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

Daniel Cirne Vilas-Boas dos Santos

Estudo comparativo entre abordagens estilométricas e textuais para atribuição de autoria em trabalhos escolares

Recife

2021

Daniel Cirne Vilas-Boas dos Santos

Estudo comparativo entre abordagens estilométricas e textuais para atribuição de autoria em trabalhos escolares

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 Mestre em Ciência da Computação.

Área de Concentração: Inteligência Computacional

Orientador (a): Cleber Zanchettin

Recife

2021

Catalogação na fonte
Bibliotecária Nataly Soares Leite Moro, CRB15-861

S237e Santos, Daniel Cirne Vilas-Boas dos
Estudo comparativo entre abordagens estilométricas e textuais para atribuição de autoria em trabalhos escolares / Daniel Cirne Vilas-Boas dos Santos. – 2021.
153 f.: il., fig., tab.

Orientador: Cleber Zanchettin.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2021.
Inclui referências.

1. Inteligência computacional. 2. Estilometria. 3. Atribuição de autoria. 4. Classificação de atividades pedagógicas. I. Zanchettin, Cleber (orientador). II. Título

006.31 CDD (23. ed.) UFPE - CCEN 2021 – 169

DANIEL CIRNE VILAS-BOAS DOS SANTOS

“Estudo Comparativo entre Abordagens Estilométricas e Textuais para Atribuição de Autoria em Trabalhos Escolares”

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

Aprovado em: 13/08/2021.

BANCA EXAMINADORA

_____

Profa. Dra. Flávia de Almeida Barros
Centro de Informática / UFPE

_____

Prof. Dr. George Gomes Cabral
Departamento de Estatística e Informática / UFRPE

_____

Prof. Dr. Cleber Zanchettin
Centro de Informática/ UFPE
(Orientador)

Dedico este trabalho à minha família. Em especial a meu pai e minha mãe, que me forneceram condições ideais para meu desenvolvimento acadêmico que culmina neste mestrado.

Também dedico este trabalho a minha companheira Camila, que me ajudou bastante e sempre me apoiou ao longo deste curso. Tenho certeza que ela comemora essa realização com a mesma intensidade que eu.

Agradeço a todos os meus amigos e colegas de trabalho, que me apoiaram de inúmeras formas ao longo deste dura jornada, permitindo que o caminho fosse menos árduo.

Por fim, agradeço àquele que está acima de tudo e todos, e misteriosamente parece conduzir o rumo de nossas vidas de maneiras inesperadas. Obrigado por me dar forças para atravessar esta longa jornada.

AGRADECIMENTOS

Gostaria de agradeçer primeiramente ao professor Cléber Zanchetin por ter me aceitado como seu orientando de braços abertos num momento crítico durante o percurso do mestrado. Seus ensinamentos, pensamento crítico e abertura foram fundamentais para conclusão deste trabalho. Também agradeço à professora Flávia Barros, que abriu a portas dessa jornada e compartilhou ensinamentos relevantes em versões iniciais desta pesquisa.

Agradeço ao Centro de Informática e à UFPE por forneceram um excelente curso de mestrado, amplo acesso a materiais didáticos e infra-estrutura, e flexibilidade para entrega deste trabalho.

Agradeço também aos professores parceiros, que disponibilizaram material investigativo que originou a pesquisa.

Por fim, agradeço a todos meus amigos e colegas de trabalho e mestrado, que exerceram contribuições e ajudaram no debate de ideias e confecção dessa dissertação.

RESUMO

O aumento no volume de documentos digitais associado ao seu uso em várias áreas de conhecimento demandam recursos computacionais para sua compreensão e análise. Em casos de verificação ou atribuição de autoria, é necessário confirmar ou identificar os autores do texto. A literatura propõe promissoras abordagens que associam aprendizagem de máquina e processamento de linguagem natural para distinguir os autores pelo seu estilo de escrita. Estes trabalhos envolvem majoritariamente contextos literários ou jornalísticos e textos em inglês. Por outro lado, no contexto educacional, poucos trabalhos exploram a análise de autoria como ferramenta de apoio durante a verificação de aprendizagem, especialmente na língua portuguesa. Tal cenário é desafiador, pois apresenta um baixo volume de documentos por autor, um conjunto de autores com estilo de escrita homogêneo e restrições de formato, tema e idioma. Este trabalho explora técnicas e abordagens reconhecidas na literatura, como modelos de aprendizagem de máquina, técnicas para representação de documentos e extração de características estilométricas, com propósito de apoiar a análise de autoria em uma base de dados composta por atividades pedagógicas de estudantes de graduação. Devido ao baixo volume de exemplos, utilizamos bases de dados jornalísticas mais robustas como referência. Por meio dos experimentos, foi verificado que em domínios restritos, representações baseadas em características de estilo são superiores às abordagens meramente textuais, que sofrem maior influência do tópico em *corpora* mais abrangentes. Este trabalho revelou que o modelo *Extremelly Randomized Trees* foi superior na atribuição de autoria aos demais modelos, (como *Naive Bayes*, SVM, *Random Forest*, Regressão logística e Redes neurais) em todas as bases utilizadas, alcançando uma média de 70% de taxa de acerto e AUC 0,81. Além disso, o trabalho detalha sua metodologia para extração de características de estilo por meio do processamento de linguagem natural e quais destas mais se destacaram durante os experimentos de acordo com seus valores Shapley.

Palavras-chaves: estilometria; atribuição de autoria; classificação de atividades pedagógicas; extração de características estilométricas.

ABSTRACT

The growth of digital documents, associated with their usage in several knowledge areas requires computational resources for its comprehension and analysis. In authorship attribution and verification cases, it is crucial to verify or identify the documents' authors. The literature proposes promissing approaches that associate machine learning and natural language processing to distinguish the authors by their writing style. Those studies mainly involve literary and journalistic contextS, and texts in English. On the other hand, in the educational context, small amount of research explored authorship analysis to support learning checks within the Portuguese language. Such scenario is challenging, because it has a lower volume of documents per author, a set of homogeneous authors, and restrictions in the formatting, theme, and idiom. This work explored known techniques and approaches from the literature, such as ML models, document representation techniques, and stylometric feature extraction to help authorship analysis in a dataset derived from this research composed of pedagogical activities done by undergraduate students. Due to the sample volume, we used more robust journalistic datasets as references. Throughout the experiments, we verified that stylometric representations overcome merely textual representations in restricted domains, who suffer greater impacts from the document subject in broader corpora. This study reveals that Extremely Randomized Trees are superior to the others models (Naive Bayes, SVM, Random Forest, Logistic Regression, Neural networks) for all the datasets used, reaching an average of 70% of accuracy and 0.81 AUC. Furthermore, this survey describes methodological steps for stylometric feature extraction through natural language processing, and which features were highlighted during the experiments according to Shapley values.

Keywords: stylometry; authorship attribution; scholar document classification; stylometric features extraction.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A comunicação é o principal meio pelo qual as informações são transferidas. As principais formas de comunicação são a oral, simbólica e escrita (BLIKSTEIN, 1985). A sociedade e a ciência exerceram forte influência no modo como a comunicação evoluiu. Nos últimos anos, com o desenvolvimento avassalador dos computadores, *smartphones* e da internet, a comunicação escrita migrou da tinta dos livros para *pixels* nas telas. Na verdade, a revolução tecnológica não apenas influenciou a escrita, mas também a forma das pessoas se relacionarem em segmentos como saúde, educação e economia.

No âmbito educacional, há um amplo crescimento no número de instituições de ensino e estudantes, bem como a disponibilidade de informações na internet. Isso está vinculado ao aumento populacional, desenvolvimento tecnológico e aderência a modalidades de Educação a Distância (EaD) que permitem que pessoas em localidades afastadas dos centros urbanos tenham acesso à educação personalizada ou coletiva, funcionando como válvula de escape em tempos de pandemia e inchaço urbano (ALBINO; AZEVEDO; BITTENCOURT, 2020).

Apesar das vantagens proporcionadas pela aderência da tecnologia na educação, também existem pontos de atenção causados por essa inserção. Neste trabalho abordamos alguns desses pontos, que são a ausência de meios para assegurar que uma atividade foi realmente escrita pelo autor designado e a reprodução de informações pré-existentes na internet (plágio) durante a realização de atividades de verificação da aprendizagem (MAURER; KAPPE; ZAKA, 2006).

Infelizmente, o número de práticas ilícitas como a venda de trabalhos acadêmicos e a divisão paralela de atividades entre alunos têm se tornado muito comuns (CURTIS; TREMAYNE, 2019) (SINGH; REMENYI, 2016). Os recursos na *web* são excelentes fontes para pesquisa, porém compreender as informações e retratá-las com seu próprio entendimento é uma importante parte do processo de aprendizado. A facilidade de comprar ou reproduzir conteúdos, além de sedutora, é difícil de ser identificada durante a correção das atividades pelo avaliador, o que prejudica professores, instituições de ensino e os próprios alunos.

Dentro da Ciência da Computação existem áreas de estudo voltadas para a compreensão e classificação automática de documentos utilizando Processamento de Linguagem Natural (PLN) e Aprendizagem de Máquina (AM). O aumento quantitativo de documen-

tos textuais digitais reforça casos de uso da classificação de documentos, que pode abarcar aspectos como tópico, idioma, autoria, sentimento e legibilidade (BEKKERMAN; GAVISH, 2011) (CHASKI, 2005).

A análise de autoria, uma subárea da classificação de documentos, teve seu objeto de estudo modificado em conjunto com a evolução da sociedade. Esta foi inicialmente utilizada para textos literários, como livros e cartas, posteriormente intensificou-se para textos jornalísticos e artigos científicos e mais recentemente em textos curtos, como e-mails, fóruns e postagens na internet (LAGUTINA et al., 2019). Atualmente, contribui para a computação forense, combate ao plágio e solução de casos com autoria contestada ou anônima (FRANTZESKOU et al., 2006) (VARELA, 2017).

A correlação dos pontos de atenção sobre o uso massivo de tecnologia para fins educacionais nos leva à primeira questão que nos propomos a responder neste trabalho: *É possível criar mecanismos de averiguação de autoria dentro do contexto educacional a partir da análise de autoria?*

Na análise de autoria, se destacam as tarefas de atribuição e verificação de autoria, que validam ou identificam a autoria a partir do estilo de escrita. As abordagens mais utilizadas em problemas semelhantes envolvem uso de sistemas baseados em regras ou algoritmos de AM com apoio do PLN (TEMPESTT et al., 2017).

A aplicação de AM em atividades de autoria pode ser fundamentada tanto no conteúdo textual dos documentos, como em representações numéricas dos mesmos. As representações numéricas têm propósitos mais específicos, como o de retratar características relativas ao estilo de escrita. Abordagens baseadas em texto são eficazes em bases de dados amplas e diversificadas, pois o vocabulário e a frequência de palavras é suficiente para distinguir não só a autoria, mas também o assunto dos documentos (GAMON, 2004). Esse fenômeno também pode ser visto como um enviesamento, pois os classificadores conseguem solucionar um problema de autoria a partir do assunto dos textos, o que é suficiente em casos gerais, mas pode falhar em atividades que tratam de único assunto.

Por outro lado, a estilometria defende o uso de métricas para quantificar e definir o estilo de escrita. Segundo Stamatatos (2009) e Varela (2017), cada autor possui um estilo único de escrita, que é composto por múltiplos fatores, como os vícios de linguagem, uso e construção de palavras, sentenças e parágrafos, pontuação, legibilidade, concordância e riqueza de vocabulário. Essa abordagem é voltada especificamente para a captura de estilo de escrita, o que pode ser benéfico em domínios restritos, entretanto a construção de tais

características exige esforço humano para extração, construção e avaliação das métricas obtidas. Além disso, a seleção de características é apontada como um dos maiores desafios da estilometria, pois não há consenso sobre quais são mais efetivas, variando bastante de acordo com o problema tratado e o classificador utilizado (NEAL et al., 2017) (JUOLA, 2007).

Tais discussões abrem um leque de opções para a análise do problema e nos incentivam a explorar as técnicas dispostas na literatura, incluindo representações textuais e numéricas, uma vasta quantidade de classificadores, desde os tradicionais até os mais recentes, baseados em aprendizagem profunda. Propomo-nos a comparar diferentes abordagens e classificadores durante os experimentos, para compreender as nuances de cada um e traçar paralelos entre o problema e a área de pesquisa com intuito de responder à nossa segunda questão de pesquisa: *Qual a combinação de técnicas de representação e classificadores que é mais eficaz na resolução de atividades de análise de autoria no domínio do problema de pesquisa?*

Os principais desafios na identificação de autores neste estudo são: (*i*) o pequeno número de documentos por autor, tanto na nossa base de dados como em cenários cotidianos, pois é provável que só exista um volume significativo de trabalhos para cada estudante do meio para o final do curso, considerando possível a integração de atividades entre as disciplinas cursadas; (*ii*) domínio restrito ao conteúdo do curso ou disciplina, que colabora com a presença de muitos documentos similares, tanto a nível de vocabulário como no campo das ideias, dificultando distingui-los; e (*iii*) uso indiscriminado de ferramentas de busca online, que leva a construção de textos compostos por excertos de frases e parágrafos de outros autores, dificultando a construção e análise de estilo.

Como alternativa para tais limitações, propomos uma metodologia que inclui a obtenção de bases comparativas compostas por textos jornalísticos. Ao usar essas bases, aumentamos a representatividade dos exemplos, confiança na originalidade dos documentos e viabilidade para experimentos com cenários de domínio restrito ou abrangente mediante agrupamento dos documentos por tema. Assim, pretendemos responder nossa última questão de pesquisa: *Existe unanimidade sobre quais características de estilo são mais importantes para verificação de autoria variando o contexto e o volume de dados? Além disso, será que a partir dessas características podemos identificar padrões de comportamento específicos de cada grupo de autores?*

Avaliamos o presente trabalho como inovador numa perspectiva investigativa, pois não

foram encontrados outros trabalhos na literatura com foco na análise de autoria no nicho educacional dentro da língua portuguesa. Apesar das semelhanças, a pesquisa se distingue de estudos relacionados, como o de Varela (2017) pelo seu nicho e natureza do *corpora*, assim como de Stavngaard et al. (2019) pelo idioma, classificador e tamanho do *dataset*. Ressaltamos que, apesar da conexão com estudos de detecção de plágio, este trabalho se distingue por não utilizar mecanismos de recuperação da informação, identificação de paráfrases ou similaridade textual (CLOUGH et al., 2003), uma vez que seu propósito é a análise de autoria por meio do estilo de escrita.

1.1 HIPOTÉSES

A partir das questões de pesquisa e levantamento bibliográfico estabelecemos nossas conjecturas. Nossa primeira hipótese sugere ser possível construir modelos de AM baseados em estilometria, com capacidade para identificar corretamente a autoria de documentos escritos por estudantes em atividades pedagógicas.

A segunda hipótese é que o uso de características de estilo é mais eficaz que abordagens baseadas em texto quando utilizadas em atividades de atribuição de autoria dentro do contexto do problema proposto.

Por fim, acreditamos que as características estilométricas de maior influência sobre os classificadores são uma alternativa interessante para conhecer melhor os autores e suas obras, não se limitando ao estilo de escrita, considerando também a composição e a estrutura dos documentos.

1.2 OBJETIVOS

Essa dissertação tem como objetivo geral investigar, discutir e propor soluções para atividades de atribuição de autoria, com foco no processo de ensino e verificação de aprendizagem, utilizando a língua portuguesa. O alcance desta meta está condicionado ao cumprimento dos seguintes objetivos específicos:

- Construir e adquirir bases de dados textuais que possam ser utilizadas em atividades de atribuição de autoria no contexto do problema de pesquisa;

- Implementar um sistema para extração de características de estilo a partir de documentos textuais;
- Investigar, executar e avaliar algoritmos e técnicas para resolução de atividades de atribuição de autoria;
- Comparar o uso de características de estilo a abordagens textuais em atividades de atribuição de autoria dentro do domínio proposto;
- Relacionar os resultados obtidos em bases de dados dentro e fora do contexto educacional, visando identificar características de estilo unicamente relacionadas aos estudantes;
- Colaborar com o estado da arte por meio dos resultados e discussões acerca do problema de pesquisa.

1.3 ESTRUTURA DA DISSERTAÇÃO

No Capítulo 2 encontramos as principais definições e fundamentação teórica que embasam esse trabalho. Partimos de uma breve discussão sobre o desenvolvimento da linguagem escrita até as necessidades contemporâneas, que se beneficiam de recursos de AM e PLN. Exploramos os principais recursos de PLN e nos debruçamos sobre uma variedade de classificadores de AM, desde abordagens tradicionais até as mais modernas, como a aprendizagem profunda. Logo depois, estudamos a associação entre as duas áreas de estudo para resolução de atividades relacionadas. No fim do capítulo, discutimos sobre análise de autoria e o estudo das características de estilo (estilometria).

Em seguida, no Capítulo 3 discutimos sobre conceitos dentro da análise de autoria, tal como verificação, atribuição e *profiling*. Além disso, trouxemos diversos trabalhos relacionados demonstrando as principais estratégias e resultados para solução de tais problemas.

No Capítulo 4 está descrita a metodologia da pesquisa, passando pela investigação inicial do problema, obtenção e construção das bases de dados, transformação dos documentos e extração de *features* estilométricas. A construção de cada grupo de características é detalhada em subseções. Encerramos com uma breve descrição da estrutura dos experimentos.

O Capítulo 5 compreende os experimentos desta pesquisa distribuídos por etapas. Iniciamos pela visualização dos dados auxílio de análise de agrupamento (*clustering*). Em seguida, temos a aplicação de classificadores de AM diante das bases de dados e suas representações. Descrevemos os critérios usados durante a avaliação e otimização dos classificadores, resultando na definição das soluções propostas. Concluímos com um estudo de interpretabilidade das soluções.

Por fim, no Capítulo 6 encerramos o trabalho trazendo as considerações finais. Também ilustramos as principais contribuições da dissertação, assim como suas limitações. Abrimos espaço para extensão desta pesquisa, sugerindo oportunidades de trabalhos futuros para complementar tais limitações.

2 REVISÃO DA LITERATURA

O propósito deste capítulo é apresentar as bases teóricas sobre o assunto, apresentando um breve histórico da evolução da comunicação e linguagem escrita, discutir sobre o Processamento de Linguagem Natural e a classificação de documentos, além de realizar um levantamento do estado da arte na área de identificação de autoria e estilometria.

2.1 DEFINIÇÕES

Para melhor compreendermos como se dá a comunicação humana nos dias atuais, é necessário primeiramente entender o que é a comunicação, linguagem e escrita, assim como seu desenvolvimento ao longo do tempo.

Existe uma variedade de conceitos sobre o que é a comunicação, que pode ser vista por diferentes perspectivas. De acordo com definições formais "1. ato de comunicar; informação, aviso; 2. passagem, caminho, ligação". (ROCHA; PIRES, 1996). Na perspectiva etimológica, a palavra vem do latim *'communis'*, comum, o que introduz a ideia de comunhão, comunidade" (MELO, 1975).

Já na visão linguística, acredita-se que o homem é um ser social, e que a comunicação sustenta a sociabilidade, por meio da qual os homens interagem, possibilitando a transmissão de experiências, conhecimentos e apelos (MELLO, 1973).

Partindo do que entendemos como comunicação, podemos expandir o conceito para a compreensão do que é linguagem, que nada mais é do que a forma como a comunicação é realizada, seja ela escrita, falada, gestual ou de programação. Desta forma, a escrita é um tipo de linguagem que permite a comunicação entre seres humanos.

2.2 DO SURGIMENTO DA ESCRITA ATÉ AS NECESSIDADES ATUAIS

A partir do surgimento do ser humano, no período pré-histórico, acredita-se que já existiam formas iniciais de comunicação. Ainda que muito cedo para a comunicação escrita, historiadores afirmam que poderia existir comunicação primordialmente de maneira visual (PELTZER; SILVA; RIBEIRO, 1991).

Acredita-se que inicialmente o homem comunicava as atividades cotidianas na mesma

ordem em que elas aconteciam, como o passo a passo de seu dia ou uma rotina de caça por meio de gestos e pinturas. Alguns vestígios desta época são os pictogramas, ou pinturas rupestres, que retratam os acontecimentos dos tempos pré-históricos.

Considera-se que o primeiro conjunto de símbolos surgiu no Egito, há cerca de 3.000 anos antes de Cristo. Descobertas arqueológicas revelaram a existência de desenhos e gravuras, conhecidos como hieróglifos, que eram capazes de expressar de maneira mais elaborada a sociedade egípcia (PERLES, 2007). Vestígios semelhantes também foram encontrados em outras sociedades históricas, como a asteca e mesopotâmica.

Entretanto, foi apenas no século IV antes de Cristo que uma das maiores invenções da humanidade foi criada, a escrita. Segundo Sampson (1996), a escrita surgiu muito depois da linguagem, no período chamado de "revolução neolítica", e que apresenta três fases não cronológicas: pictórica, ideográfica e alfabética.

A fase pictórica se refere aos desenhos, que estão mais associados à imagem do objeto que se deseja representar. É o caso da linguagem pré-histórica e alguns povos antigos. Na fase ideográfica encontramos os ideogramas, que são símbolos que representam ideias no lugar de sons ou imagens. Neste grupo estão as placas de trânsito atuais, hieróglifos e a comunicação chinesa (ANDRADE, 2010). Já fase alfabética se distingue pelo uso de letras, que surgiram a partir de ideogramas, mas perderam esse valor ideográfico para assumir função fonográfica por meio da escrita.

Alguns autores corroboram com a ideia de que a escrita alfabética foi uma "descoberta" do ser humano, que passou a fazer associações entre sons e símbolos para construir palavras, frases e ideias (ANDRADE, 2010). Especialistas também afirmam que é nesse momento que se inicia a história, visto que antes disso a maioria das assertivas é cabível de interpretação e bastante genérica (GONTIJO, 2004).

A escrita surgiu do entendimento humano de que palavras e nomes podiam ser compostos por unidades sonoras mínimas, os fonemas, e por meio da combinação de fonemas era possível representar objetos e coisas. A escrita fonográfica era constituída por símbolos gráficos que representavam os sons; estes, combinados em sequências de tamanho variável, eram capazes de descrever não apenas objetos, mas também ideias e ações (PERLES, 2007).

Os símbolos, ou signos gráficos da escrita fonográfica, foram posteriormente fragmentados em unidades sonoras ainda menores, as letras. A partir das letras, consequentemente surgiram os alfabetos que passaram por um longo processo de evolução até os moldes

atuais. Segundo Gontijo (2004), primeiro surgiram silabários, que eram conjuntos de sinais para representar sílabas. Os silabários evoluíram com o tempo para alfabetos mais conhecidos por nós, como o romano e greco-latino.

Enquanto a linguagem escrita se desenvolvia, algumas outras descobertas do homem facilitaram sua disseminação. O papel, por exemplo, substituiu a pedra e facilitou o manuseio e disseminação da linguagem escrita. Entretanto, mesmo após o surgimento das letras e dos alfabetos, grande parte da população não tinha acesso à linguagem escrita e permanecia utilizando a linguagem oral e visual. A linguagem escrita era restrita aos mais ricos, pessoas letradas e membros da igreja, e assim permaneceu por toda a idade antiga (4000 a.C a 476 d.C) e idade média (476 d.C a 1453 d. C) (NUNES, 1979).

Posteriormente, no século XV, as primeiras máquinas de tipografia e prensa foram criadas por Johann Gutenberg, isso possibilitou a produção em larga escala de livros e também o surgimento do jornal impresso, primeiro meio de comunicação em massa que possibilitou a democratização da linguagem escrita (GONTIJO, 2004).

Nos anos seguintes a esse período, o que se observou foi a expansão da linguagem escrita, uma vez que ela já havia se tornado parte intrínseca da interação humana. Além dos fatores históricos previamente citados, também houve uma adoção da leitura como alicerce do processo de educação, tornando-se fundamental que as pessoas soubessem ler e escrever para gozar de todos os seus direitos como cidadão, pois a escrita se faz presente em todos os aspectos da interação humana.

A invenção de mecanismos para geração de energia , por volta de 1880, foi um dos fatores propulsores não só da escrita, mas do potencial de produção humana como um todo. Com a energia elétrica, houve a revolução industrial e a humanidade foi capaz de fornecer a energia necessária para alimentar máquinas para as mais distintas tarefas (FARIAS; SELLITTO, 2011). Algumas dessas máquinas acabariam por influenciar completamente a forma dos humanos se comunicarem.

Ao nos aproximarmos de anos mais recentes, temos o surgimento das primeiras "máquinas de computação". Acredita-se que inicialmente seu uso era militar e que foram bastante utilizadas durante a segunda guerra mundial para tentar decifrar mensagens de países inimigos. Tratavam-se de máquinas enormes movidas a válvulas. A partir de 1945, diversos pesquisadores, como Alan Turing, John von Neumann e Arthur Burks, se dedicaram ao desenvolvimento e aprimoramento dessas máquinas (ROJAS; HASHAGEN, 2002).

Por volta de 1954 a IBM conseguiu transformar o computador num produto comercial,

que ainda estava longe de ser uma máquina pequena e fácil de utilizar. Encabeçado pela IBM, Microsoft e Apple, o conceito de computadores pessoais, que pudessem ser usados por pessoas que não fossem especialistas, surge na década de 80. Isso só foi possível devido a diversas outras invenções, como os microprocessadores, sistemas operacionais, interfaces gráficas (GUI) e periféricos como teclado e mouse.

Nosso último marco histórico, antes de adentrarmos em como a linguagem escrita se dá no século XXI, é a invenção da internet. A internet também foi construída com fins militares (LEVINE, 2018) e pode ser definida como uma rede composta por todos os dispositivos conectados a ela, estes são capazes de se comunicar por meio de protocolos bem estruturados (RYAN, 2010). As informações são empacotadas e então transferidas por meios físicos, como cabos coaxiais, fibra óptica ou sinais de onda rádio até seu endereço de destino. Isso permitiu quebrar barreiras para troca de informação e comunicação em escala mundial.

A evolução dos computadores continua acontecendo, aumentando seu poder de processamento e armazenamento, enquanto diminuem de tamanho. Telefones celulares se transformaram em *smartphones*, ou computadores pessoais de bolso; geralmente conectados por meio da internet, e que entre suas funcionalidades, está a realização de chamadas telefônicas. Estima-se que existam 3.5 bilhões de *smartphones* no mundo (STATISTA, 2020), o que representa aproximadamente 30% da população mundial em 2020 (WORLDOMETERS, 2020).

Importante citar que a evolução dos computadores até os modelos atuais é muito maior e repleta de detalhes técnicos do que foi discutido acima, mas foge do escopo desta pesquisa.

O uso massivo de dispositivos computacionais tem exercido forte influência sobre muitos aspectos da sociedade, dentre estes, a comunicação. Esta se tornou instantânea, global e com uma forte tendência migratória dos meios físicos para os digitais, do papel para as telas.

Fazendo-se valer da necessidade de comunicação para a interação humana e a onipresença de dispositivos computacionais, é inevitável que a linguagem escrita, composta por palavras, frases e textos atinja valores incomensuráveis. Essas palavras representam ideias em várias áreas do saber para diversos fins, incluindo a execução dos direitos básicos de todos os cidadãos: educação, saúde, trabalho, segurança, lazer e transporte (BRASIL, 1988).

Portanto, podemos observar que além da maior adoção da comunicação digital para manutenção das atividades que compõem a sociedade, a quantidade de textos continuará aumentando em proporções humanamente impossíveis de serem processadas, exigindo a criação de meios computacionais para a extração, processamento, compreensão e execução deste conteúdo por meio dos computadores.

Por isso, ao longo das últimas décadas vemos grandes avanços em áreas da computação que estão relacionadas as novas necessidades. Dentre estas, podemos citar a área de extração de informação, para aquisição e organização das informações disponíveis na internet, que podem estar num formato estruturado ou não; O Processamento de Linguagem Natural, para o tratamento de textos e compreensão da linguagem humana do ponto de vista computacional, e a Aprendizagem de Máquina, para tentar observar padrões, auxiliar na compreensão dos dados e apoiar a tomada de decisão por meio de modelos matemáticos e estatísticos sobre um conjunto de dados.

O processamento e compreensão de textos de maneira automática se tornou primordial para atender as necessidades de uma sociedade que segue rapidamente num caminho digital. Mesmo observando esforços interdisciplinares para alcançar esses objetivos, são necessários maiores investimentos nos âmbitos científico, educacional e de engenharia para construção de mecanismos capazes de atender tanto às demandas atuais, como aquelas que surgirão nos próximos anos. Neste trabalho nos propomos a explorar problemas reais que se enquadram no cenário descrito acima, com o intuito contribuir com a área por meio duma análise investigativa sobre o problema, possíveis soluções, experimentos e resultados encontrados.

2.3 PROCESSAMENTO DE LINGUAGEM NATURAL

A área de pesquisa de Processamento de Linguagem Natural (PLN) (JURAFSKY; MANNING, 2012) se propõe a entender como os computadores podem se tornar capazes de compreender e manipular a linguagem, seja ela em forma de texto ou áudio, para realização de diversas tarefas. Para isso, os pesquisadores buscam entender como a linguagem se estrutura e como os humanos fazem para se comunicar, baseando-se em diversas áreas do conhecimento, como Linguística, Psicologia, Ciência da Computação e Inteligência Artificial (IA).

Dentro do PLN, as palavras são chamadas de *tokens* e a sua combinação permite a

criação de frases. O conjunto de frases compõe um texto, que também pode ser chamado de documento, e o grupo de documentos constitui um *corpora*.

As principais abordagens utilizadas nas tarefas de PLN podem se basear em: *i*) métodos estatísticos baseados em corpus, *ii*) métodos baseados em sistemas de referência léxica, como o *WordNet* (MILLER et al., 1988), um dos maiores bancos de dados de léxicos multi-idioma, *iii*) métodos baseados em expressões regulares e máquinas de estado e *iv*) métodos baseados na engenharia, análise e geração de regras gramaticais, como o Lingo (http://lingo.stanford.edu/) e o *Linguistic Data Consortium* (http://www.ldc.upenn.edu/) (CHOWDHURY, 2003).

Podemos dividir a compreensão da linguagem a nível computacional em três níveis. Primeiramente é preciso entender os *tokens* que compõem aquela frase e seu papel morfológico, como por exemplo a classe gramatical. Em seguida, expandimos a análise a nível de frase, buscando compreender o papel que cada um daqueles *tokens* representa dentro da mesma e as principais entidades envolvidas. Por fim, através da compreensão de um conjunto de frases, é possível entender melhor em que contexto se inserem todas os *tokens* e frases a nível de documento, e que ideias e sentimentos o mesmo expressa (CHOWDHURY, 2003).

De acordo com Liddy (1998) e Feldman (1999), existem níveis independentes de compreensão que são usados pelos humanos para entendimento do texto, dentre estes, podemos citar: fonético para distinção de pronúncia; morfológicos para tratamento de sufixos e prefixos; léxicos para entendimento de palavras e classes gramaticais; sintático para entendimento da estruturas das frases; semântico para distinção do sentido das palavras dentro de um contexto; e pragmático, baseado em conhecimento prévio sobre o universo (CHOWDHURY, 2003); Dependendo do objetivo a alcançar por meio do PLN, torna-se necessário analisar um ou mais níveis de compreensão.

Ao nos aprofundarmos no PLN, observamos que existe uma série de tarefas específicas que podem ser combinadas para compreensão da linguagem. A execução dessas tarefas pode derivar inúmeras aplicações. Alguns exemplos são: tradução de máquina automática, análise de sentimentos, extração e recuperação da informação, processamento, geração e sumarização de textos e classificação de documentos. Neste trabalho iremos nos aprofundar em um caso específico da classificação de documentos.

2.4 PLN COMO FERRAMENTA DE APOIO PARA CLASSIFICAÇÃO DE DOCUMENTOS

2.4.1 Pré-processamento

Dentro de atividades de classificação de documentos, observa-se que o PLN é utilizado principalmente como ferramenta de apoio na etapa de pré-processamento. O pré-processamento é composto pela extração e seleção de características é sucedida pela transformação e representação do documento (BRÜCHER; KNOLMAYER; MITTERMAYER, 2002a).

O propósito da seleção de características é escolher um conjunto de termos capazes de representar o documento de maneira eficiente. Entretanto, devido ao grande número de palavras que existem e suas diversas formas de flexão, geralmente indicando variações de tempo, grau, número ou gênero, foram criadas técnicas de inflexão e remoção de palavras para diminuir a variabilidade do documento. Algumas das técnicas de inflexão e redução de palavras mais conhecidas são o *stemming* e a lematização.

O *stemming* é uma estratégia mais rígida para redução de palavras que ocorre por meio da remoção de sufixos comuns dentro do idioma (LARKEY; BALLESTEROS; CONNELL, 2002). Na língua portuguesa, temos por exemplo diversas palavras que terminam com "indo", "ando", "ão", "a", "o" e "s". Já no inglês, podemos exemplificar com os verbos que terminam em *"ing"* e adjetivos terminados em *"able"*.

Por outro lado, a lematização utiliza conhecimentos morfológicos para realizar a inflexão de palavras flexionadas para sua palavra de origem, também chamado de lema. Desta forma, é um mecanismo mais robusto que o *stemming* pois não gera palavras que não existem no vocabulário, porém exige maior conhecimento do idioma para ser implementado e por causa disso é menos disponível. Na Tabela 1 temos exemplos que ilustram as diferenças entre as técnicas.

Para redução da quantidade de termos, são frequentemente utilizadas listas de palavras de baixo poder semântico, chamadas de *stopwords*. As listas de *stopwords* geralmente são construídas manualmente, e contam com palavras de classes gramaticais fechadas. Mesmo após sua remoção, o sentido principal da frase deve ser preservado. As *stopwords* da língua portuguesa são compostas principalmente por preposições, artigos e conjunções, que podem ser filtradas do conteúdo textual original durante o pré-processamento (BRÜCHER; KNOLMAYER; MITTERMAYER, 2002b). Algumas das *stopwords* encontradas para portu-

Tabela 1 – Exemplo de palavras do português e inglês por meio das técnicas de *stemming* e lematização

Palavra	*Stemming*	Lematização
quilo	quil	quilo
caminhando	caminh	caminhar
sorrindo	sorr	sorrir
studying	*studi*	*studying*
learned	*learn*	*learn*

Fonte: Elaborada pelo autor através dos resultados obtidos por meio das ferramentas LEMMATIZER e CSTOOLS

guês dispostas na biblioteca NLTK são: "a", "ao", "aquela", "aquele", "dele", "do", "fosse", "houve", "isto", "no", "para", "pelo"e "você".

Uma outra técnica denominada *n-grams* é bastante utilizada tanto para redução da quantidade de termos, como para agrupamento destes. A técnica leva em consideração não apenas *tokens* isolados, mas a combinação sequencial de n termos. Quando n=1, para fins de remoção, não há distinção se comparado a *stopwords*. Na literatura é comum encontrarmos trabalhos que usam *n-grams* num intervalo entre 2 e 5 *tokens*. Por exemplo, dada a frase "ser ou não ser", os $bigrams$=$[' serou', ' ounão', ' nãoser']$ e $trigrams$=$[' serounão', ' ounãoser']$

Apesar de grande parte dos trabalhos que envolvem PLN usarem majoritariamente artifícios de pré-processamento como a tokenização, *stemming* ou lematização e *n-grams*, é importante mencionar que em casos específicos, pode-se optar pela não realização de algumas destas tarefas para preservar o conteúdo original dos documentos.

2.4.2 Anotação sintática e entidades nomeadas

Além de auxiliar no pré-processamento, o PLN também é usado para enriquecimento do conteúdo textual por meio da aplicação de algoritmos e técnicas de rotulação. O principal algoritmo de anotação sintática é o *POS-Tagger* (*Part Of Speech*) que tem como objetivo associar automaticamente cada *token* do documento à sua respectiva classe gramatical, tais como substantivos, verbos e adjetivos (TOUTANOVA; MANNING, 2000). Também existem *taggers* capazes de anotar os *tokens* com maior granularidade, indicando por exemplo tipos de verbos e conjunções e flexão de substantivos.

Este processo de rotulação dos elementos textuais é essencial para compreender a estrutura gramatical de sequências de palavras, sendo útil para atividades de extração de informação, remoção de ambiguidade, sintetização de texto e pesquisa lexicográfica. Diversos autores confirmam que existe uma baixa quantidade de *POS-Taggers* para português do Brasil, devido a alguns fatores como menor quantidade de *corpora*, maior complexidade gramatical e menor número de pesquisas e pessoas falantes do idioma (FILHO, 2006) (PIROVANI, 2019) (ALENCAR, 2010).

Existem três principais estratégias para criação de *POS-Taggers*: *i*) baseado em regras e contexto; *ii*) baseado na estocástica; *iii*) baseado em redes neurais.

As abordagens baseadas em regras podem se valer de regras gramaticais e máquinas de estado com auxílio de amplo armazenamento em memória. São técnicas mais antigas que perderam força ao longo do tempo. Desde 1992, Cutting et al. (1992) já pontuaram que as abordagens estocásticas se mostraram superiores em relação às baseadas em regras (CUTTING et al., 1992). As abordagens estocásticas visam usar modelos probabilísticos para anotar os *tokens* o mais corretamente possível, dentre estas abordagens podemos destacar as baseadas em maximização da entropia (RATNAPARKHI, 1996) e no modelo oculto de Markov(BRANTS, 2000). Em ambos casos, é necessário obter um corpus de treinamento sintaticamente anotado.

O modelo baseado entropia se propõe a encontrar uma distribuição em que a entropia é minimizada, respeitando um conjunto de restrições que podem ser previamente definidas, tal como uma regra que define que um conjunto de palavras deve ser anotado com uma *tag* (TOUTANOVA; MANNING, 2000). Ao fim, é selecionada a distribuição que apresentar maior grau de verossimilhança entre as *tags* atribuídas e o valor-verdade das mesmas levando em conta as restrições. Já o modelo oculto de Markov, que também é probabilístico, leva em consideração principalmente a temporalidade das *tags* dentro das frases, esta abordagem é usada em diversos problemas de reconhecimento de padrões. Sua aplicação para construção de *POS-Taggers* leva em conta a probabilidade conjunta das sequências de *tags* que aparecem dentro do documento, respeitando um tamanho máximo para limitar as sequências (TOUTANOVA; MANNING, 2000).

De maneira similar aos classificadores probabilísticos, ou seja, utilizando aprendizagem supervisionada, também podem ser treinados *POS-Taggers* por meio de diversas arquiteturas de RNAs. Estas redes necessitam de *corpora* mais robustos, visto que as redes neurais, quanto mais profundas, mais dados precisam para generalizar. Em geral,

é realizada alguma operação de *encoding* no texto de entrada para valores numéricos a fim de se tornar compatível com a RNA. Vale ressaltar que existem diferentes abordagens para construção destes vetores, com representações a nível de caracteres, palavras ou documentos.

Mesmo diante de arquiteturas mais simples, como a adaptação da MLP proposta por Milidiú (2016), podemos observar taxas de até 96% de acurácia em atividades de *POS-Tagging* diante de bases de dados relevantes, como a *MAC MORPHO* (TAYLOR; MARCUS; SANTORINI, 2003). Alguns autores afirmam ter alcançado o estado da arte para a atividade de *POS-Tagging* ao usar variações de RNAs de arquitetura profunda, tal como CNN (SANTOS; ZADROZNY, 2014) e LSTM (HUANG; XU; YU, 2015) (REIMERS; GUREVYCH, 2017). Os autores destacam que é possível alcançar ótimas taxas de acerto de maneira direta, ou seja, sem necessidade de processos manuais para extração de características.

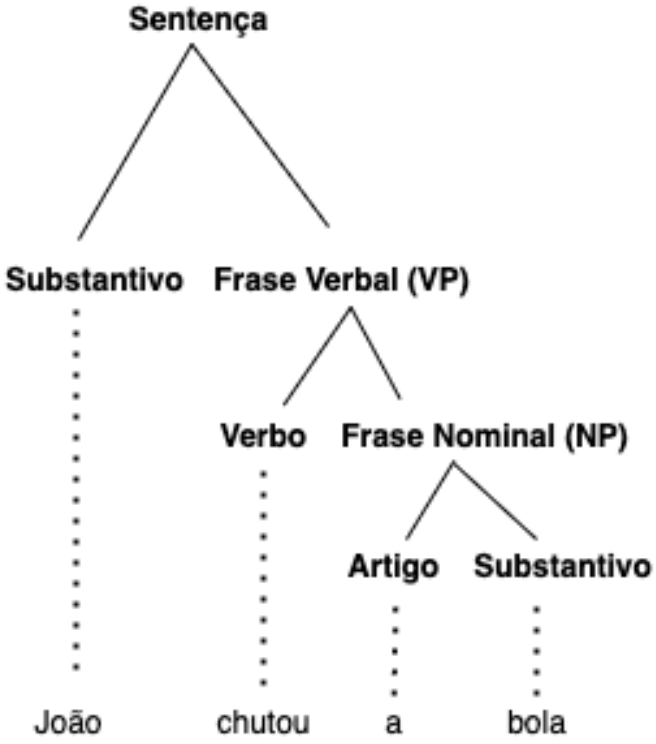
A anotação sintática aumenta o número de possibilidades dentro do PLN, uma vez que diante das classes gramaticais obtidas pela anotação e a separação do texto em frases, é possível inferir mais informações sobre a estrutura das frases. As palavras que compõem as frases neste contexto são chamadas de constituintes e a combinação destes pode representar funções sintáticas específicas numa frase, tal como sujeito, predicado ou sintagmas (BRUCKSCHEN et al., 2008).

Para transformar uma frase em um conjunto de segmentos que não se sobrepõem, é possível aplicar a técnica de *Chunking*. O *Chunking* objetiva analisar a frase e identificar seus constituintes. Esta técnica geralmente se baseia num conjunto de regras gramaticais e expressões regulares ou em modelos estatísticos treinados baseado em *corpora* (RAMSHAW; MARCUS, 1999).

Uma representação comum de frases, constituintes e suas respectivas funções é por meio de uma estrutura de árvore. A árvore é construída de maneira hierárquica, onde na sua raiz encontramos a frase original completa, que vai se expandindo através de nós, associando conjuntos de palavras (*chunks*) aos seus respectivos papéis sintáticos (Figura 1). Existem tarefas dentro do PLN que podem ser resolvidas ao utilizar tal estrutura de dados para busca, casamento de padrões e geração de sub-árvores (NAMIUTI-TEMPONI; COSTA, 2014).

