN and PT). The blue edges refer to the Portuguese language feedback and the red edges refer to the English language feedback. We can highlight in this network the clear division between Portuguese and English. The nodes and edges on the right are the most present in the Portuguese language feedback, that is, the FT, FP and GP6 good practice levels, highlighting the link between the FT level and the GP6 good practice, showing that there is a solid correlation between these two descriptors. Looking at the left side of the image, it is possible to see the predominant presence of the red color,

that is, the feedbacks from the English language are more present in the FR, FS levels and in all the good practices, except the GP6 good practice.

10.6 DISCUSSION

Answering our research question (Section 10.3), Figure 27 provided insights into the relationship between the feedback models evaluated. This graph shows that the FT and FP levels were present in many feedback for both languages. This result indicates the main goal of feedback messages was to show the student where they were wrong (FT level) and bring information that makes the student reflect on their response (FP level). In terms of good practices, the predominance of GP6 and GP3 demonstrates the effort to assist the students to close the gap between the desired performance and the activity. In both cases, the main indicators revealed by our analysis show and emphasize providing feedback on the activity (task and process), which is the most common type of feedback (DUNWORTH; SANCHEZ, 2016; AL-HAMAD; MOHIELDIN, 2013; MAITRA et al., 2018; CAVALCANTI et al., 2019).

The results presented for the Portuguese language showed a strong connection between the FT and FP levels with good practice GP6. This connection shows that the instructors' main objective concerning feedback is to present where the student was wrong, suggesting to reflect and close the learning gap. However, the literature shows that it is also interesting to pass on other information that encourages self-regulation (HENDERSON et al., 2019a). In order to provide more information and motivation for students, the instructors should not focus only on task-related feedback (WISNIEWSKI; ZIERER; HATTIE, 2020).

On the other hand, the English written feedback messages had a strong relationship between the FT and FP levels with the good practices GP3 and GP5. In other words, the English language feedback, in addition to showing where the student was wrong on their tasks, provide high-quality information with positive beliefs and self-esteem to keep students motivated. The average number can see one indicator of this multi-facet of English feedback of words (mean=437). An average is almost 15 times higher than that of feedback in the Portuguese language (mean=30).

It is essential to highlight that the feedback messages did not include many messages of FR and FS levels in both datasets. However, these levels of feedback could have a higher impact on students' self-regulation and satisfaction (WISNIEWSKI; ZIERER; HATTIE, 2020).

The ENA charts clearly showed that there is a difference in the style of providing feed-

back between the two languages. That is, the levels and best practices most present in one language are not the most present in the other language. However, this difference could have occurred due to the feedback texts being extracted from different environments or courses or the difference in the number of messages between the two sets.

Finally, in summary, the main similarities between the two models were between the FT level and the GP6 good practice, the FP level and the GP3 good practice, and the FS and FR levels with the GP5 good practice, which shows, based on the analysis of this work, that there are relationships between the two models. A combination of the use of models by the teacher can be interesting, for example, to separate the feedback into different structures, with each structure being a level of feedback and within each level using the principles of good feedback practices that best suit.

10.7 LIMITATIONS

The main limitation of this work is the datasets. The Portuguese dataset had 1000 feedback messages obtained from only two different undergraduate courses (Biology and Literature). On the other hand, the English dataset has 272 feedback messages, a small amount to perform a fully comparative analysis. Thus, further analysis would be required to provide more insights on the topic. However, this paper focused on the potential of using ENA to compare different feedback models. In future work, we intend to evaluate a dataset with more feedback messages and different theories of good quality feedback using the same approach. The final goal is to develop a feedback quality measure using ENA that encompasses traditional models proposed in the literature.

11 CONCLUSÃO

Esta seção tem como objetivo apresentar as considerações finais sobre os principais tópicos abordados nesta tese, incluindo as contribuições alcançadas e indicações para trabalhos futuros.

11.1 CONSIDERAÇÕES FINAIS

O objetivo desta tese foi treinar modelos de Inteligência Artificial para avaliar automaticamente a qualidade do feedback enviado por professores em AVAs. Os artigos publicados e submetidos mostram os resultados obtidos em vários experimentos realizados. Inicialmente foi criado um conjunto de dados com feedbacks de 2 cursos a distância e em seguida especialistas rotularam o conjunto de dados verificando se os feedbacks seguiam as boas práticas (NICOL; MACFARLANE-DICK, 2006) ou os níveis (HATTIE; TIMPERLEY, 2007). Esse conjunto de dados pode ser utilizado para treinar modelos de aprendizagem de máquina para avaliar a qualidade do feedback.

No Capítulo 4 foi realizada uma revisão sistemática da literatura sobre feedback automático em ambientes virtuais de aprendizado entre os anos de 2009 e 2018. Foram analisadas 4 questões de pesquisas para os 63 artigos encontrados depois do processo de seleção de artigos e extração de informações. Essa revisão apresenta o estado da arte sobre feedback automático em ambientes de aprendizagem online.

Em uma primeira avaliação (Capítulo 5) utilizando o classificador *Random Forest* no conjunto de dados binário das boas práticas (se o feedback tem ou não uma ou mais boas práticas) foi possível obter 0,75 de acurácia para os dados de teste. Em uma segunda avaliação (Capítulo 6) foi analisado novamente o conjunto de dados binário das boas práticas, mas utilizando classificadores baseados em árvores de decisão mais recentes (AdaBoost e XGBoost). Nessa avaliação foi obtido nos dados de teste as acurácias: 0,89 para o AdaBoost e 0,91 para o XGBoost; e também foi realizada uma análise das características mais importantes para cada classificador. Foram utilizados algoritmos baseados em árvores de decisão com o objetivo de entender e explicar as saídas que o modelo nos fornece. Esses algoritmos são amplamente utilizados na área educacional (FERREIRA-MELLO et al., 2019).

Em uma terceira avaliação (Capítulo 7) utilizando também o classificador *Random Forest*, mas dessa vez para o conjunto de dados dos níveis de feedback (quando o feedback pode

pertencer a um dos níveis propostos por (HATTIE; TIMPERLEY, 2007)) foi possível obter uma acurácia no conjunto de testes de 0,75 para o nível FT, 0,64 para o nível FP e 0,87 para o nível FS. Também foi realizada uma análise das características mais importantes dos melhores modelos obtidos para os níveis FT, FP e FS. Como forma complementar as boas práticas de (NICOL; MACFARLANE-DICK, 2006), o trabalho de (HATTIE; TIMPERLEY, 2007) tem como objetivo classificar o feedback em um dos 4 níveis. Como mostra o trabalho desenvolvido no Capítulo 10, a combinação de ambas as teorias de feedback educacional podem ser usadas em conjunto em um mesmo sistema, pode ser interessante para o professor, por exemplo, para separar o feedback em diferentes estruturas, onde cada estrutura seja um nível de feedback e dentro de cada nível termos os princípios de boas práticas de feedback que melhor se adequam.

Em seguida, no Capítulo 8 foram analisados os classificadores AdaBoost e XGBoost nos conjuntos de dados de níveis e boas práticas de feedback, ou seja, foram analisados diversos classificadores binários para cada boa prática ou nível. Foram obtidas as acurácias de 0,87, 0,84, 0,88 e 0,81 para as boas práticas BP1, BP3, BP5 e BP6, respectivamente, e as acurácias de 0,89, 0,77 e 0,89 para os níveis FT, FP e FS, respectivamente. O principal objetivo desse trabalho foi analisar um novo recurso para extração de mais características do texto. Como podemos perceber pelos resultados, esse novo recurso trouxe melhorias para cada classificador quando comparado com os estudos anteriores (Capítulo 7 e 6).

Na última avaliação (Capítulo 9) foi analisado o algoritmo de *deep learning* BERT novamente nos conjuntos de dados de níveis e boas práticas de feedback, entretanto, com a adição de uma aumentação nos dados de treinamento e uma análise usando o XAI. Nesse trabalho foi possível obter as acurácias de 0,88, 0,87, 0,90 e 0,83 para as boas práticas BP1, BP3, BP5 e BP6, respectivamente, e as acurácias de 0,90, 0,77 e 0,93 para os níveis FT, FP e FS, respectivamente, mostrando assim ser uma abordagem que trouxe resultados significativamente melhores comparada à abordagem utilizando o XGBoost (Capítulo 8).

No Capítulo 10 é realizado uma análise utilizando a técnica ENA entre as duas teorias ((NICOL; MACFARLANE-DICK, 2006) e (HATTIE; TIMPERLEY, 2007)) utilizando os 2 conjuntos de dados criados. As principais semelhanças entre os dois modelos foram entre o nível FT com a boa prática GP6, o nível FP com a boa prática GP3, e os níveis FS e FR com a boa prática GP5, o que mostra, com base na análise deste trabalho, que existem relações entre os dois modelos. Uma combinação do uso de modelos pelo professor pode ser interessante, por exemplo, para separar o feedback em diferentes estruturas, sendo cada estrutura um nível de feedback e dentro de cada nível utilizando os princípios de boas práticas de feedback que

melhor se adequam.

