Pós-Graduação em Ciência da Computação

Luís Fred Gonçalves de Sousa

Uso de aprendizado supervisionado multivisão para atribuição automática de autoria de textos

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
http://cin.ufpe.br/~posgraduacao

Recife
2020

Luís Fred Gonçalves de Sousa

Uso de aprendizado supervisionado multivisão para atribuição automática de autoria de textos

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Área de Concentração: inteligência computacional
Orientador: Renato Vimieiro

Recife
2020

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

S725u Sousa, Luís Fred Gonçalves de
Uso de aprendizado supervisionado multivisão para atribuição automática de autoria de textos / Luís Fred Gonçalves de Sousa. – 2020.
65 f.: il., fig., tab.

Orientador: Renato Vimieiro.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2020.
Inclui referências.

1. Inteligência computacional. 2. Aprendizagem de máquina. I. Vimieiro, Renato (orientador). II. Título.

006.31 CDD (23. ed.) UFPE - CCEN 2020 - 182

Luís Fred Gonçalves de Sousa

"**Uso de aprendizado supervisionado multivisão para atribuição automática de autoria de textos**"

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.

Aprovado em: 20/08/2020.

Orientador: Renato Vimieiro

BANCA EXAMINADORA

Prof. Dr. George Darmiton da Cunha Cavalcanti
Centro de Informática / UFPE

Prof. Dr. Anisio Mendes Lacerda
Departamento de Ciência da Computação / UFMG

Prof. Dr. Renato Vimieiro
Departamento de Ciência da Computação / UFMG
(Orientador)

Em memória de Maria Juscely, que teria ficado feliz com essa conquista.

AGRADECIMENTOS

Ao meu pai, pelo apoio incondicional que me prestou durante este percurso. Grato também pelo apoio dos meus irmãos. A minha esposa, por toda a paciência e compreensão que teve durante esta jornada. Ao professor Renato Vimieiro (Orientador), por sua zelosa orientação, conselhos, e por ter me dado a visão que eu precisava e não tinha. Também sou grato aos demais professores do Cin pelos ensinamentos úteis na condução deste estudo. A CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pelo incentivo financeiro fornecido durante os meses finais deste trabalho. Por fim, a todos os demais, que de alguma forma me ajudaram a tornar esse projeto possível.

RESUMO

Atribuição de autoria é o problema de identificar o autor de um ou mais textos com base no estilo de escrita do autor. Normalmente, a tarefa assume que o estilo de escrita dos autores conserva traços que são inacessíveis à manipulação consciente. Dessarte, tal poderia ser seguramente usado para identificar o autor de um texto. Os pesquisadores têm investigado um grande número de características textuais com o objetivo de validar a habilidade destas de revelar mecanismos subconscientes de variação de linguagem, os quais podem, consequentemente, refletir autoria. Muitos marcadores de estilo autorial já foram propostos na literatura. Não obstante, permanece a falta de consenso sobre qual é o melhor para representar as escolhas dos autores. Esta dissertação assume um ponto de vista neutro na disputa pelo melhor conjunto de características de texto capaz de representar estilos de escrita. No lugar disso, é investigado como diferentes fontes de informação podem relevar diferentes aspectos do estilo de um autor, complementando-se, assim, para aprimorar o processo geral de atribuição de autoria. Com esse propósito, o problema de atribuição de autoria é modelado nesse estudo como uma tarefa de aprendizado de máquina multivisão. A eficácia da abordagem proposta é avaliada em quatro conjuntos de dados com número variado de autores e obras. A performance do método é comparada ao estado da arte em abordagens de aprendizado de máquina para atribuição de autoria. No decorrer do estudo, foi analisado como o método multivisão aprimora as abordagens tradicionais que usam uma única fonte de informação para atribuir autoria, os quais foram chamados de métodos univisão. Os resultados confirmam a relevância de algumas características individuais de texto para a tarefa, mas também mostram como essas características se complementam com outros tipos de recursos linguísticos para melhorar a consistência e a precisão da atribuição de autoria. Ademais, foi verificado que os classificadores treinados com dados multivisão consistentemente concordam sobre os rótulos verdadeiros dos textos. O estudo ainda discute como essas melhorias, tanto na acurácia quanto na concordância de classificação, são benéficas para linguistas e outros especialistas.

Palavras-chaves: Atribuição de autoria. Categorização de textos. Linguística computacional. Aprendizagem de máquina supervisionada. Aprendizagem multivisão. Aprendizagem multimodal.