Outro algoritmo que nos permite adquirir mais informações sobre os *tokens* que compõem documentos é o reconhecimento de entidade nomeada (*Named Entity Recognition* - NER). O termo entidade nomeada surgiu por volta de 1996 (GRISHMAN; SUNDHEIM,

Figura 1 – Exemplo de árvore sintática



Fonte: Elaborado pelo autor

1996) e a princípio foi utilizado para auxiliar em atividades de extração da informação. Percebeu-se que era importante para estas atividades a identificação de unidades de informação, tal como nomes de pessoas, lugares e organizações e medidas numéricas, como moedas, datas, e expressões matemáticas (NADEAU; SEKINE, 2007). Um classificador de NER (NERC) tem como objetivo a identificação destas unidades de informação dentro do texto.

De maneira análoga aos *POS-Taggers*, os primeiros sistemas de NER foram criados usando técnicas baseadas em regras manualmente construídas e com o passar do tempo vêm evoluindo para técnicas pautadas em AM. As abordagens fundamentadas em regras utilizam padrões ortográficos, regras gramaticais e recursos específicos da linguagem, tal como dicionários (*gazetteers*) compostos por listas de palavras e suas respectivas entidades nomeadas. A *tag* à qual uma palavra está mapeada dentro do dicionário pode variar de acordo com o contexto e idioma em questão. Apesar de bastante eficaz em determinados problemas, a criação de regras e dicionários específicos para diversos idiomas e domínios esbarra na demanda de tempo para sua construção. Por isso, tem se tornado cada vez mais comum a utilização de algoritmos de aprendizagem para reconhecimento de entidades

nomeadas (LAMPLE et al., 2016).

Nadeau e Sekine (2007) observou que, a partir dos anos 2000, houve um aumento substancial no número de algoritmos de NERC baseados em técnicas de aprendizagem em comparação às soluções baseadas em regras dentro da competição MUC-7(*Messaging Understanding Conference*). Essas abordagens visam substituir a criação de regras manuais por um processo de indução automático a partir de exemplos rotulados. Esse processo de indução visa observar as características dos exemplos positivos e negativos associados a determinada entidade nomeada (classe) e desenvolver regras que possam capturar novas instâncias dentro de uma das classes. Na literatura é possível encontrar diversos autores que realizaram NERC usando desde modelos simples, como Modelos Ocultos de Markov, AD, SVM, Naive Bayes multinomial até soluções mais complexas, como redes neurais e modelos estatísticos baseados em *Conditional Random Fields*(CRF) (LAFFERTY; MCCALLUM; PEREIRA, 2001).

Conforme dito anteriormente, o principal requisito para abordagens supervisionadas é um corpus anotado e representativo, o que nem sempre está disponível. Por causa disso, também existem NERC desenvolvidos usando técnicas de aprendizagem não supervisionadas. Destacamos trabalhos que aplicaram técnicas de agrupamento para organizar palavras a partir das suas características léxicas em busca de possíveis entidades nomeadas (SHINYAMA; SEKINE, 2004).

O estado da arte para NERC é composto por redes neurais profundas, onde alguns autores utilizaram CNN e outros utilizaram variações da LSTM, tal como sua versão bidirecional, baseada em pilha ou associada a camadas de aleatoriedade condicionais (LSTM-CRF) (HUANG; XU; YU, 2015). Estudos reportam que após diversos ajustes na representação do texto na camada de entrada e estrutura das redes neurais é possível obter taxas de medida $f1$ (CHICCO; JURMAN, 2020) acima de 0,85 diante de bases de dados reconhecidas como CoNLL 2002 e CoNLL 2003 (LAMPLE et al., 2016).

No que se refere ao idioma português, uma das bases mais conhecidas é o HAREM, que foi inicialmente desenvolvido para competições de NER em 2004 (SANTOS; ZADROZNY, 2014) que posteriormente teve uma segunda versão publicada em 2008 (MOTA; SANTOS, 2008). Na versão mais recente, existem 9 entidades principais (categorias) que se dividem em mais de 43 entidades secundárias (tipos).

2.4.3 Representação de documentos

Após o pré-processamento, é bastante comum ocorrer a representação do conteúdo textual dos documentos para um formato inteligível pelos computadores e algoritmos de AM. Essa etapa visa maximizar características com maior importância, e minimizar as menos importantes. Isso geralmente se dá de maneira estatística, com exceção de sistemas baseados em conhecimento.

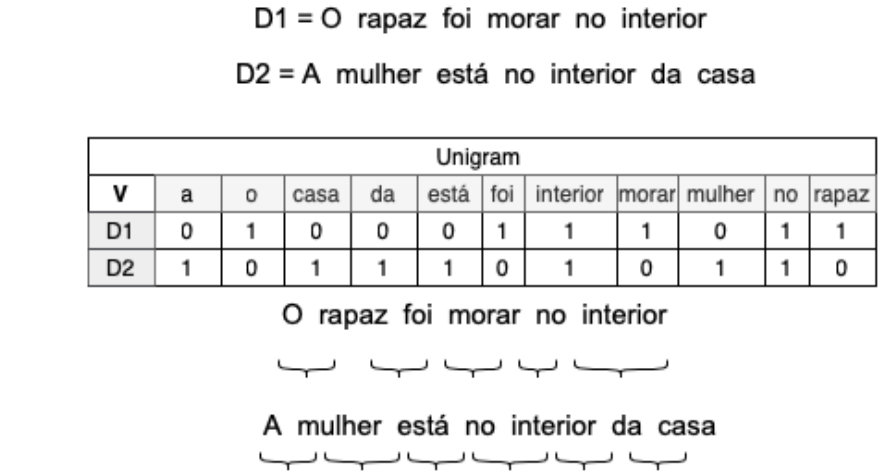
2.4.3.1 Bag of Words

Uma das principais técnicas de representação de texto é o *Bag of Words* (BoW). O termo se refere a um "saco de palavras" pois não considera a ordem de aparição das mesmas, mas apenas seu número de ocorrências no documento. O tamanho desses vetores geralmente é fixo e igual ao tamanho do vocabulário $|V|$, que é a quantidade de palavras únicas que compõem o *corpora* (GOLDBERG, 2017). A incidência e ausência de termos dentro do documento pode ser representada de várias maneiras, assim como a contagem de aparições dos termos ou a presença ou ausência. Na abordagem binária cada documento é representado por um vetor de tamanho $|V|$ e cada índice deste vetor representa uma palavra do vocabulário. Se a palavra está presente naquele documento, o valor associado é 1, caso contrário é 0. Como consequência, os vetores gerados são bastante esparsos devido a inferioridade de vocabulário ao compararmos qualquer documento isoladamente com o vocabulário do *corpora*. Isso significa que temos vetores com mais indicações de ausência (0's) do que presença de termos (1's).

Uma outra abordagem para construção de BoW se baseia em *n-grams*. Isso nos permite representar não apenas um *token* (unigram) isolado, mas uma sequência de n *tokens* contíguos, sendo os mais comuns de tamanho 2 (*bigram*), 3 (*trigram*) e 4 (*4-gram*) (Figura 2).

Uma deficiência dessas abordagens é que elas utilizam valores absolutos para representar *tokens* ou *n-grams*, porém podem existir termos que ocorrem com maior frequência e se sobrepõem sobre palavras menos frequentes. Por causa disso, a técnica de frequência de termos (*Term Frequency* - TF) e frequência inversa do documento (*Inverse Document Frequency* - IDF), oriunda da área de recuperação da informação (RI), é comumente aplicada na construção destes vetores.

Figura 2 – Ilustração do funcionamento do BoW utilizando unigram e bigram



Unigram											
V	a	o	casa	da	está	foi	interior	morar	mulher	no	rapaz
D1	0	1	0	0	0	1	1	1	0	1	1
D2	1	0	1	1	1	0	1	0	1	1	0

Bigram										
V	a mulher	da casa	está no	foi morar	interior da	morar no	mulher está	no interior	o rapaz	rapaz foi
D1	0	0	0	1	0	1	0	1	1	1
D2	1	1	1	0	1	0	1	1	0	0

Fonte: Elaborado pelo autor

O TF é calculado para cada palavra w com relação ao documento D, ou seja, a razão entre o número de ocorrências de w com relação a D podendo haver a aplicação de alguma técnica de mudança de escala a partir do tamanho do documento ou palavras mais frequentes $tf(t, w) = \frac{f_{t,w}}{\sum_{t' \in w} f_{t',w}}$. Desta forma, a variabilidade de tamanho dentre documentos distintos exerce menor influência.

Já o IDF é uma medida que indica o quão raro aquele termo w é dentre todos os documentos D, permitindo que termos pouco frequentes exerçam maior influência. Seu cálculo é realizado por meio de uma função logarítmica entre o inverso da razão do total de documentos do *corpora* e o total de documentos que apresentam o termo em questão $idf(w, D) = \log \frac{N}{|\{d \in D : w \in d\}|}$. TF-IDF é o produto entre os índices TF e IDF e pode ser utilizado para representar os modelos BoW de maneira mais robusta (BRÜCHER; KNOLMAYER; MITTERMAYER, 2002a).

2.4.3.2 Word embeddings

A representação de documentos por meio de BoW apesar de simples e bastante disseminada apresenta algumas limitações, dentre as quais podemos destacar a esparsidade

dos vetores e a perda de informações estruturais, semânticas e contextuais inerentes ao processo de construção da BoW (SARKAR, 2018). As informações perdidas no BoW são essenciais durante a comunicação humana oral e escrita; consequentemente, também são importantes no processo de contextualização, entendimento e desambiguação. Por causa disso, surgiram os *word embeddings*, que é uma maneira mais sofisticada de representar a relação entre as palavras e documentos através de valores numéricos.

2.4.3.2.1 Word2Vec

Os principais modelos para geração de *word embeddings* são baseados em modelos de *deep learning* que se iniciaram após o lançamento do *Word2Vec* (MIKOLOV et al., 2013) . A ideia é alimentar redes neurais artificiais profundas com massas de dados não rotuladas em abundância na internet, aplicando técnicas de aprendizagem não supervisionada. Este processo permite que características semânticas e sintáticas sejam extraídas automaticamente e transformadas em valores numéricos representativos (NASCIMENTO, 2019). Nos últimos anos a utilização destes *embeddings* têm se mostrado benéfica, suportando ótimos resultados em diversas tarefas de PLN (ZOU et al., 2013) (CHEN; MANNING, 2014) (STEIN; JAQUES; VALIATI, 2019).

Para o *Word2Vec*, a Google utilizou sua base de notícias (6B) e treinou usando sua enorme estrutura computacional (MIKOLOV et al., 2013). Os *embeddings* produzidos podem variar de acordo com: estrutura da RNA utilizada, o tamanho do vetor de saída, *corpora* de treinamento e técnicas de otimização aplicadas. A partir dos *embeddings* é possível observar que relações foram aprendidas, o que é constatado ao realizar operações algébricas entre palavras relacionadas e obter resultados significativos.

Mikolov, Le e Sutskever (2013) ilustram a capacidade de representação de palavras por meio de *embeddings* bem treinados mediante operações algébricas. Os autores computaram o vetor $V(X) = V(\text{"}maior\text{"}) - V(\text{"}grande\text{"}) + V(\text{"}pequeno\text{"})$; surpreendentemente, dentre todas as palavras do vocabulário, o valor de $V(X)$ apresentou a menor distância de cosseno para $V(\text{"}menor\text{"})$, Indicando uma relação lógica entre as palavras e seus *embeddings*. Os autores vão além e ilustram diversas correlações entre pares de palavras com valores semânticos similares, tal como os pares e tipos de relação dispostos na Tabela 2.

Na Tabela 2 observamos relações entre palavras de grafia distinta, mas de valor semântico similar. Por exemplo, as palavras "Estocolmo" e "Suécia" apresentam uma relação

Tabela 2 – Exemplos de palavras com alta proximidade e pares correlacionados através do *Word2Vec*

Tipo de relação	Pares de palavras
Cidade/Capital	(Estocolmo - Suécia), (Cairo - Egito)
Gênero	(irmão - irmã), (neto - neta)
Oposição	(possível - impossível), (ético - antiético)
Adjetivo/Advérbio	(aparente - aparentemente), (rápido - rapidamente)
Tempo verbal	(ter - teve), (era - é)

Fonte: Traduzido de Mikolov, Le e Sutskever (2013)

similar a "Egito" e "Cairo". Analogamente, é possível observar uma relação entre os verbos "ter" e "teve" de "era" e "é" na flexão verbal que é refletida nos *embeddings* (ALAMMAR, 2018).

Para possibilitar o entendimento da relação entre as palavras por meio das redes neurais, é preciso fornecer informações contextuais para o modelo. Em PLN, as informações contextuais são definidas pelo termo principal e aqueles que o cercam. A quantidade de termos contextuais a ser utilizada é definida por um parâmetro nomeado janela de tamanho variável N. Dito isso, existem duas técnicas principais para geração de *word embeddings* que se distinguem por seus objetivos: *Continuous Bag Of Words*(CBOW) e *Skip-gram*.

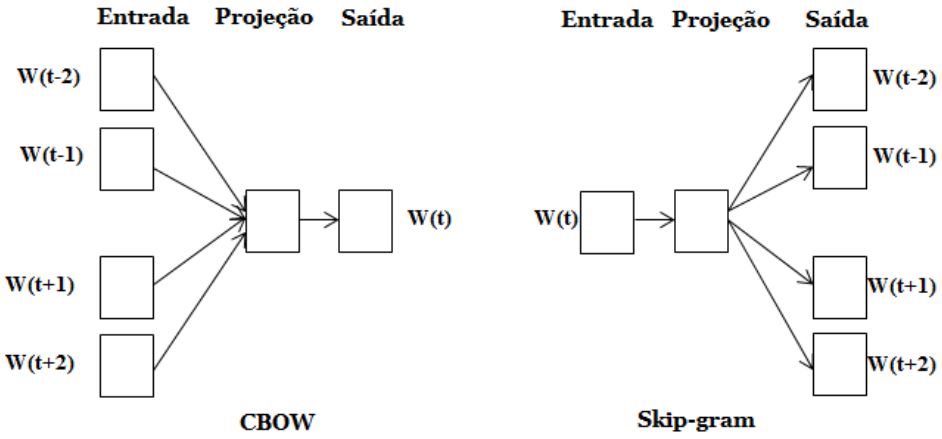
O CBOW tem como objetivo predizer uma palavra $W(t)$ a partir das palavras que lhe antecedem e sucedem, ou seja, seu contexto U'. O tamanho de U' é delimitado pela janela N. Desta forma, dada a palavra $W(t)$ e $N = 2$, temos $U(t) = W(t-2), W(t-1), W(t+1), W(t+2)$. A rede neural é treinada usando como entrada a matriz U e tem como objetivo predizer o vetor de palavras alvo W.

De outro modo, o *Skip-gram* tem objetivo oposto ao CBOW, ou seja, a partir da palavra principal $W(t)$, predizer as palavras ao seu redor U' (contexto). Esta abordagem também leva em conta a janela N, resultando num problema um pouco mais complexo, uma vez que a arquitetura precisa predizer N palavras do contexto a partir de $W(t)$ (MIKOLOV et al., 2013). Para simplificar este modelo, alguns autores sugerem utilizar um tamanho de janela fixo N=1, resultando em dois grupos de pares de entrada, os positivos: compostos por $W(t)$ combinada à uma palavra de seu contexto; e os negativos que são compostos por uma palavra de interesse $W(t)$, combinado a uma palavra aleatória que

não pertence ao seu contexto (SARKAR, 2018). A partir dos pares de entrada, o modelo consegue identificar quais palavras são relevantes ou não e gerar *embeddings* para palavras similares.

Dentro das duas arquiteturas, os *embeddings* representam o vetor de pesos das redes neurais, que é inicializado aleatoriamente e vai sendo ajustado por meio da retropropagação e medição da função de perda ao longo do treinamento (MIKOLOV et al., 2013).

Figura 3 – Arquitetura para geração de embeddings baseada em CBOW e Skip-gram



Fonte: Traduzido de Mikolov et al. (2013)

Devido às nuances das arquiteturas demonstradas na Figura 3, existem algumas diferenças a serem destacadas: No que se refere ao treinamento, o *Skip-gram* necessita de mais tempo para treinamento do que o CBOW, uma vez que a quantidade de termos que ele tenta predizer é maior. A aplicação da função de média nos vetores de contexto do CBOW leva este modelo a apresentar piores resultados para palavras mais raras se comparado ao *Skip-gram*, que não realiza a mesma operação. Entretanto, a mesma operação permite que o CBOW apresente melhores resultados em documentos com palavras mais comuns.

Em um estudo comparativo entre as arquiteturas realizado por Jang, Kim e Kim (2019), os autores afirmam obter melhores taxas de acerto com o CBOW numa base de dados de matérias de jornal. Já o *Skip-gram* apresentou melhor performance numa base de *tweets* (JANG; KIM; KIM, 2019) . Eles acreditam que isso pode estar relacionado à maior conformidade do conteúdo utilizado em textos jornalísticos se comparado a *tweets*.

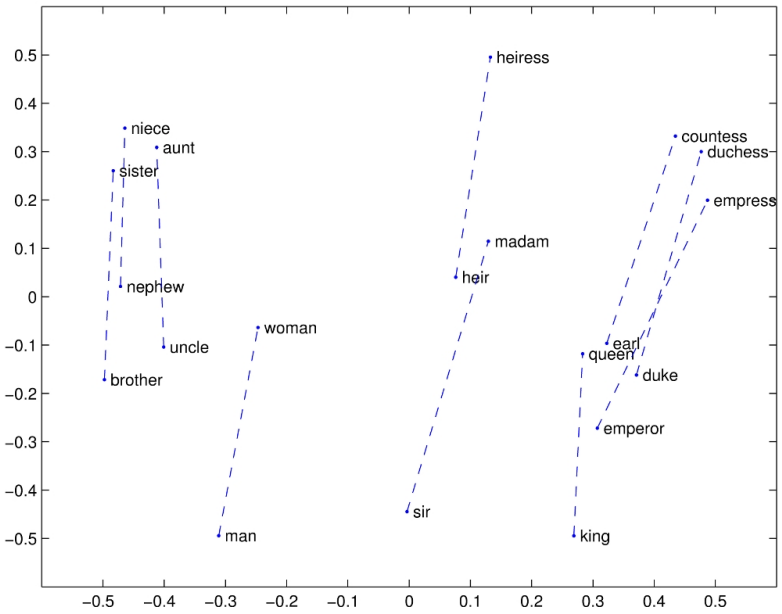
2.4.3.2.2 GloVe

Inspirado nos resultados obtidos através de *word embeddings* gerados por Word2Vec, um outro algoritmo para construção de *embeddings* chamado *Global Vectors for Word Representation* (GloVe) foi proposto (PENNINGTON; SOCHER; MANNING, 2014). O GloVe é uma abordagem que combina duas famílias de modelos para treinar os vetores: 1) fatorização global de matriz, tal como o LSA (*Latent Semantic Analysis*); e 2) métodos baseados em janela de contexto, tal como o *Skip-gram*.

De acordo os autores, ambas abordagens apresentam pontos para melhoria. No LSA, a boa representação estatística do modelo não é suficiente para algumas tarefas de PLN como analogia de palavras (LUONG; SOCHER; MANNING, 2013) (MIKOLOV et al., 2013), resultando em vetores de palavras subótimos. No *Skip-gram*, o principal problema é a baixa utilização de representações estatísticas das palavras, tal como a contagem global de co-ocorrências. Desta forma, os autores propõem a criação de matrizes globais que relacionam as palavras do texto por meio de co-ocorrências considerando seu contexto.

Similarmente ao apresentado sobre os *embeddings* produzidos por meio de CBOW e *Skip-gram*, o GloVe também é capaz de criar relações interessantes entre os vetores de palavras com grafia distinta mas valor semântico aproximado (Figura 4).

Figura 4 – Representação de palavras relacionadas através de vetores produzidos pelo GloVe



Fonte: Pennington, Socher e Manning (2014)

Na figura acima, podemos observar que os vetores que representam os pares de palavras

"irmão" e "irmã"se localizam no espaço de características numa posição muito próxima ao par "tio" e "tia", que também são relações familiares com gêneros distintos. Também é interessante observar a posição de outros pares no espaço, tal como "homem" e "mulher", que é um pouco mais afastada do par "irmão" e "irmã", porém a distância entre os dois pares de palavras é similar, o que pode ser um indicador da relação entre a distância e o gênero das palavras.

Os resultados apresentados por Pennington, Socher e Manning (2014) demonstram que o GloVe é capaz de obter resultados superiores à CBOW e *Skip-gram* na tarefa de analogia de palavras e alcança resultados similares na atividade de entidade nomeada, porém utilizando vetores com menos características e treinado em *corpora* menores (PENNINGTON; SOCHER; MANNING, 2014).

2.4.3.2.3 FastText

O *FastText* (2016) é uma outra técnica para criação de *word embeddings* proposta alguns anos após o GloVe pelo laboratório de IA do Facebook. Seu nome se dá à capacidade de treinamento mais rápido mesmo em grandes *corpora* se comparado a outras abordagens. A velocidade está vinculada ao uso de técnicas de otimização, como poda e funções hash (JOULIN et al., 2016). A estratégia utilizada pelos autores se assemelha ao *Skip-gram*, pois também visa definir o contexto a partir de termos principais. A diferença é que cada palavra do texto é tratada como um *n-gram*, havendo uma separação a nível de caracteres. Os autores definem o *FastText* como um modelo baseado em subpalavras (BOJANOWSKI et al., 2017). Ressaltamos que além dos caracteres, a palavra original também é usada durante o treinamento, o que permite que não ocorra nenhuma perda de informação com relação às abordagens anteriores.

As abordagens que consideram palavras inteiras como uma única entidade podem apresentar problemas em idiomas com vocabulário vasto e muitas palavras raras, pois existe maior probabilidade destas não estarem presentes nos *corpora* usados durante a construção dos *embeddings*. Além disso, esses modelos não consideram a estrutura morfológica das palavras, que também fornece informações relevantes sobre elas. Sendo assim, este modelo se propõe a separar as palavras em porções menores de acordo com o número de *n-grams*, o que permite ao modelo aprender também sobre a estrutura morfológica da língua, bem como sufixos, prefixos, radicais e combinações de caracteres frequentes

(BOJANOWSKI et al., 2017).

Por exemplo, dada a palavra "aonde"com n=3, além da palavra original, o *FastText* utilizaria as sequências de caracteres $[' ao', ' aon', ' ond', ' nde', ' de']$ durante o treinamento. O algoritmo faz distinção do que são palavras completas ou *n-grams* ao envolvê-las entre os símbolos <>. Isso permite a distinção entre a sequência $' de'$ da palavra $<' de' >$. Esta estratégia permite a criação de vetores de palavras mesmo que estas não estejam presentes no corpus de treinamento.

Para facilitar a utilização dos *word embeddings* pela comunidade, os vetores de palavras foram disponibilizados na Web nos mais diversos tipos, variando a estratégia utilizada para sua construção (CBOW, *Skip-gram*, GloVe, *FastText*), tamanho do vetor (50, 100, 300) e idioma. Com isso, a adoção de *word embeddings* tem se tornado mais comum e diversos trabalhos demonstram que sua utilização auxiliou a alcançar melhores resultados em pesquisas envolvendo conteúdo textual.

2.4.4 Variações linguísticas, estilo e estilometria

A linguagem é um rico instrumento de comunicação usado pelos seres humanos para transmitir ideias e pensamentos. As diversas formas que a linguagem pode ser utilizada são determinadas pelas variações linguísticas. Historicamente, essas variações ocorrem a partir do contato entre grupos distintos de pessoas e povos, estando em constante evolução e receptiva a novas variações (BAILEY et al., 1991). Algumas destas variações podem ocorrer a nível de vocabulário, ortografia e gramática.

Por isso, a variabilidade linguística pode ser capaz de representar indivíduos ou grupos de indivíduos inseridos em determinado contexto. Por exemplo, dialetos regionais podem representar que o autor pertence a uma região ou grupo social; a forma de escrita e vocabulário podem indicar a idade do autor ou época em que o documento foi escrito e até mesmo se um texto foi originalmente escrito naquele idioma ou traduzido (VARELA, 2017).

A forma de escrever de cada autor é definida por seu estilo, que é composto por características contextuais, tal como o tipo de obra, público alvo, tempo e idioma em que é escrito, e as variantes, que são características pessoais do autor do documento. Segundo Varela apud Statamatos 2017:

> "O estilo é um conjunto único de padrões gramaticais aplicados por autores durante a sua escrita, por meio da utilização habitual ou sistêmica de formas de escrita, e são conhecidos como marcadores de estilo ou características estilométricas". VARELA 2017

O estilo de cada autor é algo que é desenvolvido ao longo do tempo e muitas vezes aplicado de forma inconsciente, como se fosse parte da própria pessoa, refletindo hábitos e características do subconsciente do autor e é capaz de diferenciá-lo (KJELL, 1994). Os métodos e técnicas para análise das variações linguísticas e estilo são objeto de estudo da estilística.

A aplicação de estilística para problemas de atribuição de autoria são datados do século 19 para identificar textos literários com autoria indefinida ou contestada. As principais abordagens utilizadas pela estilística são técnicas qualitativas e quantitativas que se complementam em muitos casos.

O ponto de vista qualitativo é mais subjetivo pois se refere à forma como determinado autor escreve, para então classificar determinado texto dentro daquele estilo ou não. Um caso que ilustra este tipo de análise ocorreu no Alaska, onde uma contestação de autoria de testamento foi levada a julgamento. Os especialistas analisaram outras cartas da suposta autora que havia falecido e a partir de características qualitativas, como omissão de parágrafos, uso de dialeto inglês crioulo e baixa concordância ordinal, afirmaram que a carta de testamento era fraudulenta (MCMENAMIN, 2002).

As características quantitativas são aquelas que visam medir a frequência de características dentro do texto, não se restringindo a palavras e caracteres, mas também considerando papéis semânticos e sintáticos exercidos pelas mesmas palavras. A análise dos métodos quantitativos também é conhecida como estilometria.

A palavra estilometria tem origem do idioma grego e significa estilo e medição (*stylos* + *metron*) (BELAK; PESA; BELAK, 2008). A área de estudo visa representar a maneira de escrever de um autor por meio de características e medidas estatísticas. Sua origem está relacionada ao pesquisador Augustus de Morgan que sugeriu que a autoria de um texto poderia ser resolvida através da comparação entre o comprimento das palavras usadas pelo autor, pois autores distintos deveriam escrever de formas diferentes (HOLMES, 1998). Desde então, diversos estudos se aprofundaram na área em busca da compreensão de características estilométricas capazes de diferenciar estilos de escrita.

2.4.4.1 Características estilométricas

As características estilométricas são atributos que quantificam o estilo de escrita e podem ser divididos por grupos. Os primeiros grupos de características foram criados fundamentalmente baseados em abordagens linguísticas, das quais podemos destacar: *a*) léxicas, *b*) baseadas em caracteres, *c*) sintáticas, *d*) semânticas, *e*) específicas para uma aplicação ou domínio e *f*) baseadas em coerência e legibilidade (TEMPESTT et al., 2017) (STAMATATOS, 2009) (VARELA, 2017).

Características léxicas: Estão relacionadas ao léxico, o conjunto de caracteres e símbolos que formam as palavras e frases no documento. Este grupo pode ser dividido em subcategorias, sendo estas: *i*) baseado em *tokens* e frases, *ii*) erros gramaticais, *iii*) frequência de palavras, *iv*) frequência de *n-grams* e *iv*) medidas oriundas da riqueza de vocabulário. Um ponto positivo dessas características é sua independência do idioma, sendo viável de aplicar na maioria das línguas.

As medidas baseadas em *tokens* e frases dependem da utilização de um *tokenizer* e indicam características de estilo de escrita dos autores, que por vezes tendem a utilizar palavras, frases e parágrafos mais curtos ou compridos. Essas características podem ser mensuradas pela frequência, média e desvio padrão.

Para verificação de erros gramaticais, geralmente é empregado um corretor ortográfico automático, porém abordagens mais sofisticadas são capazes de capturar os tipos de erros cometidos pelo autor, que podem estar relacionados à pontuação e concordância.

Para medir a frequência de palavras e *n-grams* é necessário selecionar quais palavras ou sequências são mais relevantes dentro do texto, o que pode variar de um problema para o outro. As palavras que apresentam significado especial e são capazes de distinguir autores são chamadas de palavras de função. Sua identificação é relativa ao *corpora* e domínio do problema, por isso demanda expertise, conhecimento linguístico e tempo, o que muitas vezes é inviável. Como alternativa a essa restrição, alguns autores sugerem a seleção de um conjunto de N palavras ou *n-grams* (dependendo do tamanho da base) mais frequentes dentro do *corpora* e sua respectiva representação por meio de algumas das abordagens discutidas na Seção 2.4.3 (STAMATATOS, 2008). Os *n-grams* proveem uma alternativa simples de manutenção de informação contextual ao utilizar representações que desconsideram a ordem das palavras.

A riqueza de vocabulário é uma medida que varia bastante de acordo com o compri-

mento do texto, por isso diversos autores propõem medidas para mensurar este atributo de maneira compensatória, tal como as medidas K (YULE, 2014), R (HONORÉ, 1979) V (HERDAN, 1964) e U (DUGAST, 1979) (descritos na Seção 4.2.3.5). Além destas, outras métricas mais comuns são a razão entre *tokens* únicos e a quantidade de *tokens* no texto (*type-token ratio*) e a quantidade de palavras que só ocorrem uma vez no documento ou no *corpora* (*hapax legomena*) (POPESCU; ALTMANN, 2008).

Uma vez que essas características dependem da separação de frases e *tokens* que precisam estar padronizados, a etapa de pré-processamento é fundamental para a extração dessas características.

Baseada em caracteres: Esse grupo de características considera o texto como uma sequência de caracteres, sendo capaz de mensurar diversas características, tais como: contagem de letras, números, pontuação, maiúsculo, minúsculo, vogais e consoantes. Alguns autores também sugerem a utilização de *n-grams* de caracteres para obtenção de maiores informações sobre o estilo (KEŠELJ et al., 2003) (GRIEVE, 2007). De maneira geral, essas características não exigem técnicas elaboradas de PLN, visto que podem ser utilizadas técnicas de manipulação de strings e expressões regulares para encontrar e contabilizar as características dentro do texto. Alguns autores consideram *n-grams* como parte das características lexicais (TEMPESTT et al., 2017), entretanto optamos por separá-las tanto aqui como na metodologia para aumentar a granularidade da análise de cada grupo.

Apesar de bastante simples, a abordagem baseada em *n-grams* vem sendo usada para resolução de diversos problemas, pois adiciona pouco custo computacional e é menos sensível a erros de digitação. Isso porque os *n-grams* gerados por uma palavra com erro ortográfico ainda apresentam muitos trechos em comum em comparação à palavra em sua grafia correta (STAMATATOS, 2008). Um desafio à parte é a definição do tamanho ideal de n ao usar *n-grams*, por um lado um valor elevado de n é capaz de capturar informações contextuais, mas aumenta bastante a dimensionalidade do vetor de entrada. Já um valor menor pode não ser suficiente para capturar informações contextuais, mas captura subpalavras e causa menor impacto na dimensionalidade (STAMATATOS, 2009).

Características Sintáticas: Esse grupo de características visa extrair informações e padrões sintáticos dentro do texto. Pelo fato dessas informações geralmente serem transmitidas de maneira inconsciente pelos autores e ser mais difícil de manipulá-las se comparado a características léxicas, seu uso é bastante comum em casos de disputa de autoria (LUYCKX; DAELEMANS, 2008).

As informações sintáticas possuem maior dependência do idioma, uma vez que cada língua é regida por diferentes regras gramaticais. Por causa disso, é necessário utilizar ferramentas de PLN para auxiliar nessa atividade. Apesar de existirem ferramentas de PLN bastante confiáveis atualmente, a adição de um viés pela ferramenta escolhida é inevitável. Dentre as ferramentas necessárias, podemos destacar os *POS-Taggers*, que podem ser baseados em palavras ou *n-grams* e são responsáveis por definir a classe gramatical dos *tokens*s; e os *Chunkers*, que atuam extraindo e classificando porções de *tokens* e frases no texto.

Em suma, nesta categoria observamos medidas de incidência de classes gramaticais, razão entre grupos relacionados, tipos de sintagmas (nominais, verbais) e os erros gramaticais sintáticos, que estão relacionados a problemas de estruturação das frases ou sequências de *tokens* incompatíveis.

Características Semânticas: Essas características visam extrair o papel semântico das palavras e frases do texto. O papel semântico representa o significado das palavras dentro de uma gramática ou contexto. Essas características apresentam ainda maior complexidade para identificação e necessitam de mecanismos de PLN suportados por dicionários léxicos ou de sinônimos (*thesaurus*), como o *WordNet*. Esses recursos são capazes de agrupar palavras similares e prover significado para elas, auxiliando em casos de ambiguidade e coerência textual (KILGARRIFF, 2003).

Quanto maior a complexidade da atividade de PLN, maior o viés. Tarefas simples como *tokenização* e *stemming* introduzem pouco viés, entretanto tarefas complexas como as análises sintática e semântica apresentam maior enviesamento. Alguns dos exemplos mais comuns para este tipo de característica são associações funcionais, sinônimos e hiperônimos. As associações funcionais correspondem a grupos de características com funções que se assemelham. Um bom exemplo para esses grupos funcionais são as entidades nomeadas. Segundo Gamon (2004), a associação de características léxicas, sintáticas e semânticas melhora a precisão da classificação.

Específicas para aplicações/domínio: Diferente de todas as características anteriores, que eram independentes da aplicação em questão, essas características são altamente dependentes do contexto e precisam de uma análise do problema para sua confecção. Também têm como propósito a compreensão do estilo de escrita do autor por meio de aspectos que podem estar relacionados à estrutura do documento ou conhecimento prévio sobre o assunto (TEMPESTT et al., 2017).

Dentre as características relacionadas à estrutura do texto, citamos a contagem de termos-chave (saudação no começo ou fim do texto), espaçamento entre parágrafos, formatação de fonte (cor, tamanho, tipo e família de fonte, alinhamento), uso de assinaturas e indentação. Vale ressaltar que muitas dessas características não são aplicáveis em todos os gêneros textuais e as possibilidades de características foram bastante resumidas neste trabalho.

Com relação a características contextuais, a incidência de palavras-chave contidas em dicionários específicos do domínio ou tema, podem ser aferidas e originar diversas métricas (TEMPESTT et al., 2017).

Baseadas em coerência e legibilidade: Alguns autores consideram este como um subgrupo dentro de um grupo abrangente de características adicionais, porém consideramos que tal conjunto apresenta especificidades que necessitam de maior análise (TEMPESTT et al., 2017). Tais características visam capturar informações sobre coerência, coesão e legibilidade do texto, aspectos fundamentais para que o texto faça sentido para o leitor. Acredita-se que tais características podem estar relacionadas ao estilo de escrita de um autor e que autores distintos podem apresentar diferentes níveis de coesão, coerência e legibilidade.

A coesão pode ser definida pela utilização correta de palavras que conectam frases, orações e parágrafos, conforme um conjunto de regras gramaticais, o que colabora com a organização do texto(GRAESSER et al., 2004). Por outro lado, a coerência está relacionada à relação lógica das ideias apresentadas no texto, levando em conta aspectos macrotextuais, tais como o desenvolvimento de ideias e presença de erros, como redundância, superficialidade e contradições (SILVA, 2006). Finalmente, a legibilidade é o que qualifica um texto como mais fácil de ler do que outro.

Por meio de estudos linguísticos, foi possível quantificar essas medidas através de fórmulas matemáticas. Entre as mais conhecidas, estão o índice Flesch (KINCAID et al., 1975), fórmulas de Dale-Chall (CHALL; DALE, 1995) e o *framework* Lexile (STENNER, 1996) (descritos na Seção 4.2.3.5). Essas equações levam em conta diversos fatores, tal como: proporção de palavras "fáceis"e "difíceis", quantidade e tamanho médio das frases, quantidade de frases por parágrafo, tamanho médio de sílabas por palavra, quantidade de palavras monossilábicas e quantidade de pronomes na primeira, segunda e terceira pessoa (DUBAY, 2004).

Graesser et al. (2004) foram responsáveis por compilar um conjunto com mais de 200

índices relacionados à coesão e coerência denominado *Coh-metrix*. Esse sistema inspirou diversos outros autores, que criaram adaptações da versão inicialmente baseada na língua inglesa para outros idiomas, como o português (ALUÍSIO; CUNHA; SCARTON, 2016) (QUISPESARAVIA et al., 2016). Esses índices também são considerados os precursores de pesquisas voltadas para análise automática da qualidade textual (MCNAMARA; CROSSLEY; MCCARTHY, 2010).

2.5 CLASSIFICAÇÃO DE DOCUMENTOS

A classificação de documentos é uma atividade dentro da área de classificação de dados que está amplamente envolvida com o PLN. Seu objetivo é categorizar ou classificar automaticamente determinado documento dentro de uma ou mais classes a partir de suas características. Em algumas situações, pode ser fácil para um humano identificar o tipo ou tema de determinado documento, porém nem sempre a tarefa é facilmente realizada por um computador, variando de complexidade de acordo com o objetivo e formato dos dados.

A área vem ganhando notoriedade nos âmbitos científico e econômico proporcionalmente ao volume de dados textuais na internet. Essa atividade permite compreender os principais assuntos e que sentimentos eles estão proporcionando a partir de conversas na internet. Além disso, o gerenciamento do conhecimento na internet depende bastante da classificação dos documentos para sua recuperação, organização, visualização e troca de conhecimentos (BRÜCHER; KNOLMAYER; MITTERMAYER, 2002a).

As atividades de classificação de padrões se dividem em três categorias quanto a rótulo das classes: binária, múltiplas classes ou *multiclasse* e múltiplos rótulos ou *multilabel*, podemos organizá-las em vista da classificação de documentos da seguinte forma:

- **Binários**: são problemas de classificação nos quais só existem duas possibilidades de classificação, geralmente num problema de verdadeiro ou falso, tal como a caracterização de um documento como *spam* (CRAWFORD et al., 2015), ou se uma obra foi escrita por determinado autor ou não (BROCARDO et al., 2013). Os problemas de classificação binário são teoricamente menos complexos de resolver, visto que há uma chance de 50% do classificador predizer a classe correta, entretanto em casos que se deseja identificar a classe minoritária, pode ser necessário aplicar algoritmos

e técnicas para lidar com o desbalanceamento dos dados.

- **Multiclasse (Categóricos)**: refere-se aos problemas em que existem mais de duas classes, ou seja, existe uma lista de possíveis classes (categorias), e cada documento só pode ser classificado exclusivamente em uma dessas possibilidades. Expandindo os exemplos binários acima, temos a classificação de e-mails em determinada aba da caixa de entrada (principal, social, promoções, *spam*) ou definir dentre uma lista com mais de 2 autores, por qual daqueles autores determinada obra foi escrita (STAMATATOS, 2009). Esses problemas podem apresentar uma maior complexidade e dependem de uma amostra de exemplos representativa por cada autor para obter boas taxas de acerto.

- **Multilabel**: é composto por problemas em que determinado documento pode ser classificado dentro de uma ou mais classes simultaneamente. Digamos que um e-mail pode ser ao mesmo tempo classificado dentro de duas ou mais categorias (e.g. promocional e redes sociais), ou que determinada obra pode ter sido escrita por vários autores. De acordo com o número de rótulos ou *labels* que os documentos podem assumir, aumenta-se a complexidade do problema. Tem-se observado uma ampla utilização de redes neurais para solução de problemas recentes dessa natureza (CHALKIDIS et al., 2019) (YOU et al., 2018).

O processo de classificação de documentos se iniciou por meio de análises estatísticas sobre as palavras que compunham o texto, o que levou pesquisadores a observarem que uma maior frequência de determinados termos são indicativos sobre qual seria o tema abordado dentro de um conjunto pré-determinado de assuntos-chave (LUHN, 1957). Posteriormente, esse conjunto foi expandido com valores dinâmicos derivados dos próprios *corpora*, por meio de uma matriz de frequência de termos (BORKO; BERNICK, 1963).

Em seguida, observamos esforços na construção de agentes inteligentes capazes de classificar documentos por meio de diferentes estratégias, como reconhecimento de padrões (HAMILL; ZAMORA, 1980), Árvores de Decisão (CLACK et al., 1997), seleção de termos principais associados à regras de decisão (FUKETA et al., 2000) e aplicação de ontologias em conjunto com dicionários de sinônimos (*thesaurus*) para construção de mapas de conhecimento (BORGES et al., 2004).

A partir do século XXI, observamos grandes avanços na área devido ao uso de PLN associado a algoritmos de AM supervisionados e não supervisionados.

2.5.1 Abordagens não supervisionadas

Os algoritmos de AM não supervisionados lidam com exemplos que não apresentam uma classe pré-existente, desta forma buscam observar a partir da distribuição dos exemplos se existem padrões previamente desconhecidos. Para classificação de documentos, a abordagem não supervisionada baseada em agrupamento (*clustering*) é uma das mais comuns.

O algoritmo K-médias é um exemplo de algoritmo que visa criar grupos (*clusters*) usando o conceito de centroides, que são pontos médios inicializados aleatoriamente. Estes vão sendo atualizados a cada iteração a partir dos exemplos mais próximos de cada centro. O objetivo é reduzir a distância entre os centroides e exemplos daquele mesmo *cluster* até convergir a uma distância média mínima entre cada centro e respectivos exemplos próximos a ele, ou atingir um critério de parada predeterminado.

O calculo da distância é geralmente feito por meio de uma função geométrica derivada de Minkowski, como a distância euclidiana, Manhattan ou Chebyshev (SINWAR; KAUSHIK, 2014).

Na literatura, existem diversos trabalhos que reportam bons resultados para classificação de texto por meio de abordagens de agrupamento, como por exemplo usando KNN supervisionado e K-médias (SOUCY; MINEAU, 2001) (SINGH; TIWARI; GARG, 2011) (WANG; WANG, 2007).

Existem também estratégias de agrupamento que se baseiam em modelos probabilísticos, criando agrupamentos difusos (*fuzzy clustering*) em que cada exemplo não está contido num dos grupos, mas apresenta uma probabilidade de pertencer a cada um destes.

Um popular algoritmo de partição difusa é o *fuzzy* C-médias (FCM). Ele segue uma abordagem similar ao K-médias, porém produz uma matriz que representa a probabilidade de pertencimento do exemplo a cada um dos *clusters*. É realizado um cálculo a partir da matriz de probabilidades U de todos os elementos N por todos os *clusters* K. Há uma iteração por todos os elementos de U, visando aplicar uma função de decisão capaz de realocar os exemplos nos *clusters* mais prováveis. O processo se repete com objetivo de minimizar a função de decisão e atingir a convergência, o que significa não haver mais

mudanças significativas em U.

No estudo comparativo conduzido por Singh, Tiwari e Garg (2011), foi observado que o FCM se mostrou como um mecanismo robusto para agrupamento de documentos levando em conta uma variedade de bases de dados e formas de representação dos documentos.