Os artigos publicados mostram o potencial da pesquisa sobre feedback educacional e o avanço no estado da arte que este trabalho trouxe. Todas as análises realizadas e os resultados obtidos trazem para literatura de feedback educacional um direcionamento inicial de como analisar textos de feedback com base em teorias educacionais. Os resultados demonstraram que é possível criar modelos de IA para avaliar automaticamente a qualidade do feedback utilizando os níveis (HATTIE; TIMPERLEY, 2007) e as boas práticas de feedback (NICOL; MACFARLANE-DICK, 2006). Esses modelos podem ser integrados em Ambientes Virtuais de Aprendizagem para ajudar o professor no fornecimento de feedback efetivo.

11.2 PRINCIPAIS CONTRIBUIÇÕES

Entre as principais contribuições dessa tese estão: (1) a criação dos 2 conjuntos de dados com os dados rotulados; (2) geração dos modelos de aprendizagem de máquina para classificação dos feedbacks; (3) revisão sistemática da literatura que mostra o estado da arte sobre feedback automático.

11.3 LIMITAÇÕES

Como enfatizado nos artigos, esse trabalho possui algumas limitações, sendo a principal delas relacionada ao desbalanceamento das classes, seja para os níveis de (HATTIE; TIMPERLEY, 2007) ou para as boas práticas de (NICOL; MACFARLANE-DICK, 2006). O desbalanceamento pode afetar negativamente a performance de um algoritmo de aprendizado de máquina. Isso ocorre porque muitos algoritmos de aprendizado de máquina são projetados para maximizar a acurácia global do modelo. No entanto, em conjuntos de dados desbalanceados, a acurácia global pode ser enganosa, pois o modelo pode simplesmente prever a classe majoritária em todas as instâncias, ignorando completamente a classe minoritária. Contudo, existem diversas técnicas que podem ser utilizadas para lidar com esse problema e melhorar a precisão do modelo. Nesse trabalho foram utilizadas técnicas de *oversampling*, onde essa técnica envolve a replicação de instâncias da classe minoritária, aumentando sua presença no conjunto de dados, bem como a geração de novas instâncias, que envolve a criação de novas instâncias sintéticas da classe minoritária.

Outra limitação é que a amostra estudada pode não ser suficiente para generalizar todos

os tipos de feedback existentes em ambientes virtuais de aprendizagem. Os datasets foram criados com base em textos de apenas 2 cursos a distância. Para que os modelos treinados conseguissem generalizar bem a análise do feedback, seria necessário o treinamento com mais dados e que esses dados fossem de vários cursos diferentes. Esse tipo de limitação não é tão simples de resolver quando estamos falando de dados educacionais. A pesquisa no campo educacional é dificultada pela falta de bases de dados disponíveis para realizar experimentos. Nesse trabalho foram utilizados dados de um curso a distância concedidos por uma universidade pública, sendo necessário o pré-processamento e anotação dos dados para conseguirmos avançar na pesquisa. Mesmo assim, os dados concedidos não podem ser divulgados por conter informações de alunos e professores e para preservar a Lei Geral de Proteção de Dados (LGPD).

Esse trabalho também apresenta limitações relacionadas aos algoritmos utilizados. Foram analisados apenas 4 algoritmos, sendo 3 deles baseados em árvores de decisão (XGBoost, AdaBoost e Random Forest) e 1 baseado em aprendizagem profunda. Para o trabalho se tornar mais completo e representativo, seria necessário a análise de outros algoritmos de aprendizagem de máquina em conjunto com testes estatísticos para determinar qual algoritmo possui o melhor desempenho. Devido a essa limitação, é possível que as conclusões estejam limitadas aos algoritmos escolhidos, e não reflitam o conjunto completo de opções disponíveis em aprendizagem de máquina.

Por fim, outra limitação desse trabalho está relacionada a não validação dos modelos treinados em um ambiente real. Para que o ciclo de pesquisa seja completo é necessário a validação da ferramenta com os *stakeholders*, nesse caso, professores e alunos de um ambiente virtual de aprendizagem, para que possam confirmar que as análises feitas pelos modelos treinados tiveram impacto no fornecimento de novos feedbacks pelo professor e também que o aluno apresente melhorias nos seus estudos com base nos novos feedbacks recebidos.

11.4 TRABALHOS FUTUROS

Como trabalhos futuros ao trabalho de pesquisa descrito nesta tese, lista-se, nesta seção, propostas de trabalhos futuros a serem realizadas.

- Análise de outros algoritmos de aprendizagem de máquina;
- Aumento da base de dados com feedbacks de outros cursos;

- Análise estatística do desempenho dos algoritmos de aprendizagem de máquina;
- Integração dos modelos de aprendizagem de máquina em uma ferramenta online;
- Análise da ferramenta em um ambiente real.

11.5 PUBLICAÇÕES

Esta seção apresenta toda a lista de publicações desenvolvidas durante o doutorado. A primeira lista mostra a publicação diretamente relacionada à tese.

- Automatic feedback in online learning environments: A systematic literature review (CAVALCANTI et al., 2021).
- An Analysis of the use of Good Feedback Practices in Online Learning Courses (CAVALCANTI et al., 2019).
- Análise Automática de Feedback em Ambientes de Aprendizagem Online (CAVALCANTI et al., 2020a).
- Uma Análise entre Boas Práticas de Feedback em Ambientes Virtuais de Aprendizagem (CAVALCANTI et al., 2020)
- How good is my feedback?: a content analysis of written feedback (CAVALCANTI et al., 2020).
- Utilização de Recursos Linguísticos para Classificação Automática de Mensagens de Feedback (CAVALCANTI et al., 2021).

Outros dois artigos foram submetidos:

- Automatic analysis of feedback messages using deep learning and explainable artificial intelligence (CAVALCANTI et al., 2022a).
- A Comparative Analysis Between Good Feedback Descriptors on Online Courses (CAVALCANTI et al., 2022b).

Além desses, outros artigos foram publicados em colaboração, derivados do nosso estudo:

- Towards automated content analysis of educational feedback: A multi-language study (OSAKWE et al., 2022)
- Feedback Analyzer: Uma interface para análise de feedback (BARROS et al., 2021)
- Classificação Multi-classe para Análise de Qualidade de Feedback (BATISTA et al., 2022)

REFERÊNCIAS

AKÇAPINAR, G. How automated feedback through text mining changes plagiaristic behavior in online assignments. *Computers & Education*, Elsevier, v. 87, p. 123–130, 2015.

AL-HAMAD, B.; MOHIELDIN, T. E-assessment as a tool to augment face-to-face teaching and learning environment. In: IEEE. *2013 Fourth International Conference on e-Learning Best Practices in Management, Design and Development of e-Courses: Standards of Excellence and Creativity*. [S.l.], 2013. p. 348–359.

AL-YAHYA, M.; GEORGE, R.; ALFARIES, A. Ontologies in e-learning: review of the literature. *International Journal of Software Engineering and Its Applications*, v. 9, n. 2, p. 67–84, 2015.

ALEMÁN, J. L. F.; PALMER-BROWN, D.; DRAGANOVA, C. Evaluating student response driven feedback in a programming course. In: IEEE. *2010 10th IEEE International Conference on Advanced Learning Technologies*. [S.l.], 2010. p. 279–283.

ALENCAR, M.; NETTO, J. F. Tutor collaborator using multi-agent system. In: SPRINGER. *International Conference on Collaboration Technologies*. [S.l.], 2014. p. 153–159.

ALI, L.; ASADI, M.; GAŠEVIĆ, D.; JOVANOVIĆ, J.; HATALA, M. Factors influencing beliefs for adoption of a learning analytics tool: An empirical study. *Computers & Education*, Elsevier, v. 62, p. 130–148, 2013.

ALONSO, D. R.; CORTÉS, C. Z.; ZACATELCO, H. C.; CARRANZA, J. L. C. Hyperparameter tuning for multi-label classification of feedbacks in online courses. *Journal of Intelligent & Fuzzy Systems*, IOS Press, n. Preprint, p. 1–9, 2022.

ANDRÉ, M.; 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.

ARENDS, H.; KEUNING, H.; HEEREN, B.; JEURING, J. An intelligent tutor to learn the evaluation of microcontroller i/o programming expressions. In: ACM. *Proceedings of the 17th Koli Calling International Conference on Computing Education Research*. [S.l.], 2017. p. 2–9.

ARRIETA, A. B.; DÍAZ-RODRÍGUEZ, N.; SER, J. D.; BENNETOT, A.; TABIK, S.; BARBADO, A.; GARCÍA, S.; GIL-LÓPEZ, S.; MOLINA, D.; BENJAMINS, R. et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, Elsevier, v. 58, p. 82–115, 2020.

BANERES, D.; CLARISÓ, R.; JORBA, J.; SERRA, M. Experiences in digital circuit design courses: A self-study platform for learning support. *IEEE Transactions on Learning Technologies*, IEEE, v. 7, n. 4, p. 360–374, 2014.