ABSTRACT

Authorship attribution is the problem of identifying the author of texts based on the author's writing style. Usually, in this task it is assumed that the authors' writing style contains traits inaccessible to conscious manipulation and can thus be safely used to identify the author of a text. Researchers have investigated a large number of text characteristics to assess their ability to reveal subconscious mechanisms of language variation, which may consequently reflect authorship. Several author style markers have been proposed in the literature, nevertheless, there is still no consensus on which best represent the choices of authors. This work assume an agnostic viewpoint on the dispute for the best set of features that represent an author's writing style. Rather is investigated how these different sources of information may unveil different aspects of an author's style, thus complementing each other to improve the overall process of authorship attribution. For this purpose, the problem of authorship attribution is modeled as a multi-view/multimodal supervised machine learning task. The effectiveness of the proposal is assessed in four corpora with different number of authors. The performance of the proposal is compared to the state-of-the-art supervised machine learning approaches for authorship attribution. In the study is thoroughly analyzed how the multi-view approach improves on traditional methods that use a single source of data (single-view) for assigning authorship. The results confirms the relevance of some features for the task, but also show how they are complemented with other types of features to improve both on consistency and accuracy. Moreover, it was verified that classifiers trained with multi-view data consistently agree on the labels of texts. The study discusses how these improvements in both accuracy and agreement are beneficial for linguists and domain specialists.

Keywords: Authorship attribution. Text categorization. Computational linguistics. Supervised machine learning. Multi-view learning. Multimodal learning.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AA	Atribuição de Autoria
AM	Aprendizagem de Máquina
AMV	Aprendizagem de Máquina Multivisão
DRF	Dissimilaridade Random Forest
IIQ	Intervalo Interquartil
MKL	Multiple kernel learning
PLN	Processamento de Linguagem Natural
SVM	Support Vector Machines
TTR	Type Token Ratio

SUMÁRIO

1 INTRODUÇÃO

Atribuição de Autoria (AA) é a tarefa que trata de identificar o autor de um dado texto em meio a um conjunto de autores candidatos, baseando-se no pressuposto de que cada autor tem seu próprio estilo de escrita que funciona como uma *impressão digital*. Isso é possível porque vários recursos mensuráveis no texto escrito por um determinado autor permanecem-se inalterados ao longo do tempo (EBRAHIMPOUR et al., 2013).

O estilo de escrita representa precisamente as escolhas linguísticas feitas por um autor, as quais diferenciam sua escrita em relação à escrita de outros autores. É pertinente mencionar que o problema de atribuição de autoria possui estreita relação tanto com o estudo dos estilos de escrita individuais, quanto com o campo de categorização de textos. Trata-se de uma área científica que tem se desenvolvido de maneira substancial concomitante aos avanços nas pesquisas em áreas como Aprendizagem de Máquina (AM), recuperação de informação e Processamento de Linguagem Natural (PLN), das quais tira proveito. Inicialmente, a maior parte dos pesquisadores na área de AA compunha-se exclusivamente de linguistas que concentraram esforços na descoberta do conjunto ideal de características de textos capazes de representar o estilo de um autor. Foi o trabalho pioneiro de Mosteller e Wallace (1964) que marcou a transição para uma nova era em AA assistida por computador.

O estudo de métodos estatísticos e de AM mais sofisticados, combinados com o crescimento acentuado de diversos tipos de atividades em meio online, possibilitaram o surgimento de novas aplicações de AA em domínios como detecção de plágio, detecção de ameaças em mensagens eletrônicas (ZHENG et al., 2006), e investigação de crimes cibernéticos (CHEN et al., 2011). Outrossim, a intensa pesquisa no campo de AM focada no desenvolvimento de métodos mais robustos para classificação de documentos resultaram em abordagens extremamente acuradas que beneficiaram, em grande medida, os métodos de AM voltados para atribuição de autoria. Parte expressiva desses métodos consiste em treinar um modelo de aprendizagem do tipo classificador com documentos pré-categorizados para depois categorizar documentos desconhecidos.

A criação desses modelos classificadores é baseada na extração de características textuais a partir do conteúdo dos documentos disponíveis, tal como ocorre em atribuição de autoria. A título de exemplo, alguns trabalhos exploram a frequência de palavras em uma tentativa de treinar um modelo capaz de categorizar documentos. Essa abordagem, felizmente, pode ser facilmente generalizada para o domínio de AA, atribuindo textos a autores no lugar de categorias (STAMATATOS, 2009; JOCKERS; WITTEN, 2010). De fato, conforme discutido por Stamatatos (2009), toda uma gama de métodos, os quais são conhecidos como *abordagens baseadas em instâncias*, operam de tal maneira.