Nos problemas de agrupamento, é necessário que exista alguma similaridade entre o conteúdo ou as características dos documentos para que documentos similares possam ser associados corretamente aos *clusters*. Outra limitação é a capacidade de armazenamento, pois os exemplos são armazenados em memória para realização do cálculo da distância. Caso o conjunto de dados (*dataset*) seja muito extenso, pode ser inviável a utilização de tais abordagens (SLONIM; FRIEDMAN; TISHBY, 2002).

2.5.2 Abordagens supervisionadas

Ao mesmo passo que as abordagens não supervisionadas ganharam espaço, algoritmos de Aprendizagem de Máquina supervisionados cresceram ainda mais nos últimos anos. As abordagens supervisionadas são aquelas em que usamos exemplos rotulados para realizar predições de exemplos futuros não rotulados. Existe uma variedade de abordagens supervisionadas tradicionais, dentre estas, podemos separá-las em classificadores lineares e não lineares.

Os classificadores lineares se fundamentam na criação de uma fronteira de separação (linear) entre os exemplos de classes distintas. Num problema binário representado num espaço de duas dimensões, essa fronteira pode ser representada por uma linha reta. Quando o espaço de características aumenta de dimensionalidade, é necessário construir um hiperplano de separação para cada classe. A fronteira de separação é construída por meio de uma função discriminante linear F que é calculada com auxílio de um vetor de pesos W de tamanho n igual ao comprimento do vetor de características X e uma constante b. Tanto W como b serão obtidos através do processo de treinamento. Dado que X_0 e W_0 são os valores iniciais, a função objetivo é calculada por meio da multiplicação de todas as características dos exemplos de entrada X_n e do vetor de pesos W_n, distinguindo os exemplos dentro das possíveis classes por meio de um limiar T (CHEN et al., 2004).

$$Y = b + W_o * X_o + ... + W_n * X_n$$

$$F = Y > T \quad então \quad classe \quad c1 : se \quad não \quad c_0$$

Num problema multiclasse com N classes, são gerados N modelos lineares. Cada um desses modelos é responsável por classificar os exemplos de uma determinada classe W_i criando uma fronteira de separação entre os exemplos da classe W_i e os exemplos de todas as outras classes.

Os modelos lineares se subdividem entre generativos e discriminativos. Os generativos se baseiam na probabilidade conjunta $p(x, y)$. Dado que uma entrada é x e ela pertence à classe y, temos $p(x, y) = p(x|y) * p(y)$. Os principais exemplos deste grupo são as redes bayesianas.

Por outro lado, os modelos discriminativos levam em conta a probabilidade a posteriori $p(y|x)$ diretamente (RAINA et al., 2003). Esses modelos visam maximizar a qualidade da saída a partir dos dados de treinamento, o que aumenta sua dependência por dados de boa qualidade. Por causa disso, fazem-se valer de funções de otimização ou perda para alcançar os melhores resultados diante do conjunto de testes. Exemplos típicos de algoritmos de classificação com abordagens discriminativas a serem discutidos neste trabalho são a Regressão Logística (RL), Árvores de Decisão (AD) e máquinas de vetor de suporte (*Support Vector Machines* - SVM).

Existem discussões no meio científico sobre as duas abordagens. Muitos autores afirmam que obtiveram melhores resultados usando modelos discriminativos (JEBARA; PENTLAND, 1998) (NIGAM; LAFFERTY; MCCALLUM, 1999) (VAPNIK, 1999). Entretanto, modelos generativos também proveem métodos para suprir a ausência de dados, com melhor desempenho em casos em que o conjunto de treinamento é pequeno (NG; JORDAN, 2002). Alguns autores também sugerem a aplicação de abordagens híbridas tentando unir o melhor dos dois mundos (RAINA et al., 2003).

2.5.2.1 Redes Bayesianas

As redes Bayesianas são um conjunto de modelos baseados na aplicação do teorema de Bayes para classificação. O teorema parte da prerrogativa que há uma independência condicional entre todas as características que representam os exemplos visando definir a qual classe um exemplo pertence. No processo de decisão, é calculado o somatório da probabilidade a posteriori das características de um exemplo x pertencer à classe y $p(y|x)$ (D'AGOSTINI, 1995).

Existem variações das redes Bayesianas pautadas em funções de densidade e distri-

buições estatísticas, tal como a distribuição Gaussiana, Bernoulli ou multinomial (STORK et al., 2001). O algoritmo *Naive Bayes* multinomial (*Multinomial Naive Bayes* - MNB) é uma versão especializada para classificação de texto. Enquanto a abordagem tradicional levaria em conta a presença ou ausência de termos dentro do documento, a versão multinomial leva em conta a frequência dos termos (RAINA et al., 2003). Para um modelo MNB assumimos que um documento é composto por uma lista ordenada de eventos de ocorrência de palavras, que o tamanho dos documentos não tem relação com sua classe e que a probabilidade de cada evento de palavra é independente de contexto e posição de aparição (MCCALLUM; NIGAM et al., 1998).

2.5.2.2 Regressão Logística

A regressão logística, apesar de seu nome, é mais utilizada em problemas de classificação do que regressão. O algoritmo é fundamentado num conceito similar ao de classificadores lineares, ou seja, por meio da obtenção do vetor de pesos W e valor constante b, porém ao invés de construir uma função linear, é aplicada a função logística L.

$$Y = L(b + W_o * X_o + ... + W_n * X_n)$$
$$Y = (1/1 + exp[-(b + W_o * X_o + ... + W_n * X_n)])$$

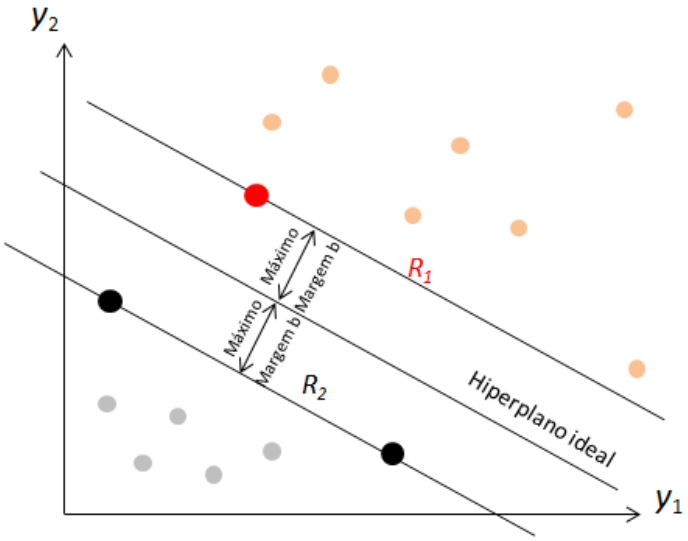
Essa função transforma valores reais de entrada em valores de saída Y entre 0 e 1, que por sua vez são interpretados como a probabilidade de uma entrada pertencer à classe positiva dado o vetor de características $(X_o, \ldots, X_n)$. Durante o treinamento, o algoritmo pode se valer de um parâmetro de regularização inversa C, que pode usar alguma técnica de regularização, tal qual as normas $L1$ e $L2$, como mecanismo de penalidade. Devido à aplicação da função logística, a fronteira de separação gerada por este classificador se assemelha ao de uma função sigmoide logística (LEE; LIU, 2003).

A regressão logística, mesmo sendo um algoritmo de baixa complexidade, tem um largo histórico de aplicações em problemas de classificação de texto. Os resultados apresentados por Pranckevičius e Marcinkevičius (2017) demonstram que para uma base de avaliações de produtos com múltiplas classes, o algoritmo se mostrou superior com relação à SVMs, *Naive Bayes* e árvores de decisão.

2.5.2.3 *Support Vector Machines (SVM)*

O algoritmo de máquina de vetor de suporte (SVM) criado por Vapnik (1998) foi inicialmente concebido para resolução de problemas linearmente separáveis binários. SVMs também podem ser expandidos para problemas multi-classe, utilizando a mesma técnica de criação de N classificadores de acordo com o número de classes. A sua principal característica consiste em obter o melhor hiperplano de separação possível durante o treinamento. Para isto, é necessário obter a maior margem possível, ou seja, maior distância entre o hiperplano e os exemplos de treinamento mais próximos.

Figura 5 – Treinamento da Support Vector Machine (SVM)



Fonte: Nascimento (2019)

Podemos observar na Figura 5 que apesar de haver espaço para construção de diversos hiperplanos de separação entre os elementos das classes representadas nas cores cinza e laranja, o hiperplano ideal possui a maior margem possível para os exemplos de referência. O exemplo acima representa a versão linear do algoritmo SVM, que é capaz de solucionar problemas linearmente separáveis, contudo existem diversos problemas que apresentam natureza não linear e por causa disso, foi introduzido o conceito de *kernel tricks* em associação a SVMs (BENNETT; MANGASARIAN, 1999).

A aplicação de um *kernel* nada mais é que utilizar uma função no conjunto de dados com o propósito de aumentar a dimensionalidade do vetor de características. Entre as funções mais utilizadas, podemos destacar a polinomial, gaussiana e sigmoide. A partir

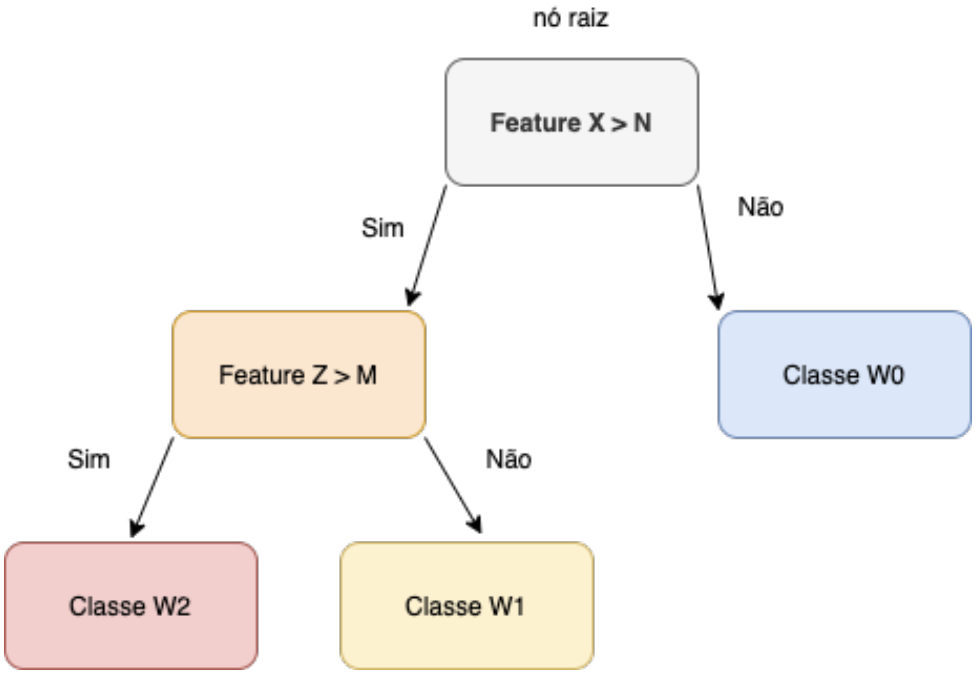
daí, é possível observar que alguns problemas não lineares podem se tornar, teoricamente, linearmente separáveis num espaço com mais dimensões, sendo possível aplicar um SVM e obter o hiperplano de separação ideal neste novo espaço de características.

No que se refere a sua utilização para classificação de texto, podemos destacar o trabalho de Joachims et al. (1999), que propôs uma versão otimizada do algoritmo, denominada TSVM (SVM com transdução) que utiliza parte dos exemplos não rotulados para cálculo da margem e definição do hiperplano de separação (JOACHIMS, 2002). A partir daí, surgiram diversos estudos visando aplicar e adaptar SVMs para dentro da classificação de texto (VARELA, 2017)

2.5.2.4 Árvores de decisão e Random Forest

As árvores de decisão (AD) são um conjunto de algoritmos de AM representados numa estrutura de árvore. Uma árvore é composta por nós, cada nó pode representar o nome da classe, chamado de nó de resposta, ou uma condição de teste para particionar o conjunto de exemplos de acordo com seu resultado. As características dos exemplos de entrada são avaliadas em cada nó de condição, até que seja alcançado algum nó de resposta (Figura 6).

Figura 6 – Ilustração de uma árvore de decisão com nós de condição e resposta



Fonte: Elaborado pelo autor

A estrutura da árvore é construída a partir de um processo automático de indução sobre o conjunto de treinamento e vetor de características (*features*). Existem variações que podem ser levadas em conta durante a construção das árvores, tais como sua topologia e critérios para poda para evitar sobreajuste sobre os dados. Por isso, existem diversos tipos de AD como a ID3 (QUINLAN, 1986), C4.5 (QUINLAN, 1993) e CART (BREIMAN et al., 1984).

Para definição das características mais importantes, podem ser utilizados alguns critérios como por exemplo o ganho de informação (DALEY; VERE-JONES, 2004), que é medido pela diferença da entropia entre o nó-pai e possíveis nós filhos; razão de ganho, que é uma versão ponderada do ganho de informação (QUINLAN, 1993) e Gini que utiliza um índice de dispersão estatística (GINI, 1912). Esse processo se repete até que todas as características sejam avaliadas ou até atingir algum critério de parada. Ao fim do processo, teremos uma AD em que as características com maior poder de separação dos exemplos farão parte dos nós mais próximos da raiz, e aquelas com menor poder de decisão nos nós-folha ficarão mais afastados.

Ao longo dos últimos anos, foram desenvolvidas técnicas de otimizações sobre ADs em busca de classificadores mais eficientes, os processos de *bagging* (BREIMAN, 1996) e *boosting* (SCHAPIRE; SINGER; SINGHAL, 1998) são exemplos. O processo de *bagging* se refere ao treinamento de um classificador a partir de um subconjunto aleatório da base de dados, criando diversos subconjuntos dessa base (*bootstrap*) e vários classificadores, combinando-os no fim para obter o melhor resultado possível (SANTOS; FALCÃO, 2017).

O algoritmo de floresta aleatória (*Random Forest* - RF) proposto por Breiman (2001) é considerado um classificador que se vale da técnica de comitê (*ensemble*), ou seja, a tomada de decisão é feita não por apenas um, mas sim por um grupo de classificadores do tipo AD. A geração diversas ADs associadas com o processo de *bagging* além da adição de uma camada extra de aleatoriedade nomeiam o algoritmo de floresta aleatória.

Ao compararmos RF com árvores comuns, onde cada nó é dividido a partir do melhor subconjunto de resultados, observamos que as árvores do RF dividem os nós partindo dos melhores resultados de um subconjunto escolhido aleatoriamente, que por sua vez contempla apenas um subconjunto, também aleatório de características do problema, o que acaba aumentando ligeiramente o bias, mas melhora o desempenho do classificador (SANTOS; FALCÃO, 2017) (LIAW; WIENER et al., 2002).

Fatores que podem influenciar na construção do modelo são o número de árvores que

compõem a floresta e o tamanho máximo do subconjunto aleatório de características selecionadas para construção das árvores. Devido a essas características, é possível observar que o algoritmo exige maior poder computacional para treinamento, mas fornece classificadores mais robustos em muitos casos (BREIMAN, 2001). A tomada de decisão é feita por meio do voto majoritário das árvores, que pode ser feito de forma unânime, maioria simples (mais da metade) ou maior número de votos (NASCIMENTO, 2019).

Um modelo similar ao RF é o comitê de árvores extremamente randomizadas (*Extra randomized Trees* - ET) (GEURTS; ERNST; WEHENKEL, 2006), ele também é um ensemble de ADs construídas de maneira *top-down*. Entretanto, em contraste com o RF, este algoritmo separa os nós escolhendo pontos de corte aleatórios ao invés de por busca gulosa. Um outro ponto importante é que o ET usa todo o conjunto de dados para construção das árvores, e não uma amostragem (bootstrap) do mesmo (GEURTS; ERNST; WEHENKEL, 2006). Ao final temos uma maior diversidade de árvores mais complexas e profundas se compararmos ao RF.

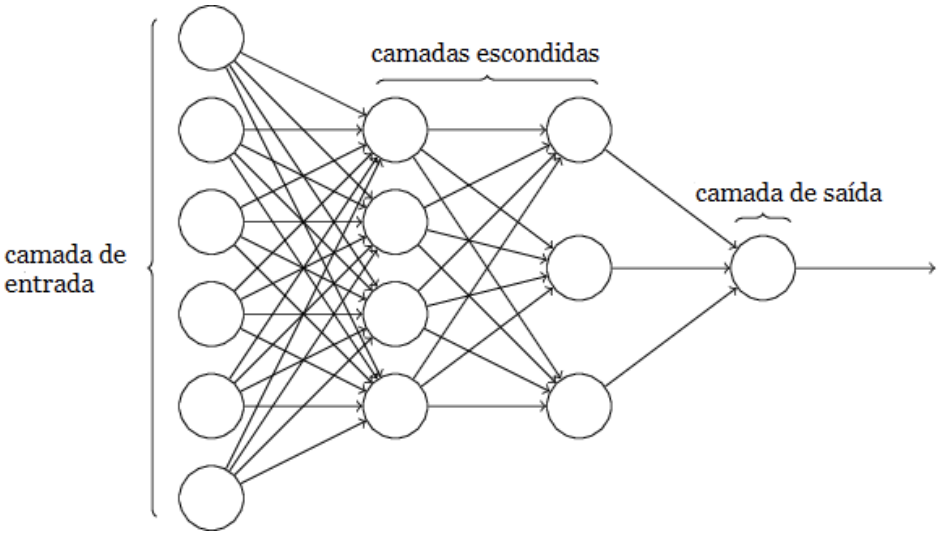
2.5.2.5 Perceptron multicamada

O Perceptron é um antigo modelo de classificação binária que funciona de maneira similar aos sistemas lineares, ou seja, através de uma função polinomial associada a um vetor de pesos. O algoritmo Perceptron multicamada (*Multi Layer Perceptron* - MLP) é a combinação de diversas unidades do Perceptron, que são os nós, dispostos em camadas e conectados por arestas, compondo um classificador capaz de solucionar problemas não linearmente separáveis. Esta arquitetura é um tipo de rede neural artificial (RNA) com nós totalmente conectados, que não possui conexões cíclicas entre nós e de alimentação direta.

Esta RNA é composta por três camadas distintas: entrada, saída e escondida. Nas camadas de entrada e saída, encontramos apenas uma camada de nós, porém na camada escondida pode haver mais de uma. A camada de entrada é por onde o vetor de características é imputado na rede, sendo assim, o número de nós nesta camada deve ser igual ou maior que o tamanho do vetor de entrada. Como a RNA é totalmente conectada, para cada nó numa camada anterior, existe um peso associado a cada um dos neurônios da camada posterior, inicando a força da conexão entre esses nós (Figura 7).

Entre as camadas de entrada e escondida, é inicializado e treinado o vetor de pesos.

Figura 7 – Estrutura de uma MLP com duas camadas escondidas



Fonte: Adaptado de Hastie e Tibshirani (2016)

Este vetor é inicializado aleatoriamente e otimizado durante o treinamento por meio de retropropagação (*backpropagation*). O algoritmo de *backpropagation*, também usado para outras arquiteturas de RNA, visa aproximar a relação entre entradas e saídas por meio do ajuste dos pesos internos. O cálculo dos pesos é iterativo e aferido por meio da correção do erro, usando uma estratégia de gradiente descendente até alcançar um critério de parada. Alguns fatores que podem ser decisivos durante o processo são a taxa de aprendizagem, valor dos pesos iniciais e função de ativação (CHOI; CHOI, 1992).

Por fim, na camada de saída, os pesos calculados nas camadas anteriores a partir do vetor de entrada são transformados em uma única saída por meio da aplicação de uma função de ativação, que em muitos casos é sigmóide. Este processo também é chamado de achatamento da rede (WILAMOWSKI; JAEGER, 1996). O valor final de saída pode representar tanto a predição de uma classe como um valor aproximado em problemas de regressão.

2.5.3 Aprendizagem Profunda

Nos últimos anos, como estado da arte na classificação de documentos com AM, são discutidas as redes neurais com várias camadas escondidas, objeto de estudo da área de aprendizado profundo (*Deep Learning* - DL). A área de DL se desenvolveu a partir da necessidade por um maior desempenho de algoritmos tradicionais para generalização de

tarefas de IA mais complexas, tais como o reconhecimento de som e imagens (GOODFELLOW et al., 2016).

De acordo com a literatura, são consideradas como objeto de estudo da área de DL as redes neurais multicamadas (*Multi-layer Neural Network* - MLNN) que possuem mais de duas camadas escondidas (GOODFELLOW et al., 2016). Com o crescimento do número de camadas na rede neural, há um crescimento do número de neurônios que compõem a rede, representando um aumento na quantidade de parâmetros a serem ajustados durante o treinamento. Por causa disso, se faz necessário um maior poder de processamento durante o treinamento e maior volume de dados para generalização (FERREIRA, 2017).

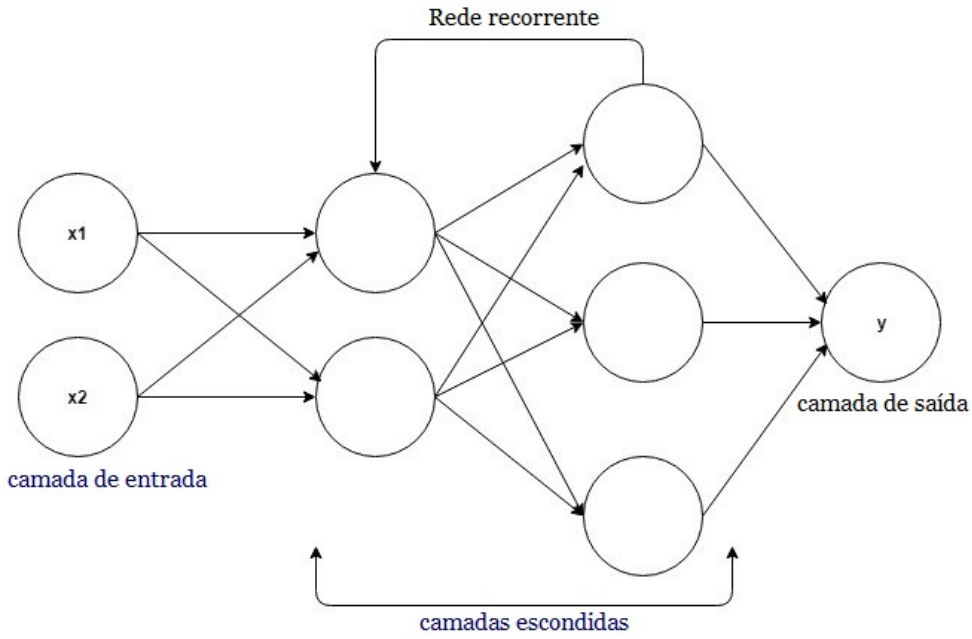
Entre estas redes, podemos destacar duas principais arquiteturas: as de alimentação direta (*feedforward*), e as redes neurais recorrentes (*Recurrent Neural Network* - RNN).

Dentro da arquitetura *feedforward*, se destacam as redes convolucionais (*Convolutional Neural Network* - CNN), originalmente criadas para tarefas de visão computacional, mas que se mostraram eficazes para alguns problemas de PLN, como na análise semântica (ZHANG; WALLACE, 2015). A CNN é composta por camadas convolucionais, responsáveis pela extração de características principais dentro de uma janela através do processo de convolução. A extração automática de características é uma das vantagens das arquiteturas profundas com relação aos modelos clássicos, que demandam maiores esforços para extração de características importantes. Intercaladas às camadas convolucionais, encontramos as camadas de *pooling*. Estas atuam na redução da dimensionalidade, o que ajuda a diminuir o custo computacional e prevenir o *overfitting*. Por último, estão as camadas de saída, que se assemelham a MLPs totalmente conectadas, fazendo uso de uma função de ativação para predição (ARAÚJO et al., 2017).

Na arquitetura recorrente, algumas limitações enfrentadas pelas redes neurais comuns são superadas, com destaque para o tamanho dos vetores de entrada e saída, que apresentam tamanho fixo na versão não-recorrente. Nas RNNs é possível receber uma série de vetores de tamanho variável. Além disso, a RNN possui um estado interno que é atualizado e armazenado a cada passo durante o processamento da série de vetores, isto permite que a RNN pondere considerando não apenas a entrada atual, mas todo o histórico de estados da rede (OLAH, 2015) (KARPATHY, 2015). É essa característica que nomeia a topologia da rede, pois seu estado interno permite que vetores anteriores interfiram no comportamento atual de maneira recorrente.

Para classificação de texto, um dos grandes benefícios dessa arquitetura é que na

Figura 8 – Arquitetura de rede neural recorrente com duas camadas escondidas



Fonte: Elaborado pelo autor

medida que esses vetores são transmitidos entre as camadas da rede neural, há uma noção de temporalidade devido ao histórico de estados, permitindo memorizar palavras mais recentes dentro de um tamanho de janela, o que pode ajudar em problemas de ambiguidade e contextualização, melhorando a acurácia das predições (KOSTADINOV, 2017).

Por outro lado, em alguns casos, essa característica também pode causar problemas de enviesamento, considerando que palavras mais recentes podem influenciar mais nas predições. Como solução para este problema, foram desenvolvidas as redes *Gated Recurrent Units* (GRU) e *Long short-term memory* (LSTM), que são arquiteturas específicas de RNR com mecanismos mais criteriosos para memorização de informação (NGUYEN, 2018).

A LSTM tem como principal característica a presença de células de memória, capazes de guardar informações por um longo período de tempo. Para decidir quais informações armazenar, há um mecanismo interno de portões de entrada e saída de informações que são reguladas por meio de redes neurais internas com função de ativação sigmóide (KARPATHY, 2015).

Já a GRU, criada por Chung et al. (2014) é uma variação da LSTM que combina os portões de entrada e saída num único portão chamado de portão de atualização. Também existem outras simplificações da GRU com relação a LSTM, como a célula de armazena-

mento interno que foi unificada ao estado escondido (KARPATHY, 2015). O resultado final é um modelo mais simples que a LSTM, mas que mantém a capacidade de armazenamento de palavras importantes a médio e longo prazo.

Pesquisadores têm buscado formas de combinar e adaptar CNNs e RNNs, construindo novas arquiteturas de redes neurais. Essas arquiteturas representam o estado da arte em diversas tarefas do PLN, incluindo a classificação de documentos. Resultados significativos foram obtidos em problemas de classificação multi-label dentro da área de análise de sentimentos (ZHOU et al., 2016) e classificação de perguntas (ZHOU et al., 2015).

3 ANÁLISE E ATRIBUIÇÃO DE AUTORIA

A análise de autoria é uma atividade dentro da classificação de documentos que visa relacionar os textos aos seus autores, dividindo-se em 3 principais áreas de acordo com seu objetivo: *i*) verificação de autoria, *ii*) atribuição de autoria e *iii*) caracterização de autoria. A verificação de autoria é um problema de classificação binário, quando se deseja confrontar a autoria de determinada obra com relação a um autor. Na atribuição de autoria temos um problema categórico, visando apontar qual dentre os possíveis autores tem maior probabilidade de ser o verdadeiro autor da obra. Por fim, a caracterização de autoria visa definir o perfil do autor (gênero, idade, raça, classe social) através de sua forma de escrever (BROCARDO et al., 2013). Esta pesquisa, devido ao problema e objetivos propostos, se limita a abordar a atribuição de autoria.

Segundo (STAMATATOS, 2009), a estratégia principal para atribuir ou verificar autoria é por meios estatísticos e computacionais para diferenciar estilos de escrita que potencialmente podem pertencer a grupos de autores distintos. Um dos primeiros estudos relacionados foi desenvolvido por Mendenhall (1887), no qual foi observado que algumas das obras de Shakespeare poderiam ter sido escritas por outros autores devido às características de escrita diferentes da maior parte de suas obras.

Poucos anos depois, seria publicado um dos trabalhos mais conhecidos na área, a pesquisa de (MOSTELLER; WALLACE, 1963), que visava identificar os verdadeiros autores de 85 artigos e cartas (*The Federalist Papers*), que exerceram forte influência política na consolidação da constituição nos Estados Unidos. Nesta pesquisa, foram usados métodos estatísticos bayesianos para observar a frequência de palavras comuns, o que conseguiu produzir resultados significativos para diferenciar o estilo de escrita dos autores envolvidos. A partir dos resultados desta pesquisa, outros autores foram influenciados a investir na tentativa de definição de características que pudessem "quantificar" estilos de escrita, criando uma área de estudo conhecida como a estilometria (HOLMES, 1998).

A partir deste ponto, pesquisadores buscaram desenvolver novas características, como vieram a se popularizar os tamanhos e frequências de frases, palavras, parágrafos, além de outras métricas derivadas do vocabulário. Estima-se que aproximadamente 1.000 métricas diferentes foram propostas na década de 90, e que a maior parte dos estudos visava uma solução em que o computador pudesse auxiliar na tomada de decisão, e não decidir

automaticamente (STAMATATOS, 2009).

Algumas soluções ganharam notoriedade na década de 90, como a técnica CUSUM Morton e Michaelson (1990), que chegou a ser utilizada como artifício forense para atribuição de autoria em documentos, porém anos depois foi abandonada após duras críticas pela comunidade científica principalmente voltadas à sua metodologia de avaliação (HOLMES; TWEEDIE, 1995).

Os pesquisadores deste período sofriam com limitações, como suas bases de dados serem compostas por textos muito grandes e estilisticamente homogêneos (livros), pequeno número de possíveis autores em questão, bases de dados não controladas por tópico e inexistência de outros métodos ou técnicas comparativas (STAMATATOS, 2009).

A partir dos anos 2000, houve um aumento na quantidade de textos disponíveis em meios eletrônicos com a difusão da internet, que demandou métodos mais eficientes para lidar com esse tipo de informação, causando impacto em diversas áreas de pesquisa, tais como aprendizagem de máquina, recuperação da informação e PLN que se combinaram para resolução de problemas de análise de autoria em diversos tipos de documentos, como mensagens, cartas, contratos, postagens, artigos para fins de inteligência, direitos civil, penal ou autoral (CHASKI, 2005)(GRANT, 2007) e computação forense (FRANTZESKOU et al., 2006).

3.1 TRABALHOS RELACIONADOS

Stamatatos (2009) afirma que a última década pode ser considerada como a nova era do desenvolvimento de tecnologias para análise e atribuição de autoria, predominantemente dominada pelo desenvolvimento de aplicações que utilizam e visam resolver problemas cotidianos, ao invés de literários. Também se observou maior ênfase na avaliação e comparação dos métodos propostos para resolução do problema em diferentes *corpora*, além da criação de novos modelos e técnicas variantes do número de autores, tamanho do texto (HIRST; FEIGUINA, 2007) e distribuição dos textos entre autores (STAMATATOS, 2008).

A seguir, iremos discutir sobre alguns dos trabalhos compreendidos neste período de grandes avanços, bem como as técnicas aplicadas e resultados obtidos.

No estudo conduzido por Luyckx e Daelemans (2008) são exaltados os efeitos colaterais do uso de pequenas bases de dados em problemas de autoria. Os autores defendem o uso da estilometria e alertam sobre projeções acerca da eficácia dos classificadores e ca-

racterísticas, que podem não representar a realidade. O trabalho reforça decisões tomadas na nossa metodologia, ao avaliar características de estilo perante bases de dados maiores.

Hadjidj et al. (2009) usaram árvores de decisão e SVMs para atribuição de autoria com documentos da base *Enron* e obtiveram taxas de até 83% de acurácia. Ressaltamos que neste estudo os autores limitaram-se a um subconjunto com apenas 3 autores, o que é um bom resultado, mas não reflete a realidade de muitos problemas práticos.

Chen et al. (2011) extraíram 150 características linguísticas a partir de mensagens de email para verificação de autoria. Nos seus experimentos, foram usados 40 autores da base Enron, e foram obtidas taxas de acerto de 84% por meio de abordagens não supervisionadas, como análise de componente principal (PCA) e agrupamento.

Já no trabalho de Canales et al. (2011) observamos um desafio interessante. Com objetivo de autenticar estudantes durante a realização de exames online, foram extraídos padrões de digitação e características do estilo de escrita, totalizando 82 características. Utilizando o algoritmo dos k-vizinhos mais próximos (*K-Nearest Neighbors* - KNN), obtendo uma taxa de erro igual (TEI) de 30% num grupo de estudo com 40 estudantes.

Na pesquisa de BROCARDO et al. foi relatada uma maior dificuldade na identificação, verificação e caracterização de autoria de e-mails, pois estes documentos além de geralmente serem mais curtos, também são pouco estruturados em comparação a obras literárias. Como alternativa, Brocardo et al. (2013) sugerem uma técnica inovadora de aprendizagem supervisionada baseada na análise de *n-grams*, obtendo TEI de 14,35% também em um subconjunto da base de dados Enron.

Neste mesmo ano, (SAVOY, 2013) fez uma análise detalhada de abordagens probabilísticas generativas por meio do *Latent Dirichlet Allocation* - LDA (BLEI; NG; JORDAN, 2003) em atividades de atribuição de autoria. Usando como base um conjunto de 10.000 artigos publicados em jornais europeus, o autor criou um modelo para cada autor a partir das palavras contidas em todos documentos escritos pelo mesmo, com exceção daquele sob análise, para extrair características. Savoy (2013) afirma que o LDA apresentou resultados superiores à regra Delta e qui-quadrado na maioria dos casos. Já em comparação a *Naives Bayes* e KLD houve muitas oscilações. Ainda assim, o autor afirma que quanto maior o número de termos comuns por autor, melhor a performance do LDA (SAVOY, 2013).

Seroussi, Zukerman e Bohnert (2014) preconizam trabalhar com textos online, pois consideram mais desafiadores que romances e redações devido ao seu tamanho e número de autores candidatos. Para superar essas limitações os autores propõem a aplicação de

LDA para caracterização de autoria através de modelagem e distribuição probabilística de tópicos. O estudo sugere ter alcançado o estado da arte naquela época para uma série de bases de dados (PAN 11, IMDb1M , Blog) além de prover fortes indícios em sua capacidade de obter informações demográficas e traços de personalidade dos autores (SEROUSSI; ZUKERMAN; BOHNERT, 2014).

Ainda em 2014, foi realizada a primeira edição do desafio de atribuição de autoria proposto pelo PAN/CLEF [1]. Nesta competição, foram usados textos de autores em inglês, espanhol e grego. As melhores taxas de precisão alcançadas pelos competidores foram 93% na língua espanhola, 84% na inglesa e 82% em grego (VARELA, 2017).

Khonji, Iraqi e Jones (2015) também investigaram o uso de *ensembles* de ADs diante da base PAN12' (JUOLA, 2012). Por meio de um modelo RF composto por 1000 árvores, os autores foram capazes de acertar mais de 87,5% dos exemplos propostos. Apesar do bom resultado, os autores destacam outros modelos promissores, como o *Extra Trees*, que não foi explorado em seu trabalho.

Maitra, Ghosh e Das (2016) também aplicaram RF para verificação de autoria diante da base PAN 15, obtendo o terceiro melhor resultado para o idioma dinamarquês, com 0,75 de área abaixo da curva (*Area Under Curve* - AUC) usando características de estilo e frequência de palavras. Ainda com relação ao PAN 15, encontramos o trabalho de Pacheco, Fernandes e Porco (2015), onde foi realizada atribuição de autoria e definição de perfis (*profiling*) por meio da extração de características linguísticas, distribuição de palavras e tópicos abordados no texto. Para classificação, foi usado o algoritmo RF e uma técnica de *encoding* universal baseada em todos os trabalhos escritos pelos autores. A solução proposta apresentou 0,822 e 0,96 AUC como melhores resultados para dinamarquês e espanhol respectivamente.

Halvani, Winter e Pflug (2016) criaram um método para verificação de autoria capaz de ser utilizado em diversos idiomas. Para isso construíram um corpus de treinamento em mais de 6 idiomas e propuseram um método para verificação de autoria universal com características estilométricas estáticas. Tal método não se fez valer de técnicas elaboradas de AM ou PLN para construção de características e ainda assim demonstra ser eficaz. Algumas das categorias de atributos utilizadas foram pontuação *n-grams*, caracteres *n-grams*, palavras mais frequentes, prefixos, sufixos, prefixos e sufixos *n-grams*, prefixos-sufixos e sufixos-prefixos. Seu trabalho reporta uma taxa de acurácia média de 75%.

[1] https://pan.webis.de

Em 2017, VARELA inovou ao desenvolver uma abordagem de atribuição de autoria multilíngue que englobasse a língua portuguesa. Ele ressalta em seu trabalho que grande parte das pesquisas usam mecanismos de PLN específicos ou características de estilo relacionadas a um conjunto de idiomas, restringindo a portabilidade de características de um idioma para o outro. Por isso, o mesmo construiu um conjunto de características universais, as quais foram testadas em 5 idiomas (português, espanhol, inglês, francês e alemão), tanto para verificação como atribuição de autoria. Em sua pesquisa, o autor construiu duas bases de dados, uma com textos literários e outra com textos jornalísticos, separados por assunto e idioma.

Sua pesquisa utilizou o classificador SVM para criar modelos dependentes e independentes de autor. Ao fim, foram obtidas taxas de acerto médias entre 86-93% na atribuição e 95-98% para verificação de autoria usando uma abordagem *top-list* (1, 3, 5 e 10). Esta pesquisa se mostra altamente relacionada ao estudo de Diederich et al. (2003), pois se enquadram no mesmo domínio e usam técnicas semelhantes. Os autores deste trabalho propuseram o uso de SVM com diversos tipos de *kernel* e fizeram uso de características estilométricas para identificar autores de matérias jornalísticas em alemão.

Yang et al. (2018) também propõem um novo algoritmo para atribuição de autoria denominado *Topic Drift Modeling* - TDM, essa abordagem tenta solucionar o problema a partir dos tópicos principais do texto levando em conta fatores temporais, como a ordem em que as palavras aparecem. Em seus resultados, eles comparam o algoritmo proposto com abordagens clássicas, como SVM, variações do LDA e RNA de alimentação direta diante de 3 renomadas bases de dados. O modelo proposto por YANG et al. foi superior aos modelos base de referência nas bases PAN11' e IMDb62.

Também em 2018, CUSTÓDIO; PARABONI construíram um *ensemble* baseado na regressão logística multinomial para resolver o problema de atribuição de autoria na competição PAN-CLEF 2018 (POTTHAST et al., 2017). O modelo proposto foi capaz de alcançar uma taxa média de acerto de 71% considerando todos os idiomas (inglês, francês, italiano, polonês e espanhol), superando os modelos de referência da própria organização, que se baseiam em *n-grams* e SVMs.

Um dos trabalhos recentes no campo de estudo desta pesquisa foi desenvolvido por Stavngaard et al. (2019). Os autores se propõem a verificar se determinada redação foi escrita pelo aluno ou um escritor fantasma, serviço que tem crescido bastante nos últimos anos. Neste trabalho, foi construída uma base de dados composta por mais de 130 mil

redações de escolas dinamarquesas, que serviu para treinamento de uma rede neural de arquitetura similar a CNN. O trabalho apresenta como resultado uma acurácia de 87,5% e AUC de 0,947 após balanceamento manual da base.

A última edição da competição PAN/CLEF (2020) ganhou novas categorias, incluindo *profiling* de celebridades, verificação de autoria com múltiplos domínios e detecção de alterações de estilo (BEVENDORFF et al., 2020a), recebendo um recorde de 81 submissões aceitas. No compilado de submissões reportado são enumeradas diferentes propostas e modelos base de referência interessantes para a comunidade científica (BEVENDORFF et al., 2020b).

Em outro panorama, também destacamos o trabalho de Halvani, Graner e Regev (2020) que ao invés de tentar desbancar o estado da arte, procurou investigar a interpretabilidade de modelos e utilização de características independentes de tópico, porque em muitos casos os resultados da verificação não se fundamentam no estilo de escrita, mas no assunto do texto.

Como se pode observar, há uma vasta quantidade de trabalhos relacionados ao tópico nos últimos anos devido ao destaque e importância que o tema têm obtido. Dentre os diversos algoritmos e abordagens propostas, destacamos os mais frequentes, que são a aplicação e criação de características de estilo, por meio do uso massivo do PLN e aplicação de algoritmos de AM como KNN, SVM, RF, LDA e RNAs para classificação.

Entretanto, ressaltamos a pequena quantidade de trabalhos na língua portuguesa, que se afunila ao tratarmos de pesquisas com fins educacionais. Por isso, acreditamos que há uma lacuna a ser preenchida por essa dissertação de mestrado.

4 METODOLOGIA

Neste capítulo serão apresentadas as etapas metodológicas utilizadas para o desenvolvimento desta pesquisa. O mesmo está dividido em seções que descrevem desde a análise inicial do problema até a aquisição, tratamento e análise dos dados.

4.1 CARACTERIZAÇÃO DO PROBLEMA

A área de atribuição de autoria tem sido impulsionada pelo volume de dados textuais na internet, especialmente em aplicações voltadas para identificação de autoria de postagens em blogs e redes sociais, matérias de jornal e trabalhos acadêmicos. Mesmo com esse crescimento, durante a revisão da literatura notamos uma escassez de pesquisas voltadas para o desenvolvimento e utilização de mecanismos de apoio às atividades educacionais por meio da atribuição de autoria. Se nos restringirmos a documentos na língua portuguesa, há um volume ainda menor de trabalhos, especialmente comparado aos idiomas inglês e espanhol. Isso está relacionado não só a fatores econômicos e demográficos, mas também a um maior volume de bases de dados e ferramentas de PLN (OLIVEIRA; GOMES, 2014).

Do ponto de vista educacional, observamos mudanças nas metodologias de avaliação de aprendizagem, que têm se voltado para maior utilização de atividades práticas e resolução de problemas, que se concretizam por meio de exercícios, pesquisas, trabalhos e apresentações. Acredita-se que esse tipo de atividade está mais alinhada a fatores centrados no aprendizado do aluno, tal como a valorização da experiência de aprendizado, empregabilidade, inclusão social e desenvolvimento de competências (KRAEMER, 2005).

Como foi visto na evolução da escrita, os trabalhos e atividades pedagógicas antes escritos à mão no papel, nos quais havia uma possibilidade maior de identificação da grafia e reconhecimento de autoria,migraram para arquivos dentro dos computadores e perderam tais características (YOUNAS et al., 2018). Isso significa que movemos rapidamente para um cenário preocupante, abordagens pedagógicas solicitam essas atividades, que por sua vez aumentam em quantidade e retroalimentam a internet. As atividades têm perdido suas características originais e são contaminadas pela abundância de recursos online e facilidade de reprodução.

Esses fatores podem ser sintetizados em: *a*) a perda das características relacionadas à grafia *b*) facilidade de reprodução do vasto conteúdo disponível na web e *c*) desenvolvimento de mecanismos de comunicação e busca capazes de prover respostas pré-fabricadas.