BARBOSA, A.; FERREIRA, M.; MELLO, R. F.; LINS, R. D.; GASEVIC, 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.

BARROS, A. N.; JÚNIOR, E. A. de A.; CAVALCANTI, A.; NASCIMENTO, A.; MIRANDA, P.; MELLO, R. F. Feedback analyzer: Uma interface para análise de feedback. In: SBC. *Anais do VI Congresso sobre Tecnologias na Educação*. [S.l.], 2021. p. 51–60.

BATISTA, H.; CAVALCANTI, A. P.; MELLO, R. F.; MIRANDA, P.; NASCIMENTO, A. Classificação multi-classe para análise de qualidade de feedback. In: SBC. *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*. [S.l.], 2022. p. 861–872.

BATISTA, R. L.; SALGADO, N.; BARRETO, R. Avaliação e feedback automático em educação apoiada por tecnologia: Um mapeamento sistemático da literatura. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2018. v. 29, n. 1, p. 178.

BECHEIKH, N.; LANDRY, R.; AMARA, N. Lessons from innovation empirical studies in the manufacturing sector: A systematic review of the literature from 1993–2003. *Technovation*, Elsevier, v. 26, n. 5-6, p. 644–664, 2006.

BELCADHI, L. C. Personalized feedback for self assessment in lifelong learning environments based on semantic web. *Computers in Human Behavior*, Elsevier, v. 55, p. 562–570, 2016.

BIRCH, G.; FISCHER, B.; POPPLETON, M. Using fast model-based fault localisation to aid students in self-guided program repair and to improve assessment. In: ACM. *Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education*. [S.l.], 2016. p. 168–173.

BLACK, P.; WILIAM, D. Assessment and classroom learning. *Assessment in Education: principles, policy & practice*, Taylor & Francis, v. 5, n. 1, p. 7–74, 1998.

BLIKSTEIN, P.; WORSLEY, M.; PIECH, C.; SAHAMI, M.; COOPER, S.; KOLLER, D. Programming pluralism: Using learning analytics to detect patterns in the learning of computer programming. *Journal of the Learning Sciences*, Taylor & Francis, v. 23, n. 4, p. 561–599, 2014.

BODILY, R.; IKAHIHIFO, T. K.; MACKLEY, B.; GRAHAM, C. R. The design, development, and implementation of student-facing learning analytics dashboards. *Journal of Computing in Higher Education*, Springer, v. 30, n. 3, p. 572–598, 2018.

BOUD, D.; FALCHIKOV, N. *Rethinking assessment in higher education: Learning for the longer term*. [S.l.]: Routledge, 2007.

BOUD, D.; MOLLOY, E. Rethinking models of feedback for learning: the challenge of design. *Assessment & Evaluation in higher education*, Taylor & Francis, v. 38, n. 6, p. 698–712, 2013.

BREIMAN, L. Bagging predictors. *Machine learning*, Springer, v. 24, n. 2, p. 123–140, 1996.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001.

BRINK, H.; RICHARDS, J.; FETHEROLF, M. *Real-world machine learning*. [S.l.]: Manning Publications Co., 2016.

BROOKHART, S. M. *How to give effective feedback to your students*. [S.l.]: ASCD, 2017.

BRYFCZYNSKI, S.; PARGAS, R. P.; COOPER, M. M.; KLYMKOWSKY, M.; DEAN, B. C. Teaching data structures with besocratic. In: ACM. *Proceedings of the 18th ACM conference on Innovation and technology in computer science education*. [S.l.], 2013. p. 105–110.

BURKE, D. Strategies for using feedback students bring to higher education. *Assessment & Evaluation in Higher Education*, Routledge, v. 34, n. 1, p. 41–50, 2009.

BUTLER, D. L.; WINNE, P. H. Feedback and self-regulated learning: A theoretical synthesis. *Review of educational research*, Sage Publications Sage CA: Thousand Oaks, CA, v. 65, n. 3, p. 245–281, 1995.

BUTLER, D. L.; WINNE, P. H. Feedback and Self-Regulated Learning: A Theoretical Synthesis. *Review of Educational Research*, v. 65, n. 3, p. 245–281, set. 1995. ISSN 0034-6543.

CAMELO, R.; JUSTINO, S.; MELLO, R. F. L. de. Coh-metrix pt-br: uma api web de análise textual para a educação. In: SBC. *Anais dos Workshops do IX Congresso Brasileiro de Informática na Educação*. [S.l.], 2020. p. 179–186.

CARLESS, D.; BOUD, D. The development of student feedback literacy: enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, Routledge, v. 43, n. 8, p. 1315–1325, 2018.

CARVALHO, P.; SILVA, M. J. Sentilex-pt: Principais características e potencialidades. *Oslo Studies in Language*, v. 7, n. 1, 2015.

CAVALCANTI, A.; FERREIRA, R.; DIONÍSIO, M.; NETO, S.; PASSERO, G.; MIRANDA, P. Uma nova abordagem para detecção de plágio em ambientes educacionais. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2017. v. 28, n. 1, p. 1177.

CAVALCANTI, A.; FERREIRA, R.; FREITAS, F.; GAŠEVIĆ, D. Automatic analysis of feedback messages using deep learning and explainable artificial intelligence. *International Journal of Artificial Intelligence in Education*, Springer, 2022.

CAVALCANTI, A.; FERREIRA, R.; FREITAS, F.; ROLIM, V. A comparative analysis between good feedback practices on online courses. *Revista Latinoamericana de Tecnología Educativa-RELATEC*, 2022.

CAVALCANTI, A.; MELLO, R.; MIRANDA, P.; FREITAS, F. Análise automática de feedback em ambientes de aprendizagem online. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2020.

CAVALCANTI, A.; ROLIM, V.; MELLO, R.; FREITAS, F. Uma análise entre boas práticas de feedback em ambientes virtuais de aprendizagem. In: *Anais do XXXI Simpósio Brasileiro de Informática na Educação*. Porto Alegre, RS, Brasil: SBC, 2020. p. 962–971. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/sbie/article/view/12852>.

CAVALCANTI, A. P.; BARBOSA, A.; CARVALHO, R.; FREITAS, F.; TSAI, Y.-S.; GAšEVIć, D.; MELLO, R. F. Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence*, v. 2, p. 100027, 2021. ISSN 2666-920X. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2666920X21000217>.

CAVALCANTI, A. P.; DIEGO, A.; MELLO, R. F.; MANGAROSKA, K.; NASCIMENTO, A.; FREITAS, F.; GAŠEVIC, D. How good is my feedback? a content analysis of written feedback. In: ACM. *Proceedings of the 10th International Conference on Learning Analytics and Knowledge - LAK*. [S.l.], 2020.

CAVALCANTI, A. P.; MELLO, R. F.; MIRANDA, P.; NASCIMENTO, A.; FREITAS, F. Utilização de recursos linguísticos para classificação automática de mensagens de feedback. In: SBC. *Anais do XXXII Simpósio Brasileiro de Informática na Educação*. [S.l.], 2021. p. 861–872.

CAVALCANTI, A. P.; MELLO, R. F. L. de; ROLIM, V.; ANDRÉ, M.; FREITAS, F.; GAŠEVIC, D. An analysis of the use of good feedback practices in online learning courses. In: IEEE. *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*. [S.l.], 2019. v. 2161, p. 153–157.

CAVALCANTI, A. P.; ROLIM, V. B.; MELLO, R. F. L. de; FREITAS, F. L. G. de. Uma análise entre boas práticas de feedback em ambientes virtuais de aprendizagem. In: SBC. *Anais do XXXI Simpósio Brasileiro de Informática na Educação*. [S.l.], 2020. p. 962–971.

CHARLEER, S.; KLERKX, J.; DUVAL, E. Learning dashboards. *Journal of Learning Analytics*, UTS ePress, v. 1, n. 3, p. 199–202, 2014.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, v. 16, p. 321–357, 2002.

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.

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*. San Francisco, California, USA: ACM, 2016. (KDD '16), p. 785–794.

CLARK, I. Formative Assessment: Assessment Is for Self-regulated Learning. *Educ. Psychol. Rev.*, v. 24, n. 2, p. 205–249, jun. 2012. ISSN 1573-336X.

COATES, H.; JAMES, R.; BALDWIN, G. A critical examination of the effects of learning management systems on university teaching and learning. *Tertiary education and management*, Springer, v. 11, p. 19–36, 2005.

COHEN, J. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, Sage Publications Sage CA: Thousand Oaks, CA, v. 20, n. 1, p. 37–46, 1960.

CORREIA, R. L.; SANTOS, J. G. dos. A importância da tecnologia da informação e comunicação (tic) na educação a distância (ead) do ensino superior (ies). *Revista Aprendizagem em EAD*, v. 2, n. 1, 2013.

CORRIGAN, O.; SMEATON, A. F.; GLYNN, M.; SMYTH, S. Using educational analytics to improve test performance. In: *Design for Teaching and Learning in a Networked World*. [S.l.]: Springer, 2015. p. 42–55.