Apesar do significativo avanço no desenvolvimento dessas abordagens, a questão sobre

quais características textuais melhor representam o estilo de escrita de um autor permanece aberta. Há pelo menos três categorias de características textuais que são usadas para representar estilos de escrita: léxica; sintática e semântica. Cada uma dessas categorias modela algum aspecto do estilo de escrita de um autor. Características léxicas podem indicar, por exemplo, a escolha particular do vocabulário de um indivíduo. Características textuais sintáticas são capazes de revelar escolhas estruturais ou gramaticais, enquanto que as semânticas podem revelar os tópicos de preferência.

Sari, Stevenson e Vlachos (2018) estudaram, recentemente, como os diferentes tipos de características textuais afetam a atribuição de autoria em conjuntos de dados diferentes. Eles verificaram que certos tipos de características são mais apropriadas quando se trata de tópicos específicos, enquanto outros são mais relevantes em cenários mais diversos. Como exemplo, consideremos os conjuntos de dados onde possa haver uma variação mais acentuada de tópicos, tais como aqueles que são construídos a partir de textos jornalísticos. Em tais cenários, características textuais como n-gramas de palavras exercem mais influência na percepção dos traços de um autor.

Esta dissertação assume uma visão agnóstica das características textuais usadas para representar o estilo de escrita de um autor. A intenção, portanto, não é comparar quais tipos de características textuais são mais oportunas que outras para um contexto em particular. No lugar disso, é levado em conta o fato de que diferentes tipos de características são capazes de contribuir para diferentes perspectivas do estilo de um autor, o que ajudaria a aprimorar o processo de atribuição de autoria. Neste trabalho, cada uma dessas características textuais de diferentes tipos é tratada como uma visão do texto — a literatura relevante em inglês refere-se a estas como *views*. A hipótese é que a integração das informações contidas nessas múltiplas visões dos dados beneficia a criação de modelos mais consistentes e acurados para a tarefa de atribuição de autoria. Portanto, nesta dissertação, a tarefa de atribuição de autoria é modelada como um problema de AM multivisão supervisionado.

A aprendizagem de máquina multivisão tem recebido considerável atenção de pesquisadores em anos recentes (XU; TAO; XU, 2013; ZHAO et al., 2017; BALTRUŠAITIS; AHUJA; MORENCY, 2019). Trata-se de um campo de estudo em crescimento, cujo objetivo é obter modelos treinados a partir de múltiplas fontes de dados. Tais métodos tentam explorar representações paralelas — e distintas — dos dados para aprimorar a performance de generalização dos modelos. Como exemplo, considere que uma página da web pode ser representada tanto pelo seu conteúdo em texto quanto por seus links e imagens. Por certo, em muitas situações, o objeto de estudo pode ser representado a partir de diferentes perspectivas dos dados. Tal se aplica ao caso dos estudos em atribuição de autoria que discutimos nos parágrafos anteriores.

O sucesso dos métodos de aprendizagem multivisão em outros contextos, tais como medicina (CAO et al., 2019; CURTIS et al., 2012), e mesmo em estudos de atribuição de

autoria com aprendizagem não supervisionada (DUQUE; CARVALHO; VIMIEIRO, 2019), é o que motivou o método proposto nesta dissertação.

Ademais, é oportuno mencionar que os resultados preliminares desse estudo foram submetidos à apreciação de uma revista científica relevante, para que possam ser devidamente publicados.

1.1 OBJETIVOS

A presente dissertação tem como objetivo apresentar uma nova abordagem para atribuição de autoria baseada em Aprendizagem de Máquina Multivisão (AMV). A hipótese levantada neste trabalho é a de que a aprendizagem multivisão oferece modelos mais acurados e consistentes. As seguintes questões de pesquisa são de interesse em nosso estudo:

QP1. Podemos aprimorar a acurácia geral dos modelos de atribuição de autoria ao utilizar dados multivisão?

QP2. Os dados multivisão permitem que os classificadores concordem mais entre si sobre a autoria dos textos? Em outros termos, podemos obter modelos mais consistentes se usarmos dados multivisão?

QP3. Podemos identificar as visões dos dados que melhoram, ou pioram, a performance geral dos modelos?

É conveniente ressaltar que quanto mais os diferentes modelos de atribuição de autoria predizem um mesmo autor para um texto, mais consistentes eles são na tarefa de atribuição. É para tal direção que apontamos ao tratar de consistência em atribuição de autoria neste trabalho.