O maior grau de incerteza sobre a autoria de trabalhos digitalizados pode acarretar numa maior tendência por parte das instituições de ensino, professores e tutores em adotarem metodologias tradicionais de verificação de aprendizagem, como exames e prova oral, o que não é de todo mal. Porém diversos educadores apontam que o conceito de "teste" ou "prova" tem raízes remotas e deve ser um meio e não o fim (DEPRESBITERIS, 2017) .

Os altos investimentos feitos em educação exigem um processo de avaliação que demonstre resultados sem deixar de considerar o componente humanista (VIANNA, 2005). Segundo Depresbiteris e Tavares (2017) a avaliação corresponde a aprendizagem dos alunos e está sob responsabilidade do professor. Consequentemente, os instrumentos de avaliação precisam ser elaborados para auxiliar os professores na tomada de decisão

Por isso, observa-se que a criação de ferramentas de verificação de autoria pode colaborar para adoção de abordagens de avaliação e construção de conhecimento mais modernas com maior segurança. Além disso, a existência de tais mecanismos pode desencorajar estudantes a praticar atos como plágio, divisão e compra de trabalhos acadêmicos.

Acreditamos que o ensino superior brasileiro supre algumas das premissas para nossa investigação, pois satisfaz dois fatores essenciais: *i*) há um volume maior de pesquisas e trabalhos escritos em cursos de graduação, se comparado ao ensino médio e fundamental e *ii*) o público alvo é mais maduro, visto que já passaram pela infância e buscam se estabelecer profissionalmente por meio do ensino superior. A partir deste cenário, iremos inicialmente nos aprofundar na investigação de trabalhos e pesquisas de alunos de nível superior, escritos na língua portuguesa, para entender se essa tarefa é solúvel por meio das atividades de análise autoria com apoio da AM e estilometria.

4.2 GERAÇÃO DA BASE DE DADOS

Através da nossa busca na web identificamos muitos trabalhos e bases de dados voltadas para atribuição de autoria, mas nenhum destes atingiu por completo os requisitos desta pesquisa. Dentre os mais relevantes, podemos citar PAN19' e PAN20', as bases dis-

poníveis no portal brasileiro de dados abertos[1] e a base de dados elaborada por Varela (2017).

O PAN19' e 20' são grandes bases de dados que compreendem várias tarefas relacionadas à análise de autoria e vêm ganhando notoriedade nos últimos anos por meio de competições anuais. Eles são compostos por documentos em idiomas estrangeiros (inglês, espanhol, francês e italiano) e nos últimos anos trouxe textos escritos por fãs que simulam os verdadeiros autores de obras literárias, também conhecidos como *fandoms* [2].

No portal de dados abertos do Brasil, existem algumas bases de dados voltadas para disseminação de informações educacionais do nosso país, porém a maioria destes está relacionada a censos, taxas de desempenho escolar e transparência no gerenciamento da educação pública no Brasil.

Já a base construída por (VARELA, 2017) é composta por textos jornalísticos em diversos idiomas, incluindo o português, separados em categorias de acordo com o assunto do texto (política, esportes, economia). Apesar desta base não estar voltada para propósitos educacionais, a mesma apresenta um bom volume de textos em português, que pode ser usado para fins de atribuição de autoria. Por isso, a mesma foi utilizada como referência em etapas posteriores desta pesquisa.

Desta forma, não encontramos nenhuma base composta por trabalhos escolares com autoria como pretendemos. Por isso, tornou-se necessária a obtenção e construção de uma base de dados própria para atender tal objetivo. Por meio de uma parceria com professores de uma instituição privada de ensino superior na cidade do Recife, foi possível obter acesso à plataforma digital de acompanhamento (*Google Classroom*) utilizada em algumas das disciplinas, nas quais se obteve acesso às atividades dos estudantes. O acesso a essa base de dados foi apenas o primeiro dos muitos desafios enfrentados.

Inicialmente foi realizada a extração e fichamento manual dos trabalhos. Os mesmos estavam espalhados entre diversas salas de aulas que representam diferentes períodos letivos e disciplinas. Cada disciplina possui atividades propostas e trabalhos enviados pelos alunos. Ao final deste fichamento, constatamos que havia um baixo volume de trabalhos, tanto de maneira geral, como por autor, com raras exceções.

[1] http://dados.gov.br

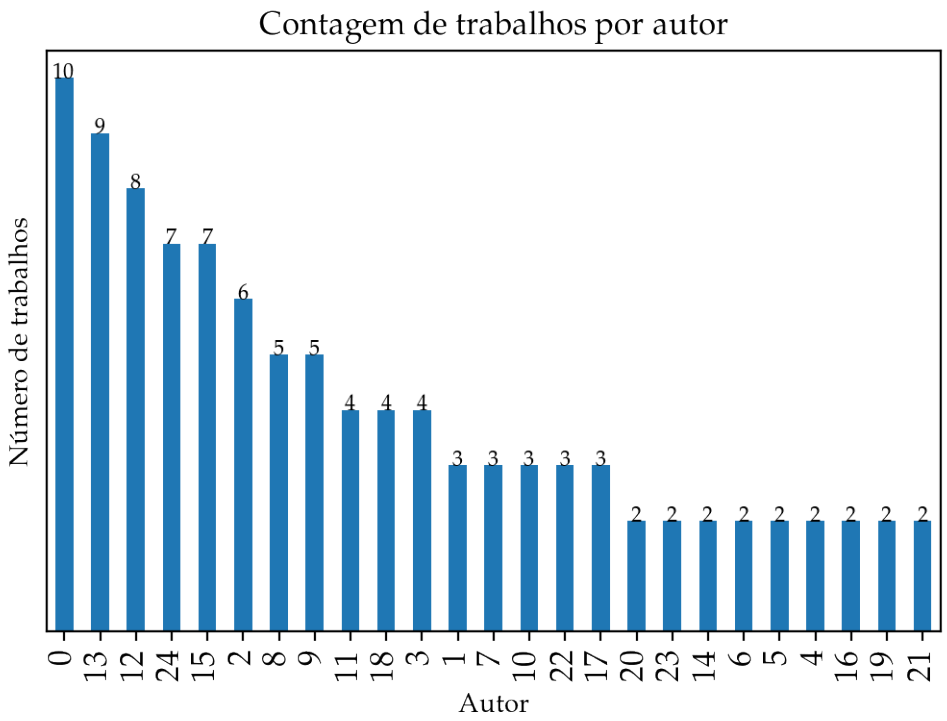
[2] *Fandoms* são comunidades ou subcultura de fãs construídas em cima da cultura *pop*, incluindo livros, filmes, jogos e músicas. Geralmente incluem consumidores desta cultura, mas também podem ser compostos por fãs que participam ativamente na criação de material criativo, como textos, musicas ou vídeos (DESIGN, 2019)

Os trabalhos, por sua vez, se encontravam em diversos formatos (pdf, doc, docx, txt) e alguns estavam em formato de imagem. As imagens que retratavam documentos digitados no computador puderam ser convertidas para arquivos de texto por meio da ferramenta de visão computacional *"Top OCR"* [3]. Trabalhos entregues no formato de foto de documentos manuscritos não puderam ser convertidos e precisaram ser excluídos do conjunto analisado. Os arquivos foram organizados de acordo com a estrutura de disciplinas e atividades, e o nome completo dos autores foi utilizado para nomear os arquivos.

Para preservar o anonimato dos autores, foi aplicado um codificador para ofuscação de identidade dos autores. Desta forma, ao longo desta pesquisa o nome dos autores desta base serão representados por valores numéricos.

Ao todo, a base de dados foi composta por 116 trabalhos distribuídos entre 25 alunos. Dentre estes, 14 trabalhos foram realizados em grupo (por mais de um autor) e foram inicialmente removidos da base de dados. Dentre os trabalhos individuais (102), observou-se também um grande desbalanceamento na distribuição de trabalhos por autor, conforme é observado na Figura 9.

Figura 9 – Contagem de trabalhos individuais por autor



Fonte: Elaborado pelo autor

[3] Disponível em http://www.tucows.com/preview/403821

Podemos observar que 35% (9) dos autores apresentam apenas 2 trabalhos escritos e que apenas 20% (5) dos autores possuem mais de 6 trabalhos escritos.

Ao entrevistar o principal professor parceiro, foram informados alguns motivos que justificam o fato observado *i*) alguns alunos desistiram de determinadas disciplinas ou do curso; *ii*) alguns alunos preferiam ou só podiam entregar os trabalhos presencialmente; *iii*) algumas atividades eram complementares ou opcionais; *iv*) algumas disciplinas tiveram mais atividades escritas que outras; *v*) alguns alunos se dedicam mais às atividades que outros; *vi*) poucos professores adotaram a ferramenta online para envio de atividades até o momento.

Para a extração do conteúdo original dos arquivos e construção da base de dados, foi desenvolvido um sistema[4] na linguagem *Python* capaz de ler arquivos em diferentes formatos, extrair o conteúdo e executar uma limpeza nos dados para eliminar termos relacionados à origem do documento, tais como nome dos estudantes, professores, disciplinas, cabeçalhos e imagens. A base de dados resultante deste processo se encontra em formato CSV e possui apenas duas colunas, o conteúdo textual pré-processado (*Text*) e nome ofuscado do autor *Author*).

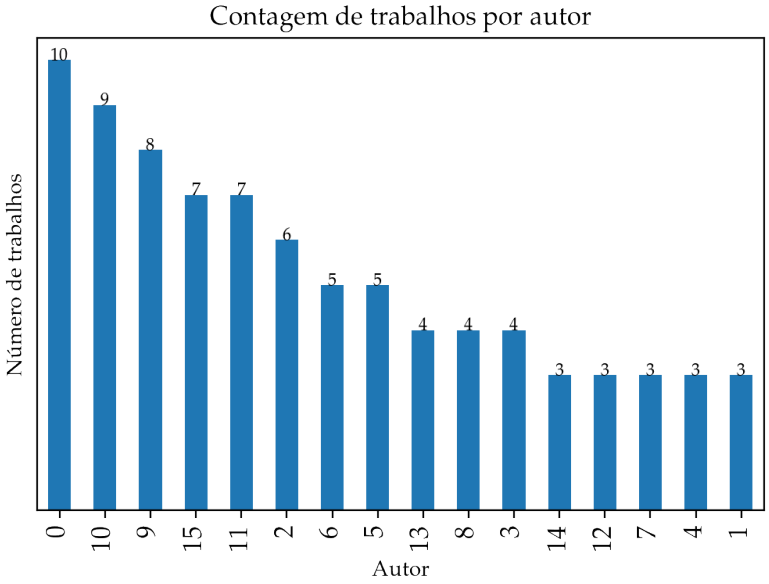
Por conta da estratégia utilizada durante a construção e avaliação dos modelos de classificação, tornou-se necessário criar um ponto de corte de pelo menos 3 trabalhos por autor para respeitar os conjuntos de treinamento, teste e validação. Além disso, conforme mencionado anteriormente, alguns trabalhos foram realizados em equipe, configurando um problema *multilabel* com múltiplos autores num mesmo documento, fugindo do escopo da pesquisa. A partir destes dois critérios, a base de dados foi reduzida para 84 documentos distribuídos entre 16 autores distintos conforme a Figura 10.

A distribuição acima representa a configuração final dos trabalhos selecionados e será referenciada como base de estudantes ao longo deste trabalho.

Ao nos aprofundarmos na base de estudantes, é possível observar a distribuição das palavras de duas formas: *a*) considerando a autoria (*tokens* por autor) ou *b*) desconsiderando a autoria (*tokens* por trabalho). Desconsiderando a autoria, temos um total de 46,477 palavras distribuídas entre 84 trabalhos, contabilizando uma média ($\overline{x}$) de 553,29 *tokens* por documento. Seu vocabulário contém apenas 8390 palavras únicas, o menor documento é composto por 49 *tokens* e o mais extenso por 5894 *tokens*. A alta variabilidade de *tokens* por autor é confirmada pelo desvio padrão (σ) de 782,08.

[4] Disponível em https://github.com/daanielvb/text-extractor

Figura 10 – Contagem de trabalhos individuais por autor com ao menos três exemplos



Fonte: Elaborado pelo autor

Levando em conta o número de autores (16), temos $\overline{x} = 782,08$ *tokens* por autor. Entretanto, observamos que existem autores que escreveram mais que outros, e tais valores médios não representam o comportamento real. Por exemplo, o autor "0" é o que possui mais *tokens* na base dos estudantes com 8655 *tokens*, enquanto que do lado oposto, o autor "4" possui apenas 567 *tokens* em obras com sua autoria. Observamos que o desvio padrão do número de *tokens* por autor é ainda mais elevado ($\sigma = 2143,41$), confirmando a discrepância de *tokens* por autor.

Uma outra medida interessante é a verificação do número de *tokens* por autor por documento. Dado que o autor "0" possui 8655 *tokens* distribuídos em 4 documentos de sua autoria, sua média de *tokens* por documentos seria de $8655/4 = 2163,75$ que é a maior média de *tokens* por documento A menor média é do autor "4" com apenas 189 *tokens* por documento. Mais uma vez calculamos $\overline{x} = 782,08$ e $\sigma = 503,52$, e observamos uma variabilidade menor na razão de *tokens* por documentos por autor.

Esta análise preliminar nos leva a concluir que, de maneira geral, a base de dados é desbalanceada, principalmente devido aos motivos citados pelos professores parceiros, destacando o diferente número de trabalhos por autor. Ainda assim, ao analisar isoladamente cada autor verificamos que a quantidade de palavras por documentos de cada autor é homogênea.

4.2.1 Geração das bases de dados comparativas

Como é amplamente discutido em trabalhos de AM, o tamanho e a representatividade dos dados são fatores críticos para o sucesso ou insucesso em atividades de classificação (CATAL; DIRI, 2009). De antemão, sabemos que a base de estudantes é extremamente pequena e passível de baixa representatividade no que se refere à autoria dos documentos, principalmente nos casos com apenas 3 trabalhos por estudante. Segundo Luyckx e Daelemans (2008), quando há uma limitação de dados sobre autores específicos, a tarefa de atribuição de autoria se torna extremamente difícil.

Por isso, decidiu-se construir uma base de dados comparativa com textos jornalísticos, visto que esses textos podem ser encontrados em nosso idioma com abundância na *web* e pelos bons resultados obtidos em problemas de atribuição e verificação de autoria na literatura (VARELA, 2017) (DIEDERICH et al., 2003). A ideia principal é comparar se os resultados dos experimentos em cima da base de estudantes é similar aos resultados em bases maiores, balanceadas e de contexto distinto.

Coletamos manualmente textos jornalísticos de um reconhecido portal de notícias brasileiro, buscando colunas dos jornalistas em destaque no portal. Foram coletados 10 textos por autor, de um total de 10 autores, totalizando 100 exemplos. O tema das matérias não foi levado em conta, mas este portal é notoriamente conhecido por textos de cunho político e econômico. Ao longo deste trabalho, iremos referenciar esta base de dados como a base de notícias.

Durante a construção da base de notícias, optamos por coletar um número similar de exemplos e classes (autores) com relação à base de estudantes, mas com um contraste intencional que foi o mesmo número de documentos por autor. O motivo é que a base de estudantes possui desbalanceamento e este é um fator importante em problemas de classificação (AKBANI; KWEK; JAPKOWICZ, 2004). Assim, a base de estudantes é altamente desbalanceada e a de notícias é completamente balanceada. Além disso, destacamos que os autores dos textos da primeira base são estudantes de graduação, enquanto os autores da segunda base são jornalistas, com pelo menos alguns anos de experiência em comunicação escrita. Tais diferenças de perfil podem afetar diretamente no vocabulário e estrutura textual de cada base.

Apesar de balanceada, a base de notícias ainda é relativamente pequena (100 exemplos) e potencialmente pouco representativa. Por causa disso, decidimos utilizar uma terceira

base de dados, ainda mais robusta, para ter mais segurança com relação aos resultados obtidos. A base escolhida foi a previamente descrita no Capítulo 4.2 e disponibilizada por Varela (2017). Esta é composta por 3.000 documentos distribuídos igualitariamente entre 100 autores, o que resulta em 30 documentos por autor. Além do texto e autor, essa base também possui o assunto do texto em questão, que se divide entre 10 tópicos distintos.

O sistema de extração de textos desenvolvido neste trabalho foi ajustado para extrair o conteúdo textual destas bases e construir bases no formato desejado (.csv). Daqui em diante, esta base será referenciada como base de Varela.

4.2.2 Transformação das bases de dados

De posse das bases de estudantes, notícias e Varela, primordialmente compostas pelo autor, conteúdo textual e assunto (apenas Varela) iniciamos uma investigação sobre as melhores formas de representação dos documentos para processamento pelos algoritmos de AM. Seguindo a literatura, selecionamos três formas de representação dos documentos *i*) conversão para um vetor de frequência de palavras com TF-IDF; *ii*) construção de uma matriz de características por meio de *word embeddings* pré-treinados (GLoVe, Word2Vec e FastText); *iii*) transformação do documento em um vetor de características estilométricas.

O sistema anteriormente concebido para extração do conteúdo textual dos documentos originais foi expandido, recebendo capacidade de conversão dos textos para as representações mencionadas. Para isso, foram utilizadas bibliotecas de ciência de dados *Pandas*[5], *Sklearn*[6], *Gensim*[7], *TensorFlow*[8] e *Keras*[9] suportando a representação e transformação das bases.

Na primeira representação, os documentos foram convertidos para vetores de frequência de termos (TF) e frequência inversa dos documentos (IDF) considerando *n-grams* de tamanho entre 1 e 4. Optou-se pela não aplicação de *stopwords* para preservar o conteúdo original dos textos.

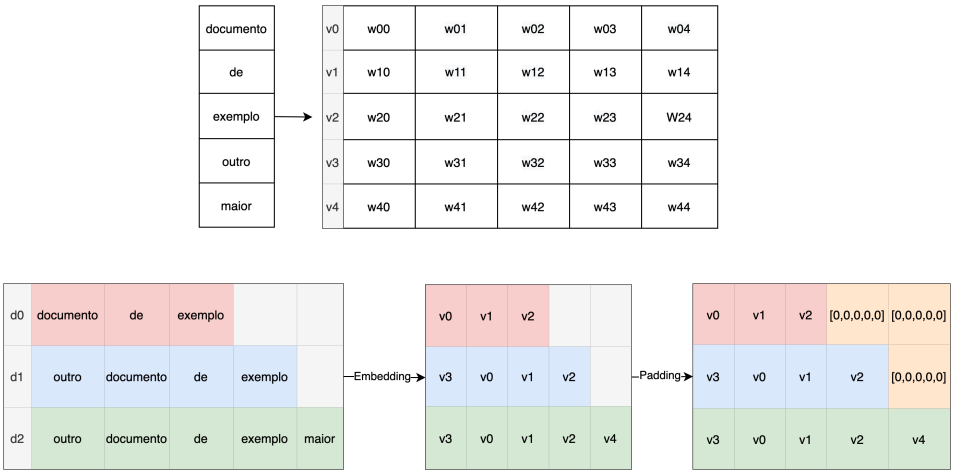
A segunda representação fez uso de *word embeddings* pré-treinados. Foram realizados experimentos usando Word2Vec (CBOW), GloVe, e FastText (*Skip-gram*) treinados na

[5] https://pandas.pydata.org
[6] https://scikit-learn.org/stable/
[7] https://github.com/RaRe-Technologies/gensim
[8] https://www.tensorflow.org
[9] https://keras.io

língua portuguesa com o *corpora* STIC 2017 (HARTMANN et al., 2017) nos tamanhos de 50 e 100 dimensões por palavra. Não utilizamos *embeddings* com mais dimensões devido ao alto custo computacional associado. Os *embedding* utilizados foram disponibilizados no site do Núcleo Interinstitucional de Linguística Computacional (NILC) da USP[10].

A partir de um documento D, cada *token* presente no mesmo foi substituído por um vetor de tamanho igual ao número de características (dimensões) do *embedding* D. Nesta abordagem, palavras presentes nos documentos, mas ausentes nos *word embeddings* foram descartadas por não haver vetores ponderados que pudessem representá-las. Outro detalhe é que devido à presença de documentos de tamanhos distintos, foi necessário uniformizar seu tamanho, realizando uma operação de *padding*. Considerando que o maior documento do nosso *corpora* tem tamanho T', caso D seja menor que T', o *padding* preencherá D com matrizes de zeros em seu final até que o mesmo obtenha o tamanho T'. Esse procedimento aumenta a esparsidade dos nossos vetores porém é necessário para uniformizar as matrizes para utilização pelos algoritmos de AM.

Figura 11 – Transformação de documentos numa matriz numérica por meio de word embeddings com 5 dimensões e padding



Fonte: Elaborado pelo autor

$$TF\text{-}IDF = (V * n)$$

$$Embeddings = (T^{max} * D) * n$$

[10] http://nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc

O processo de transformação de documentos em vetores ponderados é ilustrado na Figura 11 na qual observamos que os documentos são transformados em matrizes numéricas através de um *embedding* estático pré-treinado e em seguida têm seu tamanho uniformizado pelo *padding*.

Por fim, a última representação utilizada neste trabalho se baseou na criação de características estilométricas. Conforme vimos na Seção 2.4.4.1, existem diversos trabalhos que utilizaram características específicas para representar características associadas ao estilo de escrita de seus autores com sucesso. Com isso, o sistema de extração de textos foi estendido para realizar a conversão de documentos para vetores estilométricos. A lista completa de características desenvolvidas ou aplicadas neste trabalho está descrita na Seção 4.2.3.

Para a conversão, foi desenvolvido um algoritmo que recebe como entrada um conjunto de documentos e computa os valores das características de estilo para cada um dos documentos a partir do seu texto e todo *corpora*. O algoritmo utilizou diversas técnicas de PLN, como tokenização, separação de palavras em sílabas, criação e utilização de *stopwords* e *keywords* relativas ao domínio, *pos-tagging*, reconhecimento de entidade nomeada, *chunking*, verificação ortográfica e uso de expressões regulares (*Regex*). Ao fim deste processo, cada documento foi transformado em um vetor de 74 características estilométricas.

Esses vetores foram exportados em formato CSV e transformados em uma nova base de dados. Desta forma, todas as bases textuais passaram a ter além da original, uma versão de estilo baseada em suas características.

4.2.3 Características estilométricas

A identificação de autoria por meio do estilo de escrita é uma abordagem frequente na literatura. Nesta atividade é importante preservar ao máximo o conteúdo original do texto, pois ele pode destacar características relevantes daquele autor em suas obras. De acordo com a literatura, essas características podem ser agrupadas entre os grupos lógicos explicados na Seção 2.4.4.1 e terão sua implementação explanada a seguir.

Ressaltamos que durante esta fase, construímos e exploramos outras características além das 74 aqui selecionadas. Optamos pela remoção de algumas características para excluir elementos não relacionados ao estilo de escrita ou pela possibilidade de enviesamento dos classificadores. Por exemplo, características influenciadas por um volume maior

de documentos escritos por um autor em comparação a outros poderiam gerar viés nos classificadores, visto que temos bases desbalanceadas.

4.2.3.1 Características lexicais

O grupo de características lexicais representa a forma de escrever dos autores. A forma de escrever compreende características como a frequência de parágrafos, média e variância do tamanho de parágrafos e frases, média de sílabas por palavra e a frequência de palavras monossilábicas no documento. Segundo Lex, Juffinger e Granitzer (2010) esse grupo de características ajuda na análise de autoria pois são simples de extrair e refletem traços únicos do estilo de escrita dos autores, como a distribuição de palavras e parágrafos. Os autores sugerem selecionar características independentes do tópico para resultados mais generalistas.

Figura 12 – Características Lexicais

Grupo	Feature	Observação
Lexicais	Tamanho médio das palavras	
	Tamanho médio das sentenças	
	Desvio padrão do tamanho das sentenças	
	Tamanho médio dos paragráfos	
	Paragráfos	
	Média de sílabas por palavra	Algoritmo desenvolvido a partir da combinações de soluções disponíveis na Web
	Monossilabas	

Fonte: Elaborado pelo autor

Para construir as características lexicais, utilizamos a distribuição de frequência de palavras e caracteres especiais. Para identificação de parágrafos, foram analisados caracteres de quebra de linha (`\n`, `\t` e `\r`) em sequência. A separação do texto em frases se deu por meio do caractere de pontuação. Através do número de ocorrências derivamos outras medidas estatísticas, como média, frequência, desvio padrão e razão entre frases e palavras.

Para separar as palavras em sílabas, encontrou-se um maior desafio visto que essa funcionalidade não estava disponível nas principais bibliotecas de PLN na língua portuguesa. Analisamos duas alternativas para essa limitação: a biblioteca *Pyphen*, que se ampara em um dicionário de palavras, e o algoritmo desenvolvido por SILVA (2011) que se baseia em regras gramaticais. Através de uma análise exploratória nas palavras e sílabas geradas pelos dois algoritmos, observou-se que o algoritmo de SILVA foi capaz de separar as sílabas com mais precisão do que o *Pyphen*, entretanto, o primeiro não conseguia separar corretamente palavras com "ss"e "rr". Desta forma, as duas abordagens foram combinadas para alcançar uma separação de sílabas mais precisa.

A capacidade de separação de palavras em sílabas possibilitou a construção de uma característica de frequência de monossílabos e permitiu calcular o índice de inteligibilidade *Flesch-Kincaid* do grupo de características relacionadas a riqueza de vocabulário.

4.2.3.2 Características baseadas em caracteres e palavras-chave

Para criação das características baseadas em caracteres e palavras-chave, mensuramos a incidência de termos e palavras ao longo do texto por meio do módulo *FreqDist* [11] do NLTK. Caracteres muito comuns, como as letras do alfabeto e pontuações frequentes, como pontos e vírgulas, foram removidas dada sua baixa variabilidade entre os documentos e resultados preliminares durante os experimentos. Em contrapartida, símbolos de pontuação menos comuns como o travessão, exclamação e ponto e vírgula foram amplamente adotadas por exaltarem peculiaridades sobre os autores. Complementando este grupo, também foram capturados operadores lógicos, como "e", "ou", "nem", "se", "desde que" de maneira semelhante ao que foi sugerido por Scarton e Aluísio (2010) para construção de características.

As características baseadas em *n-grams* também foram enquadradas neste grupo. O objetivo foi observar os T *n-grams* mais frequentes em cada *corpora* e calcular a sua frequência de aparição em cada documento como proposto por Stamatatos (2013). Após experimentos iniciais, decidimos utilizar apenas os três *n-grams* (T=3) mais frequentes em cada *corpora*. Além disso, também limitamos os tamanhos dos *n-grams* para $n = [1, 3, 4, 5]$. Por exemplo, imaginando que os T=3 *unigrams* identificados foram [”*a*”, ”*que*”, ”*o*”], a característica *top-unigram* é calculada por meio da frequência de aparições desses termos

[11] Documentação disponível em: http://www.nltk.org/api/nltk.html?highlight=freqdist

Figura 13 – Características baseadas em caracteres

Grupo	Feature	Observação
Baseada em caractere	Ponto	Frequência de aparição
	Vírgula	
	Exclamação	
	Dois pontos	
	Citações	Presença de aspas duplas ou simples
	Operadores lógicos	e, ou, nunca, sem que, jamais, desde que...
	Top 1-gram	Frequência no documento dos n-grams mais comuns
	Top 3-gram	
	Top 4-gram	
	Top 5-gram	
	Palavras capitalizadas	Palavras iniciadas com a primeira letra maíuscula
	Dígitos	[0-9]
	Quebra de linha	Frequência de caracteres especiais (/n /r /t)

Fonte: Elaborado pelo autor

no documento dividido por 3. Optamos pela remoção de *bigrams* após observar grande redundância com relação aos *trigrams* e *collocations* de tamanho 2.

4.2.3.3 Características sintáticas

As características sintáticas indicam o papel morfológico dos *tokens* e frases. Por serem características mais elaboradas, é necessário utilizar algum mecanismo de PLN para extração destas características. Nosso primeiro objetivo nesta direção foi obter as classes gramaticais de cada *token* dentro do documento, que pode ser feito recorrendo a análise sintática da frase com o uso de um *POS-Tagger*. Como não foi encontrado um *POS-Tagger* pré-treinado na língua portuguesa dentro da biblioteca NLTK, optamos por treinar um *POS-Tagger* próprio.

O *POS-Tagger* foi treinado em cima da base de dados MACMORPHO (ALUÍSIO et al., 2003), que é composto por mais de um milhão de palavras em português do Brasil,

extraídos de matérias do jornal Folha de São Paulo em 1994. O fato desta base apresentar documentos com palavras sintaticamente anotadas nos permitiu criar um *tagger* próprio. Durante o treinamento foi usada uma proporção de aproximadamente 70% para treinamento (35.000 frases) e 30% para testes (16.397 frases), o *tagger* treinado obteve 89% de acurácia diante do conjunto de testes. Também foram usados *taggers* de *trigram*, *bigram* e *unigram* como mecanismo alternativo para termos desconhecidos pelo classificador.

De posse do *POS-Tagger* treinado, anotamos os *tokens* dentro das seguintes classes gramaticais: adjetivos, advérbios, artigos, conjunções, pronomes, preposições, substantivos e verbos. Também criamos uma característica de correlação entre pronomes e preposições, que é sugerida como medida de complexidade sintática por (FRANÇOIS; FAIRON, 2012). Os termos que não conseguiram ser classificados pelo *tagger* foram anotados com um valor especial que também foi convertido em característica.

Além destas, de maneira similar ao trabalho de Scarton e Aluísio (2010), foram implementados dois grupos sintáticos de palavras: palavras de conteúdo (substantivos, adjetivos, advérbios e verbos) e palavras funcionais (artigos, preposições, pronomes e conjunções).

Outra possibilidade oriunda da anotação sintática foi a análise da estrutura sintática das frases. Para isso, foi implementado um mecanismo de busca por meio de *chunking* e expressões regulares capaz de identificar sintagmas nominais (SN) e verbais (SV). A extração dos sintagmas foi feita por meio de expressões regulares.

$$SN = (ARTIGO^{?} ADJETIVO^{*} (SUBSTANTIVO|PRONOME)^{+})$$

$$SV = (VERBO)^{?} ADVÉRBIO^{*} VERBO^{+})$$

Nestas expressões, deve-se interpretar o símbolo de interrogação como zero ou uma ocorrência do termo, o símbolo de asterisco como zero ou mais ocorrências do termo e o símbolo de adição por uma ou mais ocorrências do termo.

Características sintáticas mais detalhadas também puderam ser extraídas por meio de adaptações da biblioteca *SpaCy* (HONNIBAL; MONTANI, 2017). As características criadas representam flexões de gênero (masculino, feminino), plural e singular, pessoas do discurso (primeira e terceira pessoa) e tempos verbais mais comuns (passado, presente e futuro). A ausência de verbos na segunda pessoa se deve a não captura de *tokens* nesta categoria pela biblioteca.

Figura 14 – Características Sintáticas

Grupo	Feature	Observação
Sintáticas	Adjetivos	Obtidas por meio de POS-Tagger próprio treinado no dataset MAC_MORPHO
	Advérbios	
	Artigos	
	Substantivos	
	Preposições	
	Verbos	
	Conjunções	
	Pronomes	
	Pronomes por preposição	
	Temos não tageados	Palavras que o *POS-Tagger* não consegui classificar
	Palavras de conteúdo	Substantivos, adjetivos, advérbios e verbos
	Palavras funcionais	Artigos, preposições, pronomes e conjunções
	Frases nominais	Obtidas por meio de expressões regulares, *Tree parsing* e *Chunking*
	Frases verbais	
	Palavras no gênero masculino	Obtidas por meio da biblioteca Spacy
	Palavras no gênero feminino	
	Palavras no singular	
	Palavras no plural	
	Verbos na primeira pessoa	
	Verbos na terceira pessoa	
	Verbos no passado	
	Verbos no presente	
	Verbos no futuro	

Fonte: Elaborado pelo autor

4.2.3.4 Características semânticas

A criação de características semânticas demandou mecanismos mais sofisticados de PLN, uma vez que as palavras podem exercer diversos papéis semânticos de acordo com a frase e o contexto. O reconhecimento de entidade nomeada (*Named Entity Recognition* - NER) é um algoritmo de PLN para anotação semântica fundamentado em abordagens linguísticas ou estatísticas, que leva em conta os *tokens*, sua anotação sintática e frase em que estão inseridos.

Apesar de grandes dificuldades para identificar bibliotecas ou modelos pré-treinados de NER no idioma português, conseguimos reaproveitar o trabalho de Pires (2017). Neste

trabalho, o autor treinou classificadores usando a primeira e a segunda versão da base HAREM (SANTOS et al., 2006) (FREITAS et al., 2010). O autor disponibilizou os modelos construídos sobre diversos *frameworks*, porém devido a detalhes de implementação desta pesquisa, o modelo que utilizamos se baseou na opção provido com suporte para o NLTK. Tal modelo obteve aproximadamente 31% de precisão na medida $f1$ de acordo com o autor. O mesmo é capaz de classificar termos em português dentro das categorias da base HAREM que retratam entidades relevantes, como: "coisa", "local", "pessoa", "tempo" e "valor".

O HAREM possui uma lista de categorias que vai além das entidades da Figura 15. Optamos pela remoção de alguns destes grupos por considerá-los não relevantes para a nossa pesquisa ou pela falta de exemplos categorizados durante os experimentos.

Figura 15 – Características Semânticas

Grupo	Feature	Observação
Semânticas	Frequência de entidades nomeadas	
	Entidade abstração	Categorias de entidades nomeadas presentes no dataset HAREM
	Entidade acontecimento	
	Entidade coisa	
	Entidade local	
	Entidade organização	
	Entidade obra	
	Entidade outro	
	Entidade pessoa	
	Entidade tempo	
	Entidade valor	

Fonte: Elaborado pelo autor

4.2.3.5 Características de riqueza de vocabulário

Até o momento, foram apresentadas análises superficiais das palavras e frases em que estão inseridas. Entretanto, ao nos aprofundarmos em características mais complexas, como os índices de coerência, coesão e legibilidade, nos deparamos com uma necessidade

de maior compreensão da estrutura das frases, pois esses índices se baseiam numa investigação mais profunda sobre a construção das frases e sua complexidade, levando em conta não só rotulagens sintáticas e semânticas, mas também flexões de gênero, plural e função exercida pelas palavras dentro da frase (VARELA, 2017).

Este tipo de análise vertical demanda ferramentas de PLN mais robustas, como o *Coh-metrix* (GRAESSER et al., 2004), que não foram encontradas abertamente no idioma e sistema propostos. Por causa disso, ao invés de calcular índices de coesão e coerência, decidimos representar essa categoria de características por índices de riqueza de vocabulário. Apesar de distintas, realizamos essa substituição respaldados na premissa da necessidade de vocabulário rico para articulação de ideias e palavras corretamente, que estão intimamente ligadas à coesão e coerência textual (FREEBODY; ANDERSON, 1983). Assim, extraímos os índices de riqueza de vocabulário do conjunto lexical e associamos ao índice de legibilidade para formar um novo grupo.

No profundo estudo realizado por Tweedie e Baayen (1998) é realizado um agrupamento de medidas de riqueza de vocabulário propostas por outros autores, além de uma extensa avaliação de sua variabilidade e qualidade. Apesar de se tratar de um grupo de características relativamente antigas, segundo Baayen (2008) e Hou e Chu-Ren (2020), a utilização destas colaborou para resultados melhores.

Por causa disso, selecionamos alguns destes índices para implementados e avaliação, que resultou na aplicação de 7 medidas de riqueza de vocabulário como características: *i*) R GUIRAUD, *ii*) C e V (HERDAN, 1964), *iii*) medida K (YULE, 1944), *iv*) U (DUGAST, 1979), *v*) A (MASS, 1972) e *vi*) H (HONORÉ, 1979).

Dado que T representa o tamanho do texto (quantidade de *tokens* no texto) e V o tamanho do vocabulário (conjunto de *tokens* únicos no texto), a fórmula para cálculo dos índices está disposto na Figura 12.

Representando a legibilidade, temos o índice de inteligibilidade *Flesch-Kincaid* (FK), bastante reconhecido na literatura para classificação indicativa de obras literárias, identificação da complexidade de frases e identificação de nível de escolaridade de estudantes a partir de redações. O índice FK foi adaptado para o português Martins et al. (1996) a partir de sua versão original em inglês e é calculado a partir do número de palavras, frases e sílabas que compõem o documento. Dado que P é o número total de palavras, F o número total de frases e S o número total de sílabas, o índice pode ser calculado por meio da fórmula a seguir:

Figura 16 – Características baseadas em riqueza lexical

Grupo	Feature	Observação
Riqueza de vocabulário e legibilidade	Diversidade léxica	Frequência de palavras distintas
	Giraud R	***R = V / √ T***
	Herdan C	***C = log (V) / log(T)***
	Herdan V	***V = T ^ C***
	Medida K	***K = log(V) / log(log(T))***
	Dugast U	***U = log(T) ^ 2 / log(T) - log(V)***
	Maas A	***A = √ ((log(V) - log(V)) / log(T) ^ 2)***
	Honoré H	***H = quantidade de termos únicos (hapax local) / V***
	Índice Flesch-Kincaid Brasileiro	
	Hapax Legomena local	Presença de termos únicos em relação ao próprio documento ou *corpora*
	Hapax Legomena global	
	Frequência de palavras repetidas	

Fonte: Elaborado pelo autor

$$BRFlesch = 206,835 - (1,015 * P/S) - (84,6 * (S/P))$$

Por fim, outra medida mais simples de riqueza lexical é o cálculo de *hapax legomena*. Este índice é representado pela frequência de palavras que só aparecem uma vez dentro de um vocabulário. Sua frequência pode ser medida de maneira local, ou seja, o vocabulário é considerado como o conjunto de palavras distintas dentro de um único documento, ou de maneira global, quando o vocabulário se refere a todas as palavras pertencentes ao corpus. Para a extração dessas características, codificamos um algoritmo de identificação e contagem de palavras únicas.

4.2.3.6 Características relacionadas à aplicação

Por fim, também construímos características dentro do grupo relacionado à aplicação. Essa categoria é mais abrangente e flexível e por isso foram inseridas características relacionadas a conjuntos de palavras específicas, além de não se enquadrarem nas demais categorias. Neste grupo temos: i) quantidade de palavras escritas de maneira errada, que

implementamos usando a biblioteca *spellcheck*, pautada em dicionários da língua portuguesa; *ii*) incidência de *stopwords* dentro do texto, que utiliza um conjunto fornecido pelo pacote NLTK; *iii*); e frequência de *collocations*, termo sugerido por (ALCARAZ, 1990), que indica palavras que geralmente aparecem próximas umas das outras, como por exemplo: "bom dia", "não sei", "meu nome é". Foram contabilizadas *collocations* de tamanho 2, 3 e 4 por meio do NLTK.

Figura 17 – Características do domínio da aplicação

Grupo	Feature	Observação
Aplicação	Palavras com erro ortográfico	Obtidas por meio da biblioteca spellcheck
	Stopwords	Padrão da biblioteca NLTK
	Collocations tamanho 2	Palavras que comumente se apresentam juntas com no máximo n termos
	Collocations tamanho 3	
	Collocations tamanho 4	

Fonte: Elaborado pelo autor

Em suma, as 74 características extraídas foram divididas em 6 grupos dispostas no Quadro 1. Ressaltamos que todas as características mencionadas neste capítulo, inicialmente obtidas em valores absolutos (quantidade de observações), foram convertidas para frequência (f), visando diminuir o viés introduzido pela diferença na quantidade de *tokens* por autor. O cálculo da frequência é obtido de maneira simples, uma vez que tenham sido encontradas N ocorrências de determinado padrão dentro de um texto de tamanho T, o valor foi transformado em f por meio de N/T.

Quadro 1 – Resumo das características estilómetricas utilizado

Grupo	Quantidade	Detalhes
Lexical	8	Estrutura do documento em relação a parágrafos, sentenças, palavras únicas e silabas
Caractere e palavras-chave	13	Ocorrência de palavras, pontuações e *top grams*
Sintático	25	Classes gramaticais, flexões de gênero, plural e tempo verbal
Semântico	11	Entidades nomeadas
Riqueza de vocabulário e legibilidade	12	Índices de riqueza de vocabulário e legibilidade
Especifico da aplicação	5	Baseados em listas de palavras, como dicionários ortográficos, *stopwords* e *collocations*

Fonte: Elaborado pelo autor (2021)

5 EXPERIMENTOS

Neste capítulo abordaremos com profundidade como as etapas experimentais foram executadas, assim como os principais resultados observados em cada uma das fases propostas.

5.1 CONFIGURAÇÃO DOS EXPERIMENTOS

Os experimentos foram divididos em quatro etapas sequenciais descritas à seguir.

A primeira etapa visou explorar a distribuição dos documentos dentro das três representações por meio de técnicas de visualização de dados e técnicas de agrupamento. A identificação de agrupamentos de documentos nos ajuda a visualizar características em comum devido à proximidade dos exemplos, sejam elas relacionadas a autoria ou não. A partir das observações realizadas nesta etapa, foram selecionados diversos classificadores para experimentação na fase seguinte. Os grupos foram inspecionados em critérios relacionados à autoria e ao conteúdo textual dos documentos.

A segunda fase experimental se inicia a partir dos modelos sugeridos na etapa anterior. Não foi aplicada nenhuma técnica de normalização ou mudança de escala nos dados, visto que um dos objetivos também era compreender a contribuição de cada representação para resolução das atividades propostas. Assim, os exemplos foram separados em conjunto para treinamento e teste, respeitando a proporção 70/30%, escolhidos aleatoriamente. As partições foram mantidas por meio de *random seeds*, com propósito de preservar a integridade dos experimentos durante fases posteriores, evitando que elementos usados no treinamento fossem realocados para testes.

Para avaliar o desempenho dos modelos, levamos em consideração o alto desbalanceamento em algumas bases, ou seja, existem autores com mais documentos escritos que outros. Em bases menores, como a de estudantes, o desbalanceamento é ainda mais evidente e pode influenciar fortemente na variação dos resultados. Com isso, é preferível usar alguma métrica menos sensível ao desbalanceamento, como a área abaixo da curva ROC (*Area Under Curve* - AUC), que leva em conta a razão entre verdadeiros e falsos positivos.

Apesar de ser comumente usada em problemas binários, encontramos adaptações da AUC ROC para problemas multiclasse (HAND; TILL, 2001) e optamos pela abordagem

um contra todos. Em associação com AUC, mensuramos a acurácia dos classificadores por ser uma das medidas mais comuns em problemas de classificação. A combinação da acurácia e AUC faz parte do conjunto de métricas usadas nas principais atividades de análise de autoria, conforme observamos no relatório do PAN/CLEF 2020 (BEVENDORFF et al., 2020c).