D'ANTONI, L.; KINI, D.; ALUR, R.; GULWANI, S.; VISWANATHAN, M.; HARTMANN, B. How can automatic feedback help students construct automata? *ACM Transactions on Computer-Human Interaction (TOCHI)*, ACM, v. 22, n. 2, p. 9, 2015.

DAVIS, D.; JIVET, I.; KIZILCEC, R. F.; CHEN, G.; HAUFF, C.; HOUBEN, G.-J. Follow the successful crowd: raising mooc completion rates through social comparison at scale. In: ACM. *Proceedings of the seventh international learning analytics & knowledge conference*. [S.l.], 2017. p. 454–463.

DAWSON, P.; HENDERSON, M.; MAHONEY, P.; PHILLIPS, M.; RYAN, T.; BOUD, D.; MOLLOY, E. What makes for effective feedback: Staff and student perspectives. *Assessment & Evaluation in Higher Education*, Taylor & Francis, v. 44, n. 1, p. 25–36, 2019.

DEMAIDI, M. N.; GABER, M. M.; FILER, N. Ontopefege: Ontology-based personalized feedback generator. *IEEE Access*, IEEE, v. 6, p. 31644–31664, 2018.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

DIETTERICH, T. G. Ensemble methods in machine learning. In: SPRINGER. *International workshop on multiple classifier systems*. [S.l.], 2000. p. 1–15.

DILLENBOURG, P.; NUSSBAUM, M.; DIMITRIADIS, Y.; ROSCHELLE, J. Design for classroom orchestration. *Computers & Education*, v. 69, n. 0, p. 485–492, 2013.

DUNWORTH, K.; SANCHEZ, H. S. Perceptions of quality in staff-student written feedback in higher education: a case study. *Teaching in Higher Education*, Routledge, v. 21, n. 5, p. 576–589, 2016.

DUTCHUK, M.; MUHAMMADI, K. A.; LIN, F. Quizmaster-a multi-agent game-style learning activity. In: SPRINGER. *International Conference on Technologies for E-Learning and Digital Entertainment*. [S.l.], 2009. p. 263–272.

DWECK, C. S. *Self-theories: Their role in motivation, personality, and development*. New York, NY, US: Psychology Press, 1999. (Self-theories: Their role in motivation, personality, and development). Pages: xiii, 195.

EFSTATHIOU, C.; HOVARDAS, T.; XENOFONTOS, N. A.; ZACHARIA, Z. C.; DEJONG, T.; ANJEWIERDEN, A.; RIESEN, S. A. van. Providing guidance in virtual lab experimentation: the case of an experiment design tool. *Educational technology research and development*, Springer, v. 66, n. 3, p. 767–791, 2018.

ER, E.; DIMITRIADIS, Y.; GAŠEVIĆ, D. Collaborative peer feedback and learning analytics: theory-oriented design for supporting class-wide interventions. *Assessment & Evaluation in Higher Education*, Routledge, p. 1–22, 2020.

ESPASA, A.; MENESES, J. Analysing feedback processes in an online teaching and learning environment: an exploratory study. *Higher education*, Springer, v. 59, n. 3, p. 277–292, 2010.

EVANS, C. Making sense of assessment feedback in higher education. *Review of educational research*, Sage Publications Sage CA: Los Angeles, CA, v. 83, n. 1, p. 70–120, 2013.

FARROW, E.; MOORE, J.; GAŠEVIĆ, D. Analysing discussion forum data: a replication study avoiding data contamination. In: ASSOCIATION FOR COMPUTING MACHINERY (ACM). *International Learning Analytics & Knowledge Conference 2019*. [S.l.], 2019. p. 170–179.

FAST, E.; LEE, C.; AIKEN, A.; BERNSTEIN, M. S.; KOLLER, D.; SMITH, E. Crowd-scale interactive formal reasoning and analytics. In: ACM. *Proceedings of the 26th annual ACM symposium on User interface software and technology*. [S.l.], 2013. p. 363–372.

FERGUSON, P. Student perceptions of quality feedback in teacher education. *Assess Eval High Educ*, v. 36, n. 1, p. 51–62, jan. 2011. ISSN 0260-2938.

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.

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-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, v. 9, n. 6, p. e1332, 2019.

FERREIRA, R.; KOVANOVIĆ, V.; GAŠEVIĆ, D.; ROLIM, V. Towards combined network and text analytics of student discourse in online discussions. In: SPRINGER. *International Conference on Artificial Intelligence in Education*. [S.l.], 2018. p. 111–126.

FILHO, P. P. B.; PARDO, T. A. S.; ALUÍSIO, S. M. An evaluation of the brazilian portuguese liwc dictionary for sentiment analysis. In: *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*. [S.l.: s.n.], 2013.

FREUND, Y.; SCHAPIRE, R. E. et al. Experiments with a new boosting algorithm. In: CITESEER. *icml*. [S.l.], 1996. v. 96, p. 148–156.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. *The elements of statistical learning*. [S.l.]: Springer series in statistics New York, NY, USA:, 2001. v. 1.

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.

GAsEVIć, D.; TSAI, Y.-S.; DAWSON, S.; PARDO, A. How do we start? an approach to learning analytics adoption in higher education. *The International Journal of Information and Learning Technology*, Emerald Publishing Limited, v. 36, n. 4, p. 342–353, 2019.

GAšEVIć, D.; MIRRIAHI, N.; DAWSON, S.; JOKSIMOVIć, S. Effects of instructional conditions and experience on the adoption of a learning tool. *Computers in Human Behavior*, v. 67, p. 207–220, 2017.

GIBBS, G.; SIMPSON, C. Conditions under which assessment supports students' learning. *Learning and teaching in higher education*, University of Gloucestershire, n. 1, p. 3–31, 2005.

GULWANI, S.; RADIČEK, I.; ZULEGER, F. Feedback generation for performance problems in introductory programming assignments. In: ACM. *Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering*. [S.l.], 2014. p. 41–51.

GUNNING, D. Explainable artificial intelligence (xai). *Defense advanced research projects agency (DARPA), nd Web*, v. 2, n. 2, p. 1, 2017.

HARVEY, L. Student feedback. *Quality in higher education*, Taylor & Francis, v. 9, n. 1, p. 3–20, 2003.

HATTIE, J.; GAN, M. Instruction based on feedback. In: *Handbook of research on learning and instruction*. [S.l.]: Routledge, 2011. p. 263–285.

HATTIE, J.; TIMPERLEY, H. The power of feedback. *Review of educational research*, Sage Publications Sage CA: Thousand Oaks, CA, v. 77, n. 1, p. 81–112, 2007.

HELMINEN, J.; MALMI, L. Jype-a program visualization and programming exercise tool for python. In: ACM. *Proceedings of the 5th international symposium on Software visualization*. [S.l.], 2010. p. 153–162.

HENDERSON, M.; AJJAWI, R.; BOUD, D.; MOLLOY, E. (Ed.). *The Impact of Feedback in Higher Education: Improving Assessment Outcomes for Learners*. Cham, Switzerland: Springer International Publishing, 2019. Google-Books-ID: WyxQxgEACAAJ. ISBN 978-3-030-25111-6.

HENDERSON, M.; PHILLIPS, M.; RYAN, T.; BOUD, D.; DAWSON, P.; MOLLOY, E.; MAHONEY, P. Conditions that enable effective feedback. *Higher Education Research & Development*, Taylor & Francis, v. 38, n. 7, p. 1401–1416, 2019.

HENTEA, M.; SHEA, M. J.; PENNINGTON, L. A perspective on fulfilling the expectations of distance education. In: *Proceedings of the 4th conference on Information technology curriculum*. [S.l.: s.n.], 2003. p. 160–167.

HERNÁNDEZ-GARCÍA, Á.; GONZÁLEZ-GONZÁLEZ, I.; JIMÉNEZ-ZARCO, A. I.; CHAPARRO-PELÁEZ, J. Applying social learning analytics to message boards in online distance learning: A case study. *Computers in Human Behavior*, Elsevier, v. 47, p. 68–80, 2015.

HU, Y.; MELLO, R. F.; GAŠEVIĆ, D. Automatic analysis of cognitive presence in online discussions: An approach using deep learning and explainable artificial intelligence. *Computers and Education: Artificial Intelligence*, Elsevier, v. 2, p. 100037, 2021.

IHANTOLA, P.; AHONIEMI, T.; KARAVIRTA, V.; SEPPÄLÄ, O. Review of recent systems for automatic assessment of programming assignments. In: *Proceedings of the 10th Koli calling international conference on computing education research*. [S.l.: s.n.], 2010. p. 86–93.

IRONS, A. *Enhancing learning through formative assessment and feedback*. [S.l.]: Routledge, 2007.

JEREMIĆ, Z.; JOVANOVIĆ, J.; GAŠEVIĆ, D. Student modeling and assessment in intelligent tutoring of software patterns. *Expert Systems with Applications*, Elsevier, v. 39, n. 1, p. 210–222, 2012.