Ademais, o propósito com a primeira pergunta é investigar quando a abordagem multivisão resulta em modelos que são mais acurados comparados com estudos tradicionais que aplicam uma única visão dos dados. Mais especificamente, o interesse é investigar quando os múltiplos tipos de características dos dados se complementam, tornando o estilo de escrita latente do autor mais entendível pelos classificadores, de maneira que estes o possam explorar mais facilmente. A recompensa mais direta disso seria uma melhoria na performance de generalização. Com a segunda pergunta, o objetivo é isentar linguistas e outros especialistas do encargo envolvido na escolha do melhor classificador para atribuição de autoria, posto que tal tarefa frequentemente requer diligência e remete a um esforço mais exaustivo.

O vasto número de algoritmos de aprendizagem de diferentes tipos, cada um com seu distinto conjunto de hiperparâmetros a ser ajustado, dificulta em grande medida seu emprego em tarefas específicas por parte dos especialistas em domínio. Assim, nesse trabalho, deseja-se investigar se a performance em AA poderia se tornar independente, ou menos dependente quanto possível, da escolha do classificador.

É importante observar que há um estreito relacionamento entre as duas primeiras questões. É mais desejável termos métodos consistentes e que sejam, ao mesmo tempo, capazes de oferecer um bom desempenho. Além disso, é apropriado mencionar que nem sempre o sucesso com um único classificador garante que este seja a melhor escolha em um contexto específico. Portanto, o principal objetivo deste trabalho é demonstrar que estudos em atribuição de autoria podem se beneficiar da modelagem multivisão, de maneira que a preocupação sobre qual tipo de características de textos leva ao melhor modelo classificador de autoria seja minimizada.

Por fim, nosso objetivo com a terceira pergunta é analisar a complementariedade das diferentes visões dos dados. Para melhor ilustrar o problema, imaginemos que ao combinar duas visões dos mesmos dados o desempenho de um modelo específico melhora em relação ao uso individual de ambas. Isso poderia indicar, portanto, que essas duas visões se complementam. Assim, pretendemos identificar quais conjuntos melhoram, ou prioram, o desempenho de algumas visões ao estudar sua complementariedade.

1.2 ESTRUTURA DO TRABALHO

Esta dissertação foi escrita em seis capítulos, incluindo esta introdução.

- O Capítulo 2 apresenta um referencial teórico sobre atribuição de autoria mencionando métodos relacionados que são importantes para este trabalho. Menciona, também, os marcadores de estilo de escrita frequentemente tratados na literatura, e como alguns pesquisadores os classificam. Esse mesmo capítulo também referencia métodos relevantes propostos na literatura que combinam várias fontes de informação com o objetivo de melhorar a performance dos modelos na tarefa de atribuição de autoria, bem como alguns métodos de fusão dessas diferentes fontes de dados.

- O Capítulo 3 apresenta a metodologia da abordagem proposta; conceitualiza um conjunto de dados multivisão; mostra como as características de textos foram categorizadas nesse estudo e de que maneira as diferentes fontes de dados foram combinadas para formar o modelo de atribuição de autoria.

- O Capítulo 4 apresenta a metodologia seguinda na execução dos experimentos; descreve os conjuntos de dados utilizados; discute os resultados experimentais obtidos em cada corpus, comparando-os com as abordagens de base; e, por fim, verifica quando o método multivisão resulta em melhor acurácia na tarefa de atribuição de autoria, bem como torna os classificadores mais consistentes na tarefa de atribuição.

- O Capítulo 5 conclui este trabalho e indica futuras direções a seguir.

2 REVISÃO BIBLIOGRÁFICA

A revisão bibliográfica abaixo está dividida em duas seções. Na primeira, são mencionados os trabalhos relevantes na área de atribuição de autoria; em seguida, apresentamos a revisão de literatura referente à aprendizagem com dados multivisão.

2.1 ATRIBUIÇÃO DE AUTORIA

Os primeiros estudos em atribuição de autoria remontam ao século XIX tendo, portanto, uma longa história. Esses estudos iniciais contavam com especialistas focados em identificar características textuais que melhor pudessem auxiliar na quantificação do estilo de escrita de um autor. Foi o trabalho pioneiro de Mosteller e Wallace (1964) — considerado o mais influente trabalho em atribuição de autoria — que marcou a transição para uma nova era em estudos de AA, na qual passou-se a utilizar métodos computacionais e estatística para identificar os contrastes nos estilos de escrita de diferentes autores em análise. Em particular, Mosteller e Wallace (1964) usaram análise estatística Bayesiana para investigar a autoria de uma série de artigos de jornal impressos em Nova York no século XVII, os quais defendiam a ratificação da constituição dos Estados Unidos. Esses artigos são amplamente conhecidos como *Federalist Papers*.

A despeito dessa transição, até o início da última década, muitos trabalhos em AA se limitaram em propor novos tipos de características textuais para identificar estilos (STAMATATOS, 2009). Esses marcadores de estilo foram classificados por Stamatatos (2009) em cinco categorias: *léxica*; *caractere*; *sintática*; *semântica*; e *específico de aplicação*.