Ao final desta etapa experimental, selecionamos os classificadores que tiveram bom desempenho em ao menos numa das representações. Com isso, estreitamos o amplo número de possibilidades para um grupo específico de classificadores, com maior probabilidade de solucionar o problema em questão.

Diante de um conjunto menor de alternativas, a etapa seguinte teve como objetivo identificar as soluções ótimas dentro das combinações entre modelos e representações. Para isso, cada um dos classificadores passou por um processo de otimização por meio do ajuste de hiperparâmetros. Também utilizamos técnicas de mudança de escala e normalização dos dados, em todas representação, pois isso diminui o espaço amostral e impacto dos *outliers*.

Visando diminuir o erro amostral e aumentar a confiabilidade dos experimentos, alteramos a configuração da separação dos conjuntos de treinamento. Essa passou a ser feita por meio da validação cruzada estratificada com *3-folds*. O número de *folds* se deu pela quantidade mínima de exemplos por classe, respeitando os três conjuntos. Ao final desta etapa, apresentamos a solução proposta neste trabalho, que foram os melhores modelos selecionados para cada uma das bases, e o modelo com maior valor médio entre as três.

Os resultados obtidos nesta etapa serviram de subsídio para justificar ou contrapor nossas hipóteses iniciais. Um estudo de interpretabilidade dos modelos por meio de SHAP *values* (SHAPLEY, 1953) foi fundamental na compreensão das melhores características e classificadores. A comparação de resultados entre as bases de dados permitiu traçar conexões entre o estilo de escrita de cada grupo de autores e as variações de formato, tamanho e tema de cada uma.

Devido à maior complexidade e alta quantidade de combinações entre as variáveis dos experimentos, tornou-se necessário construí-los numa plataforma com maiores recursos computacionais. A execução das etapas experimentais demanda de muita memória, tanto para armazenamento como processamento, inviabilizando sua execução local. Por esse motivo, o arcabouço de experimentos, o que concerne às quatro etapas experimentais descritas agora, foram implementadas dentro de uma máquina virtual hospedada

no *Google Colab*, que fornece *notebooks* online com maior capacidade computacional, e provê acesso automático a diversas das bibliotecas de AM úteis durante a implementação. Os experimentos excederam mais de 400 horas para implementação e execução, e foram disponibilizados como produto deste trabalho no formato de *Jupyter notebooks* [1].

Desta forma, as etapas experimentais podem ser resumidas da seguinte maneira:

- **Análise exploratória e agrupamento**: Nesta etapa foram usadas técnicas de visualização de dados e aprendizagem não supervisionada para entender quais classificadores poderiam ser adequados para resolução do problema e observar possíveis agrupamentos.
- **Avaliação das soluções**: Construção e avaliação de modelos como solução do problema para cada uma das bases de dados, nas três representações possíveis
- **Otimização**: Aplicação de técnicas de mudança de escala e ajuste de hiperparâmetros
- **Interpretação de resultados**: Estudo de interpretabilidade dos modelos vencedores por meio da técnica valores Shapley

5.2 VISUALIZAÇÃO DE DADOS

A visualização de dados é uma importante etapa em problemas de AM. É por meio dela que podemos entender a distribuição dos exemplos dentro de um espaço n-dimensional. Essa distribuição nos permite entender melhor diversos fatores, tais como relação entre as variáveis dos eixos, distribuição, fronteira de separabilidade e *outliers*; fornecendo *insights* para quais técnicas e classificadores podem funcionar melhor para aquele conjunto de dados.

Na visualização, cada eixo de um plano cartesiano representa uma variável, o que torna sua visualização mais difícil de acordo com o crescimento da dimensionalidade. Por exemplo, se quiséssemos representar as características estilométricas dentro de alguma das bases, necessitaríamos de uma representação com 74 dimensões.

Nas representações textuais, temos um problema ainda maior. Numa representação TF-IDF, cada documento é representado por um vetor de tamanho igual ao vocabulário

[1] Disponível em https://github.com/daanielvb/authorship-classification

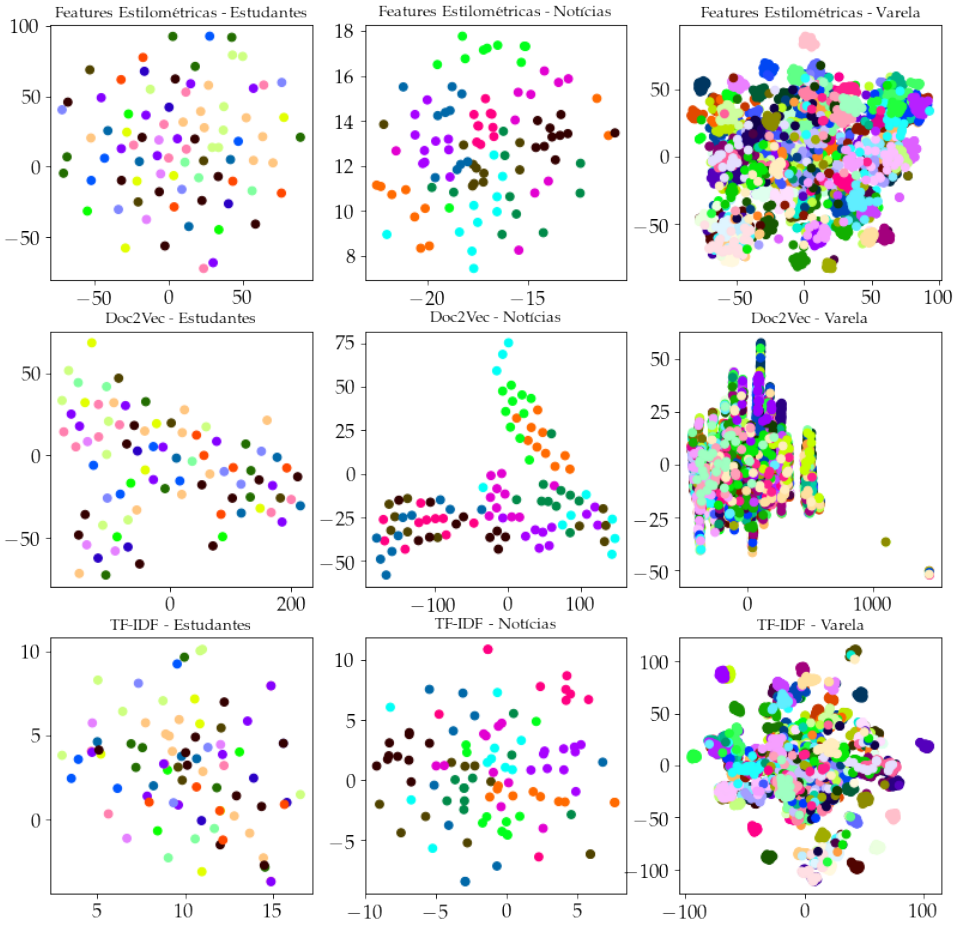
do corpus. No caso da base de estudantes, temos um vocabulário composto por 5.894 palavras, logo, cada documento é representado por um vetor de tamanho 5.894. Para representação de texto por meio de *word embeddings*, a dimensionalidade depende do tamanho do vocabulário V e do comprimento do *embedding* E, conforme discutido no Capítulo 4.2.2, na qual cada documento será representado por uma matriz composta por V vetores de tamanho E.

A aplicação de técnicas de redução de dimensionalidade é uma solução adequada para auxiliar na visualização de dados nestes casos. Nesta pesquisa, foram utilizadas as técnicas não supervisionadas de análise de componente principal (*Principal Component Analysis* - PCA) e TSNE (*T-distributed Stochastic Neighbor Embedding*) (MAATEN; HINTON, 2008). Primeiramente aplicamos o PCA, reduzindo o número de características para 20 visando reduzir o número de dimensões preservando ao máximo as características, porém para facilitar o treinamento do TSNE, que pode se tornar impraticável em casos de alta dimensionalidade (KAMBHATLA; LEEN, 1993). Em seguida, reduzimos o vetor de 20 dimensões para 2 dimensões utilizando o TSNE, que foi treinado por 3.000 iterações para cada base. Este processo resultou numa visualização humanamente interpretável das características de estilo em relação às classes.

Como os vetores TF-IDF são muito esparsos, optou-se por utilizar o algoritmo de análise semântica latente (*Latent Semantic Analysis* - LSA) (DEERWESTER et al., 1990) uma vez que o PCA e TSNE são sensíveis a esparsidade e estudos apontam que o LSA é mais eficaz para tratar vetores de contagem de palavras, como no TF-IDF; resultando em representações de menor dimensionalidade, mas com alta representatividade pois se localizam em um espaço de características latente (SALTON, 1983) (HOFMANN, 2001).

A visualização dos documentos representados por *word embeddings* não foi possível de ser realizada devido à alta dimensionalidade dos vetores, o que acarretou em problemas de falta de memória para armazenamento dos mesmos na plataforma utilizada para os experimentos. A alternativa que encontramos para representar os documentos através de vetores numéricos ponderados foi treinar um modelo *Doc2Vec* para cada base de dados. Desta forma, criamos nossos próprios vetores a partir dos documentos ao invés de palavras, o que permitiu a visualização. O modelo foi construído com auxílio da biblioteca *Gensim* e foi treinado por 100 iterações, produzindo vetores numéricos com 100 dimensões. Para a visualização, esses vetores foram reduzidos a 2 dimensões por meio da combinação do PCA e TSNE (Figura 18).

Figura 18 – Apresentação das bases de dados em 2 dimensões após redução de dimensionalidade com PCA + TSNE ou LSA (TF-IDF)



Fonte: Elaborado pelo autor

Na Figura 18 temos os documentos plotados em duas dimensões, cada exemplo representa um documento pintado numa cor única para cada um dos autores. Na representação estilométrica (RE) - linha 1 - é possível observar que a base de estudantes não apresenta separabilidade facilmente visível entre os exemplos das classes. Em contrapartida, na base de notícias da mesma linha, observamos maior separação entre os exemplos por classes. Para base Varela (linha 1) constatamos que apesar dos exemplos se agruparam numa região espacial menor, enxergamos separações claras entre algumas das classes.

A segunda linha da Figura 18, na RT com *Doc2Vec* observamos uma diferença no espaço de representação das características, na qual a base de notícias se dispersa em um espaço maior do que as outras duas bases de dados. Neste caso, é possível visualizar a distinção espacial entre exemplos das bases de notícias e em menor grau na base de estudantes. Nesta representação, destacamos a base Varela, que apresenta uma sobreposição de exemplos na região central do gráfico, porém existem regiões de maior separabilidade

de acordo com as cores.

Por fim, na representação textual (RT) com TF-IDF e LSA, linha 3, primeiramente observamos uma distribuição muito mais densa, pois os exemplos ocupam um espaço muito menor se comparado às anteriores. A base de estudantes mais uma vez apresenta exemplos entrelaçados entre as classes. Também observamos uma alta concentração de exemplos em regiões específicas para a base de Varela, mas é possível identificar fronteiras de separação entre exemplos de algumas classes. A base de notícias é a que mais se destaca nessa representação, sendo bastante visível a separação dos exemplos por suas classes.

Acreditamos que tal distribuição se dá por alguns fatores: a base de estudantes possui um vocabulário menor e restrito aos assuntos das disciplinas em que os trabalhos foram solicitados; a base de notícias é principalmente composta por textos sobre temas políticos e econômicos, havendo uma menor separabilidade ao levarmos em conta representações textuais mais genéricas, como *Doc2Vec*; e a base de Varela é composta por textos nos mais diversos assuntos, o que resulta numa grande separabilidade nas representações textuais.

Ao considerarmos o estilo de escrita, observamos que apesar de haver maior densidade de exemplos em regiões específicas quando os autores são jornalistas, há uma maior separabilidade entre os autores. Quando observamos o estilo de escrita dos estudantes, há uma dispersão muito maior dos exemplos, com autores ocupando diversas áreas no espaço, o que não ocorre tanto com os jornalistas.

5.2.1 Agrupamento de dados

Visando agregar as observações iniciais acerca dos dados, dedicamos uma etapa dos experimentos para o agrupamento de dados. O objetivo principal foi entender os critérios para agrupamento dos exemplos, tal como as principais características de cada grupo e o desempenho de algoritmos não supervisionados como solução para a atribuição de autoria. A partir das visualizações construídas na seção anterior, eliminamos a representação TF-IDF para diminuir o escopo dos experimentos e usar como comparação uma representação baseada em texto e outra baseada nas características de estilo. A escolha do *Doc2Vec* em contrapartida ao TF-IDF se deu pela maior separabilidade dos exemplos na primeira representação, conforme observado anteriormente.

Optamos por utilizar duas abordagens de agrupamento dos dados: uma *crisp* na qual cada elemento pertence totalmente a uma classe por meio do algoritmo K-médias, e uma

soft em que os exemplos podem pertencer a várias classes com valores probabilísticos de pertinência, através do algoritmo *fuzzy* c-médias (YONAMINE et al., 2002). Utilizamos as bibliotecas NLTK e *Skfuzzy* (WARNER; SEXAUER,) para obter implementações dos algoritmos citados.

O agrupamento dos exemplos em *clusters* depende principalmente do parâmetro K, que define o número de centros. A definição de K é um problema à parte, porém, existem estratégias simplistas para sua definição. Visto que este trabalho visa solucionar o problema de atribuição de autoria, decidimos utilizar duas estratégias. A primeira é baseada no conhecimento prévio do problema, definindo K de acordo com o número de classes (autores) por base de dados (16, 10 e 100). A segunda estratégia desconsidera qualquer conhecimento prévio sobre o problema, optando por observar possíveis *clusters* a partir da distribuição dos exemplos. Neste caso, o valor de K foi obtido por meio de técnicas que maximizam a separabilidade dos *clusters* de acordo com o algoritmo utilizado.

No caso do K-médias, realizamos o cálculo da silhueta variando $\{x \in \mathbb{R} \quad | 2 \leq x \leq n\}$, sendo n o total de autores possíveis, visando obter o valor de K que otimiza essa medida. A silhueta é uma medida que leva em conta a proximidade dos pontos dentro de seus próprios *clusters* (coesão) com relação à distância entre pontos dos outros *clusters* (separabilidade), servindo como indicativo de melhor separabilidade dos exemplos em K grupos (THINSUNGNOENA et al., 2015). Seu valor pode variar entre $[-1, 1]$ com função objetivo de maximização.

Para os experimentos conduzidos utilizando o *fuzzy* c-médias, obtivemos o valor de K que maximiza o coeficiente de partição difusa (*Fuzzy Partition Coefficient* - FPC). O FPC varia de $[0, 1]$, 1 é o valor máximo. Essa métrica foi desenvolvida por Bezdek (2013) e serve para mensurar o grau de sobreposição entre os *clusters* difusos. Segundo Zhang et al. (2009), um alto valor deste índice indica que há menor sobreposição entre os *clusters*, gerando consequentemente partições menos difusas. O principal problema desta abordagem é considerar apenas as distâncias entre *clusters* difusos, dando menos importância a exemplos individuais do conjunto. Mesmo assim, a abordagem é amplamente utilizada em problemas de agrupamento (HU; MENG; SHI, 2008) (RAJKUMAR; YESUBABU; SUBRAHMANYAM, 2019).

Após o cálculo dos valores de K, realizamos o agrupamento dos exemplos utilizando os dois algoritmos citados para cada uma das representações. Isto é, o algoritmo K-médias foi aplicado usando os valores de K baseado em autoria e nos valores de K obtidos pelo

cálculo da silhueta. Para o *fuzzy* c-médias, usamos K baseado em autoria e no FPC. Cada um dos dois algoritmos foi aplicado para as 3 bases nas representações estilométricas e textual.

Os *clusters* gerados foram analisados de maneira *ad-hoc*, com o propósito de entender os critérios utilizados pelos algoritmos durante o agrupamento. Em casos com número elevado de *clusters*, realizamos uma análise exploratória escolhendo alguns dos *clusters* de maneira aleatória. A análise foi feita considerando critérios como a autoria, tema e conteúdo dos exemplos. Para a análise do conteúdo textual, capturamos as palavras mais frequentes de cada *clusters* e para análise de autoria e tema, identificamos os autores e assuntos contidos em cada um dos agrupamentos.

5.2.1.1 K-médias

5.2.1.1.1 Agrupamento baseado na autoria

Na primeira análise, utilizamos valores de K iguais ao número de autores (16, 10 e 100) para cada uma das bases de dados. No que se refere a autoria, observamos que a distribuição de exemplos por *cluster* em todas as bases é similar a distribuição real na RE (estilométrica), porém na RT (textual) vemos muitos *clusters* com uma quantidade de exemplos muito diferente da distribuição real (Figura 20).

Na RE da base de estudantes tomamos como exemplo o *cluster* 14, composto pelos exemplos $c(14) = [4, 38, 49, 51, 58, 63, 67, 79, 82, 83]$. Neste grupo, não observamos dominância de autoria. Ao explorar o conteúdo textual dos exemplos, percebe-se que 80% dos exemplos são atividades realizadas na disciplina de metodologia de ensino de ciências e que as palavras-chave mais frequentes neste grupo são "disciplina" (16 incidências), "ensino"(14), "estudo"(14) e "aprendizagem"(12). Nesta mesma base, analisando o *cluster* 10 da RT composto pelos exemplos $c(10) = [58, 59, 60, 63]$, observamos que os exemplos contidos neste grupo se referem a uma atividade específica sobre as bases nacionais comum curriculares (BNCC). Podemos destacar algumas das palavras mais frequentes, como "tecnologia"(32), "aprendizagem"(19), "conceito"(18) e "base"(15) (Figura 19).

Repetindo o processo acima para a base de notícias, na RE tomamos como exemplo o *cluster* 8 composto pelos exemplos $c(8) = [41, 44, 45, 46, 47, 49, 63, 91]$ na qual é possível observar que 6 exemplos pertencem ao mesmo autor (os exem-

plos desta base estão ordenados por autor). Também observamos que outros 4 *clusters* possuem 5 ou mais exemplos de um único autor. Dentre as palavras mais frequentes no *cluster* 8, observamos "política"(14), "sociedade"(11) e "estado"(11). Na RT destacamos os *clusters* $c(5) = [31, 60, 61, 62, 63, 64, 65, 66, 68, 69, 90, 91]$ e $c(6) = [28, 72, 74, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89]$ respectivamente. Apesar de existirem exemplos de outros autores dentro dos *clusters*, destacamos que o *cluster* 5 contém 9 exemplos de um mesmo autor, e o *cluster* 6 contém todos os exemplos de um outro autor. Analisando o *cluster* 6, vemos que os trabalhos que pertencem ao mesmo autor tratam em sua maioria sobre economia e política. Já os trabalhos de outros autores (28, 72 e 74) também falam sobre assuntos relacionados, com destaque para os presidentes dos Estados Unidos e Brasil em suas manchetes. Não surpreendentemente, as palavras mais comuns deste grupo são "governo"(57), "presidente"(38), "Bolsonaro"(27) e "Trump"(22) (Figura 19).

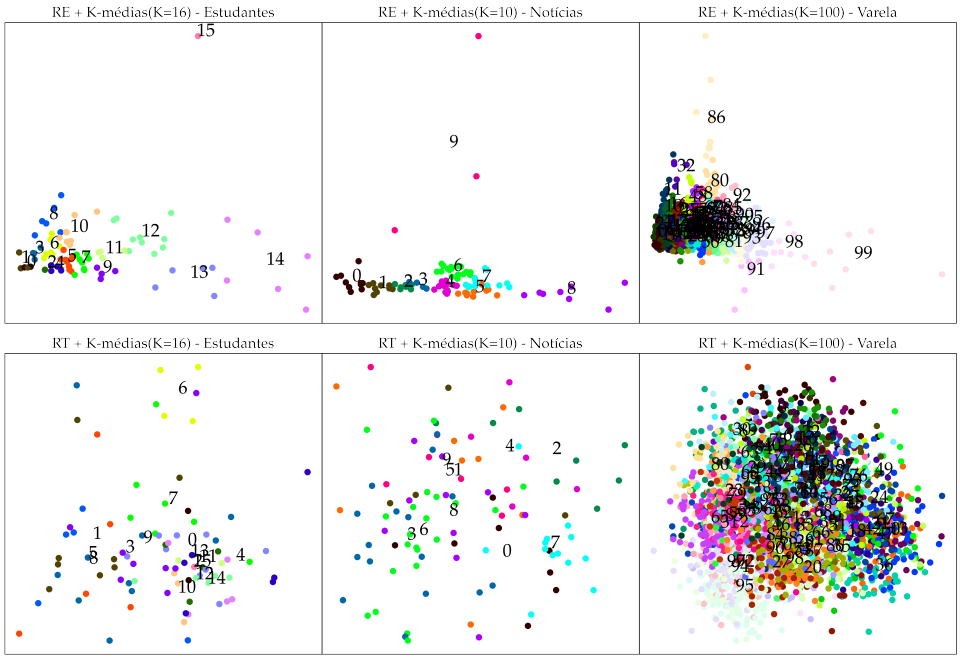
Figura 19 – Nuvens de palavras de alguns cluster das bases de estudantes e notícias após aplicação do K-médias com K derivado da autoria

Fonte: Elaborado pelo autor

Para a base Varela, devido ao grande número de autores e *clusters*, decidimos inspecionar os dados de maneira macro. Ao comparar o número de exemplos por *cluster* das representações estilométrica e *Doc2Vec*, observamos que a RE assemelha-se mais à distribuição real (30 exemplos por autor), uma vez que o menor *cluster* contém 10 exemplos e o maior 50. Na RT o menor *cluster* possui 4 exemplos e o maior 80. A disparidade é confirmada por meio do σ, que é de 12,37 e 8,39 nas representações textual e estilométrica respectivamente.

Na Figura 20 observamos que a aplicação do K-médias sugere agrupamentos de autores num espaço bem delimitado, o que não é verdadeiro conforme verificamos na distribuição real (Figura 18).

Figura 20 – Distribuição dos exemplos após aplicação do K-médias usando K relativo ao número de autores



Fonte: Elaborado pelo autor

Explorando aleatoriamente alguns *clusters* nesta mesma base, para a RE observamos que existe grande variabilidade de autores e assuntos dentro dos *clusters*, não sendo facilmente identificável o motivo de agrupamento daqueles exemplos. Ao analisar a RT vemos que desta vez, vários *clusters* possuem exemplos pertencentes ao mesmo autor. Após uma análise mais profunda nos exemplos de alguns agrupamentos, foi possível observar que há um forte agrupamento de exemplos baseado no assunto, como pode ser visto nos *clusters* 20 e 30. Para solidificar essa observação, verificamos outros *clusters*, como o 50. Nele temos 33 exemplos, sendo 12 destes pertencentes a um mesmo autor. Curiosamente, 28 dos 33 exemplos pertencem ao assunto "direito" (Quadro 2).

Diante dessas constatações, acreditamos que a convergência de autores em determinados *clusters* na RT pode estar relacionado ao tópico dos documentos.

5.2.1.1.2 Agrupamento baseado na distribuição

Na segunda estratégia de agrupamento com o K-médias, utilizamos os valores de K obtidos por meio do cálculo da silhueta usando a distância euclidiana. Para as bases de estudantes, notícias e Varela na RT obtivemos os valores de $K = [2, 3, 2]$ respectivamente.

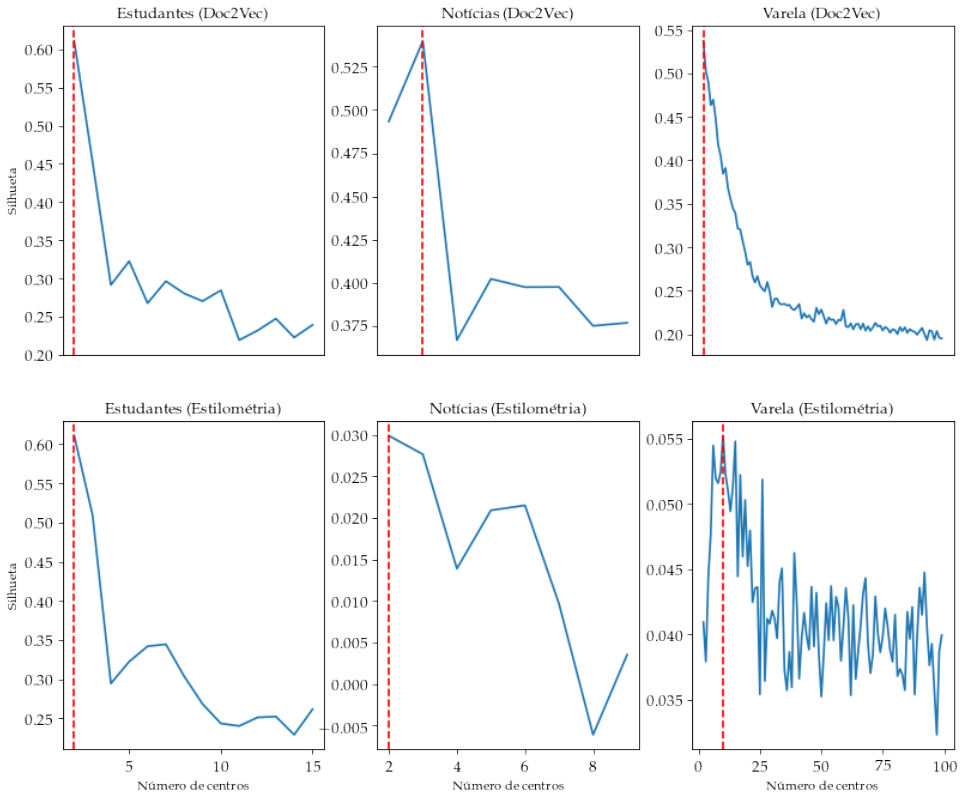
Quadro 2 – Agrupamento com K-médias para base Varela na representação textual

Assunto	$c(20)$	$c(30)$	$c(50)$
Assuntos variados	2	4	1
Economia	1	1	1
Esportes	0	1	0
Literatura	1	0	0
Politica	29	6	3
Turismo	0	21	0
Direito	0	0	28

Fonte: Elaborado pelo autor

Na RE, o valor ótimo de K foi 10 para a base Varela e 2 para as outras duas bases (Figura 21).

Figura 21 – Razão entre número de centros e valor da silhueta.

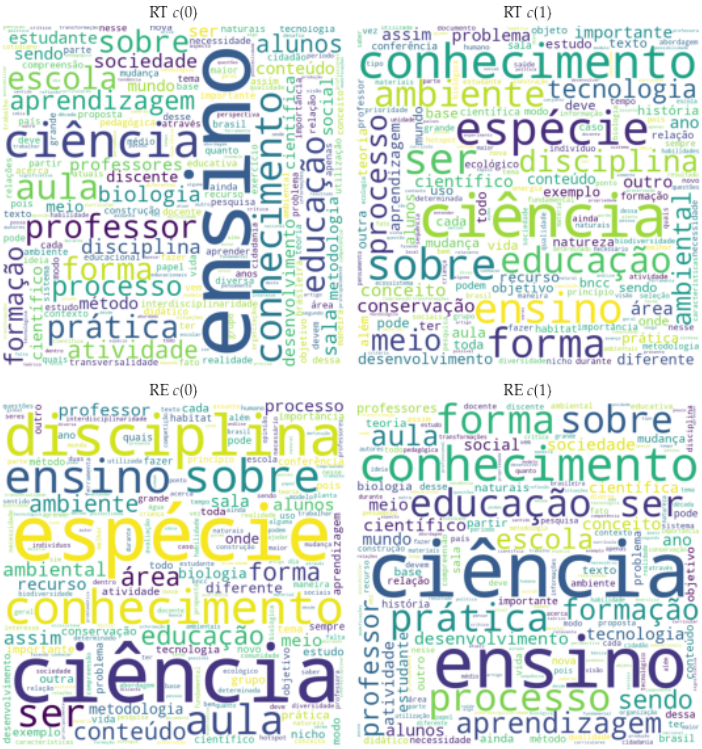


Fonte: Elaborado pelo autor

Investigando as palavras mais frequentes em cada um dos grupos, na base de estudantes temos novamente uma separação óbvia por disciplina na RT ao observarmos "ensino"(161), "professor"(89) e "ciência"(85) como mais frequentes no primeiro *cluster* e "ser"(91), "conhecimento"(91) e "espécies"(82) no segundo *cluster*. Já na RE, algumas

palavras relacionadas as disciplinas aparecem novamente, mas a separação é menor, uma vez que há repetição: no primeiro *cluster* temos "espécies"(78), "sobre"(78) e "conhecimento"(64) e no segundo "ensino"(167), "conhecimento"(105) e "ciência"(99) (Figura 22).

Figura 22 – Nuvens de palavras da base de estudantes após aplicação do K-médias com K derivado da silhueta



Fonte: Elaborado pelo autor

Para a base de notícias, observando as principais palavras dos 3 *cluster* na RT, enxergamos que dois *clusters* tratam de assuntos políticos, enquanto o outro possui palavras relacionadas à sociedade. O primeiro apresenta "ser"(66), "pessoas"(66) e "casa"(54); o segundo "ser"(67), "estado"(58) e "política"(44); o terceiro e último é composto por "governo"(112), "ser"(84) e "crise"(57). Desta análise vemos que talvez seja interessante incluir a palavra "ser" na lista de *stopwords* para esta base de dados, uma vez que se trata do verbo "ser" e não do adjetivo "ser vivo" muito frequente na base de estudantes. Na RE desta mesma base, vemos que um *cluster* possui muito mais documentos que o outro pela frequência das principais palavras e que existem duas palavras se repetem entre os *cluster* - no primeiro temos "governo"(31), "sobre"(28) e "pessoas"(27) e no segundo "ser"(202), "governo"(121) e "sobre"(103) (Figura 23).

Figura 23 – Nuvens de palavras da base de notícias após aplicação do K-médias com K derivado da silhueta

Fonte: Elaborado pelo autor

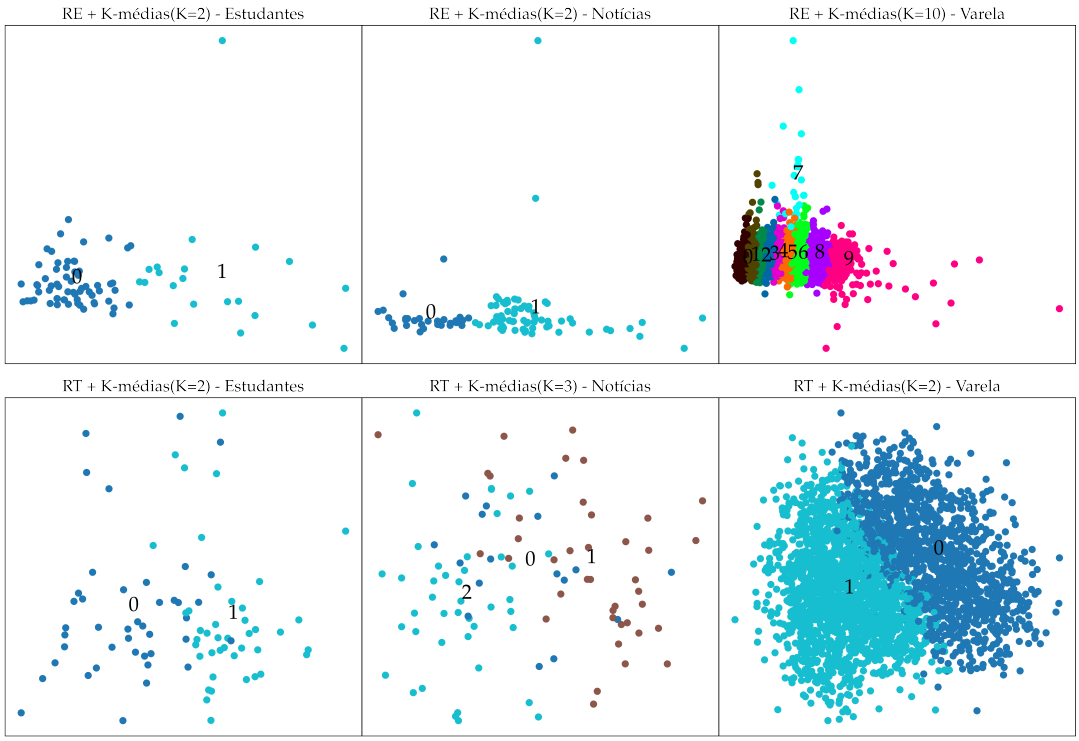
Averiguando a base Varela, a RT se mostra bastante balanceada com relação ao número de exemplos entre os grupos (1432 e 1568). No primeiro *cluster* vemos as palavras "ser"(2226), "sobre"(1265) e "brasil"(1193), no segundo *cluster* "ser"(2254), "pode"(1328) e "bem"(1197). Para a RE, optamos por destacar apenas as informações mais relevantes dos dez grupos. A palavra "ser" é a principal de 8 *clusters*, seguida pelas palavras "sobre" e "pode" que estão no *top-3* da maioria dos *cluster*. Essa análise reforça a ideia de que *stopwords* específicas podem beneficiar até mesmo representações não baseadas em palavras, como é o caso da RE (Figura 24).

Figura 24 – Nuvens de palavras da base Varela após aplicação do K-médias com K derivado da silhueta

Fonte: Elaborado pelo autor

Também realizamos uma breve análise com relação à autoria dos *clusters* (Figura 25). Para a base de estudantes, notamos que em ambas representações não há uma relação óbvia dos exemplos por *cluster*. Na base de notícias notamos que nos *clusters* da RE (estilométrica) os autores estão mais segregados dentro de um mesmo grupo do que na RT (textual). É verdade que a RT possui um *cluster* a mais que na RE e isso leva a uma maior dispersão. Mesmo assim, levando em conta apenas os casos em que temos 80% dos exemplos de um autor num mesmo *cluster*, observamos que na RE existem oito destes casos e na RT apenas cinco.

Figura 25 – Distribuição dos exemplos após aplicação do K-médias usando K relativo à silhueta



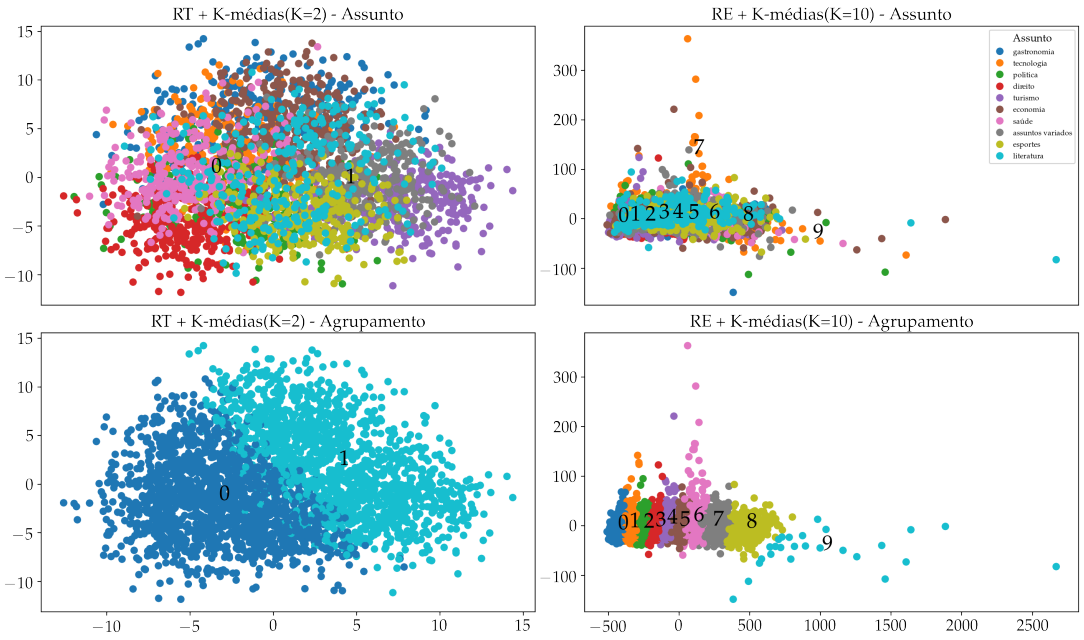
Fonte: Elaborado pelo autor

Na base de Varela investigamos se o agrupamento por autoria supera o agrupamento por assunto na RT. Para realizar essa verificação, fizemos a contagem de autores ou assuntos com pelo menos 80% do número total de documentos daquela classe num mesmo *cluster*. O valor de corte foi de 24 exemplos de um mesmo autor ou 240 exemplos de um mesmo assunto. O agrupamento por autor nos dois *cluster* da RT somou 89% e o agrupamento por assunto 70%.

Para a RE, o número de *cluster* é o mesmo número de assuntos e por isso decidimos analisar essa possível correlação. Observamos que dos 10 possíveis assuntos, apenas economia não consta como um dos dois assuntos mais frequentes de cada *cluster*. Também não há distinção clara de assunto por *cluster*, visto que há uma distribuição bastante heterogênea entre os *cluster*. Com relação a autoria, vemos que cada *cluster* possui pelo menos 3 ou 4 autores com mais de 20 exemplos por *cluster*, porém devido ao grande número de autores (300) essa observação pode ser irrelevante.

Na primeira linha da Figura 26 plotamos os centroides obtidos pelo K-médias na base Varela e destacamos os exemplos de acordo com o assunto. Observamos que a RT apresenta maior segregação por assuntos do que na RE, que por sua vez também demonstra alguns

Figura 26 – Distribuição dos exemplos da base Varela por assunto usando K relativo à silhueta



Fonte: Elaborado pelo autor

agrupamentos na região central. Para fins comparativos, na segunda linha temos as classes definidas pelo K-médias de acordo com o número de centroides.

Note que o agrupamento ou segregação por *cluster* citados nesta seção não significam que há isolamento de exemplos de um único autor por grupo. Na maior parte dos casos isso é impossível devido ao número de *cluster* ser inferior ao número de classes. A constatação que fizemos sobre os agrupamentos é que existe uma condensação de autores que pode ser causada por similaridades não óbvias ou desconhecidas.

5.2.1.2 Fuzzy C-médias

5.2.1.2.1 Agrupamento baseado na autoria

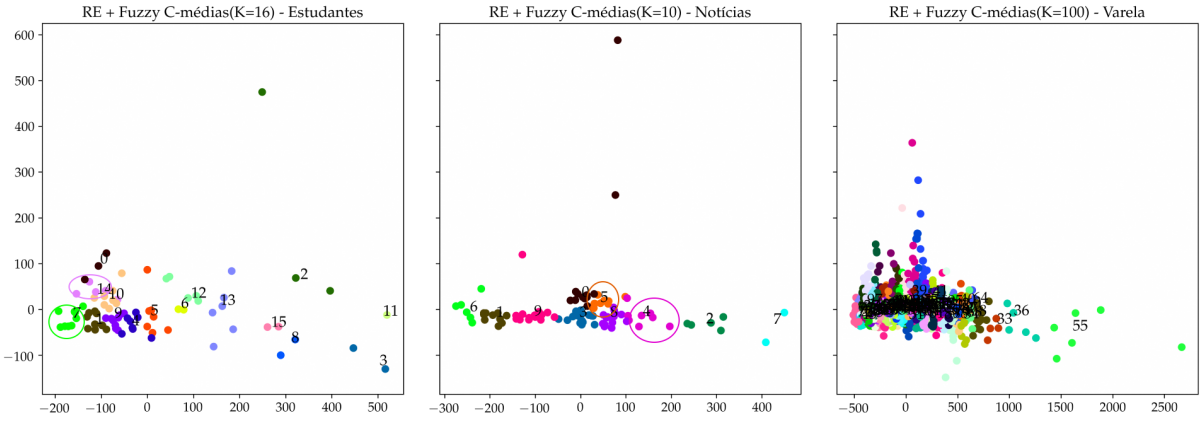
Durante o agrupamento difuso com valores de K baseados no número de autores esperados, observamos que o número total de *clusters* produzidos foi menor que K em algumas das combinações entre representação e base. Isto é, alguns grupos não tiveram exemplos associados. Isso se dá pela estratégia de enquadramento de exemplos em algum dos *clusters* que foi feita por meio daquele com maior probabilidade de pertencimento.

No que se refere à autoria na base de estudantes, os *clusters* gerados a partir da RT divergiram bastante com relação à distribuição de exemplos por autores. Observamos que

haviam *clusters* com apenas um exemplo associado e dois *clusters* principais continham 20 exemplos ao todo, divergindo completamente da distribuição original. Uma observação curiosa é que ao examinar os documentos 23 e 24, presentes em um destes *cluster* com poucos exemplos, constatamos que ambos documentos correspondem à mesma atividade e possuem conteúdo textual praticamente idênticos mas são de autores distintos. Tal observação é um forte indício de plágio.

Dentro da RE, também não observamos relação direta entre os *clusters* e a autoria, mas cabe citar que o *cluster* 1 foi capaz de agrupar 6 dos 7 exemplos existentes do autor "10". Com relação às palavras mais frequentes, constatamos uma distribuição similar aos *clusters* gerados pelo K-médias, com uma dominância de termos relacionados às disciplinas. Algo que chamou atenção no *cluster* 14 na RE foi a ocorrência da palavra "afin", escrita com um pequeno erro ortográfico seis vezes. Ao analisar os autores capturados por este *cluster*, observamos que 3 dos 5 exemplos existentes pertencem ao autor "4" e que este *cluster* pode ter sido construído a partir de informações associadas ao estilo de escrita do mesmo (Figura 27).

Figura 27 – Distribuição dos exemplos após aplicação do Fuzzy C-médias usando K relativo à autoria



Fonte: Elaborado pelo autor

Durante a análise da base jornalística, nas duas representações constatamos que na maioria dos *clusters* existe uma classe majoritária (autoria) predominante. Por exemplo, os *clusters* 4 e 5 da base estilométrica apresentam 62% e 77% dos exemplos de um único autor respectivamente. Com relação às palavras com maior incidência nestes *clusters*, não observamos muitas diferenças em comparação aos *clusters* gerados pelo K-médias, as palavras mais frequentes estão novamente relacionadas à política e economia. Ao comparar as principais palavras dos *clusters* entre a RE e RT, notamos que na textual, um dos *clusters*

possui palavras relacionadas a trabalho remoto ("trabalho", "*home*", "*office*", "casa") que não se destacaram na RE.

Na base Varela, o número de *clusters* gerados foi bastante diferente entre as representações. A RT agrupou os exemplos dentro de 59 *clusters* e a estilométrica em 90. Inspecionando os autores de alguns *clusters*, observamos que na RE há uma coesão maior de autores por *cluster*, enquanto que na RT há uma maior homogeneidade de trabalhos do mesmo assunto num *cluster*. Nesta base, o agrupamento de documentos por tema pode ser um indicativo para maior presença de documentos de um mesmo autor dentro do *cluster*, visto que nesta base, cada autor possui documentos sobre apenas um assunto.

Usando o *cluster* 10 de ambas representações como exemplo, na RE verificamos 8 trabalhos do mesmo autor e outros 8 exemplos de autores distintos. O assunto principal deste *cluster* foi "direito" com 9 aparições dentre as 16. Na RT, o *cluster* de número 40 contém 31 documentos acerca do tema "esportes" e apenas 8 de outros temas. Dentre estes documentos, um dos autores se destaca com 14 aparições.