JERIA, H.; VILLALON, J. Incorporating open education resources into computer supported marking tool to enhance formative feedback creation. In: IEEE. *Advanced Learning Technologies (ICALT), 2017 IEEE 17th International Conference on*. [S.l.], 2017. p. 256–260.

JIN, S.-H. Using visualization to motivate student participation in collaborative online learning environments. *Journal of Educational Technology & Society*, JSTOR, v. 20, n. 2, p. 51–62, 2017.

JOKSIMOVIĆ, S.; GAŠEVIĆ, D.; LOUGHIN, T. M.; KOVANOVIĆ, V.; HATALA, M. Learning at distance: Effects of interaction traces on academic achievement. *Computers & Education*, Elsevier, v. 87, p. 204–217, 2015.

JOULANI, P.; GYORGY, A.; SZEPESVÁRI, C. Online learning under delayed feedback. In: *International Conference on Machine Learning*. [S.l.: s.n.], 2013. p. 1453–1461.

JUGO, I.; KOVAČIĆ, B.; SLAVUJ, V. Using data mining for learning path recommendation and visualization in an intelligent tutoring system. In: IEEE. *2014 37th international convention on Information and communication technology, electronics and microelectronics (MIPRO)*. [S.l.], 2014. p. 924–928.

JURAFSKY, D.; MARTIN, J. H. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. [S.l.]: Citeseer, 2000. I–XXVI p.

JURAFSKY, D.; MARTIN, J. H. Speech and language processing: An introduction to speech recognition, computational linguistics and natural language processing. *Upper Saddle River, NJ: Prentice Hall*, 2008.

KARAVIRTA, V.; HELMINEN, J.; IHANTOLA, P. A mobile learning application for parsons problems with automatic feedback. In: ACM. *Proceedings of the 12th Koli Calling International Conference on Computing Education Research*. [S.l.], 2012. p. 11–18.

KASTRATI, Z.; DALIPI, F.; IMRAN, A. S.; NUCI, K. P.; WANI, M. A. Sentiment analysis of students' feedback with nlp and deep learning: A systematic mapping study. *Applied Sciences*, MDPI, v. 11, n. 9, p. 3986, 2021.

KEBODEAUX, K.; FIELD, M.; HAMMOND, T. Defining precise measurements with sketched annotations. In: ACM. *Proceedings of the Eighth Eurographics Symposium on Sketch-Based Interfaces and Modeling*. [S.l.], 2011. p. 79–86.

KEELE, S. et al. *Guidelines for performing systematic literature reviews in software engineering*. [S.l.], 2007.

KEUNING, H.; HEEREN, B.; JEURING, J. Strategy-based feedback in a programming tutor. In: ACM. *Proceedings of the Computer Science Education Research Conference*. [S.l.], 2014. p. 43–54.

KEUNING, H.; JEURING, J.; HEEREN, B. Towards a systematic review of automated feedback generation for programming exercises. In: *Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education*. [S.l.: s.n.], 2016. p. 41–46.

KHAN, I.; PARDO, A. Data2u: Scalable real time student feedback in active learning environments. In: ACM. *Proceedings of the sixth international conference on learning analytics & knowledge*. [S.l.], 2016. p. 249–253.

KHOSRAVI, H.; SHUM, S. B.; CHEN, G.; CONATI, C.; GASEVIC, D.; KAY, J.; KNIGHT, S.; MARTINEZ-MALDONADO, R.; SADIQ, S.; TSAI, Y.-S. Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, Elsevier, p. 100074, 2022.

KITCHENHAM, B. Procedures for performing systematic reviews. *Keele, UK, Keele University*, v. 33, n. 2004, p. 1–26, 2004.

KLEIJ, F. M. Van der; FESKENS, R. C.; EGGEN, T. J. Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of educational research*, Sage Publications Sage CA: Los Angeles, CA, v. 85, n. 4, p. 475–511, 2015.

KOHAVI, R. et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: STANFORD, CA. *Ijcai*. [S.l.], 1995. v. 14, n. 2, p. 1137–1145.

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: ACM. *Proceedings of the sixth international conference on learning analytics & knowledge*. [S.l.], 2016. p. 15–24.

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*. New York, NY, USA: ACM, 2016. p. 15–24.

KRUSCHE, S.; SEITZ, A. Artemis: An automatic assessment management system for interactive learning. In: ACM. *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*. [S.l.], 2018. p. 284–289.

KUHN, M. Caret: classification and regression training. *Astrophysics Source Code Library*, 2015.

KULKARNI, C.; WEI, K. P.; LE, H.; CHIA, D.; PAPADOPOULOS, K.; CHENG, J.; KOLLER, D.; KLEMMER, S. R. Peer and self assessment in massive online classes. *ACM Transactions on Computer-Human Interaction (TOCHI)*, ACM, v. 20, n. 6, p. 33, 2013.

LAI, S.; XU, L.; LIU, K.; ZHAO, J. Recurrent convolutional neural networks for text classification. In: *Twenty-ninth AAAI conference on artificial intelligence*. [S.l.: s.n.], 2015.

LAN, A. S.; VATS, D.; WATERS, A. E.; BARANIUK, R. G. Mathematical language processing: Automatic grading and feedback for open response mathematical questions. In: ACM. *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*. [S.l.], 2015. p. 167–176.

LANDIS, J. R.; KOCH, G. G. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, JSTOR, p. 363–374, 1977.

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

LANGER, P. The use of feedback in education: a complex instructional strategy. *Psychological reports*, SAGE Publications Sage CA: Los Angeles, CA, v. 109, n. 3, p. 775–784, 2011.

LAPES. *Start-state of the art through systematic review tool*. 2014. Disponível em: <http://lapes.dc.ufscar.br/tools/start_tool>. Acesso em: 02 abril 2019.

LAURILLARD, D. *Rethinking University Teaching: Rethinking University Teaching: a Framework for the Effective Use of Educational Technology*. [S.l.]: Routledge, 1993.

LEE, A.; LIM, T. M. Mining opinions from university students' feedback using text analytics. *Information Technology in Industry*, v. 4, n. 1, p. 26–33, 2016.

LEE, Y.; CHOI, J. A review of online course dropout research: Implications for practice and future research. *Educational Technology Research and Development*, Springer, v. 59, n. 5, p. 593–618, 2011.

LEWKOW, N.; FEILD, J.; ZIMMERMAN, N.; RIEDESEL, M.; ESSA, A.; BOULANGER, D.; SEANOSKY, J.; KUMAR, V.; KODE, S. et al. A scalable learning analytics platform for automated writing feedback. In: ACM. *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*. [S.l.], 2016. p. 109–112.

LIAW, A.; WIENER, M. et al. Classification and regression by randomforest. *R news*, v. 2, n. 3, p. 18–22, 2002.

LIM, L.-A.; GENTILI, S.; PARDO, A.; KOVANOVIĆ, V.; WHITELOCK-WAINWRIGHT, A.; GAŠEVIĆ, D.; DAWSON, S. What changes, and for whom? a study of the impact of learning analytics-based process feedback in a large course. *Learning and Instruction*, Elsevier, p. 101202, 2019.

LIU, M.; LI, Y.; XU, W.; LIU, L. Automated Essay Feedback Generation and Its Impact on Revision. *IEEE Trans. Learn. Technol.*, v. 10, n. 4, p. 502–513, out. 2017. ISSN 1939-1382.

LODDER, J.; HEEREN, B.; JEURING, J. Generating hints and feedback for hilbert-style axiomatic proofs. In: ACM. *Proceedings of the 2017 ACM SIGCSE Technical Symposium on Computer Science Education*. [S.l.], 2017. p. 387–392.

LOULA, A. C. *Emergência de comunicação e representações em criaturas artificiais*. [S.l.]: Angelo Conrado Loula, 2011.

MA, E. *NLP Augmentation*. 2019. Https://github.com/makcedward/nlpaug.

MA, X.; WIJEWICKREMA, S.; ZHOU, S.; ZHOU, Y.; MHAMMEDI, Z.; O'LEARY, S.; BAILEY, J. Adversarial Generation of Real-time Feedback with Neural Networks for Simulation-based Training. In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. Melbourne, Australia: International Joint Conferences on Artificial Intelligence Organization, 2017. p. 3763–3769.

MACHADO, A. P.; FERREIRA, R.; BITTENCOURT, I. I.; ELIAS, E.; BRITO, P.; COSTA, E. Mineração de texto em redes sociais aplicada à educação a distância. *Colabor@-A Revista Digital da CVA-RICESU*, v. 6, n. 23, 2010.

MAITRA, S.; MADAN, S.; KANDWAL, R.; MAHAJAN, P. Mining authentic student feedback for faculty using naïve bayes classifier. *Procedia computer science*, Elsevier, v. 132, p. 1171–1183, 2018.

MANI, I.; ZHANG, I. knn approach to unbalanced data distributions: a case study involving information extraction. In: *Proceedings of Workshop on Learning from Imbalanced Datasets*. [s.n.], 2003. Disponível em: <https://www.site.uottawa.ca/~nat/Workshop2003/workshop2003.html>.