A categoria léxica abrange frequências de palavras, riqueza de vocabulário, tamanhos de palavras e frases, bem como n-gramas de palavras. A *riqueza de vocabulário*, para exemplificar, trata de quantificar a diversidade do vocabulário de um texto. Um exemplo típico é a *taxa de tipo de token*, a qual relaciona o tamanho do vocabulário, em termos de *tokens* únicos, com o número total de *tokens* no texto. Outro exemplo é o número palavras únicas presentes no texto.

Características textuais da categoria caractere são normalmente associadas com *n-gramas* em nível caractere, dígitos, letras e *emojis*. Por exemplo, o uso de frequências de n-gramas em nível de caractere é capaz de capturar certas sutilezas do estilo de um autor, tais como erros de ortografia e uso anormal de pontuação — comuns em textos de e-mails e outros tipos de mensagens em meio eletrônico. Sendo uma característica de texto robusta contra ruídos, ela não é radicalmente afetada em tais circunstâncias.

Características sintáticas estão ligadas aos aspectos estruturais do estilo de escrita do autor. É predominante nesta categoria a etiquetagem de classes de palavras - também conhecidas como *Part-of-speech tags* (POS-tags). Além disso, as *palavras funcionais* tam-

bém desempenham um papel nesta categoria, dado que são normalmente encontradas em certas estruturas sintáticas (STAMATATOS, 2009). Palavras funcionais costumam ser empregadas de maneira inconsciente pelos autores, sendo capazes de capturar escolhas estilísticas independentemente do tópico abordado no texto.

Características semânticas incluem sinônimos e os chamados *Word Embeddings*, que são representações de palavras incorporadas em espaços vetoriais. Finalmente, as características textuais dependentes de aplicação são aquelas que só podem ser utilizadas em contextos específicos. Fazem parte desta categoria aquelas características extraídas a partir da estrutura dos documentos, tais como metadados e formatação.

Essa categorização das características textuais sugere que cada tipo de característica é capaz de manifestar diferentes aspectos do estilo de escrita de um autor. Ainda assim, há vários estudos disponíveis na literatura relevante comparando as melhores características textuais para atribuição de autoria (GRIEVE, 2007; SARI; STEVENSON; VLACHOS, 2018). Por exemplo, Sari, Stevenson e Vlachos (2018) discutiram a utilidade de alguns recursos textuais tanto em cenários limitados a tópicos específicos quanto em cenários que combinam diferentes tópicos.

Stamatatos (2009) também distinguiu os métodos de atribuição de autoria com base na maneira como são tratadas as amostras de textos, se cumulativamente por autor (no que chamou de *abordagem baseada em perfil*), ou de maneira individual (que denominou *abordagem baseada em instâncias*). Na *abordagem baseada em perfil*, frequentemente se conjuga em um único arquivo todos os textos disponíveis de um autor, de modo a representar seu estilo de escrita a partir daquele ponto. De tal modo, é composto o perfil do autor. Já a *abordagem baseada em instâncias*[1], por sua vez, necessita que múltiplas amostras de texto por autor sejam consideradas individualmente para desenvolver um modelo de atribuição preciso. Posto de tal modo, cada texto é individualmente tratado como uma instância do estilo do autor. Por fim, um terceiro método de atribuição de autoria foi mencionado naquele trabalho: *abordagem híbrida*. Na abordagem híbrida, são combinados os elementos das duas anteriores.

Esta dissertação considera a possibilidade de não haver um tipo de características de texto mais eficaz para a atribuição de autoria. No lugar disso, é levado em conta que os diferentes tipos de características textuais, juntos, podem melhorar o desempenho geral dos métodos de atribuição. De fato, é possível que cada uma dessas categorias de características textuais apontadas por Stamatatos (2009) se complementem, refletindo diferentes perspectivas do estilo de escrita de um autor.

Esta abordagem foi preliminarmente investigada por Duque, Carvalho e Vimieiro (2019) em um estudo envolvendo aprendizagem não supervisionada. Eles discutiram como o uso destas diferentes perspectivas dos dados melhoram a identificação de grupos de textos de um dado autor. Apesar de não tentarem atribuir a autoria dos textos explici-

[1] Não confundir com *aprendizagem baseada em instâncias*

tamente, eles mostraram que os textos de um mesmo autor, bem como de autores com perfis compatíveis, tendem a se agrupar. Trata-se, portanto, de uma evidência em favor da ideia de que o esquema de múltiplas visões dos dados pode refletir melhor os estilos de escrita dos autores.