Ao analisar as palavras de maior frequência destes *clusters*, enxergamos uma relação direta com os temas dos textos, na primeira representação destacamos "consumidor"(50), "art"(42), "férias"(36), "contrato"(35) e "bem"(33). No outro agrupamento temos "time"(94), "futebol"(84), "jogo"(78) e "gol"(61). Não enxergamos relação direta entre as palavras mais frequentes e o estilo de escrita dos autores (Figura 28).

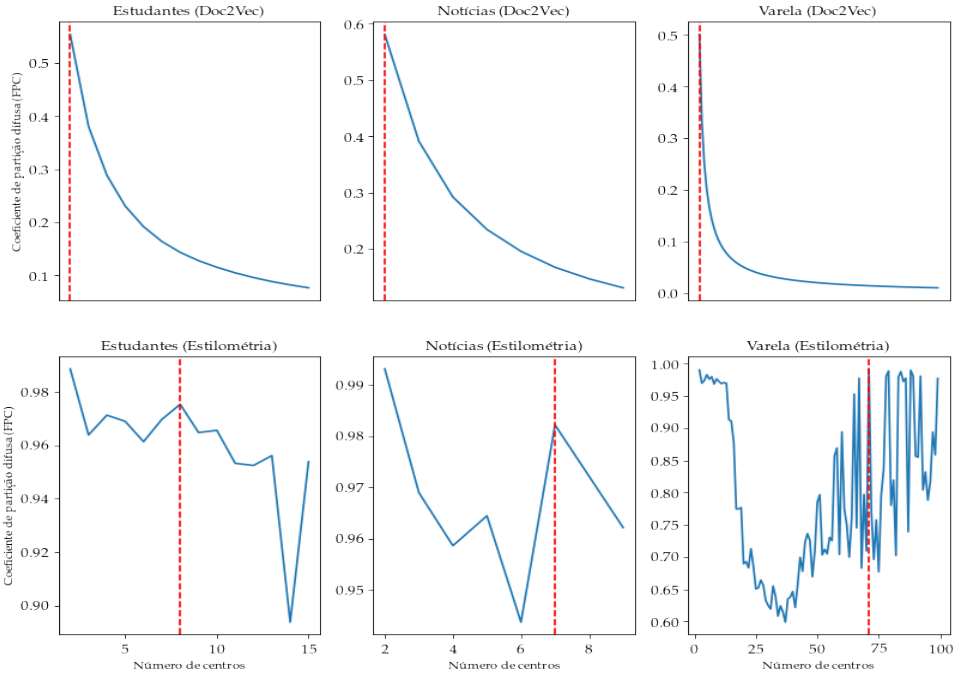
Figura 28 – Nuvens de palavras provenientes do Fuzzy C-médias com K relativo à autoria

Fonte: Elaborado pelo autor

5.2.1.2.2 Agrupamento baseado na distribuição

Um novo agrupamento dos exemplos foi realizado, desta vez utilizando os valores de K obtidos por meio do FPC. Para a RT, o valor máximo de K foi de apenas 2 *clusters* em todas as bases. Já na RE, optamos por escolher o melhor valor de K que fosse diferente de 2 para explorar melhor o problema. Para as bases de estudantes, notícias e Varela obtivemos os valores de K iguais a 8, 7 e 72 respectivamente (Figura 21).

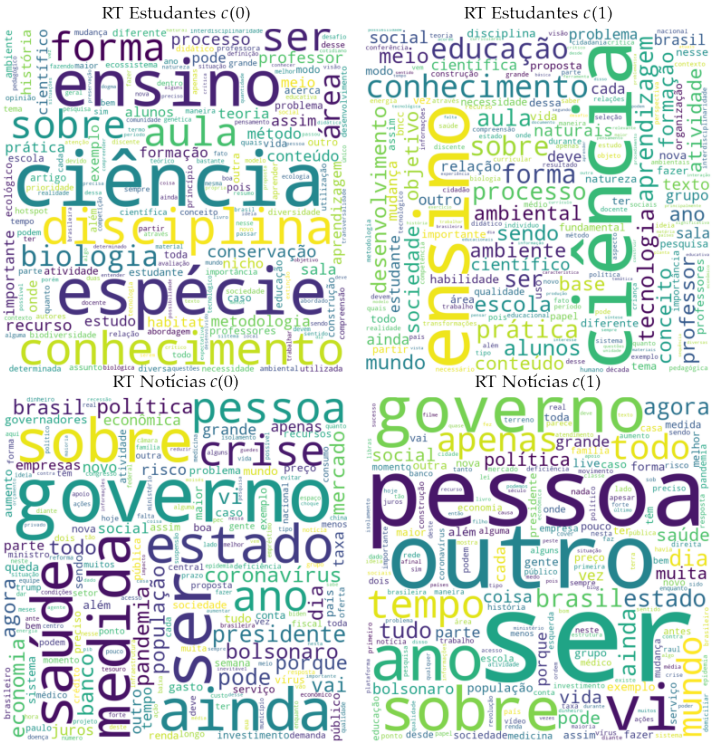
Figura 29 – Razão entre número de centros e FPC



Fonte: Elaborado pelo autor

Na RT, utilizamos o valor de K=2 para todas as bases. Na base de estudantes, não observamos padrão entre os autores de cada *cluster*. Ao analisar as palavras mais frequentes de cada grupo, vemos que três das *top-5* estão presentes em ambos os *clusters*: "conhecimento", "ciência" e "ensino". Notamos que as palavras mais frequentes de cada *cluster* estão relacionadas às disciplinas, pois em um grupo se destacam "ensino", "biologia" e "espécie"e no outro "educação" e "ambiental". Para a base de notícias, novamente não há uma relação óbvia entre a autoria e cada *cluster*. Entre as palavras mais frequentes, vemos que ambos grupos possuem muitas palavras em comum, tais como "governo", "pessoas" e "estado"(Figura 30).

Figura 30 – Nuvens de palavras dos agrupamentos após aplicação do Fuzzy c-médias com K derivado do FPC



Fonte: Elaborado pelo autor

Complementando, na base Varela temos dois grupos bem separados por assunto. No primeiro *cluster*, observamos 5 assuntos principais (economia, esportes, política, tecnologia e turismo) com no mínimo 182 documentos de cada assunto. Outros 5 assuntos distintos dominam o segundo *cluster* com no mínimo 175 documentos por assunto (assuntos variados, direito, gastronomia, literatura e saúde) (Quadro 3).

No primeiro *cluster* destacamos as palavras mais frequentes "Brasil"(1270), "anos"(1150) e "mundo"(990). No segundo encontramos "bem"(1119), "pessoas"(869) e "vida"(863). A primeira vista podemos dizer que as palavras do primeiro *cluster* estão mais próximas de temas como política e economia, por outro lado, o segundo *cluster* contém palavras mais próximas dos temas direito e saúde.

Usando os valores de K para a RE, obtivemos resultados interessantes. Na base de estudantes com $K = 8$, analisamos todos os *clusters* e constatamos que apenas um destes demonstrou concentração por autoria, com aproximadamente metade dos exemplos do autor "11". Na base de notícias, o agrupamento por autores é mais concreto: 4 dos 7 *clusters* são compostos por ao menos metade dos exemplos pertencentes a um mesmo autor (Figura 31). Por existirem muitos *clusters* na base de Varela, exploramos apenas

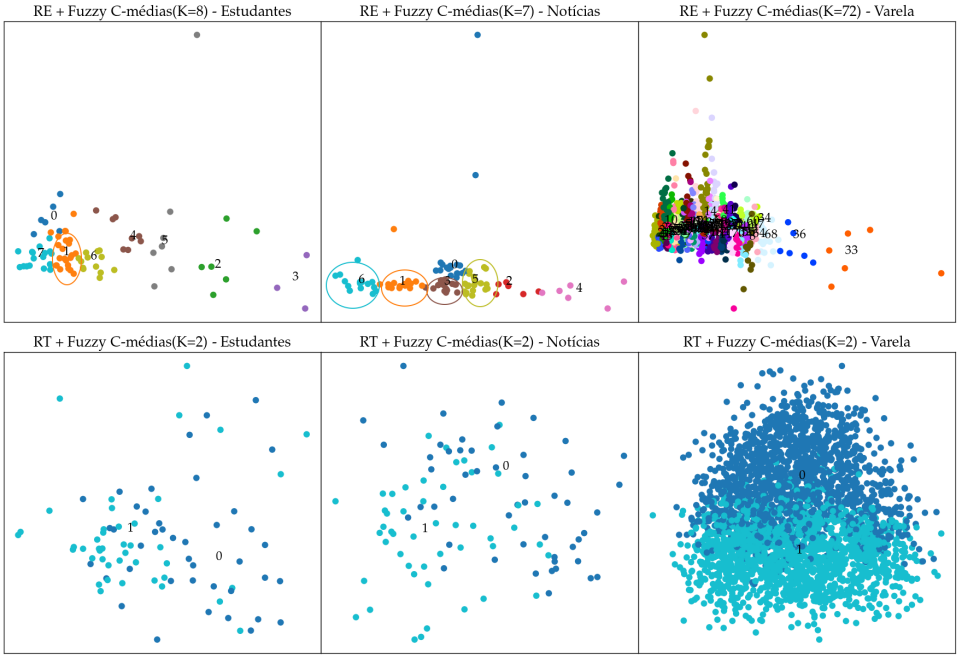
Quadro 3 – Agrupamento com Fuzzy C-médias para base Varela na representação textual

Assunto	$c(0)$	$c(1)$
Assuntos variados	125	175
Direito	41	259
Economia	266	34
Esportes	252	48
Gastronomia	81	219
Literatura	62	238
Política	182	118
Saúde	17	283
Tecnologia	236	64
Turismo	247	53

Fonte: Elaborado pelo autor

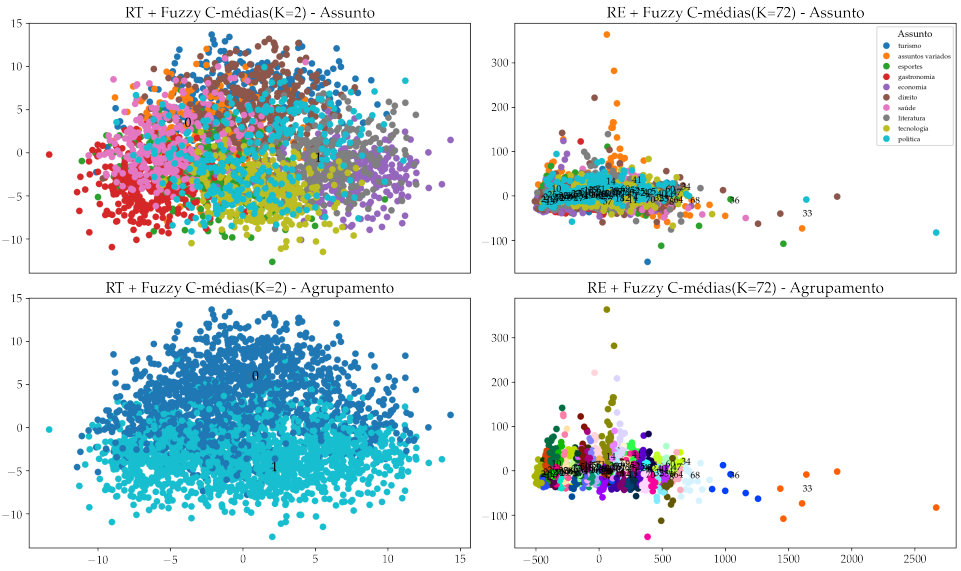
10% do total. Percebemos que não há uma dominância de exemplos de um mesmo autor nos *clusters* explorados, mas também vimos que agrupamentos por assunto são bastante presentes nos *clusters* analisados, incluindo as representações não textuais (Figura 32).

Figura 31 – Distribuição dos exemplos após aplicação do Fuzzy C-médias usando K relativo ao FPC



Fonte: Elaborado pelo autor

Figura 32 – Distribuição dos exemplos da base Varela por assunto usando K relativo ao FPC



Fonte: Elaborado pelo autor

Nesta seção exploramos os agrupamentos de dados a partir da distribuição natural dos exemplos e número de autores. Durante a análise observamos que abordagens de partição *crisp* e *soft* não foram capazes de solucionar a atividade de autoria, e por isso foram removidas dos experimentos posteriores. A análise exploratória dos *clusters* demonstrou indícios de agrupamento por tema, que foi constatado pelas palavras mais frequentes e o assunto dos documentos. Isso motivou a experimentação num subconjunto da base Varela descrito na Seção 5.6. A base Varela foi a única que demonstrou alguns grupos com maior concentração de documentos de um mesmo autor, entretanto os autores desta base escreveram apenas sobre um único assunto.

5.3 ANÁLISE INICIAL DE CLASSIFICADORES

A partir da visualização dos dados, observamos que para a base de estudantes não há fronteira de separabilidade clara entre os exemplos das classes. Por outro lado, nas duas bases jornalísticas, existem representações nas quais se pode visualizar fronteiras linearmente separáveis. Essa análise nos sugere que modelos lineares não seriam capazes de classificar corretamente os exemplos em todos os casos. Mesmo assim, devido ao possível viés introduzido por diversos fatores, como a representação do documento, extração de características, redução de dimensionalidade e grande variabilidade de combinações entre bases de dados e representações dos dados, optamos por diversificar a escolha dos modelos.

Dividimos os classificadores escolhidos em 5 categorias: 1) comitês, escolhemos dois baseados em árvores: *Random Forest* e Árvores Extra e um baseado em gradiente descendente; 2) modelos probabilísticos: *Naive bayes* multinomial e Gaussiano; 3) Entre os modelos lineares, selecionamos regressão logística e SVM com *kernel* linear; 4) Também selecionamos versões não lineares do SVM, com *kernel* polinomial e função base radial; 5) Por fim, dentro da arquitetura de redes neurais, foram escolhidos MLP, CNN, LSTM e uma RNA de alimentação direta, totalizando 14 modelos para experimentação. A escolha dos modelos em cada categoria foi realizada com base em seu uso na literatura base, sua importância histórica e efetividade em experimentos prévios.

Para a construção dos modelos clássicos (categorias 1 a 4) utilizamos majoritariamente a biblioteca *Sklearn*. Já para a implementação das redes neurais usamos os pacotes *Keras*, *TensorFlow* e *Gensim*.

5.4 AVALIAÇÃO E SELEÇÃO DOS MODELOS

A partir da lista inicial de classificadores, executamos a terceira etapa experimental, que teve como objetivo avaliar os modelos pré-selecionados da etapa anterior para as três bases de dados nas duas representações (textual e estilométrica).

Nesta etapa, decidimos por não ajustar nenhum dos hiperparâmetros de cada um dos modelos. Desta forma, o valor usado foram os valores padrão da implementação padrão da biblioteca *Sklearn*. Utilizamos como métricas a acurácia e Área Abaixo da Curva ROC. Os conjuntos de dados foram separados em treinamento (70%) e teste (30%) de maneira estratificada para todos os experimentos desta seção. Ressaltamos também que o modelo de aprendizagem profunda (CNN-LSTM) só foi utilizado na representação textual por causa da incompatibilidade de dimensão desta arquitetura e os dados da representação estilométrica. As representações textuais se apresentam em matrizes de duas dimensões, pois cada palavra é representada por um vetor, enquanto que a estilométrica é composta por vetores numéricos com apenas uma dimensão. Optamos por não utilizar a convolução 1D como alternativa para esta limitação, pois os dados de entrada não possuem natureza temporal.

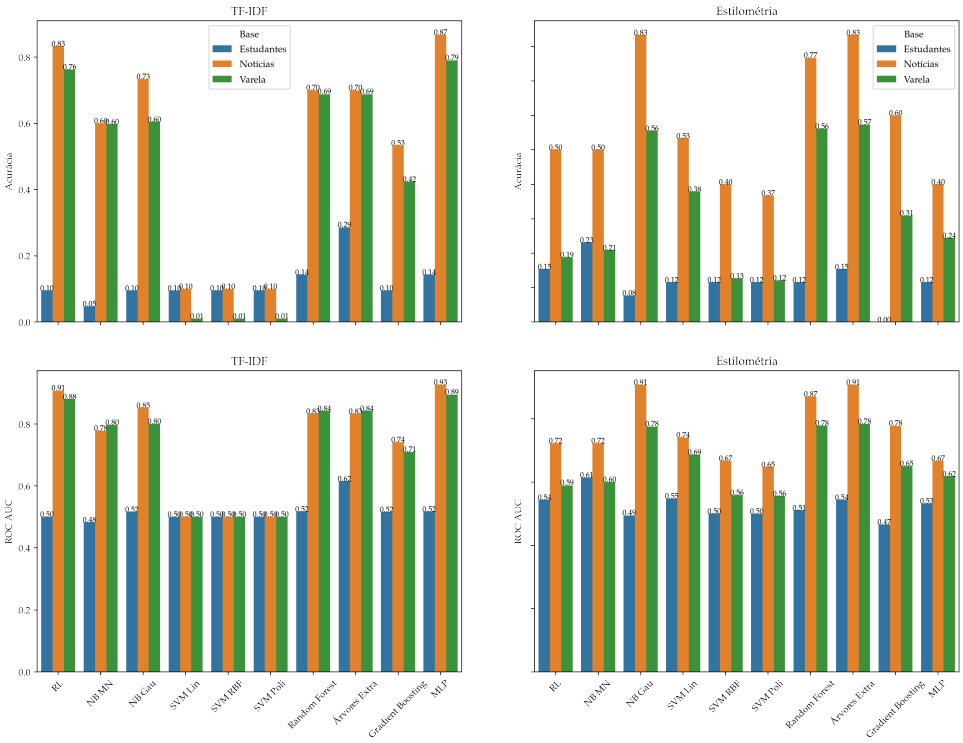
Dividimos a execução dos experimentos em três partes: *a*) modelos clássicos; *b*) rede neural de alimentação direta autoral e *c*) CNN-LSTM e *word embeddings*. Cada um dos modelos foi executado 10 vezes, nas quais mensuramos a média da acurácia e AUC. Os

experimentos com modelos clássicos levaram aproximadamente 68 horas de execução e os modelos de aprendizagem profunda aproximadamente 130 horas.

Nos resultados obtidos (Figura 33) pode ser visto que modelos baseados em árvores apresentaram bom desempenho em ambas às representações. Os modelos Bayesianos, MLP e regressão logística apresentaram resultados compatíveis com a literatura, especialmente na representação textual (BEVENDORFF et al., 2020c). Por outro lado, como destaque negativo, todos os modelos de SVM apresentaram resultados abaixo do esperado em ambas as representações, o que pode estar relacionado com a não normalização dos dados. Os algoritmos *Naive Bayes* multinomial e o aumento de gradiente baseado num comitê de AD ficaram numa faixa intermediária. Estes algoritmos atingiram resultados entre 30% e 60% de acurácia em pelo menos uma das representações.

Uma possível consequência da não normalização dos dados nessa etapa é o menor desempenho de classificadores que são altamente beneficiados dessas técnicas, como por exemplo os SVMs (AGARAP, 2018).

Figura 33 – Resultados obtidos para seleção de modelos clássicos



Fonte: Elaborado pelo autor

Para selecionar os modelos, usamos como ponto de corte o mínimo de 60% de acurácia e 70% de AUC em qualquer uma das representações. Os resultados da base de estudante

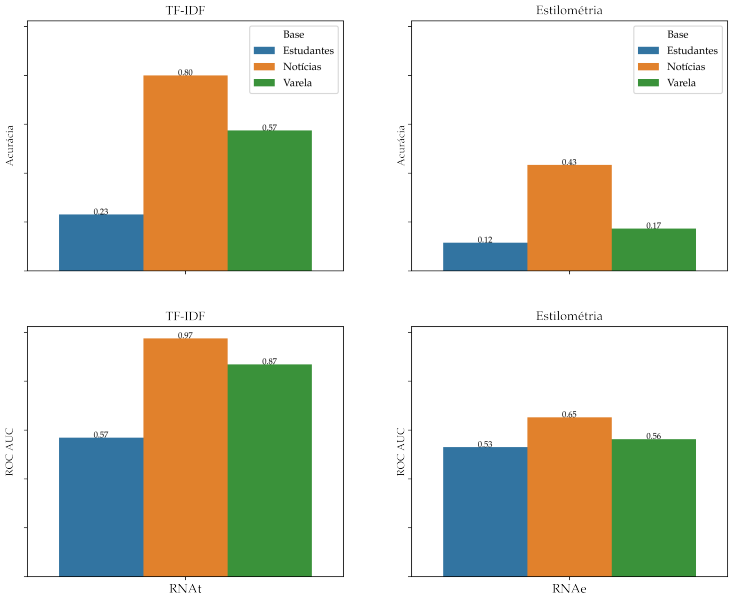
foram ignorados neste critério, pois nenhum classificador obteve resultados acima do ponto de corte. Sendo assim, dentre os modelos clássicos, selecionamos para a próxima fase: Regressão Logística, *Naive Bayes* Gaussiano, MLP, *Random Forest* e Árvores Extra (*Extra Trees* - ET).

Com relação aos experimentos com arquiteturas profundas, primeiramente avaliamos duas RNAs de alimentação direta autoral. A primeira rede foi utilizada para a representação textual (RNAt), sendo composta por 4 camadas escondidas sequenciais contendo 160, 100, 80 e 40 neurônios respectivamente, intercaladas por camadas de *dropout* de 10%. Estes valores foram escolhidos de forma experimental. Já a RNA para a representação estilométrica (RNAe) é um pouco mais simples: o tamanho da entrada é fixo de acordo com o número de características (74), internamente temos três camadas escondidas sequenciais compostas por 64, 32 e 16 neurônios sucessivamente. Novamente intercalamos camadas de *dropout* de 10%, também definidos de maneira experimental.

Em comum entre as duas RNAs temos a função de unidade linear retificada (ReLU) para ativação entre as camadas internas e função de ativação *softmax* na camada de saída e número de neurônios na saída igual ao número de classes. Como função de perda, usamos a função probabilística *categorical cross-entropy*.

Ambas as redes foram treinadas por 200 épocas, capturando a média dos valores finais na acurácia diante do conjunto de testes. Conforme podemos observar na Figura 34, os resultados da RNAt foram superiores ao RNAe em todas as bases de dados. Baseado nos resultados positivos observados na representação textual, esse modelo foi selecionado para a próxima etapa dos experimentos.

Figura 34 – Acurácia e ROC AUC da RNA nas representações estilométrica e textual



Fonte: Elaborado pelo autor

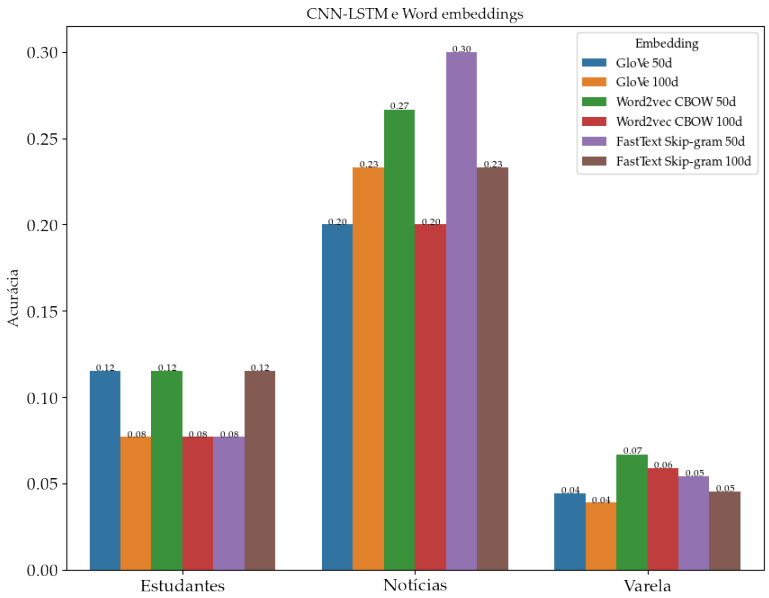
O outro experimento com redes profundas voltou-se para a utilização de arquiteturas recorrentes com camadas convolucionais, associadas a *word embeddings* pré-treinados. Nestes experimentos, além de tomar conhecimento sobre quais modelos tiveram melhor desempenho, buscamos avaliar se houve alguma diferença significativa na utilização de *word embeddings* treinados com diferentes dimensões e estratégias de construção.

Utilizamos uma rede neural profunda, mesclando camadas recorrentes (LSTM) e camadas de convolução (CNN) associadas a vetores de pesos iniciais fornecidos pelos *embeddings*. Essa arquitetura se assemelha à C-LSTM proposta por Zhou et al. (2015) e RNN LSTM bidirecional por Li et al. (2017). O modelo foi treinado por 200 épocas e novamente coletamos a média da acurácia obtida a partir do conjunto de testes.

Como podemos verificar na Figura 35, a utilização de CNN-LSTM associado a *word embeddings* não foi eficaz para atribuição de autoria e por isso não foi selecionada para a próxima etapa. Destacamos que a base de notícias foi a que apresentou melhores resultados, porém estes não ultrapassam 30% de acurácia. Curiosamente neste experimento a base de estudantes apresentou resultados superiores à base de Varela, o que destoa dos experimentos anteriores.

Também pudemos observar uma variação de performance de acordo com o grupo de *word embeddings* utilizado. Apesar da diferença não ser muito grande e não termos aplicado um teste estatístico sobre os resultados, podemos ver que ao usar *embeddings*

Figura 35 – Acurácia do CNN-LSTM para atribuição de autoria usando diferentes tipos de *word embeddings* pré-treinados



Fonte: Elaborado pelo autor

com 50 dimensões tivemos maior acurácia ao comparar com *embeddings* de 100 dimensões. Dentre estes, o *FastText* se mostrou superior ao *Word2Vec* e *GloVe* para o nosso conjunto de dados.

Analisando cada base de dados isoladamente, verificamos que a diferença de acurácia é muito pequena no uso de diferentes *word embeddings* indicando que para essa combinação de modelo e dados, a utilização de *embeddings* distintos não produz muita diferença.

Apesar da estratégia de remoção de modelos utilizada nesta seção ser bastante simplista, é preciso reconhecer que a comparação de diversos modelos e bases de dados é altamente beneficiada por testes estatísticos paramétricos e não paramétricos (DEMŠAR, 2006). Mesmo assim, como o propósito desta etapa foi apenas de diminuir a quantidade de caminhos a serem explorados, aceitamos correr o risco de remover algumas das opções em troca de limitação de escopo e foco nos resultados mais promissores.

Ao fim desta etapa, reduzimos significativamente o número de modelos a explorar, permitindo que possamos focar nos seis que foram selecionados a seguir.

5.5 AJUSTES E OTIMIZAÇÃO

Nesta etapa, buscamos nos aprofundar na otimização de parâmetros dos modelos pré-selecionados, além de aplicar técnicas de normalização nos dados para alcançar melhores resultados. Outro objetivo desta etapa é limitar o número de soluções, visando consolidar as melhores alternativas entre as variações de modelos, pré-processamento e ajuste de hiperparâmetros.

5.5.1 Normalização e mudança de escala

Iniciamos essa etapa dos experimentos avaliando o benefício da aplicação de técnicas de normalização e mudança de escala. As técnicas que utilizamos foram *i*) *Standard Scaler*, *ii*) normalização mínimo-máximo (*MinMax*) e *iii*) *Power Transformer* (WEISBERG, 2001). Mantivemos a estrutura dos experimentos, ressaltando que os exemplos foram separados entre os conjuntos de treinamento e teste antes da aplicação destas técnicas para evitar o vazamento de dados e enviesamento.

Para cada um dos modelos, verificamos qual técnica teve melhor desempenho através do aferimento da acurácia e ROC AUC considerando as três bases. Em casos de empate, usamos como critério a técnica que apresentou maior aumento percentual de acurácia em valores absolutos.

Na Tabela 3 podemos ver os resultados obtidos na RT. Constata-se que: 1) *Random Forest* e MLP se demonstraram mais eficientes sem nenhuma técnica de normalização. 2) O comitê de Árvores Extra (ET) teve melhores resultados com o *Standard scaler* (SS); 3) Regressão logística e NB Gaussiano tiveram maiores ganhos com a normalização mínimo-máximo (MM); e 4) RNAt demonstrou melhores resultados com *Power Transformer* (PT).

Já para a RE, pela Tabela 4 observamos que: 1) NB Gaussiano e *Random Forest* demonstram melhores resultados com seus respectivos valores originais; 2) MLP e RNAe apresentaram melhores resultados com SS; e 3) ET e Regressão Logística tiver melhor desempenho com o PT.

Nas duas tabelas, os itens marcados em verde claro representam a técnica de normalização com maior pontuação para aquela base de dados e modelo.

Tabela 3 – Otimização através da transformação dos dados na representação textual

Modelo		Estudantes		Notícias		Varela	
	Normalização	Acurácia	AUC	Acurácia	AUC	Acurácia	AUC
RL	Sem Normalização	0.095	0.500	0.833	0.907	0.762	0.879
	Standard Scaler	0.095	0.500	0.800	0.888	0.788	0.893
	Min Max	0.142	0.517	0.833	0.907	0.864	0.931
	Power Transformer	0.190	0.534	0.766	0.870	0.776	0.887
NB Gaus.	Sem normalização	0.952	0.516	0.733	0.851	0.604	0.800
	Standard Scaler	0.142	0.533	0.733	0.851	0.622	0.809
	Min Max	0.190	0.565	0.766	0.870	0.604	0.800
	Power Transformer	0.142	0.533	0.733	0.851	0.634	0.815
Random Forest	Sem normalização	0.142	0.517	0.700	0.833	0.687	0.842
	Standard Scaler	0.190	0.550	0.700	0.833	0.681	0.838
	Min Max	0.142	0.517	0.700	0.833	0.664	0.830
	Power Transformer	0.095	0.515	0.633	0.796	0.626	0.811
Árvores Extra	Sem normalização	0.285	0.615	0.700	0.833	0.687	0.842
	Standard Scaler	0.142	0.532	0.733	0.851	0.696	0.846
	Min Max	0.190	0.550	0.700	0.833	0.680	0.838
	Power Transformer	0.238	0.598	0.700	0.833	0.694	0.845
MLP	Sem normalização	0.142	0.517	0.866	0.925	0.790	0.893
	Standard Scaler	0.190	0.566	0.766	0.870	0.591	0.793
	Min Max	0.142	0.517	0.800	0.888	0.777	0.887
	Power Transformer	0.238	0.567	0.733	0.851	0.597	0.796
RNAt	Sem normalização	0.230	0.558	0.800	0.781	0.574	0.796
	Standard Scaler	0.115	0.534	0.633	0.752	0.543	0.769
	Min Max	0.192	0.574	0.533	0.712	0.366	0.675
	Power Transformer	0.038	0.471	0.800	0.982	0.576	0.875

Tabela 4 – Otimização através da transformação dos dados na representação estilométrica

Modelo		Estudantes		Notícias		Varela	
	Normalização	Acurácia	AUC	Acurácia	AUC	Acurácia	AUC
RL	Sem Normalização	0.230	0.598	0.500	0.722	0.187	0.589
	Standard Scaler	0.153	0.528	0.833	0.907	0.606	0.801
	Min Max	0.230	0.556	0.766	0.870	0.517	0.756
	Power Transformer	0.230	0.567	0.766	0.870	0.613	0.804
NB Gaus.	Sem normalização	0.115	0.516	0.833	0.907	0.555	0.775
	Standard Scaler	0.115	0.500	0.100	0.500	0.038	0.514
	Min Max	0.076	0.494	0.200	0.555	0.134	0.562
	Power Transformer	0.038	0.482	0.100	0.500	0.038	0.514
Random Forest	Sem normalização	0.115	0.516	0.766	0.870	0.561	0.778
	Standard Scaler	0.115	0.511	0.733	0.851	0.561	0.778
	Min Max	0.153	0.528	0.633	0.796	0.450	0.722
	Power Transformer	0.115	0.506	0.666	0.814	0.570	0.782
Árvores Extra	Sem normalização	0.076	0.500	0.833	0.907	0.572	0.783
	Standard Scaler	0.115	0.505	0.733	0.851	0.570	0.782
	Min Max	0.115	0.511	0.500	0.722	0.432	0.713
	Power Transformer	0.192	0.539	0.800	0.888	0.594	0.795
MLP	Sem normalização	0.192	0.597	0.400	0.666	0.244	0.618
	Standard Scaler	0.192	0.555	0.833	0.907	0.596	0.796
	Min Max	0.115	0.516	0.600	0.777	0.556	0.776
	Power Transformer	0.192	0.534	0.800	0.888	0.623	0.809
RNAe	Sem normalização	0.115	0.555	0.433	0.755	0.173	0.570
	Standard Scaler	0.230	0.664	0.733	0.993	0.563	0.694
	Min Max	0.076	0.711	0.666	0.865	0.382	0.570
	Power Transformer	0.192	0.631	0.733	0.995	0.557	0.717

Cada uma das combinações entre representação dos dados e modelo foram associadas a técnica de normalização vencedora e aplicada em todos os experimentos posteriores.

5.5.2 Otimização

Complementando esta etapa, realizamos experimentos voltados para ajuste dos hiperparâmetros dos classificadores. Nesta etapa, optamos por substituir a divisão dos dados entre treinamento e teste e usar validação cruzada com 3 *folds*, visando diminuir o erro amostral. O número de *folds* está relacionado ao número mínimo de exemplos de uma só classe em nossas bases de dados. Para cada um dos modelos, elencamos alguns dos

principais parâmetros e intervalos de valores para otimização:

- **Modelos baseados em árvore**: número de árvores na floresta, profundidade máxima, quantidade máxima de características utilizadas, número mínimo de nós folha e número mínimo de exemplos para separação de nó.
- **MLP**: função de ativação na camada de saída, parâmetro α, número de neurônios na camada escondida, função de aprendizagem e máximo de iterações.
- **RL**: técnica de regularização para penalidade ($L1$, $L2$), parâmetro C, máximo de iterações e algoritmo de otimização.
- ***Naive Bayes* Gaussiano**: parâmetro de suavização da variância.
- **RNA**: algoritmo de otimização do modelo, quantidade de épocas para treinamento, *kernel* de inicialização dos pesos e percentual de *dropout* entre as camadas.

Para otimizar os parâmetros, realizamos uma combinação aleatória entre os possíveis valores para os parâmetros de cada modelo, seguida por uma busca em profundidade limitada. Utilizamos o algoritmo *RandomizedSearchCV* disponível na biblioteca *Sklearn*. Optamos por essa abordagem por ser mais rápida e tão eficiente quanto a busca em *grid* (PALKOVITS, 2020). Durante o treinamento, limitamos a um máximo de até 50 combinações de parâmetros. Isso significa que para cada modelo em questão, foram treinados 50 outros modelos com parâmetros distintos. Ao final deste processo, avaliamos o aumento da acurácia e AUC ao comparar o melhor modelo dentre os 50 treinados e o modelo de base de referência (sem ajuste de parâmetros). Nos casos de melhoria, substituímos o modelo de base de referência pelo otimizado.

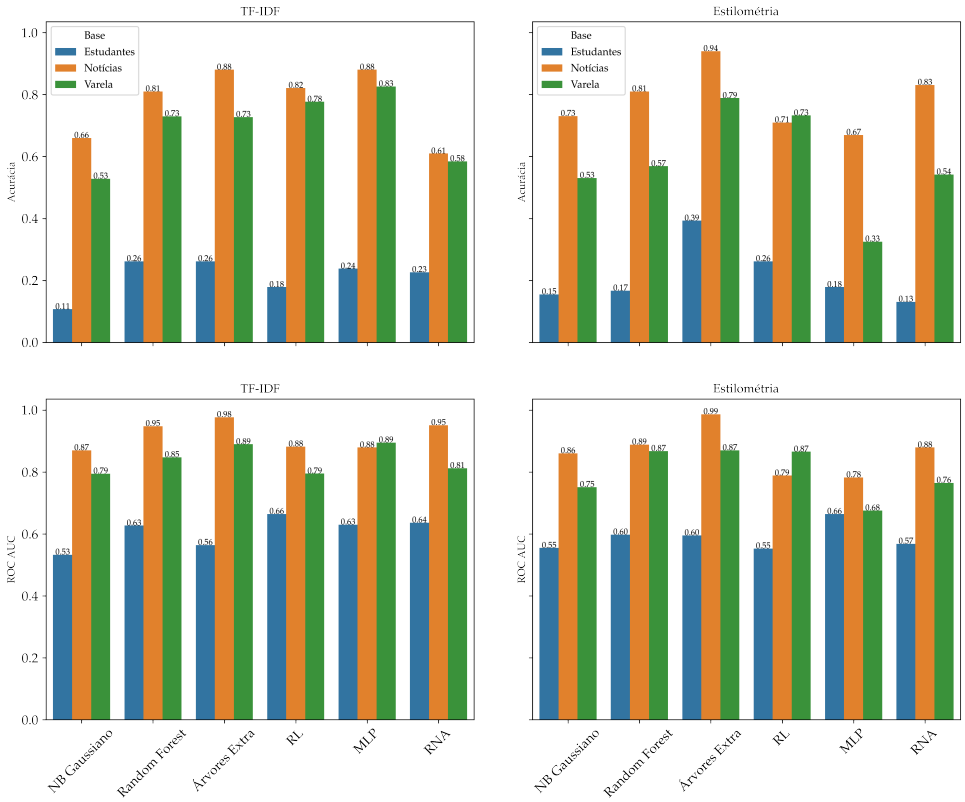
Ao final desta etapa, foram produzidos 36 modelos otimizados. Para cada uma das duas representações de dados foi criado um modelo por classificador e por conjunto de dados, como dispomos de 6 classificadores e 3 bases, temos para cada uma das representações 18 modelos.

5.6 DEFINIÇÃO DOS MELHORES MODELOS

A definição dos melhores classificadores se deu a partir dos resultados observados após a otimização (Figura 36). Ao comparar os resultados pós-otimização com os valores das

primeiras etapas de avaliação dos classificadores, pudemos observar que na maioria dos modelos, os valores absolutos melhoraram e ratificaram experimentos anteriores. Entretanto, em alguns casos como na RNAt, houve redução nos valores de acurácia. Isso pode estar relacionado à utilização da validação cruzada, reduzindo o sobreajuste sob o conjunto de treinamento da etapa anterior (MCCABE; LIN; LOVE, 2020).

Figura 36 – Resultados experimentais pós otimização agrupados por métrica, modelo e base



Fonte: Elaborado pelo autor

Para a definição dos melhores classificadores, usamos como critério aquele que foi superior aos demais em cada uma das bases de dados e o melhor classificador considerando a média dos resultados para as três bases. Ratificamos os resultados por meio do teste estatístico de análise da variância simples (*One-way* ANOVA), que foi escolhido por causa da distribuição normal dos exemplos e presença de mais de dois grupos não pareados (DEMŠAR, 2006).

Rodamos os 3 melhores classificadores (baseado na média) para cada base de dados por 30 iterações, usando partições aleatórias dos dados, e armazenamos as acurácias. Em seguida, aplicamos o teste estatístico, comparando os resultados dos classificadores dentro de cada base (90 observações) e de todos os grupos independentes para avaliação

do critério geral (270 observações).

Neste teste, a hipótese nula defende que não há diferença significativa entre os grupos observados, ou seja, que os classificadores são equivalentes e as diferenças são meramente aleatórias (DEMŠAR, 2006). Dado que m é o número de grupos sob análise e n a quantidade de observações, realizamos o cálculo de F e p, usando α=0,05, grau de liberdade no numerador $df1 = m - 1$ e grau de liberdade no denominador $df2 = m - n$. Obtendo os resultados do Quadro 4.

Quadro 4 – Análise de variância simples a partir da acurácia dos classificadores com ANOVA *one-way*

Base	$df1$	$df2$	F	p	H_0
Estudantes	2	87	40.5	$3.67 * 10^{-12}$	Rejeitada
Notícias	2	87	858	$5.21 * 10^{-58}$	Rejeitada
Varela	2	87	24855	$1.09 * 10^{-120}$	Rejeitada
Geral	8	269	5539.20	$1.76 * 10^{-286}$	Rejeitada

Fonte: Elaborado pelo autor (2021)

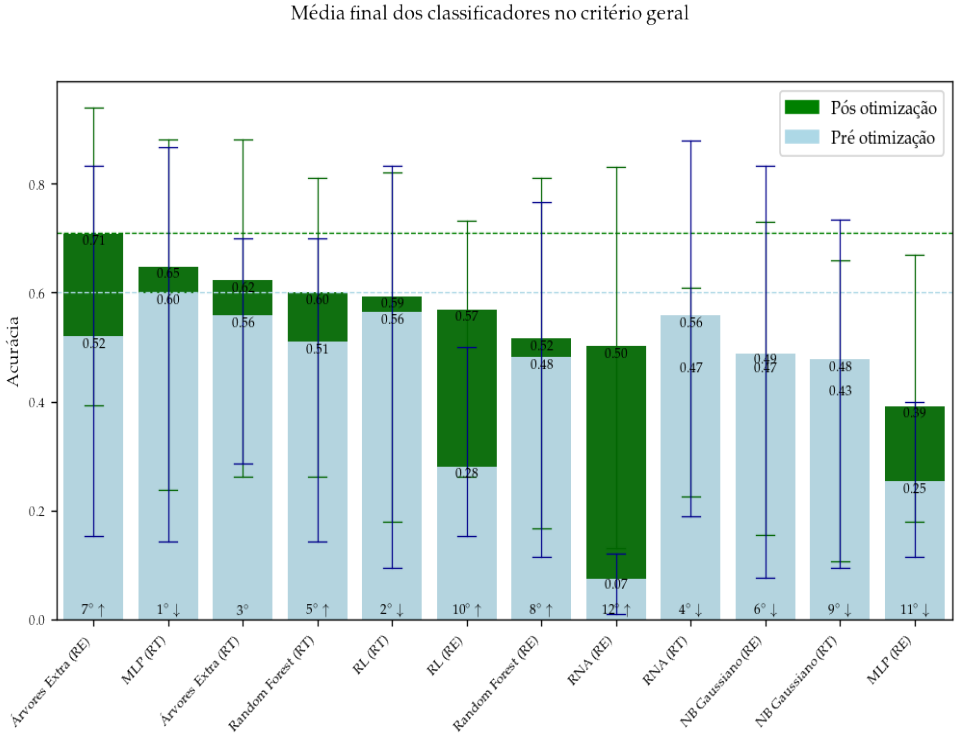
O cálculo foi realizado com auxílio da biblioteca SciPy (VIRTANEN et al., 2020) e validado por meio de calculadoras estatísticas[2]. Observamos que a hipótese nula (H_0) foi rejeitada uma vez que $p < \alpha$ e F é superior aos respectivos valores críticos perante a tabela F. A afirmativa é validada pelo teste de Tukey (TUKEY, 1949), também fornecido pela calculadora online, que confirma existirem diferenças significativas entre as observações. Desta forma, podemos definir os melhores classificadores por meio de sua média. Ao total foram selecionados apenas três classificadores, indicando que houve recorrência entre os modelos selecionados individualmente em alguma das base de dados e o modelo selecionado pela média entre todas as bases.