MANOHARAN, S. Personalized assessment as a means to mitigate plagiarism. *IEEE Transactions on Education*, IEEE, v. 60, n. 2, p. 112–119, 2016.

MARIN, V. J.; PEREIRA, T.; SRIDHARAN, S.; RIVERO, C. R. Automated personalized feedback in introductory java programming moocs. In: IEEE. *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*. [S.l.], 2017. p. 1259–1270.

MARQUÉS, A. I.; GARCÍA, V.; SÁNCHEZ, J. S. Exploring the behaviour of base classifiers in credit scoring ensembles. *Expert Systems with Applications*, Elsevier, v. 39, n. 11, p. 10244–10250, 2012.

MARTIN, M.; ALVAREZ, A.; RUIZ, S.; FERNANDEZ-CASTRO, I.; URRETAVIZCAYA, M. Helping teachers to track student evolution in a b-learning environment. In: IEEE. *2009 Ninth IEEE International Conference on Advanced Learning Technologies*. [S.l.], 2009. p. 342–346.

MATCHA, W.; GASEVIC, D.; PARDO, A. et al. A systematic review of empirical studies on learning analytics dashboards: A self-regulated learning perspective. *IEEE Transactions on Learning Technologies*, IEEE, 2019.

MATCHA, W.; GAŠEVIĆ, D.; UZIR, N. A.; JOVANOVIĆ, J.; PARDO, A. Analytics of learning strategies: Associations with academic performance and feedback. In: ACM. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*. [S.l.], 2019. p. 461–470.

MCCALLUM, A.; LI, W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*. [S.l.], 2003. p. 188–191.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, Springer, v. 5, n. 4, p. 115–133, 1943.

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.

MELLO, R. F.; GAŠEVIĆ, D. What is the effect of a dominant code in an epistemic network analysis? In: SPRINGER. *International Conference on Quantitative Ethnography*. [S.l.], 2019. p. 66–76.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

MILLER, T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, Elsevier, v. 267, p. 1–38, 2019.

MITROVIC, A.; WILLIAMSON, C.; BEBBINGTON, A.; MATHEWS, M.; SURAWEERA, P.; MARTIN, B.; THOMSON, D.; HOLLAND, J. Thermo-tutor: An intelligent tutoring system for thermodynamics. In: IEEE. *2011 IEEE Global Engineering Education Conference (EDUCON)*. [S.l.], 2011. p. 378–385.

MOREL, G. M. Evaluating student and instructor use of video feedback in an online learning environment. In: IEEE. *Advanced Learning Technologies (ICALT), 2016 IEEE 16th International Conference on*. [S.l.], 2016. p. 549–551.

MULLINER, E.; TUCKER, M. Feedback on feedback practice: perceptions of students and academics. *Assessment & Evaluation in Higher Education*, Routledge, v. 42, n. 2, p. 266–288, 2017.

MURAD, D.; WANG, R.; TURNBULL, D.; WANG, Y. Slions: A karaoke application to enhance foreign language learning. In: ACM. *2018 ACM Multimedia Conference on Multimedia Conference*. [S.l.], 2018. p. 1679–1687.

MUTCH, A. Exploring the practice of feedback to students. *Active learning in higher education*, Sage Publications, v. 4, n. 1, p. 24–38, 2003.

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.

NEUENDORF, K. A. *The content analysis guidebook*. [S.l.]: Sage, 2016.

NICOL, D. J.; MACFARLANE-DICK, D. Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in higher education*, Taylor & Francis, v. 31, n. 2, p. 199–218, 2006.

NIELSEN, M. A. *Neural networks and deep learning*. [S.l.]: Determination press San Francisco, CA, USA, 2015. v. 25.

ONO, Y.; ISHIHARA, M.; YAMASHIRO, M. Preliminary construction of instant qualitative feedback system in foreign language teaching. In: IEEE. *2013 Second IIAI International Conference on Advanced Applied Informatics*. [S.l.], 2013. p. 178–182.

ORRELL, J. Feedback on learning achievement: Rhetoric and reality. *Teaching in higher education*, Taylor & Francis Group, v. 11, n. 4, p. 441–456, 2006.

OSAKWE, I.; CHEN, G.; WHITELOCK-WAINWRIGHT, A.; GAŠEVIC, D.; CAVALCANTI, A. P.; MELLO, R. F. Towards automated content analysis of feedback: A multi-language study. In: *Proceedings of the 14th International Conference on Educational Data Mining*. [S.l.: s.n.], 2021.

OSAKWE, I.; CHEN, G.; WHITELOCK-WAINWRIGHT, A.; GAŠEVIĆ, D.; CAVALCANTI, A. P.; MELLO, R. F. Towards automated content analysis of educational feedback: A multi-language study. *Computers and Education: Artificial Intelligence*, Elsevier, v. 3, p. 100059, 2022.

PAN, B. Application of XGBoost algorithm in hourly PM2.5 concentration prediction. *IOP Conference Series: Earth and Environmental Science*, v. 113, p. 012127, fev. 2018. ISSN 1755-1315. Publisher: IOP Publishing.

PARDO, A. A feedback model for data-rich learning experiences. *Assessment & Evaluation in Higher Education*, Routledge, v. 43, n. 3, p. 428–438, 2018.

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.

PARDO, A.; JOVANOVIC, J.; DAWSON, S.; GAŠEVIĆ, D.; MIRRIAHI, N. Using learning analytics to scale the provision of personalised feedback. *British Journal of Educational Technology*, Wiley Online Library, v. 50, n. 1, p. 128–138, 2019.

PARIKH, A.; MCREELIS, K.; HODGES, B. Student feedback in problem based learning: a survey of 103 final year students across five ontario medical schools. *Medical education*, Wiley Online Library, v. 35, n. 7, p. 632–636, 2001.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V. et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, v. 12, n. Oct, p. 2825–2830, 2011.

PENNEBAKER, J. W.; BOYD, R. L.; JORDAN, K.; BLACKBURN, K. *The development and psychometric properties of LIWC2015*. [S.l.], 2015.

PERERA, R.; NAND, P. Recent advances in natural language generation: A survey and classification of the empirical literature. *Computing and Informatics*, v. 36, n. 1, p. 1–32, 2017.

PERNA, C. L.; DELGADO, H. K.; FINATTO, M. J. *Linguagens Especializadas em Corpora: modos de dizer e interfaces de pesquisa*. [S.l.]: EDIPUCRS, 2010.

PITT, E.; CARLESS, D. Signature feedback practices in the creative arts: integrating feedback within the curriculum. *Assessment & Evaluation in Higher Education*, Taylor & Francis, p. 1–13, 2021.

POLSON, M. C.; RICHARDSON, J. J. *Foundations of intelligent tutoring systems*. [S.l.]: Psychology Press, 2013.

PRICE, M.; HANDLEY, K.; MILLAR, J.; O'DONOVAN, B. Feedback : all that effort, but what is the effect? *Assess Eval High Educ*, v. 35, n. 3, p. 277–289, maio 2010. ISSN 0260-2938, 1469-297X.

PRICE, M.; HANDLEY, K.; MILLAR, J.; O'DONOVAN, B. Feedback: all that effort, but what is the effect? *Assessment & Evaluation in Higher Education*, Routledge, v. 35, n. 3, p. 277–289, 2010.

PRIETO, L. P.; SHARMA, K.; DILLENBOURG, P.; JESÚS, M. Teaching analytics: towards automatic extraction of orchestration graphs using wearable sensors. In: *Proceedings of the sixth international conference on learning analytics & knowledge*. [S.l.: s.n.], 2016. p. 148–157.

REISER, R. A.; DEMPSEY, J. V. *Trends and issues in instructional design and technology*. [S.l.]: Pearson Boston, MA, 2012.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "why should i trust you?"explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. [S.l.: s.n.], 2016. p. 1135–1144.

RIOFRÍO-LUZCANDO, D.; RAMIREZ, J.; BERROCAL-LOBO, M. Predicting student actions in a procedural training environment. *IEEE Transactions on Learning Technologies*, IEEE, v. 10, n. 4, p. 463–474, 2017.

ROBISON, J.; MCQUIGGAN, S.; LESTER, J. Evaluating the consequences of affective feedback in intelligent tutoring systems. In: IEEE. *2009 3rd international conference on affective computing and intelligent interaction and workshops*. [S.l.], 2009. p. 1–6.

ROGERS, P. L. Barriers to adopting emerging technologies in education. *Journal of educational computing research*, SAGE Publications Sage CA: Los Angeles, CA, v. 22, n. 4, p. 455–472, 2000.

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.

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.

RUSSAKOVSKY, O.; DENG, J.; SU, H.; KRAUSE, J.; SATHEESH, S.; MA, S.; HUANG, Z.; KARPATHY, A.; KHOSLA, A.; BERNSTEIN, M. et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, Springer, v. 115, n. 3, p. 211–252, 2015.

SADLER, D. R. Formative assessment and the design of instructional systems. *Instructional science*, Springer, v. 18, n. 2, p. 119–144, 1989.