Outros trabalhos investigam como subconjuntos de frequências de palavras podem ser combinados em um problema de aprendizado envolvendo um comitê de classificadores para aprimorar a eficácia da atribuição de autoria (STAMATATOS, 2006). É válido afirmar que esse estudo reflete uma tentativa inicial de combinar a informação complementar contida em diferentes visões dos documentos com o objetivo de melhorar a classificação. E, de fato, o aprendizado supervisionado de múltiplas visões pode ser implementado como um sistema de múltiplos classificadores (comitê) (ZHAO et al., 2017). Uma diferença importante entre o trabalho de Stamatatos (2006) e o nosso é que aquele considerou apenas frequências de palavras, enquanto esta dissertação considera, também, outros tipos de características de texto.

Fourkioti, Symeonidis e Arampatzis (2019) também exploraram uma combinação de múltiplas visões para identificar estilos de escrita. Eles propuseram uma combinação de recursos léxicos e sintáticos em uma abordagem baseada em perfil para identificar os autores de textos curtos, tais como *tweets* e resenhas de filmes. Naquele trabalho, todos os textos de um autor foram primeiramente combinados para compor seu perfil, antes de treinar um modelo. Em seguida, para cada visão dos dados, um modelo de linguagem foi treinado. Por fim, combinaram linearmente as características textuais por meio da combinação linear dos modelos de linguagem treinados. Os experimentos conduzidos no trabalho mostraram que, contradizendo estudos anteriores, recursos sintáticos são relevantes para AA em cenários que envolvem textos curtos.

Outra constatação mencionada naquela pesquisa é que a combinação de características textuais, de fato, produz resultados melhores do que aqueles que seriam obtidos com cada uma dessas características usadas individualmente.

É apropriado dizer que nosso estudo adota esses princípios, não obstante a abordagem acima mencionada possa não ser muito adequada para textos longos. Isso se deve ao fato de a linguagem e estrutura do texto usadas nas redes sociais e outras plataformas online serem muito peculiares a tais contextos. Portanto, é justificável que nessas situações a estrutura sintática e os recursos linguísticos baseados em caracteres se sobressaiam frente aos recursos mais tradicionais frequentemente usados na literatura, a exemplo de frequências de palavras. Por outro lado, a abordagem proposta nesta dissertação pode não alcançar um bom desempenho em textos curtos, considerando que uma ampla variedade de tipos de características textuais é explorada.

Relativamente à escolha do método de atribuição, tanto os métodos supervisionados quanto os não supervisionados são aplicados na literatura relevante. Abordagens não supervisionadas tendem a procurar um modelo que melhor possa representar o perfil de

um autor. Por exemplo, Arefin et al. (2014) propuseram aplicar teoria da informação para agrupar textos de acordo com frequências de palavras. Eles visualizaram documentos como distribuições de probabilidades. Em seguida, compararam os documentos baseando-se no valor da divergência *Jensen-Shannon*. Eles observaram que os textos de um mesmo autor ou de autores com perfis similares se agrupavam. Textos disputados, ou sem um autor determinado, se mantiveram em grupos onde havia apenas um único autor. De acordo com os autores da pesquisa, isso poderia indicar uma autoria plausível para os textos.

Kocher e Savoy (2017) também propuseram um método baseado na similaridade de documentos para atribuir autoria. Contudo, eles usaram representações de linguagem distribuídas. Especificamente, eles usaram *word-embeddings* para representar os documentos. Em seguida, usaram essas representações vetoriais para comparar os documentos e atribuir autoria. Duas abordagens foram propostas: uma delas utilizou apenas um subconjunto de palavras, o qual pode ser composto pelas palavras mais frequentes ou por palavras funcionais, as quais são normalmente utilizadas na literatura; e uma segunda abordagem na qual todas as palavras foram levadas em consideração. Outro trabalho relevante é o de Layton, Watters e Dazeley (2013), no qual propuseram uma abordagem não supervisionada baseada em n-gramas. Eles mostraram que o método proposto é capaz de encontrar grupos altamente correlacionados com a verdadeira autoria dos textos.

As abordagens acima citadas possuem algo em comum. Cada uma delas considerou um tipo específico de característica textual para representar textos e negligenciou a possível contribuição de outros tipos. Duque, Carvalho e Vimieiro (2019) foram os primeiros a propor o uso de múltiplos tipos de características textuais em uma abordagem multivisão não supervisionada. Contudo, o trabalho deles se limitou em perfilar os autores.