No critério global, o ET demonstrou ser o mais eficaz considerando todas as combinações entre modelos e representações, alcançando valores médios de acurácia de 0,70 e AUC 0,81 na representação de estilo. De acordo com os escores gerais dos classificadores, observamos que os modelos baseados em AD se sobressaem. Também observamos que a regressão logística alcançou resultados semelhantes em ambas representações. Ressaltamos que as taxas de acerto na Figura 37 são negativamente influenciadas pelos escores da base de estudantes, visível através da variância. Na imagem também é possível acompanhar alterações no *ranking* dos classificadores após a otimização, é possível constatar

[2] Disponível em https://www.statskingdom.com/180Anova1way.html

que os modelos baseados em AD foram os maiores beneficiados, como o ET que subiu da sétima posição para a primeira. De maneira oposta, os modelos *Naive Bayes* e RNAt caíram no *ranking*.

Figura 37 – Acurácia média pré e pós otimização agrupados por classificador e representação



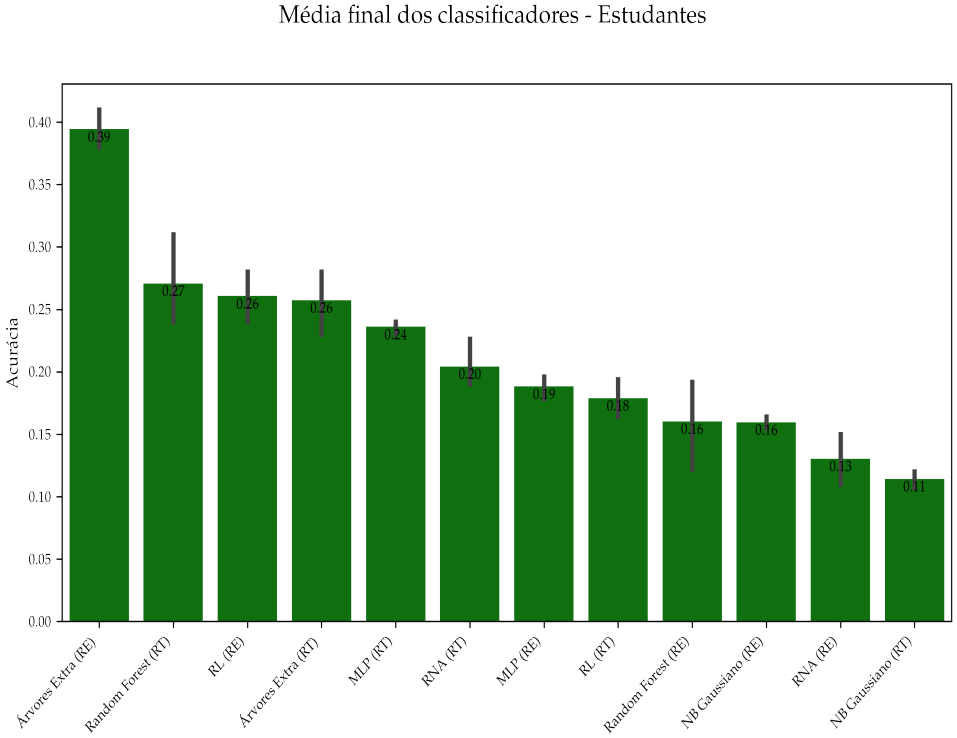
Fonte: Elaborado pelo autor

Para a base de estudantes, o melhor classificador foi o ET na representação do estilo com aplicação do PT, alcançando 39% de acurácia e 0,6 de ROC AUC. O mesmo se repete para a base de notícias, o comitê de Árvores Extra baseado em estilo com PT chegou ao máximo de 94% de acurácia de 0,98 de AUC. Já para a base de Varela, o melhor classificador foi o MLP usando a representação textual em TF-IDF com 88% de acurácia e 0,87 AUC.

Analisando os resultados da base de estudantes isoladamente, fica claro que as soluções exploradas aqui não são capazes de solucionar essa atividade de atribuição de autoria, diferentemente do presenciado nas outras duas. Ainda assim, os classificadores baseados em estilometria superam os textuais significativamente.

O fato de um algoritmo baseado na representação textual ter sido o vencedor apenas para a base Varela despertou nossa curiosidade por causa das observações durante o agrupamento e a natureza desta base, que é mais numerosa que as demais, e está distribuído

Figura 38 – Acurácia obtida na base de estudantes agrupados por classificador e representação



Fonte: Elaborado pelo autor

entre 10 assuntos diferentes. Isso colabora com nossa hipótese, pois a atividade pode estar sendo influenciada mais pelo conteúdo dos documentos do que sua autoria.

Para melhor compreensão deste fenômeno, organizamos um pequeno experimento utilizando um subconjunto de documentos da base Varela pertencentes a uma mesma categoria. Esse subconjunto é composto por 300 documentos, distribuídos igualmente entre 10 autores. Após repetir as etapas experimentais descritas neste capítulo, verificamos que o melhor classificador ainda foi a MLP na representação textual que alcançou 93,3% de acurácia e 0,98 de ROC AUC.

A novidade ficou a cargo do segundo lugar, o ET na representação de estilo, com 92% de acurácia e 0,96 ROC AUC. A diferença entre as abordagens diminui para aproximadamente 1%, contribuindo com a hipótese de que modelos baseados em texto podem ser beneficiados pela diversidade de assuntos e que características de estilo tem bastante potencial quando o escopo é reduzido.

5.7 INTERPRETAÇÃO DOS MODELOS SELECIONADOS

O campo de estudo da interpretabilidade na IA é voltado para o entendimento dos algoritmos desta área, que muitas vezes são utilizados em formato de "caixa-preta". Apesar de ser um tema em alta nos últimos anos, vale ressaltar que esse conceito existe desde os primórdios da IA (GOEBEL et al., 2018). O que tem trazido esse tema à tona é o crescimento da utilização de modelos de AM complexos, onde não há certeza sobre os fatores de maior influência sobre o mesmo.

A compreensão destes fatores pode auxiliar os usuários a entender os modelos de AM, assim como a relevância de características e funções de decisão. Além disso, há um aumento da transparência, confiança e entendimento das predições por parte dos humanos com relação aos modelos (DOŠILOVIĆ; BRČIĆ; HLUPIĆ, 2018). Como consequência, decidimos investigar os classificadores propostos como solução para o problema de pesquisa com propósito de compreender tais fatores.

Para isso, utilizamos uma tradicional técnica nomeada valores Shapley (SHAPLEY, 1953), oriunda da teoria dos jogos e amplamente utilizada na literatura por ser capaz de satisfazer diversos axiomas como: eficiência, linearidade, simetria, monotonicidade e proporcionalidade (SUNDARARAJAN; NAJMI, 2020). Interpretamos os classificadores e construímos gráficos que demonstram as características de maior influência sobre os modelos por meio da biblioteca SHAP (LUNDBERG; LEE, 2017).

Para as bases de estudantes e de notícias, conforme visto no capítulo anterior, o classificador interpretado foi o ET diante da representação estilométrica.

Já para a base Varela, o melhor classificador foi o MLP diante da representação textual com TF-IDF. Entretanto, devido à alta dimensionalidade da representação (3000 exemplos x 68.829 palavras) não foi computacionalmente possível realizar essa análise diante da infraestrutura disposta para o experimento. Em poucos minutos os 25GB de memória RAM eram ocupados, encerrando o *kernel*. Por isso, optamos por analisar o segundo melhor modelo para esta base de dados, que também foi o ET na representação de estilo.

Para simplificar a análise, optamos por selecionar aproximadamente um quarto (20) das características mais importantes de cada classificador de acordo com os valores Shapley.

Na base de estudantes, as características baseadas em caracteres, lexicais e de riqueza de vocabulário predominam igualmente entre as de maior importância. Na base de notícias

também temos dominância das características baseadas em caracteres, seguidas pelas de riqueza de vocabulário, lexicais e sintáticas.

Para a base Varela, se destacam as características lexicais e sintáticas, seguidas por riqueza de vocabulário, baseada em caracteres e de aplicação. Podemos dizer que esta base se difere das demais por possuir maior diversidade de autores, temas e vocabulário, sendo mais adequada para verificação de autoria em documentos de domínio abrangente. O destaque de características sintáticas e de aplicação são um contraponto interessante ao comparar com outras bases, que são menores e de domínio restrito, nos levando a considerar que a relevância das características na atividade de verificação é relativa ao contexto dos documentos.

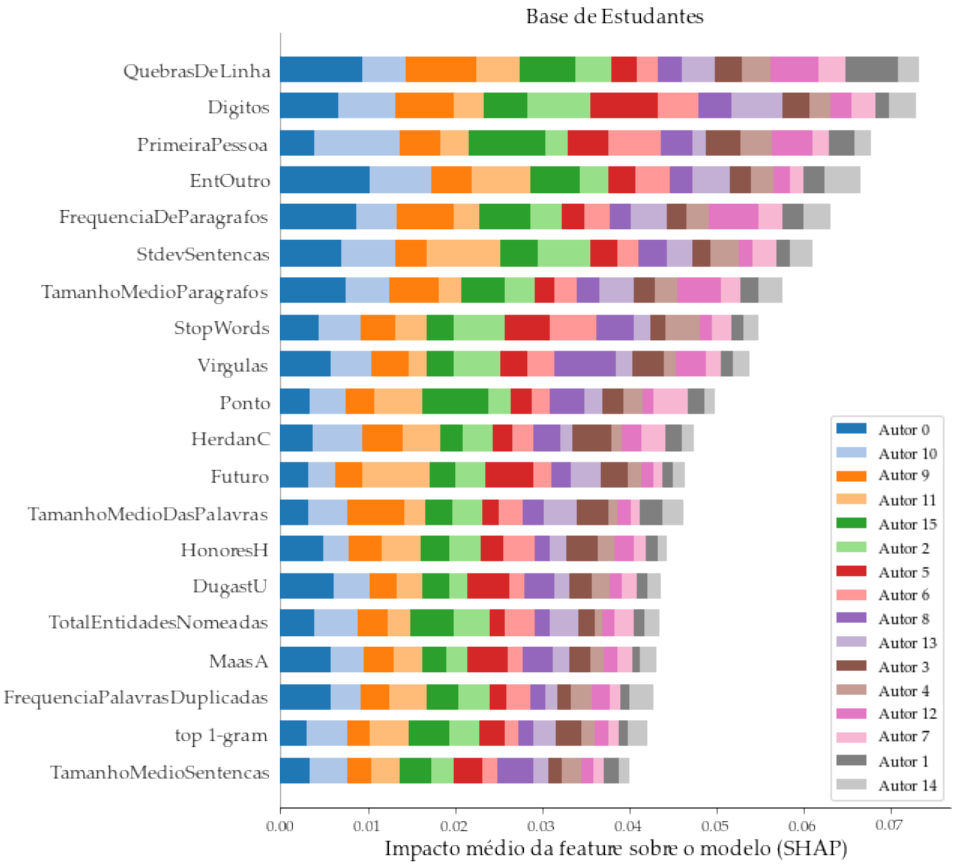
Considerando a proporção do número total de características de cada grupo de características, podemos dizer que as características lexicais são superiores às demais por pertencerem a um grupo que possui pelo menos cinco características a menos que os demais grupos. Ainda assim, a incidência de caracteres especiais e a riqueza de vocabulário também demonstraram ser importantes para os modelos. Por outro lado, as características semânticas não aparecem entre as mais importantes em nenhum dos classificadores.

A imagem 39 referente a base de estudantes nos permite observar que há uma homogeneidade entre a importância das características entre os autores. É provável que esse fator tenha colaborado com a baixa taxa de acerto observada nos diversos modelos para esta base de dados. Ainda assim, conseguimos enxergar peculiaridades de alguns autores, tais como:

- A incidência de termos na primeira pessoa escritas pelo autor 15, bem como a de termos no futuro (infinitivo) para o autor 11, são indicativos de que estes autores possuem características de escrita que os diferenciam.
- Autores 0 e 10 aparentam possuir um estilo de estruturação de texto através de parágrafos que destoam dos demais.
- A importância das *stopwords* na maioria dos autores é um indicativo de artigos e preposições no texto, que pode estar associado a esse grupo de estudantes.

Na base de notícias há uma maior heterogeneidade na importância das características, o que demonstra maior facilidade de diferenciação dos autores, confirmando os resultados

Figura 39 – Valores SHAP para base de estudantes

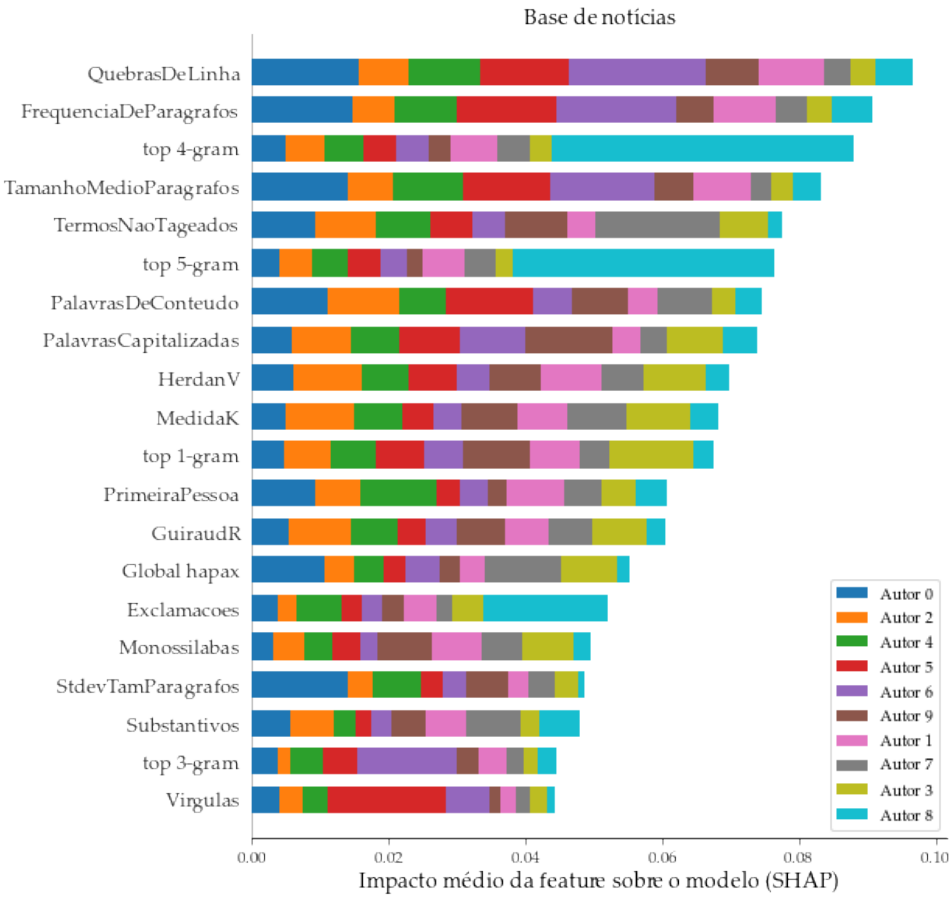


Fonte: Elaborado pelo autor

obtidos anteriormente. Analisando individualmente os autores, realizamos as seguintes observações:

- O autor 0 possui como características mais influentes aquelas relacionadas ao estilo de distribuição do documento ao longo de parágrafos.
- O impacto causado pela incidência de vírgulas indica que o autor 5 pode ter um estilo de escrita mais pausado ou sem pausas.
- Alta frequência de *top* 4-grams e 5-grams para o autor 4, que é um indicativo de uso de sequências de palavras incomuns.
- No grupo de características de riqueza de vocabulário, destacamos o autor 3 que corresponde a uma porção razoável nas quatro características deste grupo.
- Também vemos que o autor 7 se destaca na lista de termos não tagueados, que tem relação direta com o *global hapax* e pode ser indicativo de uso de palavras

Figura 40 – Valores SHAP para base de notícias

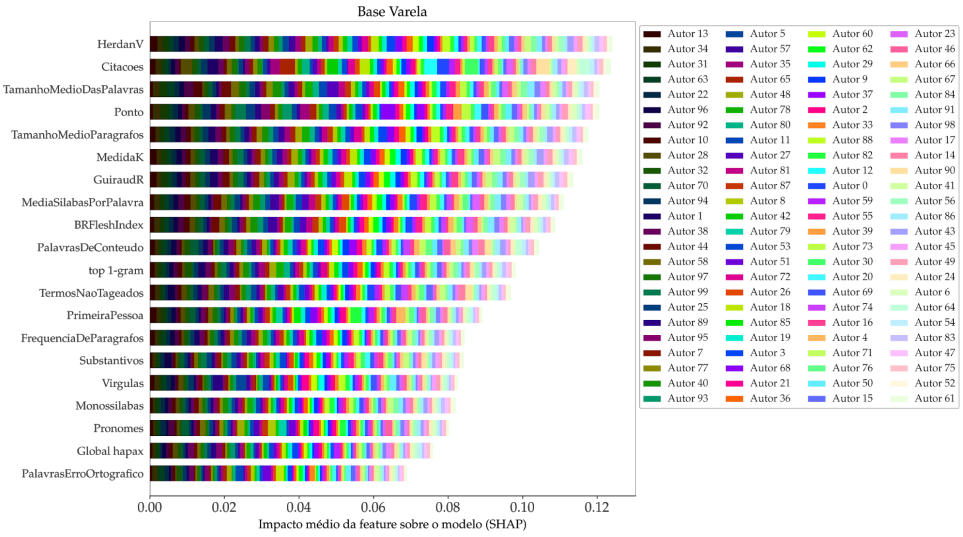


Fonte: Elaborado pelo autor

únicas ou palavras de idiomas estrangeiros. Importância das exclamações em textos jornalísticos, que a priori consideramos não ser uma prática muito comum, mas que pode ter sido introduzida neste grupo por influência do autor 8.

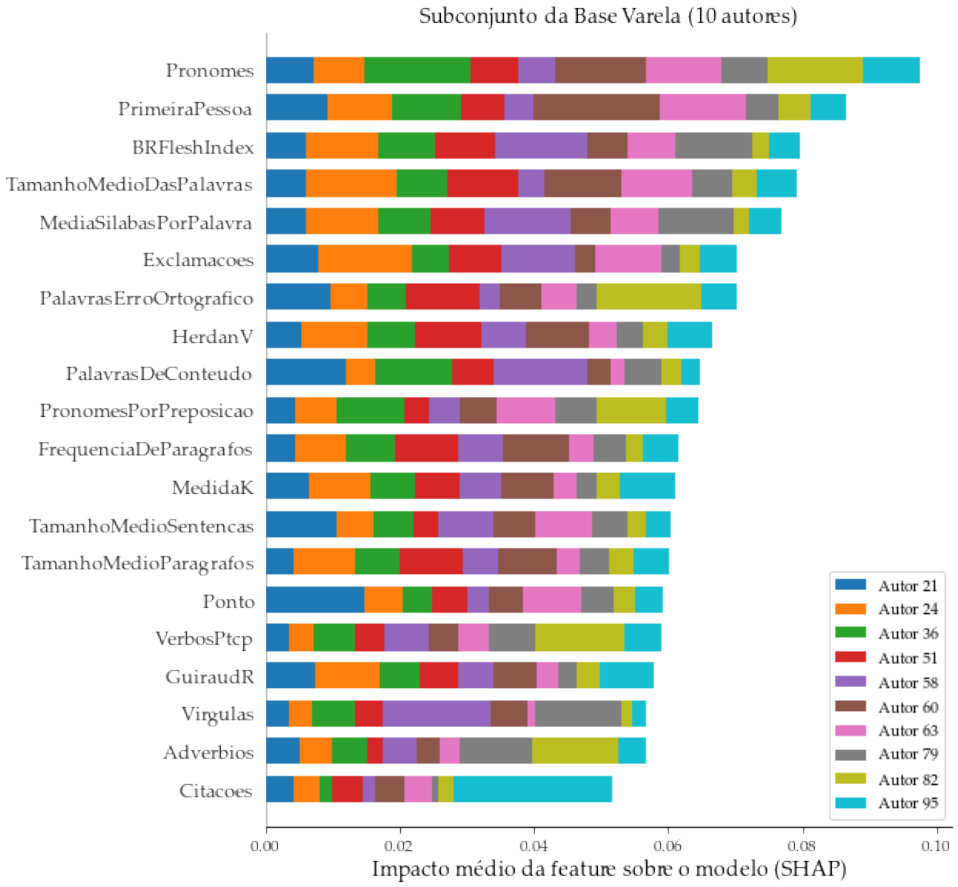
No que tange a base de Varela, vemos na Figura 41 que a distinção de características importantes para o modelo não é facilmente observável devido ao elevado número de autores. Entretanto, essa visão macro nos permitiu entender quais características se destacam no contexto geral. Para compreender como ocorreu a distinção de autoria nesta base, selecionamos aleatoriamente apenas dez autores e construímos um novo modelo passível de melhor interpretabilidade, conforme podemos ver na Figura 42

Figura 41 – Valores SHAP para a base Varela



Fonte: Elaborado pelo autor

Figura 42 – Valores SHAP para um subconjunto de 10 autores na base Varela



Fonte: Elaborado pelo autor

Analisando o subconjunto de autores, de maneira geral vemos uma maior influência

de características sintáticas, lexicais e baseadas em caracteres. Ratificando os resultados globais da base. Do ponto de vista individual, destacamos os seguintes pontos:

- Para os autores 36, 60 e 63, há um indicativo de estilo de escrita na primeira pessoa e pronomes pessoais.
- O autor 58 se distingue dos demais pelo seu índice de legibilidade (*BR-Flesch*), número de vírgulas e tamanho médio das sentenças. Analisando o conteúdo textual dos documentos escritos por esse autor, verificamos que todos estão dentro do tema saúde e há uma série de termos médicos, tais como: "arritmia", "espondilite", "hipertireoidismo" e "amiloidose". A partir daí, concluímos que o autor se distinguiu pelo cunho técnico e elaborado de seus textos.
- Os autores 51 e 82 se sobressaem pelo número de erros ortográficos; contudo ao verificar que o tema principal destes documentos é turismo, inferimos que deve haver uma quantidade de palavras estrangeiras superior aos demais documentos.
- Um valor destoante no número de citações do autor 95, inicialmente nos levou a pensar que seus textos poderiam conter citações literárias ou depoimentos. Entretanto, verificamos que os textos estavam dentro do tema economia e que estes continham diversas entrevistas com políticos e autoridades relacionadas à economia.

Como resultado da análise global, inferimos que, independente da base de dados, para atividades de atribuição de autoria, o ET se beneficia principalmente de características lexicais, baseadas em caracteres, sintáticas e de riqueza de vocabulário. Já as características semânticas não apresentaram resultados expressivos para os nossos classificadores. Contudo, ressaltamos a probabilidade deste fato estar relacionado ao uso de um *POS-Tagger* previamente treinado e possivelmente menos estável se comparado aos recursos sintáticos e lexicais fornecidos pela renomada biblioteca NLTK.

Do ponto de vista individual, foi possível enxergar padrões de estilo de escrita relacionados à autoria. Para a base de textos jornalísticos, onde temos maior segurança de que os textos são propriedade intelectual de seus autores, há uma maior heterogeneidade na relação entre autoria e características de estilo mais influentes. Já na base de estudantes, onde há uma probabilidade maior de utilização de fontes de informação na internet e colaboração entre os autores, as características são mais homogêneas. Além disso, destacamos o contexto em que os textos se inserem.

Na base de estudantes, temos um conjunto de documentos que visam demonstrar os conhecimentos acadêmicos absorvidos ao decorrer de uma disciplina de graduação, se limitando à temas pré-definidos na disciplina formatados em texto livre, questionários e análises. Seu público alvo é majoritariamente composto por professores, coordenadores, monitores e colegas de classe. Diante disto, temos *a*) muitos documentos tratando do mesmo tema e em alguns casos respondendo às mesmas perguntas - o que acarreta numa restrição ideológica e de vocabulário; e *b*) muitos documentos dentro do mesmo formato, às vezes especificado pelo professor ou vinculado à normas institucionais e acadêmicas - limitando a estrutura do texto ao formato preterido.

Por outro lado, as duas bases de textos jornalísticos, apesar de tratarem de temas semelhantes (direito, economia, política), são compostas por documentos que retratam fatos ou temas no plano linguístico, escritos por jornalistas, com propósito de democratizar o conhecimento por meio do entendimento do público (FILHO, 2006). Obviamente que há uma chance de existir mais de uma matéria sobre o mesmo assunto ou fato, porém esse não é o caso mais comum, permitindo que os documentos tratem de temas e ideias mais abrangentes que a base de estudantes. Além disso, esses documentos possuem maior flexibilidade na estruturação dos textos, uma vez que estão atrelados a veículos de comunicação online.

Algumas das observações que reforçam os pontos acima são: *a*) menor influência de *top-grams* e *hapax* na base de estudantes. Ou seja, não há uma incidência de palavras únicas no corpora suficiente para distinguir autores, provavelmente devido às limitações ideológicas e de vocabulário citadas anteriormente. *b*) alta influência de palavras duplicadas, relacionado às mesmas restrições e também a características dos autores, que ainda são estudantes e possivelmente possuem habilidades linguísticas menos desenvolvidas que os jornalistas.

Além disso, o baixo desempenho da base de estudantes em comparação às bases de textos jornalísticos pode estar associado a um problema conhecido em atividades relacionadas à autoria, que é a maior influência do tópico do documento que o estilo de escrita sobre os classificadores (GAMON, 2004). A utilização de características não relacionadas aos tópicos dos documentos é uma das abordagens sugeridas por diversos autores (HALVANI; GRANER; REGEV, 2020) (SARWAR et al., 2018). No contexto desta pesquisa, consideramos que dentre as características de estilo, apenas os *top-grams* podem estar relacionados ao tópico do documento. Por outro lado, a representação textual sofre grande influência deste

fenômeno (HALVANI; GRANER; REGEV, 2020).

6 CONSIDERAÇÕES FINAIS

Neste capítulo serão apresentadas as considerações finais deste trabalho, incluindo suas principais conclusões, contribuições, limitações e oportunidades para trabalhos futuros.

6.1 CONCLUSÕES

Este trabalhou buscou inicialmente solucionar a atividade de atribuição de autoria num contexto educacional, no qual existem mais restrições do que a grande maioria dos problemas nesta área. Exploramos diversos algoritmos de AM apoiados por recursos de PLN sofisticados, porém nenhuma das abordagens alcançou resultados satisfatórios para a base de estudantes. Esse cenário confirma o levantamento feito por Tempestt et al. (2017), em que os autores afirmam que o pequeno número de documentos em comparação ao grande número de autores é um dos principais desafios na área.

As análises exploratórias visuais demonstraram alta dificuldade de separação dos exemplos em representações de baixa dimensionalidade, obtidas por meio de técnicas para redução de dimensionalidade como PCA, TSNE e LSA. O agrupamento de exemplos por meio de técnicas não supervisionadas, como KNN e *fuzzy* c-médias não demonstraram ser eficazes para separação de autores entre os *clusters*. Inspeções ad-hoc no conteúdo dos *clusters* indicaram que o agrupamento dos exemplos apresenta maior correlação por assunto do que autoria.

Através da comparação com outras duas bases de dados, uma construída durante essa pesquisa, com documentos coletados num portal de notícias e a outra disposta na tese de VARELA, pudemos observar que para estas bases, o problema pode ser solucionado por meio da metodologia sugerida.

Propomo-nos a distinguir os autores pelo seu estilo de escrita, que pode ser quantificado pelas características estilométricas (STAMATATOS, 2008). Partindo do pressuposto que tal abordagem é superior às análises feitas sobre o conteúdo textual, que por sua vez têm maior afinidade em casos de seleção e modelagem de tópicos (SRIURAI, 2011).

A pesquisa convergiu para a construção de 74 características de estilo, em sua maioria, voltadas para a língua portuguesa e subdivididas em grupos lógicos. Tais características se inspiraram em trabalhos anteriores (STAMATATOS, 2009) e o processo de transformação

do texto original em tais características derivou um sistema de conversão.

Seguimos as tradicionais etapas de projetos de AM, passando por análise, normalização, otimização e validação dos modelos diante de representações de estilo e texto para as 3 bases de dados. Surpreendentemente, modelos clássicos superaram modelos baseados em aprendizagem profunda. Apesar dos resultados promissores provenientes da associação entre *word embeddings* e RNRs em atividades similares na literatura (CHOWDHURY; IMON; ISLAM, 2018) (SHRESTHA et al., 2017), neste trabalho tal composição ficou aquém dos modelos tradicionais. Acreditamos que isso se dá principalmente por causa do baixo volume de dados, uma vez que tais arquiteturas têm maior êxito diante de quantidades mais expressivas de exemplos para treinamento. Os modelos também podem ter sido influenciados pela perda de palavras ausentes nos *embeddings*. Uma alternativa futura pode ser o uso de aprendizagem por transferência (TORREY; SHAVLIK, 2010).

Avaliamos os classificadores através da validação cruzada, acurácia e ROC AUC, constatamos que as representações de estilo superaram as representações textuais em praticamente todos os critérios, exceto para a base Varela, onde a MLP textual supera por 5% os classificadores baseados em estilo. Também consideramos a solução como adequada para as duas bases jornalísticas, pois foi possível obter 94% (notícias) e 79% (Varela) em taxa de acerto usando características de estilo.

Analisando os resultados e a composição das bases de dados, acreditamos que algoritmos baseados em representações textuais se beneficiam em *corpora* com maior diversidade de assuntos, isso porque os modelos são capazes de distinguir entre assuntos por meio de palavras chave e sua frequência, passando a falsa impressão de separação por autoria, que não se repete em domínios mais específicos, como a base de estudantes.

Neste trabalho, concluímos que as características de estilo lexical, baseadas em caracteres, sintáticas e de riqueza de vocabulário foram as que mais colaboraram para os modelos propostos, auxiliando na atividade de atribuição de autoria. Outros grupos de características, tais quais as semânticas, apesar de não se destacarem tanto aqui quanto em trabalhos relacionados (SARWAR et al., 2018) (PAVELEC et al., 2008) podem ser justificados pela menor estabilidade do algoritmo de NER utilizado, se comparado aos oferecidos através de renomadas bibliotecas em outros idiomas.

Mesmo não solucionando a atividade proposta para a base de estudantes, o uso de características de estilo associado às árvores de decisão extremamente aleatórias (ET) demonstrou ser a melhor opção, dentre várias, para verificação de autoria. Diante dos

resultados positivos para as bases de textos jornalísticos, acreditamos que a composição da base de estudantes não é representativa o suficiente para distinção de autores. Pressupomos que o baixo volume de exemplos, a grave limitação de assuntos, possível utilização de recursos online semelhantes e a ausência de um estilo de escrita bem definido por este grupo de autores são fatores candidatos a elucidar nossas observações.

6.2 CONTRIBUIÇÕES

As principais contribuições deste trabalho são:

- Pesquisa e agrupamento de material bibliográfico, majoritariamente dispostos nos Capítulos 2 e 3, não se limitando a tópicos nas áreas de estudo (AM e PLN), mas também uma extensa discussão sobre a escrita, estilo, análise de autoria em documentos e estilometria.

- Descrição de estratégias para construção de características de estilo, assim como referência para recursos de PLN que convergiram para construção de um sistema de conversão de documentos em características de estilo na língua portuguesa, disponibilizado em código aberto, que supera as ferramentas pré-existentes (ALUÍSIO; CUNHA; SCARTON, 2016) no número de características geradas, extensibilidade e usabilidade.

- As etapas metodológicas e experimentais desta pesquisa contribuem para a literatura na área de atribuição de autoria, com destaque para atividades com dados limitados na língua portuguesa. Além disso, o trabalho cria precedentes para futuras pesquisas com propósito de mitigar o problema de pesquisa original, que é a construção de uma ferramenta de apoio no processo de ensino, aprendizagem e verificação de conhecimento.

- Devido ao uso de uma vasta combinação de modelos de AM, representações dos documentos e bases de dados, consideramos que os experimentos e observações feitas durante essa etapa podem auxiliar outros trabalhos, indicando quais combinações estão mais propensas ou não ao êxito. Dificuldades em aberto na área, como a limitação dos dados, também foram enfrentadas aqui, corroborando com trabalhos anteriores.

- Através dos experimentos pudemos confirmar nossa hipótese, que defende o uso de características de estilo em contraste a abordagens textuais. Pudemos observar o benefício de tais características, em especial para bases de pequeno porte. Além disso, para a maior base estudada existem indícios de influência do tópico sob os classificadores.

- Também evidenciamos que para a base de Varela, os resultados encontrados foram relativamente superiores aos propostos pelo trabalho original (VARELA, 2017). Principalmente porque a nossa proposta se baseia em 74 características de estilo, inferior às 132 usadas por Varela, e ainda assim obteve resultados equivalentes (79,5% de acurácia). Além disso, a MLP textual supera o SVM usado por Varela com 83% de acerto.

- Por fim, mesmo não propondo uma solução para o problema original dos estudantes, as discussões e resultados apresentados neste trabalho colaboram para a ciência, trazendo observações relevantes acerca de técnicas para atribuição de autoria por intermédio da estilometria em um nicho até então pouco discutido, especialmente na língua portuguesa.

6.3 LIMITAÇÕES

Este trabalhou buscou investigar o comportamento de vários modelos de AM para a atividade de atribuição de autoria diante de bases de dados distintas. O uso de características estilométricas se sobressai como solução final, confirmando resultados de trabalhos anteriores. Contudo, consideramos que estas são apenas etapas iniciais para solução do problema e que existem diversas limitações conhecidas nesta dissertação.

Em primeiro lugar, as dificuldades encontradas devido ao pequeno número de exemplos na base de estudantes, nos levando à utilização de bases comparativas, foi um fator limitante para construção de hipóteses generalistas. Tentativas de aumento da base de dados por geração sintética de exemplos ou separação dos documentos em sentenças e parágrafos não solucionaram o problema.

Pelo mesmo motivo, não foi possível separar um conjunto de exemplos (*holdout*) para validação final, que é uma prática recomendada para avaliar classificadores diante de dados nunca vistos anteriormente, reduzindo o enviesamento (RASCHKA, 2018).

Uma vez que dispúnhamos de três bases de dados para comparações, decidimos repetir todos os procedimentos aplicados na base de estudantes para as outras duas bases. Isso implicou numa limitação no número de *folds* (3) para a validação cruzada e a não aplicação da técnica de *holdout* mesmo para as bases maiores.

Apesar de termos explorado uma variedade significativa de modelos de AM neste trabalho, temos ciência que existem outras possibilidades que poderiam ser testadas, tal como as árvores com aumento de gradiente extremo (*XGBoost Trees*) e modelos de aprendizagem profunda mais robustos, como por exemplo CNNs encadeadas, CNN-RNN e BERT (UCHENDU et al., 2020). Os principais motivos para restrição de uso de tais modelos foram a limitação de escopo e faltas de expertise e tempo.

Outro ponto que não exploramos foi a aplicação de técnicas de extração automática de características através de redes neurais durante o pré-processamento, tal como proposto por Boenninghoff et al. (2020).

As etapas de seleção dos classificadores também apresentam restrições. Apesar de usarmos a validação cruzada e executar experimentos por múltiplas interações, a utilização de testes estatísticos só veio à tona na última etapa e de maneira não pareada. Com isso, é possível que opções interessantes tenham sido excluídas anteriormente e que os testes conduzidos poderiam ser aprimorados. Idealmente, gostaríamos de ter aplicado testes estatísticos em todas as etapas de seleção para aumentar a certeza acerca das decisões tomadas (KHOMYTSKA; TESLYUK, 2020).

Por fim, o presente trabalho não conseguiu solucionar o problema da atribuição de autoria para a base de estudantes, visto que os modelos propostos não alcançaram valores satisfatórios. Mesmo assim, tanto a metodologia usada neste trabalho, como seus resultados e discussões servem de subsídios para pesquisas futuras.

6.4 TRABALHOS FUTUROS

O trabalho apresentado pode ser expandido e aprimorado por meio de novas contribuições nos aspectos a seguir:

- Maior investigação com relação ao uso de modelos não explorados nesta pesquisa, tais como modelos de aprendizagem profunda e inclusão de comitês de classificadores híbridos dentro do arcabouço de experimentos.

- Expansão da base de estudantes através da coleta de novos documentos, ou criação de novas parcerias entre pesquisadores e instituições de ensino, para construção de bases de dados maiores.

- Melhor interpretação da importância de *tokens* em modelos baseados na representação textual, especialmente para a base Varela, pois não conseguimos interpretar o seu melhor classificador.

- Análise de viabilidade e desempenho da utilização de modelos baseado em características estilométricas e textuais simultaneamente.

- Execução da metodologia proposta em bases de dados similares, visando contrastar características demográficas com o estilo de escrita dos autores.

- Exploração de outras estratégias para avaliação dos classificadores, tal como o uso de *top-n* autores (VARELA, 2017), medida $f1$ e adaptações desta medida, como a "c@1" proposta por Kestemont et al. (2020).

- Extensão da metodologia proposta neste trabalho para bases de dados comumente utilizadas em competições de atribuição de autoria, com propósito de comparar os resultados apresentados aqui com o estado da arte em outros domínios. A extensão para outros idiomas, como inglês e espanhol também pode melhorar a qualidade das características de estilo, uma vez que estas dependem das ferramentas de PLN.

- Investigação entre causa efeito da relação entre a perda de palavras raras ao usar *word-embeddings* e sua influência na atividade de classificação.

- Verificação da abordagem proposta em problemas *multilabel*, com documentos escritos por mais de um autor (SARWAR et al., 2018) (SAVOY, 2020).

- Condução de um estudo de campo com professores e responsáveis pela avaliação das atividades, a fim de validar projeções sobre o público alvo e refinar a utilidade deste estudo durante a tomada de decisão.

- Maiores esforços são necessários para permitir o uso prático da solução proposta pelo público alvo. Tanto no que se refere ao refinamento dos classificadores como na

criação de sistemas com interface de usuário ou integração com plataformas educacionais, permitindo aos professores e tutores analisar as principais características de estilo dos alunos, podendo ser uma ferramenta útil num futuro próximo.

Não obstante, a maior parte dos pontos citados como limitações desta pesquisa também são opções de extensão e melhoria desta obra.

REFERÊNCIAS

AGARAP, A. F. M. A neural network architecture combining gated recurrent unit (gru) and support vector machine (svm) for intrusion detection in network traffic data. In: *Proceedings of the 2018 10th International Conference on Machine Learning and Computing*. [S.l.: s.n.], 2018. p. 26–30.

AKBANI, R.; KWEK, S.; JAPKOWICZ, N. Applying support vector machines to imbalanced datasets. In: SPRINGER. *European conference on machine learning*. [S.l.], 2004. p. 39–50.

ALAMMAR, J. The illustrated bert, elmo, and co.(how nlp cracked transfer learning). *Dec*, v. 3, p. 1–18, 2018.

ALBINO, J. P.; AZEVEDO, M. L. de; BITTENCOURT, P. A. S. A evolução do ead no ensino superior e suas tendências na educação brasileira. *Brazilian Journal of Development*, v. 6, n. 5, p. 28146–28155, 2020.

ALCARAZ, R. C. Expressões idiomáticas e convencionais de stella ortweiler tagnin. *Fragmentos: Revista de Língua e Literatura Estrangeiras*, v. 3, n. 1, p. 157–159, 1990.

ALENCAR, L. F. de. *Aelius: uma ferramenta para anotação automática de corpora usando o NLTK*. [S.l.]: ELC, 2010.

ALUÍSIO, S.; CUNHA, A.; SCARTON, C. Evaluating progression of alzheimer's disease by regression and classification methods in a narrative language test in portuguese. In: SPRINGER. *International Conference on Computational Processing of the Portuguese Language*. [S.l.], 2016. p. 109–114.

ALUÍSIO, S.; PELIZZONI, J.; MARCHI, A. R.; OLIVEIRA, L. de; MANENTI, R.; MARQUIAFÁVEL, V. An account of the challenge of tagging a reference corpus for brazilian portuguese. In: SPRINGER. *International Workshop on Computational Processing of the Portuguese Language*. [S.l.], 2003. p. 110–117.

ANDRADE, L. M. A escrita, uma evolução para a humanidade. *Linguagem em (Dis) curso*, v. 1, n. 1, 2010.

ARAÚJO, F. H.; CARNEIRO, A.; SILVA, R. R.; MEDEIROS, F. N.; USHIZIMA, D. M. Redes neurais convolucionais com tensorflow: Teoria e prática. *SOCIEDADE BRASILEIRA DE COMPUTAÇÃO. III Escola Regional de Informática do Piauí. Livro Anais-Artigos e Minicursos*, Sociedade Brasileira de Computação, v. 1, p. 382–406, 2017.

BAAYEN, R. H. *Analyzing linguistic data: A practical introduction to statistics using R*. [S.l.]: Cambridge University Press, 2008.

BAILEY, G.; WIKLE, T.; TILLERY, J.; SAND, L. The apparent time construct. *Language variation and change*, v. 3, n. 3, p. 241–264, 1991.

BEKKERMAN, R.; GAVISH, M. High-precision phrase-based document classification on a modern scale. In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.: s.n.], 2011. p. 231–239.

BELAK, S.; PESA, A. R.; BELAK, B. Stylometry–definition and development. *Annals of DAAAM & Proceedings*, DAAAM International Vienna, p. 85–87, 2008.

BENNETT, K.; MANGASARIAN, O. Combining support vector and mathematical programming methods for induction. *Advances in Kernel Methods-SV Learning*, p. 307–326, 1999.

BEVENDORFF, J.; GHANEM, B.; GIACHANOU, A.; KESTEMONT, M.; MANJAVACAS, E.; POTTHAST, M.; RANGEL, F.; ROSSO, P.; SPECHT, G.; STAMATATOS, E. et al. Shared tasks on authorship analysis at pan 2020. In: SPRINGER. *European Conference on Information Retrieval*. [S.l.], 2020. p. 508–516.

BEVENDORFF, J.; GHANEM, B.; GIACHANOU, A.; KESTEMONT, M.; MANJAVACAS, E.; POTTHAST, M.; RANGEL, F.; ROSSO, P.; SPECHT, G.; STAMATATOS, E. et al. Shared tasks on authorship analysis at pan 2020. In: SPRINGER. *European Conference on Information Retrieval*. [S.l.], 2020. p. 508–516.