SADLER, D. R. Beyond feedback: Developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education*, Taylor & Francis, v. 35, n. 5, p. 535–550, 2010.

SAMEK, W.; MÜLLER, K.-R. Towards explainable artificial intelligence. In: *Explainable AI: interpreting, explaining and visualizing deep learning*. [S.l.]: Springer, 2019. p. 5–22.

SCARTON, C.; GASPERIN, C.; ALUISIO, S. Revisiting the readability assessment of texts in portuguese. In: SPRINGER. *Ibero-American Conference on Artificial Intelligence*. [S.l.], 2010. p. 306–315.

SCARTON, C. E.; ALUÍSIO, S. M. Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh-metrix para o português. *Linguamática*, v. 2, n. 1, p. 45–61, 2010.

SCHWENDIMANN, B. A.; RODRIGUEZ-TRIANA, M. J.; VOZNIUK, A.; PRIETO, L. P.; BOROUJENI, M. S.; HOLZER, A.; GILLET, D.; DILLENBOURG, P. Perceiving learning at a glance: A systematic literature review of learning dashboard research. *IEEE Transactions on Learning Technologies*, IEEE, v. 10, n. 1, p. 30–41, 2016.

SHAFFER, D. W. *Quantitative ethnography*. Madison, WI: Cathcart Press, 2017.

SHAFFER, D. W.; COLLIER, W.; RUIS, A. A tutorial on epistemic network analysis: Analyzing the structure of connections in cognitive, social, and interaction data. *Journal of Learning Analytics*, v. 3, n. 3, p. 9–45, 2016.

SHAFFER, D. W.; HATFIELD, D.; SVAROVSKY, G. N.; NASH, P.; NULTY, A.; BAGLEY, E.; FRANK, K.; RUPP, A. A.; MISLEVY, R. Epistemic Network Analysis: A Prototype for 21st-Century Assessment of Learning. *International Journal of Learning and Media*, v. 1, n. 2, p. 33–53, 2009.

SHERIDAN, R. Reducing the online instructor's workload. *Educause Quarterly*, Educause, v. 29, n. 3, p. 65, 2006.

SIMON, H. A. Why should machines learn? In: *Machine learning*. [S.l.]: Springer, 1983. p. 25–37.

SIMONSON, M.; ZVACEK, S. M.; SMALDINO, S. *Teaching and Learning at a Distance: Foundations of Distance Education 7th Edition*. [S.l.]: IAP, 2019.

SIROTHEAU, S.; SANTOS, J.; FAVERO, E.; FREITAS, S. N. de. Avaliação automática de respostas discursivas curtas baseado em três dimensões linguísticas. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2019. v. 30, n. 1, p. 1551.

SMITHIES, A.; BRAIDMAN, I.; BERLANGA, A.; HALEY, D.; WILD, F. Using language technologies to support individual formative feedback. 2010.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: pretrained bert models for brazilian portuguese. In: SPRINGER. *Brazilian Conference on Intelligent Systems*. [S.l.], 2020. p. 403–417.

SUNG, E.; MAYER, R. E. Five facets of social presence in online distance education. *Computers in Human Behavior*, Elsevier, v. 28, n. 5, p. 1738–1747, 2012.

TAN, P.-N. et al. *Introduction to data mining*. [S.l.]: Pearson Education India, 2007.

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.

TEMPELAAR, D.; NGUYEN, Q.; RIENTIES, B. Learning feedback based on dispositional learning analytics. In: *Machine Learning Paradigms*. [S.l.]: Springer, 2020. p. 69–89.

TENÓRIO, T.; BITTENCOURT, I. I.; ISOTANI, S.; SILVA, A. P. Does peer assessment in on-line learning environments work? a systematic review of the literature. *Computers in Human Behavior*, Elsevier, v. 64, p. 94–107, 2016.

THEODORIDIS, S.; KOUTROUMBAS, K. Pattern recognition and neural networks. *Machine Learning and Its Applications*, Springer, p. 169–195, 2001.

TOMEK, I. Two Modifications of CNN. *IEEE Transactions on Systems, Man, and Cybernetics*, v. 6, n. 11, p. 769–772, 1976.

TOSHNIWAL, S.; DEY, P.; RAJPUT, N.; SRIVASTAVA, S. Vibrein: An engaging and assistive mobile learning companion for students with intellectual disabilities. In: ACM. *Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction*. [S.l.], 2015. p. 20–28.

TRAUSAN-MATU, S.; DASCALU, M.; REBEDEA, T. Polycafe—automatic support for the polyphonic analysis of cscl chats. *International Journal of Computer-Supported Collaborative Learning*, Springer, v. 9, n. 2, p. 127–156, 2014.

TSIAKMAKI, M.; KOSTOPOULOS, G.; KOTSIANTIS, S.; RAGOS, O. Implementing automl in educational data mining for prediction tasks. *Applied Sciences*, Multidisciplinary Digital Publishing Institute, v. 10, n. 1, p. 90, 2020.

ULLMANN, T. D. Automated analysis of reflection in writing: Validating machine learning approaches. *International Journal of Artificial Intelligence in Education*, Springer, v. 29, n. 2, p. 217–257, 2019.

USENER, C. A. Easy-dsbuilder: automated assessment of tree data structures in computer science teaching. In: ACM. *Proceedings of the 30th Annual ACM Symposium on Applied Computing*. [S.l.], 2015. p. 220–226.

UTOMO, A. Y.; SANTOSO, H. B. Development of gamification-enriched pedagogical agent for e-learning system based on community of inquiry. In: ACM. *Proceedings of the International HCI and UX Conference in Indonesia*. [S.l.], 2015. p. 1–9.

VERBERT, K.; GOVAERTS, S.; DUVAL, E.; SANTOS, J. L.; ASSCHE, F. V.; PARRA, G.; KLERKX, J. Learning dashboards: an overview and future research opportunities. *Personal and Ubiquitous Computing*, Springer, v. 18, n. 6, p. 1499–1514, 2014.

VERIKAS, A.; GELZINIS, A.; BACAUSKIENE, M. Mining data with random forests: A survey and results of new tests. *Pattern recognition*, Elsevier, v. 44, n. 2, p. 330–349, 2011.

VIEIRA, R.; LOPES, L. Processamento de linguagem natural e o tratamento computacional de linguagens científicas. *EM CORPORA*, p. 183, 2010.

VILLALóN, J.; KEARNEY, P.; CALVO, R. A.; REIMANN, P. Glosser: Enhanced Feedback for Student Writing Tasks. In: *2008 Eighth IEEE International Conference on Advanced Learning Technologies*. Santander, Cantabria, Spain: IEEE, 2008. p. 454–458.

WANG, K.; SINGH, R.; SU, Z. Search, align, and repair: data-driven feedback generation for introductory programming exercises. In: ACM. *Proceedings of the 39th ACM SIGPLAN Conference on Programming Language Design and Implementation*. [S.l.], 2018. p. 481–495.

WASIK, S.; ANTCZAK, M.; BADURA, J.; LASKOWSKI, A.; STERNAL, T. A survey on online judge systems and their applications. *ACM Computing Surveys (CSUR)*, ACM New York, NY, USA, v. 51, n. 1, p. 1–34, 2018.

WEAVER, M. R. Do students value feedback? Student perceptions of tutors' written responses. *Assess Eval High Educ*, v. 31, n. 3, p. 379–394, jun. 2006. ISSN 0260-2938.

WEAVER, M. R. Do students value feedback? student perceptions of tutors' written responses. *Assessment & Evaluation in Higher Education*, Taylor & Francis, v. 31, n. 3, p. 379–394, 2006.

WEI, J.; ZOU, K. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*, 2019.

WEYTEN, L.; ROMBOUTS, P.; CATTEAU, B.; BOCK, M. D. Validation of symbolic expressions in circuit analysis e-learning. *IEEE Transactions on Education*, IEEE, v. 54, n. 4, p. 564–568, 2010.

WEYTEN, L.; ROMBOUTS, P.; MAEYER, J. D. Web-based trainer for electrical circuit analysis. *IEEE Transactions on Education*, IEEE, v. 52, n. 1, p. 185–189, 2008.

WHITELOCK, D.; TWINER, A.; RICHARDSON, J. T.; FIELD, D.; PULMAN, S. Openessayist: a supply and demand learning analytics tool for drafting academic essays. In: ACM. *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*. [S.l.], 2015. p. 208–212.

WINSTONE, N.; CARLESS, D. *Designing effective feedback processes in higher education: A learning-focused approach*. [S.l.]: Routledge, 2019.

WISE, A. F. Designing pedagogical interventions to support student use of learning analytics. In: *Proceedings of the fourth international conference on learning analytics and knowledge*. [S.l.: s.n.], 2014. p. 203–211.

WISE, A. F.; JUNG, Y. Teaching with analytics: Towards a situated model of instructional decision-making. *Journal of Learning Analytics*, v. 6, n. 2, p. 53–69, 2019.