Vários métodos supervisionados também têm sido usados para atribuição de autoria. Da mesma forma que acontece para os tipos de características de texto, não há um consenso sobre qual desses métodos é o melhor (JOCKERS; WITTEN, 2010). Para citar exemplos da falta de anuência neste caso, mencionemos o trabalho de Juola, Sofko e Brennan (2006), no qual foram discutidos os resultados de um concurso de atribuição de autoria que foi realizado em 2004. Foi verificado que *Support Vector Machines (SVM)* com *kernel* linear eram a abordagem mais eficaz. De fato, SVM é um algoritmo largamente adotado como método de base em muitos estudos na literatura em AA (DIEDERICH et al., 2003; KOPPEL; SCHLER; BONCHEK-DOKOW, 2007; STAMATATOS, 2009; JOCKERS; WITTEN, 2010). A falta de um consenso é mais patente com a comparação apresentada por Jockers e Witten (2010), onde o algoritmo SVM foi considerado o menos eficaz.

Juola, Sofko e Brennan (2006) comenta que a escolha do classificador não é uma tarefa trivial. Isso vale, principalmente, para linguistas ou outros especialistas em domínio. A presente dissertação está de acordo com tal argumento e assume que esta falta de consenso se deve à escolha de um único tipo de característica textual para representar o estilo de escrita do autor. A abordagem multivisão, por outro lado, ao ser capaz de oferecer

múltiplas perspectivas do estilo de um autor, poderia resultar em classificadores mais consistentes, os quais concordam mais frequentemente uns com os outros a respeito da autoria de um texto.

Por fim, com respeito à AMV, trata-se de uma área de pesquisa em franco crescimento, a qual considera o problema de AM a partir de dados representados por múltiplos conjuntos de características (chamados de visões ou, às vezes, multimodalidades em alguns estudos) (SUN, 2013). Tratamos desse assunto na próxima seção.

2.2 APRENDIZAGEM MULTIVISÃO

De acordo com Zhao et al. (2017), o principal objetivo da AMV é ajustar uma função capaz de modelar cada visão e otimizar todas elas conjuntamente, com o objetivo de melhorar a performance de generalização do modelo. A ideia principal é que os diferentes tipos de representação dos dados podem conter informações complementares as quais, se combinadas corretamente, podem ajudar a aumentar a capacidade preditiva dos modelos.

Um grande desafio em AMV está em como fundir em uma única representação multivisão as diferentes representações dos dados. Em outras palavras, como lidar com dados vindos de fontes heterogêneas com diferentes níveis de ruídos (FU et al., 2008; CAO et al., 2019).

Baltrušaitis, Ahuja e Morency (2019) dividiram os métodos de fusão em duas principais categorias: agnósticos de modelo e baseados em modelo. Abordagens baseadas em modelos são implementadas diretamente em algum algoritmo de aprendizagem de máquina. O principal representante de tal abordagem é a família de algoritmos *Multiple kernel learning (MKL)* (GÖNEN; ALPAYDIN, 2011). Em tal caso, *kernels* distintos são ajustados em cada visão, os quais são, em seguida, combinados para formar um único modelo.

Abordagens agnósticas de modelo, como sugere o nome, são esquemas de fusão independentes de modelo. Elas podem ser subdivididas em três sub-categorias: fusão precoce; fusão tardia; e híbrido. Nos métodos de fusão precoce, as diferentes visões são diretamente concatenadas para formar uma única visão, a qual pode ser dada como entrada para um algoritmo de aprendizagem. Métodos de visão precoce seguem, portanto, uma abordagem tradicional de aprendizagem baseada em visão única. Nos métodos de fusão tardia, diferentes modelos são separadamente treinados em cada visão e combinados posteriormente. Métodos de fusão tardia incluem os sistemas de múltiplos classificadores e outros métodos baseados em comitê, de acordo com Baltrušaitis, Ahuja e Morency (2019). Por último, os métodos híbridos combinam fusão precoce e tardia.

Apesar de o esquema de fusão precoce ter a vantagem de ser facilmente implementado, a alta dimensionalidade resultante do processo de concatenação degrada consideravelmente a performance geral dos modelos em alguns conjuntos de dados, em virtude do sobre-ajuste causado (CHAO; SUN, 2016; ZHAO et al., 2017). Posto que a maioria das categorias de recursos linguísticos anteriormente citadas resultam em dados com alta di-

mensionalidade, o esquema de fusão precoce poderia, assim, levar a um sério problema em tais contextos. Por outro lado, a fusão tardia pode levar a um custo computacional elevado, dado que múltiplos modelos devem ser treinados.