BEVENDORFF, J.; GHANEM, B.; GIACHANOU, A.; KESTEMONT, M.; MANJAVACAS, E.; POTTHAST, M.; RANGEL, F.; ROSSO, P.; SPECHT, G.; STAMATATOS, E. et al. Shared tasks on authorship analysis at pan 2020. In: SPRINGER. *European Conference on Information Retrieval*. [S.l.], 2020. p. 508–516.

BEZDEK, J. C. *Pattern recognition with fuzzy objective function algorithms*. [S.l.]: Springer Science & Business Media, 2013.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. *the Journal of machine Learning research*, JMLR. org, v. 3, p. 993–1022, 2003.

BLIKSTEIN, I. *Técnicas de comunicação escrita*. [S.l.]: Ática, 1985. v. 12.

BOENNINGHOFF, B.; RUPP, J.; NICKEL, R. M.; KOLOSSA, D. Deep bayes factor scoring for authorship verification. *arXiv preprint arXiv:2008.10105*, 2020.

BOJANOWSKI, P.; GRAVE, E.; JOULIN, A.; MIKOLOV, T. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, MIT Press, v. 5, p. 135–146, 2017.

BORGES, T. B.; KICKHÖFEL, R. B.; LICHTNOW, D.; RODRIGUES, R.; LOH, S.; SALDAÑA, R. G. Classificação automática de documentos em uma biblioteca digital com o uso de ontologias. In: *IV Congresso Brasileiro de Computação–CBCOMP 2004, Itajaí*. [S.l.: s.n.], 2004.

BORKO, H.; BERNICK, M. Automatic document classification. *Journal of the ACM (JACM)*, ACM New York, NY, USA, v. 10, n. 2, p. 151–162, 1963.

BRANTS, T. Tnt-a statistical part-of-speech tagger. *arXiv preprint cs/0003055*, 2000.

BRASIL, S. F. do. Constituição da república federativa do brasil. *Brasília: Senado Federal, Centro Gráfico*, 1988.

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.

BREIMAN, L.; FRIEDMAN, J.; STONE, C. J.; OLSHEN, R. A. *Classification and regression trees*. [S.l.]: CRC press, 1984.

BROCARDO, M. L.; TRAORE, I.; SAAD, S.; WOUNGANG, I. Authorship verification for short messages using stylometry. In: IEEE. *2013 International Conference on Computer, Information and Telecommunication Systems (CITS)*. [S.l.], 2013. p. 1–6.

BRÜCHER, H.; KNOLMAYER, G.; MITTERMAYER, M.-A. Document classification methods for organizing explicit knowledge. Institute of Information Systems, 2002.

BRÜCHER, H.; KNOLMAYER, G.; MITTERMAYER, M.-A. Document classification methods for organizing explicit knowledge. Institute of Information Systems, 2002.

BRUCKSCHEN, M.; MUNIZ, F.; SOUZA, J.; FUCHS, J.; INFANTE, K.; MUNIZ, M.; GONÇALVES, P.; VIEIRA, R.; ALUISIO, S. Anotaçao lingüstica em xml do corpus pln-br. *Série de relatórios do NILC, ICMC-USP*, 2008.

CANALES, O.; MONACO, V.; MURPHY, T.; ZYCH, E.; STEWART, J.; CASTRO, C. T. A.; SOTOYE, O.; TORRES, L.; TRULEY, G. A stylometry system for authenticating students taking online tests. *P. of Student-Faculty Research Day, Ed., CSIS. Pace University*, 2011.

CATAL, C.; DIRI, B. Investigating the effect of dataset size, metrics sets, and feature selection techniques on software fault prediction problem. *Information Sciences*, Elsevier, v. 179, n. 8, p. 1040–1058, 2009.

CHALKIDIS, I.; FERGADIOTIS, M.; MALAKASIOTIS, P.; ANDROUTSOPOULOS, I. Large-scale multi-label text classification on eu legislation. *arXiv preprint arXiv:1906.02192*, 2019.

CHALL, J. S.; DALE, E. *Readability revisited: The new Dale-Chall readability formula*. [S.l.]: Brookline Books, 1995.

CHASKI, C. E. Who's at the keyboard? authorship attribution in digital evidence investigations. *International journal of digital evidence*, Citeseer, v. 4, n. 1, p. 1–13, 2005.

CHEN, D.; MANNING, C. D. A fast and accurate dependency parser using neural networks. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. [S.l.: s.n.], 2014. p. 740–750.

CHEN, X.; HAO, P.; CHANDRAMOULI, R.; SUBBALAKSHMI, K. Authorship similarity detection from email messages. In: SPRINGER. *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. [S.l.], 2011. p. 375–386.

CHEN, X.; TIAN, J.; CHENG, J.; YANG, X. Segmentation of fingerprint images using linear classifier. *EURASIP Journal on Advances in Signal Processing*, Springer, v. 2004, n. 4, p. 1–15, 2004.

CHICCO, D.; JURMAN, G. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, Springer, v. 21, n. 1, p. 1–13, 2020.

CHOI, J. Y.; CHOI, C.-H. Sensitivity analysis of multilayer perceptron with differentiable activation functions. *IEEE Transactions on Neural Networks*, IEEE, v. 3, n. 1, p. 101–107, 1992.

CHOWDHURY, G. G. Natural language processing. *Annual review of information science and technology*, Wiley Online Library, v. 37, n. 1, p. 51–89, 2003.

CHOWDHURY, H. A.; IMON, M. A. H.; ISLAM, M. S. A comparative analysis of word embedding representations in authorship attribution of bengali literature. In: IEEE. *2018 21st International Conference of Computer and Information Technology (ICCIT)*. [S.l.], 2018. p. 1–6.

CHUNG, J.; GULCEHRE, C.; CHO, K.; BENGIO, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

CLACK, C.; FARRINGDON, J.; LIDWELL, P.; YU, T. Autonomous document classification for business. In: *Proceedings of the first international conference on Autonomous agents*. [S.l.: s.n.], 1997. p. 201–208.

CLOUGH, P. et al. Old and new challenges in automatic plagiarism detection. In: CITESEER. *National Plagiarism Advisory Service, 2003; http://ir. shef. ac. uk/cloughie/index. html*. [S.l.], 2003.

CRAWFORD, M.; KHOSHGOFTAAR, T. M.; PRUSA, J. D.; RICHTER, A. N.; NAJADA, H. A. Survey of review spam detection using machine learning techniques. *Journal of Big Data*, SpringerOpen, v. 2, n. 1, p. 1–24, 2015.

CSTOOLS. *center for sprog teknologi*. Disponível em: <https://cst.dk/tools/index.php#output>. Acesso em: out. 2020.

CURTIS, G. J.; TREMAYNE, K. Is plagiarism really on the rise? results from four 5-yearly surveys. *Studies in Higher Education*, Taylor & Francis, p. 1–11, 2019.

CUSTÓDIO, J. E.; PARABONI, I. Each-usp ensemble cross-domain authorship attribution. In: *Working notes of CLEF 2018—conference and labs of the evaluation forum*. [S.l.: s.n.], 2018.

CUTTING, D.; KUPIEC, J.; PEDERSEN, J.; SIBUN, P. A practical part-of-speech tagger. In: *Third Conference on Applied Natural Language Processing*. [S.l.: s.n.], 1992. p. 133–140.

D'AGOSTINI, G. A multidimensional unfolding method based on bayes' theorem. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, Elsevier, v. 362, n. 2-3, p. 487–498, 1995.

DALEY, D. J.; VERE-JONES, D. Scoring probability forecasts for point processes: The entropy score and information gain. *Journal of Applied Probability*, JSTOR, p. 297–312, 2004.

DEERWESTER, S.; DUMAIS, S. T.; FURNAS, G. W.; LANDAUER, T. K.; HARSHMAN, R. Indexing by latent semantic analysis. *Journal of the American society for information science*, Wiley Online Library, v. 41, n. 6, p. 391–407, 1990.

DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, JMLR. org, v. 7, p. 1–30, 2006.

DEPRESBITERIS, L. *Avaliação educacional em três atos*. [S.l.]: Senac, 2017.

DEPRESBITERIS, L.; TAVARES, M. R. *Diversificar é preciso...: instrumentos e técnicas de avaliação de aprendizagem*. [S.l.]: Senac, 2017.

DESIGN, H. L. *Fandoms: What They Are And How To Get Into It (Or Not)*. 2019. Disponível em: <https://www.hundredlifedesign.com/fandoms-what-they-are-and-how-to-get-into-it-or-not/>. Acesso em: 24 Maio 2021.

DIEDERICH, J.; KINDERMANN, J.; LEOPOLD, E.; PAASS, G. Authorship attribution with support vector machines. *Applied intelligence*, Springer, v. 19, n. 1, p. 109–123, 2003.

DOŠILOVIĆ, F. K.; BRČIĆ, M.; HLUPIĆ, N. Explainable artificial intelligence: A survey. In: IEEE. *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*. [S.l.], 2018. p. 0210–0215.

DUBAY, W. H. The principles of readability. *Online Submission*, ERIC, 2004.

DUGAST, D. *Vocabulaire et stylistique*. [S.l.]: Slatkine, 1979. v. 8.

FARIAS, L. M.; SELLITTO, M. A. Uso da energia ao longo da história: evolução e perspectivas futuras. *Revista Liberato*, v. 12, n. 17, p. 07–16, 2011.

FELDMAN, S. Nlp meets the jabberwocky: Natural language processing in information retrieval. *ONLINE-WESTON THEN WILTON-*, Citeseer, v. 23, p. 62–73, 1999.

FERREIRA, A. E. T. Estimação do ângulo de direção por vídeo para veículos autônomos utilizando redes neurais convolucionais multicanais. 2017.

FILHO, C. B. Elementos fundamentais para a prática do jornalismo científico. *Biblioteca on-line de ciências da comunicação*, 2006.

FRANÇOIS, T.; FAIRON, C. An "ai readability" formula for french as a foreign language. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. [S.l.: s.n.], 2012. p. 466–477.

FRANTZESKOU, G.; STAMATATOS, E.; GRITZALIS, S.; KATSIKAS, S. Effective identification of source code authors using byte-level information. In: *Proceedings of the 28th international conference on Software engineering*. [S.l.: s.n.], 2006. p. 893–896.

FREEBODY, P.; ANDERSON, R. C. Effects of vocabulary difficulty, text cohesion, and schema availability on reading comprehension. *Reading research quarterly*, JSTOR, p. 277–294, 1983.

FREITAS, C.; CARVALHO, P.; OLIVEIRA, H. G.; MOTA, C.; SANTOS, D. Second harem: advancing the state of the art of named entity recognition in portuguese. In: EUROPEAN LANGUAGE RESOURCES ASSOCIATION. *quot; In Nicoletta Calzolari; Khalid Choukri; Bente Maegaard; Joseph Mariani; Jan Odijk; Stelios Piperidis; Mike*

Rosner; Daniel Tapias (ed) Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)(Valletta 17-23 May de 2010) European Language Resources Association. [S.l.], 2010.

FUKETA, M.; LEE, S.; TSUJI, T.; OKADA, M.; AOE, J.-i. A document classification method by using field association words. *Information Sciences*, Elsevier, v. 126, n. 1-4, p. 57–70, 2000.

GAMON, M. Linguistic correlates of style: authorship classification with deep linguistic analysis features. In: *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*. [S.l.: s.n.], 2004. p. 611–617.

GEURTS, P.; ERNST, D.; WEHENKEL, L. Extremely randomized trees. *Machine learning*, Springer, v. 63, n. 1, p. 3–42, 2006.

GINI, C. Variabilità e mutabilità. *Reprinted in Memorie di metodologica statistica (Ed. Pizetti E*, 1912.

GOEBEL, R.; CHANDER, A.; HOLZINGER, K.; LECUE, F.; AKATA, Z.; STUMPF, S.; KIESEBERG, P.; HOLZINGER, A. Explainable ai: the new 42? In: SPRINGER. *International cross-domain conference for machine learning and knowledge extraction*. [S.l.], 2018. p. 295–303.

GOLDBERG, Y. Neural network methods for natural language processing. *Synthesis lectures on human language technologies*, Morgan & Claypool Publishers, v. 10, n. 1, p. 1–309, 2017.

GONTIJO, S. *Livro de ouro da comunicação*. [S.l.]: Ediouro Publicações, 2004.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A.; BENGIO, Y. *Deep learning*. [S.l.]: MIT press Cambridge, 2016. v. 1.

GRAESSER, A. C.; MCNAMARA, D. S.; LOUWERSE, M. M.; CAI, Z. Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, Springer, v. 36, n. 2, p. 193–202, 2004.

GRANT, T. Quantifying evidence in forensic authorship analysis. *International Journal of Speech, Language & the Law*, v. 14, n. 1, 2007.

GRIEVE, J. Quantitative authorship attribution: An evaluation of techniques. *Literary and linguistic computing*, Oxford University Press, v. 22, n. 3, p. 251–270, 2007.

GRISHMAN, R.; SUNDHEIM, B. M. Message understanding conference-6: A brief history. In: *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*. [S.l.: s.n.], 1996.

GUIRAUD, P. *Les caractères statistiques du vocabulaire: essai de méthodologie*. [S.l.]: Presses universitaires de France, 1954.

HADJIDJ, R.; DEBBABI, M.; LOUNIS, H.; IQBAL, F.; SZPORER, A.; BENREDJEM, D. Towards an integrated e-mail forensic analysis framework. *digital investigation*, Elsevier, v. 5, n. 3-4, p. 124–137, 2009.

HALVANI, O.; GRANER, L.; REGEV, R. A step towards interpretable authorship verification. *arXiv preprint arXiv:2006.12418*, 2020.

HALVANI, O.; WINTER, C.; PFLUG, A. Authorship verification for different languages, genres and topics. *Digital Investigation*, Elsevier, v. 16, p. S33–S43, 2016.

HAMILL, K. A.; ZAMORA, A. The use of titles for automatic document classification. *Journal of the American Society for Information Science*, Wiley Online Library, v. 31, n. 6, p. 396–402, 1980.

HAND, D. J.; TILL, R. J. A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine learning*, Springer, v. 45, n. 2, p. 171–186, 2001.

HARTMANN, N.; FONSECA, E.; SHULBY, C.; TREVISO, M.; RODRIGUES, J.; ALUISIO, S. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. *arXiv preprint arXiv:1708.06025*, 2017.

HASTIE, T.; TIBSHIRANI, R. *Statistical Learning and Data Mining IV*. 2016. Disponível em: <ttps://github.com/ledell/sldm4-h2o>. Acesso em: Ago. 2020.

HERDAN, G. Quantitative linguistics. Butterworth, 1964.

HIRST, G.; FEIGUINA, O. Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, Oxford University Press, v. 22, n. 4, p. 405–417, 2007.

HOFMANN, T. Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, Springer, v. 42, n. 1, p. 177–196, 2001.

HOLMES, D. I. The evolution of stylometry in humanities scholarship. *Literary and linguistic computing*, Oxford University Press, v. 13, n. 3, p. 111–117, 1998.

HOLMES, D. I.; TWEEDIE, F. J. Forensic stylometry: A review of the cusum controversy. *Revue Informatique et Statistique dans les Sciences Humaines*, University of Liege Belgium, v. 31, n. 1, p. 19–47, 1995.

HONNIBAL, M.; MONTANI, I. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, v. 7, n. 1, p. 411–420, 2017.

HONORÉ, A. Some simple measures of richness of vocabulary. *Association for literary and linguistic computing bulletin*, v. 7, n. 2, p. 172–177, 1979.

HOU, R.; CHU-REN, H. Robust stylometric analysis and author attribution based on tones and rimes. *Natural Language Engineering*, Cambridge University Press, v. 26, n. 1, p. 49–71, 2020.

HU, C.; MENG, L.; SHI, W. Fuzzy clustering validity for spatial data. *Geo-spatial information science*, Taylor & Francis, v. 11, n. 3, p. 191–196, 2008.

HUANG, Z.; XU, W.; YU, K. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

JANG, B.; KIM, I.; KIM, J. W. Word2vec convolutional neural networks for classification of news articles and tweets. *PloS one*, Public Library of Science San Francisco, CA USA, v. 14, n. 8, p. e0220976, 2019.

JEBARA, T.; PENTLAND, A. Maximum conditional likelihood via bound maximization and the cem algorithm. In: CITESEER. *NIPS*. [S.l.], 1998. v. 1, n. 2, p. 7.

JOACHIMS, T. *Learning to classify text using support vector machines*. [S.l.]: Springer Science & Business Media, 2002. v. 668.

JOACHIMS, T. et al. Transductive inference for text classification using support vector machines. In: *Icml*. [S.l.: s.n.], 1999. v. 99, p. 200–209.

JOULIN, A.; GRAVE, E.; BOJANOWSKI, P.; DOUZE, M.; JÉGOU, H.; MIKOLOV, T. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016.

JUOLA, P. Future trends in authorship attribution. In: SPRINGER. *IFIP International Conference on Digital Forensics*. [S.l.], 2007. p. 119–132.

JUOLA, P. An overview of the traditional authorship attribution subtask. In: CITESEER. *CLEF (Online Working Notes/Labs/Workshop)*. [S.l.], 2012.

JURAFSKY, D.; MANNING, C. Natural language processing. *Instructor*, v. 212, n. 998, p. 3482, 2012.

KAMBHATLA, N.; LEEN, T. K. Fast nonlinear dimension reduction. In: IEEE. *IEEE International Conference on Neural Networks*. [S.l.], 1993. p. 1213–1218.

KARPATHY, A. The unreasonable effectiveness of recurrent neural networks. *Andrej Karpathy blog*, v. 21, p. 23, 2015.

KEŠELJ, V.; PENG, F.; CERCONE, N.; THOMAS, C. N-gram-based author profiles for authorship attribution. In: *Proceedings of the conference pacific association for computational linguistics, PACLING*. [S.l.: s.n.], 2003. v. 3, p. 255–264.

KESTEMONT, M.; MANJAVACAS, E.; MARKOV, I.; BEVENDORFF, J.; WIEGMANN, M.; STAMATATOS, E.; POTTHAST, M.; STEIN, B. Overview of the cross-domain authorship verification task at pan 2020. In: *CLEF*. [S.l.: s.n.], 2020.

KHOMYTSKA, I.; TESLYUK, V. The multifactor method applied for authorship attribution on the phonological level. In: *COLINS*. [S.l.: s.n.], 2020. p. 189–198.

KHONJI, M.; IRAQI, Y.; JONES, A. An evaluation of authorship attribution using random forests. In: IEEE. *2015 International Conference on Information and Communication Technology Research (ICTRC)*. [S.l.], 2015. p. 68–71.

KILGARRIFF, A. Thesauruses for natural language processing. In: IEEE. *International Conference on Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003*. [S.l.], 2003. p. 5–13.

KINCAID, J. P.; JR, R. P. F.; ROGERS, R. L.; CHISSOM, B. S. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. [S.l.], 1975.

KJELL, B. Authorship attribution of text samples using neural networks and bayesian classifiers. In: IEEE. *Proceedings of IEEE International Conference on Systems, Man and Cybernetics*. [S.l.], 1994. v. 2, p. 1660–1664.

KOSTADINOV, S. *How Recurrent Neural Networks work*. 2017. Disponível em: <https://towardsdatascience.com/learn-how-recurrent-neural-networks-work-84e975feaaf7>. Acesso em: Junho 2020.

KRAEMER, M. E. P. Avaliação da aprendizagem como construção do saber. INPEAU, 2005.

LAFFERTY, J.; MCCALLUM, A.; PEREIRA, F. C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

LAGUTINA, K.; LAGUTINA, N.; BOYCHUK, E.; VORONTSOVA, I.; SHLIAKHTINA, E.; BELYAEVA, O.; PARAMONOV, I.; DEMIDOV, P. A survey on stylometric text features. In: IEEE. *2019 25th Conference of Open Innovations Association (FRUCT)*. [S.l.], 2019. p. 184–195.

LAMPLE, G.; BALLESTEROS, M.; SUBRAMANIAN, S.; KAWAKAMI, K.; DYER, C. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*, 2016.

LARKEY, L. S.; BALLESTEROS, L.; CONNELL, M. E. Improving stemming for arabic information retrieval: light stemming and co-occurrence analysis. In: *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. [S.l.: s.n.], 2002. p. 275–282.

LEE, W. S.; LIU, B. Learning with positive and unlabeled examples using weighted logistic regression. In: *ICML*. [S.l.: s.n.], 2003. v. 3, p. 448–455.

LEMMATIZER, L. *NLX-Grupo de Fala e Linguagem Natural*. Disponível em: <http://lxcenter.di.fc.ul.pt/services/pt/LXServicesLemmatizerPT.html>. Acesso em: out. 2020.

LEVINE, Y. *Surveillance valley: The secret military history of the Internet*. [S.l.]: PublicAffairs, 2018.

LEX, E.; JUFFINGER, A.; GRANITZER, M. A comparison of stylometric and lexical features for web genre classification and emotion classification in blogs. In: IEEE. *2010 Workshops on Database and Expert Systems Applications*. [S.l.], 2010. p. 10–14.

LI, S.; YAN, Z.; WU, X.; LI, A.; ZHOU, B. A method of emotional analysis of movie based on convolution neural network and bi-directional lstm rnn. In: IEEE. *2017 IEEE Second International Conference on Data Science in Cyberspace (DSC)*. [S.l.], 2017. p. 156–161.

LIAW, A.; WIENER, M. et al. Classification and regression by randomforest. *R news*, v. 2, n. 3, p. 18–22, 2002.

LIDDY, E. D. Enhanced text retrieval using natural language processing. *Bulletin of the American Society for Information Science and Technology*, v. 24, n. 4, p. 14–16, 1998.

LUHN, H. P. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development*, IBM, v. 1, n. 4, p. 309–317, 1957.

LUNDBERG, S.; LEE, S.-I. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

LUONG, M.-T.; SOCHER, R.; MANNING, C. D. Better word representations with recursive neural networks for morphology. In: *Proceedings of the seventeenth conference on computational natural language learning*. [S.l.: s.n.], 2013. p. 104–113.

LUYCKX, K.; DAELEMANS, W. Authorship attribution and verification with many authors and limited data. In: *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. [S.l.: s.n.], 2008. p. 513–520.

MAATEN, L. Van der; HINTON, G. Visualizing data using t-sne. *Journal of machine learning research*, v. 9, n. 11, 2008.

MAITRA, P.; GHOSH, S.; DAS, D. Authorship verification-an approach based on random forest. *arXiv preprint arXiv:1607.08885*, 2016.

MARTINS, T. B.; GHIRALDELO, C. M.; NUNES, M. d. G. V.; JUNIOR, O. N. de O. *Readability formulas applied to textbooks in brazilian portuguese*. [S.l.]: Icmsc-Usp, 1996.

MASS, H.-D. Über den zusammenhang zwischen wortschatzumfang und länge eines textes. *Zeitschrift für Literaturwissenschaft und Linguistik*, Vandenhoeck und Ruprecht, v. 2, n. 8, p. 73, 1972.

MAURER, H. A.; KAPPE, F.; ZAKA, B. Plagiarism-a survey. *J. UCS*, v. 12, n. 8, p. 1050–1084, 2006.

MCCABE, S. D.; LIN, D.-Y.; LOVE, M. I. Consistency and overfitting of multi-omics methods on experimental data. *Briefings in bioinformatics*, Oxford University Press, v. 21, n. 4, p. 1277–1284, 2020.

MCCALLUM, A.; NIGAM, K. et al. A comparison of event models for naive bayes text classification. In: CITESEER. *AAAI-98 workshop on learning for text categorization*. [S.l.], 1998. v. 752, n. 1, p. 41–48.

MCMENAMIN, G. R. *Forensic linguistics: Advances in forensic stylistics*. [S.l.]: CRC press, 2002.

MCNAMARA, D. S.; CROSSLEY, S. A.; MCCARTHY, P. M. Linguistic features of writing quality. *Written communication*, Sage Publications Sage CA: Los Angeles, CA, v. 27, n. 1, p. 57–86, 2010.

MELLO, E. B. de S. Comunicação humana: importância da voz na educação. *Curriculum*, v. 12, n. 1, p. 45–54, 1973.

MELO, J. M. D. Comunicação, opinião, desenvolvimento. In: _____. [S.l.]: Editôra Vozes, 1975. v. 1, p. 209.

MENDENHALL, T. C. The characteristic curves of composition. *Science*, JSTOR, v. 9, n. 214, p. 237–249, 1887.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

MIKOLOV, T.; LE, Q. V.; SUTSKEVER, I. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013.

MILIDIÚ, R. L. *Anotação morfossintática a partir do contexto morfológico*. Tese (Doutorado) — PUC-Rio, 2016.

MILLER, G.; FELLBAUM, C.; KEGL, J.; MILLER, K. Wordnet: An electronic lexical reference system based on theories of lexical memory. *Revue quebecoise de linguistique*, Université du Québec à Montréal, v. 17, n. 2, p. 181–212, 1988.

MORTON, A. Q.; MICHAELSON, S. *The qsum plot*. [S.l.]: Edinburgh: University of Edinburgh, Dept. of Computer Science, 1990.

MOSTELLER, F.; WALLACE, D. L. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, Taylor & Francis Group, v. 58, n. 302, p. 275–309, 1963.

MOTA, C.; SANTOS, D. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*. [S.l.]: Linguateca, 2008.

NADEAU, D.; SEKINE, S. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, John Benjamins, v. 30, n. 1, p. 3–26, 2007.

NAMIUTI-TEMPONI, C.; COSTA, A. S. Reflexões sobre anotação sintática e ferramentas de busca-uso da linguagem xml para anotação sintática no corpus digital dovic. *Letras & Letras*, v. 30, n. 2, p. 82–103, 2014.

NASCIMENTO, P. d. A. *Aplicando Ensemble para classificação de textos curtos em português do Brasil*. Dissertação (Mestrado) — universidade Federal de Pernambuco, 2019.

NEAL, T.; SUNDARARAJAN, K.; FATIMA, A.; YAN, Y.; XIANG, Y.; WOODARD, D. Surveying stylometry techniques and applications. *ACM Computing Surveys (CSUR)*, ACM New York, NY, USA, v. 50, n. 6, p. 1–36, 2017.

NG, A. Y.; JORDAN, M. I. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2002. p. 841–848.

NGUYEN, M. Illustrated guide to lstm's and gru's: A step by step explanation. *Online] Towards Data Science*, 2018.

NIGAM, K.; LAFFERTY, J.; MCCALLUM, A. Using maximum entropy for text classification. In: STOCKHOLOM, SWEDEN. *IJCAI-99 workshop on machine learning for information filtering*. [S.l.], 1999. v. 1, n. 1, p. 61–67.

NUNES, R. A. da C. *História da educação na Idade Média*. [S.l.]: Editora Pedagógica e Universitária, 1979.

OLAH, C. Understanding lstm networks. 2015.

OLIVEIRA, H. G.; GOMES, P. Eco and onto. pt: a flexible approach for creating a portuguese wordnet automatically. *Language resources and evaluation*, Springer, v. 48, n. 2, p. 373–393, 2014.

PACHECO, M. L.; FERNANDES, K.; PORCO, A. Random forest with increased generalization: A universal background approach for authorship verification. In: *CLEF (Working Notes)*. [S.l.: s.n.], 2015.

PALKOVITS, S. A primer about machine learning in catalysis-a tutorial with code. *ChemCatChem*, Wiley Online Library, 2020.

PAVELEC, D.; OLIVEIRA, L. S.; JUSTINO, E. J.; BATISTA, L. V. Using conjunctions and adverbs for author verification. *J. UCS*, v. 14, n. 18, p. 2967–2981, 2008.

PELTZER, G.; SILVA, A. P. da; RIBEIRO, G. *Jornalismo iconográfico*. [S.l.: s.n.], 1991.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. [S.l.: s.n.], 2014. p. 1532–1543.

PERLES, J. B. Comunicação: conceitos, fundamentos e história. *Biblioteca on-line de Ciências da Comunicação*, 2007.

PIRES, A. R. O. Named entity extraction from portuguese web text. 2017.

PIROVANI, J. P. C. *CRF+ LG: uma abordagem híbrida para o reconhecimento de entidades nomeadas em português*. Tese (Doutorado) — Universidade Federal do Espírito Santo, Vitória (Brésil), 2019.

POPESCU, I.-I.; ALTMANN, G. Hapax legomena and language typology. *Journal of Quantitative Linguistics*, Taylor & Francis, v. 15, n. 4, p. 370–378, 2008.

POTTHAST, M.; RANGEL, F.; TSCHUGGNALL, M.; STAMATATOS, E.; ROSSO, P.; STEIN, B. Overview of pan'17. In: SPRINGER. *International Conference of the Cross-Language Evaluation Forum for European Languages*. [S.l.], 2017. p. 275–290.

PRANCKEVIČIUS, T.; MARCINKEVIČIUS, V. Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Baltic Journal of Modern Computing*, University of Latvia, v. 5, n. 2, p. 221, 2017.

QUINLAN, J. R. Induction of decision trees. *Machine learning*, Springer, v. 1, n. 1, p. 81–106, 1986.

QUINLAN, J. R. Combining instance-based and model-based learning. In: *Proceedings of the tenth international conference on machine learning*. [S.l.: s.n.], 1993. p. 236–243.

QUISPESARAVIA, A.; PEREZ, W.; CABEZUDO, M. S.; ALVA-MANCHEGO, F. Coh-metrix-esp: A complexity analysis tool for documents written in spanish. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. [S.l.: s.n.], 2016. p. 4694–4698.

RAINA, R.; SHEN, Y.; NG, A. Y.; MCCALLUM, A. Classification with hybrid generative/discriminative models. In: CITESEER. *NIPS*. [S.l.], 2003. v. 3, p. 545–552.

RAJKUMAR, K. V.; YESUBABU, A.; SUBRAHMANYAM, K. Fuzzy clustering and fuzzy c-means partition cluster analysis and validation studies on a subset of citescore dataset. *International Journal of Electrical & Computer Engineering (2088-8708)*, v. 9, n. 4, 2019.

RAMSHAW, L. A.; MARCUS, M. P. Text chunking using transformation-based learning. In: *Natural language processing using very large corpora*. [S.l.]: Springer, 1999. p. 157–176.

RASCHKA, S. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*, 2018.

RATNAPARKHI, A. A maximum entropy model for part-of-speech tagging. In: *Conference on empirical methods in natural language processing*. [S.l.: s.n.], 1996.

REIMERS, N.; GUREVYCH, I. Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. *arXiv preprint arXiv:1707.09861*, 2017.

ROCHA, R.; PIRES, H. da S. *Minidicionário enciclopédico escolar*. [S.l.]: Scipione, 1996.

ROJAS, R.; HASHAGEN, U. *The first computers: History and architectures*. [S.l.]: MIT press, 2002.

RYAN, J. *A History of the Internet and the Digital Future*. [S.l.]: Reaktion Books, 2010.

SALTON, G. Modern information retrieval. McGraw-Hill, 1983.

SAMPSON, G. Sistemas de escrita: tipologia, história e psicologia. *São Paulo: Ática*, 1996.

SANTOS, C. D.; ZADROZNY, B. Learning character-level representations for part-of-speech tagging. In: PMLR. *International Conference on Machine Learning*. [S.l.], 2014. p. 1818–1826.

SANTOS, D.; SECO, N.; CARDOSO, N.; VILELA, R. Harem: An advanced ner evaluation contest for portuguese. In: *quot; In Nicoletta Calzolari; Khalid Choukri; Aldo Gangemi; Bente Maegaard; Joseph Mariani; Jan Odjik; Daniel Tapias (ed) Proceedings of the 5 th International Conference on Language Resources and Evaluation (LREC'2006)(Genoa Italy 22-28 May 2006)*. [S.l.: s.n.], 2006.

SANTOS, D. C. V.-B. dos; FALCÃO, T. P. Acompanhamento de alunos em ambientes virtuais de aprendizagem baseado em sistemas tutores inteligentes. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2017. v. 28, n. 1, p. 1267.

SARKAR, D. *Traditional Methods for Text Data*. 2018. Disponível em: <https://towardsdatascience.com/understanding-feature-engineering-part-3-traditional-methods-for-text-data-f6f7d70acd41>. Acesso em: Set 2020.

SARWAR, R.; YU, C.; NUTANONG, S.; URAILERTPRASERT, N.; VANNABOOT, N.; RAKTHANMANON, T. A scalable framework for stylometric analysis of multi-author documents. In: SPRINGER. *International Conference on Database Systems for Advanced Applications*. [S.l.], 2018. p. 813–829.

SAVOY, J. Authorship attribution based on a probabilistic topic model. *Information Processing & Management*, Elsevier, v. 49, n. 1, p. 341–354, 2013.

SAVOY, J. Advanced models for stylometric applications. In: *Machine Learning Methods for Stylometry*. [S.l.]: Springer, 2020. p. 153–187.

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.

SCHAPIRE, R. E.; SINGER, Y.; SINGHAL, A. Boosting and rocchio applied to text filtering. In: *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. [S.l.: s.n.], 1998. p. 215–223.

SEROUSSI, Y.; ZUKERMAN, I.; BOHNERT, F. Authorship attribution with topic models. *Computational Linguistics*, MIT Press, v. 40, n. 2, p. 269–310, 2014.

SHAPLEY, L. S. A value for n-person games. *Contributions to the Theory of Games*, v. 2, n. 28, p. 307–317, 1953.

SHINYAMA, Y.; SEKINE, S. Named entity discovery using comparable news articles. In: *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*. [S.l.: s.n.], 2004. p. 848–853.

SHRESTHA, P.; SIERRA, S.; GONZÁLEZ, F. A.; MONTES, M.; ROSSO, P.; SOLORIO, T. Convolutional neural networks for authorship attribution of short texts. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. [S.l.: s.n.], 2017. p. 669–674.

SILVA, C. A. d. Coesão e coerência na produção escrita na língua estrangeira: uma investigação da influência da língua materna. 2006.

SILVA, D. d. C. Algoritmos de processamento da linguagem e síntese de voz com emoções aplicados a um conversor texto-fala baseado em hmm. *Doutorado, Programa de Engenharia Elétrica, Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia (COPPE/UFRJ), Rio de Janeiro*, 2011.

SINGH, S.; REMENYI, D. Plagiarism and ghostwriting: The rise in academic misconduct. *South African Journal of Science*, Academy of Science of South Africa, v. 112, n. 5-6, p. 1–7, 2016.

SINGH, V. K.; TIWARI, N.; GARG, S. Document clustering using k-means, heuristic k-means and fuzzy c-means. In: IEEE. *2011 International Conference on Computational Intelligence and Communication Networks*. [S.l.], 2011. p. 297–301.

SINWAR, D.; KAUSHIK, R. Study of euclidean and manhattan distance metrics using simple k-means clustering. *Int. J. Res. Appl. Sci. Eng. Technol*, v. 2, n. 5, p. 270–274, 2014.

SLONIM, N.; FRIEDMAN, N.; TISHBY, N. Unsupervised document classification using sequential information maximization. In: *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. [S.l.: s.n.], 2002. p. 129–136.

SOUCY, P.; MINEAU, G. W. A simple knn algorithm for text categorization. In: IEEE. *Proceedings 2001 IEEE international conference on data mining*. [S.l.], 2001. p. 647–648.

SRIURAI, W. Improving text categorization by using a topic model. *Advanced Computing*, Academy & Industry Research Collaboration Center (AIRCC), v. 2, n. 6, p. 21, 2011.

STAMATATOS, E. Author identification: Using text sampling to handle the class imbalance problem. *Information Processing & Management*, Elsevier, v. 44, n. 2, p. 790–799, 2008.

STAMATATOS, E. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, Wiley Online Library, v. 60, n. 3, p. 538–556, 2009.

STAMATATOS, E. On the robustness of authorship attribution based on character n-gram features. *Journal of Law and Policy*, v. 21, n. 2, p. 421–439, 2013.

STATISTA. *Number of smartphone users worldwide from 2016 to 2021*. 2020. Disponível em: <https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/>. Acesso em: maio 2020.

STAVNGAARD, M.; SØRENSEN, A.; LORENZEN, S.; HJULER, N.; ALSTRUP, S. Detecting ghostwriters in high schools. *arXiv preprint arXiv:1906.01635*, 2019.

STEIN, R. A.; JAQUES, P. A.; VALIATI, J. F. An analysis of hierarchical text classification using word embeddings. *Information Sciences*, Elsevier, v. 471, p. 216–232, 2019.

STENNER, A. J. Measuring reading comprehension with the lexile framework. ERIC, 1996.

STORK, D. G.; DUDA, R. O.; HART, P. E.; STORK, D. Pattern classification. *A Wiley-Interscience Publication*, 2001.

SUNDARARAJAN, M.; NAJMI, A. The many shapley values for model explanation. In: PMLR. *International Conference on Machine Learning*. [S.l.], 2020. p. 9269–9278.

TAYLOR, A.; MARCUS, M.; SANTORINI, B. The penn treebank: an overview. *Treebanks*, Springer, p. 5–22, 2003.

TEMPESTT, N.; SUNDARARAJAN, A. F. K.; YAN, Y.; XIANG, Y.; WOODARD, D. Surveying stylometry techniques and applications. *ACM Computing Surveys*, v. 50, n. 6, 2017.

THINSUNGNOENA, T.; KAOUNGKUB, N.; DURONGDUMRONCHAIB, P.; KERDPRASOPB, K.; KERDPRASOPB, N. The clustering validity with silhouette and sum of squared errors. *learning*, v. 3, n. 7, 2015.

TORREY, L.; SHAVLIK, J. Transfer learning. In: *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. [S.l.]: IGI global, 2010. p. 242–264.

TOUTANOVA, K.; MANNING, C. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In: *Proceedings of the 2000 Joint SIGDAT Conference EMNLP/VLC, 63-71, 2000*. [S.l.: s.n.], 2000.

TUKEY, J. W. Comparing individual means in the analysis of variance. *Biometrics*, JSTOR, p. 99–114, 1949.

TWEEDIE, F. J.; BAAYEN, R. H. How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities*, Springer, v. 32, n. 5, p. 323–352, 1998.

UCHENDU, A.; LE, T.; SHU, K.; LEE, D. Authorship attribution for neural text generation. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. [S.l.: s.n.], 2020. p. 8384–8395.

VAPNIK, V. N. An overview of statistical learning theory. *IEEE transactions on neural networks*, IEEE, v. 10, n. 5, p. 988–999, 1999.

VARELA, P. J. *UMA ABORDAGEM COMPUTACIONAL BASEADA EM ANÁLISE SINTÁTICA MULTILÍNGUE NA ATRIBUIÇÃO DA AUTORIA DE DOCUMENTOS DIGITAIS*. Tese (Doutorado) — Pontifícia Universidade Católica do Paraná, 2017.

VIANNA, H. M. Fundamentos de um programa de avaliação educacional. Líber Livro Editora, 2005.

VIRTANEN, P.; GOMMERS, R.; OLIPHANT, T. E.; HABERLAND, M.; REDDY, T.; COURNAPEAU, D.; BUROVSKI, E.; PETERSON, P.; WECKESSER, W.; BRIGHT, J.; van der Walt, S. J.; BRETT, M.; WILSON, J.; MILLMAN, K. J.; MAYOROV, N.; NELSON, A. R. J.; JONES, E.; KERN, R.; LARSON, E.; CAREY, C. J.; POLAT, İ.; FENG, Y.; MOORE, E. W.; VanderPlas, J.; LAXALDE, D.; PERKTOLD, J.; CIMRMAN, R.; HENRIKSEN, I.; QUINTERO, E. A.; HARRIS, C. R.; ARCHIBALD, A. M.; RIBEIRO, A. H.; PEDREGOSA, F.; van Mulbregt, P.; SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, v. 17, p. 261–272, 2020.

WANG, Y.; WANG, Z.-O. A fast knn algorithm for text categorization. In: IEEE. *2007 International Conference on Machine Learning and Cybernetics*. [S.l.], 2007. v. 6, p. 3436–3441.

WARNER, J.; SEXAUER, J. scikit-fuzzy, twmeggs, alexandre ms, aishwarya unnikrishnan,... himanshu mishra.(2017, october 6). *JDWarner/scikit-fuzzy: Scikit-Fuzzy 0.3*, v. 1.

WEISBERG, S. Yeo-johnson power transformations. *Department of Applied Statistics, University of Minnesota. Retrieved June*, v. 1, p. 2003, 2001.

WILAMOWSKI, B. M.; JAEGER, R. C. Implementation of rbf type networks by mlp networks. In: IEEE. *Proceedings of International Conference on Neural Networks (ICNN'96)*. [S.l.], 1996. v. 3, p. 1670–1675.

WORLDOMETERS. *Current world population*. 2020. Disponível em: <https://www.worldometers.info/world-population/>. Acesso em: maio 2020.

YANG, M.; CHEN, X.; TU, W.; LU, Z.; ZHU, J.; QU, Q. A topic drift model for authorship attribution. *Neurocomputing*, Elsevier, v. 273, p. 133–140, 2018.

YONAMINE, F. S.; SPECIA, L.; CARVALHO, V.; NICOLETTI, M. Aprendizado não supervisionado em domínios fuzzy–algoritmo fuzzy c-means. *São Carlos: UFSCAR*, 2002.

YOU, R.; DAI, S.; ZHANG, Z.; MAMITSUKA, H.; ZHU, S. Attentionxml: Extreme multi-label text classification with multi-label attention based recurrent neural networks. *arXiv preprint arXiv:1811.01727*, v. 137, p. 138–187, 2018.

YOUNAS, J.; FRITSCH, S.; PIRKL, G.; AHMED, S.; MALIK, M. I.; SHAFAIT, F.; LUKOWICZ, P. What am i writing: Classification of on-line handwritten sequences. In: *Intelligent Environments (Workshops)*. [S.l.: s.n.], 2018. p. 417–426.

YULE, C. U. *The statistical study of literary vocabulary*. [S.l.]: Cambridge University Press, 2014.

YULE, G. *The statistical study of literary vocabulary. Cambridge, Cambridge [Eng.]*. [S.l.]: University Press. Journal of the Royal Statistical Society, 1944.

ZHANG, M.; ZHANG, W.; SICOTTE, H.; YANG, P. A new validity measure for a correlation-based fuzzy c-means clustering algorithm. In: IEEE. *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. [S.l.], 2009. p. 3865–3868.

ZHANG, Y.; WALLACE, B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*, 2015.

ZHOU, C.; SUN, C.; LIU, Z.; LAU, F. A c-lstm neural network for text classification. *arXiv preprint arXiv:1511.08630*, 2015.

ZHOU, P.; QI, Z.; ZHENG, S.; XU, J.; BAO, H.; XU, B. Text classification improved by integrating bidirectional lstm with two-dimensional max pooling. *arXiv preprint arXiv:1611.06639*, 2016.

ZOU, W. Y.; SOCHER, R.; CER, D.; MANNING, C. D. Bilingual word embeddings for phrase-based machine translation. In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. [S.l.: s.n.], 2013. p. 1393–1398.