WISNIEWSKI, B.; ZIERER, K.; HATTIE, J. The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, Frontiers, v. 10, p. 3087, 2020.

WISSEN, L. V.; BOOT, P. An electronic translation of the liwc dictionary into dutch. In: LEXICAL COMPUTING. *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference*. [S.l.], 2017. p. 703–715.

WONG, S.; TAYLOR, J. E.; BEAUMONT, T. Enhancing student learning experience through a novel electronic coursework assessment and feedback management system. Higher Education Academy, 2012.

WULFF, P.; MIENTUS, L.; NOWAK, A.; BOROWSKI, A. Utilizing a pretrained language model (bert) to classify preservice physics teachers' written reflections. *International Journal of Artificial Intelligence in Education*, Springer, p. 1–28, 2022.

XIAO, Z.; WANG, Y.; FU, K.; WU, F. Identifying Different Transportation Modes from Trajectory Data Using Tree-Based Ensemble Classifiers. *ISPRS International Journal of Geo-Information*, v. 6, n. 2, p. 57, fev. 2017. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

XIE, W.; ZHANG, W. Negative emotion enhances mnemonic precision and subjective feelings of remembering in visual long-term memory. *Cognition*, Elsevier, v. 166, p. 73–83, 2017.

XIE, X.; LI, X. Research on personalized exercises and teaching feedback based on big data. In: ACM. *Proceedings of the 3rd International Conference on Intelligent Information Processing*. [S.l.], 2018. p. 166–171.

YING, M.-H.; HONG, Y. The development of an online sql learning system with automatic checking mechanism. In: IEEE. *The 7th International Conference on Networked Computing and Advanced Information Management*. [S.l.], 2011. p. 346–351.

YING, M.-H.; YANG, K.-T.; DENG, G.-H. Development of a multiplayer online game-based learning system based on arcs motivation model. In: IEEE. *2012 Sixth International Conference on Genetic and Evolutionary Computing*. [S.l.], 2012. p. 589–594.

YOUNG, T.; HAZARIKA, D.; PORIA, S.; CAMBRIA, E. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, IEEE, v. 13, n. 3, p. 55–75, 2018.

YPSILANDIS, G. Feedback in distance education. *Computer Assisted Language Learning*, Taylor & Francis, v. 15, n. 2, p. 167–181, 2002.

YU, Y.-C. Teaching with a dual-channel classroom feedback system in the digital classroom environment. *IEEE Transactions on Learning Technologies*, IEEE, v. 10, n. 3, p. 391–402, 2016.

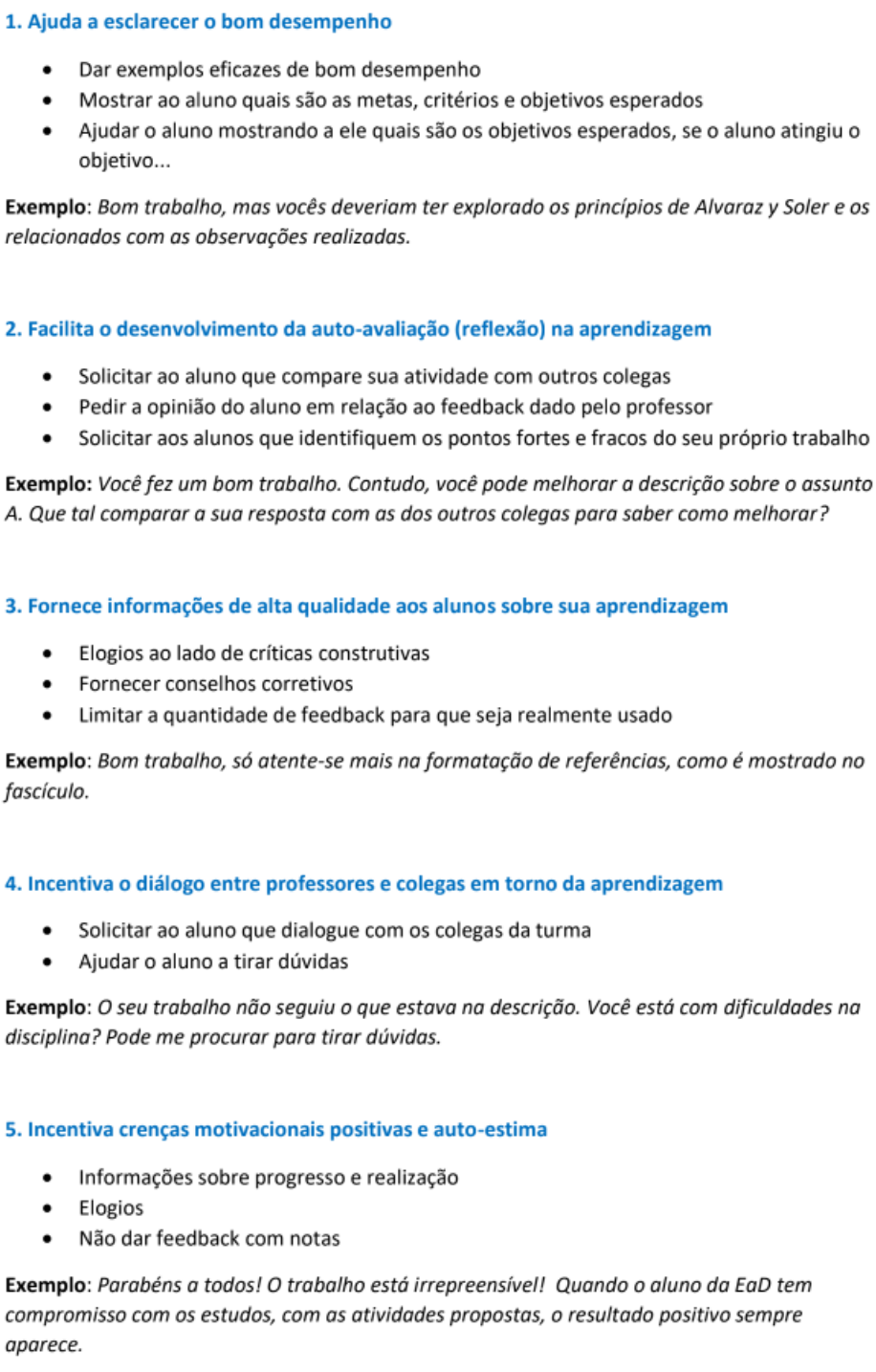
ZHAO, Y.; CZIKO, G. A. Teacher adoption of technology: A perceptual control theory perspective. *Journal of technology and teacher education*, Society for Information Technology & Teacher Education, v. 9, n. 1, p. 5–30, 2001.

ZHOU, W.; PAN, Y.; ZHOU, Y.; SUN, G. The framework of a new online judge system for programming education. In: ACM. *Proceedings of ACM Turing Celebration Conference-China*. [S.l.], 2018. p. 9–14.

APÊNDICE A – DATASET DE FEEDBACK

A.1 DOCUMENTO PARA ANOTAÇÃO DO DATASET

Figura 28 – Documento para ajudar na anotação do dataset.



1. Ajuda a esclarecer o bom desempenho

- Dar exemplos eficazes de bom desempenho
- Mostrar ao aluno quais são as metas, critérios e objetivos esperados
- Ajudar o aluno mostrando a ele quais são os objetivos esperados, se o aluno atingiu o objetivo...

Exemplo: *Bom trabalho, mas vocês deveriam ter explorado os princípios de Alvaraz y Soler e os relacionados com as observações realizadas.*

2. Facilita o desenvolvimento da auto-avaliação (reflexão) na aprendizagem

- Solicitar ao aluno que compare sua atividade com outros colegas
- Pedir a opinião do aluno em relação ao feedback dado pelo professor
- Solicitar aos alunos que identifiquem os pontos fortes e fracos do seu próprio trabalho

Exemplo: *Você fez um bom trabalho. Contudo, você pode melhorar a descrição sobre o assunto A. Que tal comparar a sua resposta com as dos outros colegas para saber como melhorar?*

3. Fornece informações de alta qualidade aos alunos sobre sua aprendizagem

- Elogios ao lado de críticas construtivas
- Fornecer conselhos corretivos
- Limitar a quantidade de feedback para que seja realmente usado

Exemplo: *Bom trabalho, só atente-se mais na formatação de referências, como é mostrado no fascículo.*

4. Incentiva o diálogo entre professores e colegas em torno da aprendizagem

- Solicitar ao aluno que dialogue com os colegas da turma
- Ajudar o aluno a tirar dúvidas

Exemplo: *O seu trabalho não seguiu o que estava na descrição. Você está com dificuldades na disciplina? Pode me procurar para tirar dúvidas.*

5. Incentiva crenças motivacionais positivas e auto-estima

- Informações sobre progresso e realização
- Elogios
- Não dar feedback com notas

Exemplo: *Parabéns a todos! O trabalho está irrepreensível! Quando o aluno da EaD tem compromisso com os estudos, com as atividades propostas, o resultado positivo sempre aparece.*

Fonte: Elaborada pelo autor (2022)