Recentemente, Cao et al. (2019) propuseram uma nova abordagem híbrida bastante apropriada para dados em alta dimensionalidade, na qual projeta-se as visões em um espaço comum, valendo-se do conceito de matrizes de dissimilaridade supervisionadas (CAO et al., 2020). Essas matrizes são, então, combinadas e dadas como entrada para um classificador que será treinado com base em uma visão única das representações dos dados. Os autores propuseram, especificamente, o uso do algoritmo *Random Forest* em seu método. Nesta dissertação, entretanto, é investigado o uso de dados projetados com vários classificadores. Ainda, aqueles autores aplicaram seu método na investigação de um problema na área da medicina (radiologia). Nesta dissertação, o método foi adaptado para uso em atribuição de autoria.

É pertinente mencionar que, especificamente no ponto onde as diferentes matrizes de dissimilaridade supervisionadas são projetadas em um espaço comum, há métodos alternativos. Por exemplo, Cruz et al. (2013) utilizaram as saídas de diferentes classificadores, treinados em diferentes representações dos dados, para compor uma matriz simétrica de dissimilaridades usando a medida Double Fault (GIANCINTO; ROLI, 2001). Em tal caso, a referida matriz poderia ser utilizada diretamente como um conjunto de aprendizagem para treinar classificadores.

3 METODOLOGIA

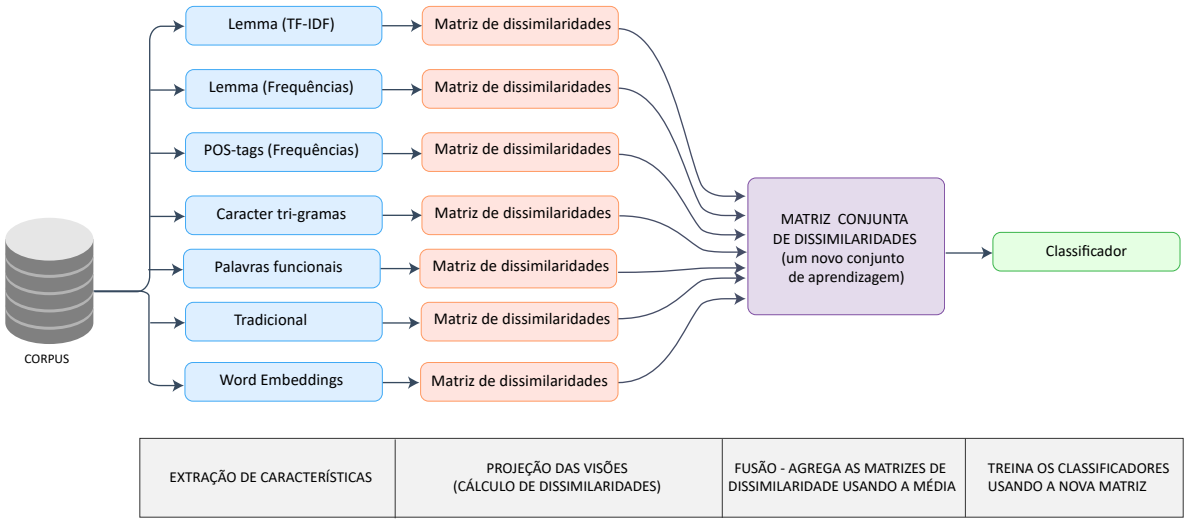
Esta dissertação propõe um novo método baseado em instâncias para atribuição de autoria, no qual cada texto é utilizado individualmente para representar o estilo autoral. Especificamente, o problema de AA é tratado como uma tarefa de classificação de textos, onde os documentos são rotulados de acordo com um conjunto predefinido de autores. Múltiplas categorias de características de texto são utilizadas para representar o conteúdo dos documentos e, assim, o estilo de escrita dos autores.

A metodologia multivisão adotada neste trabalho considera a informação complementar contida nas múltiplas representações de cada documento no conjunto de aprendizagem. O método proposto, cujo esquema é apresentado na Figura 1, é agnóstico de modelo com uma abordagem de fusão híbrida sendo adotada.

A abordagem é dividida em quatro etapas: extração de características; projeção de *visões*; fusão; e, por fim, aprendizagem. Primeiro, diferentes conjuntos de características (*visões*) são extraídos a partir do *corpus*. Em seguida, um método de aprendizagem baseada em dissimilaridades é utilizado para projetar cada *visão* em um espaço comum. Nesse espaço, cada documento é descrito em termos de sua dissimilaridade com todas as outras instâncias. Cada *visão* é separadamente projetada no mesmo espaço e, assim, tais projeções são comparáveis diretamente. Depois, as *visões* projetadas são agregadas pela média em uma única matriz. Por último, a matriz única é enviada para um algoritmo de classificação para obter um modelo de atribuição de autoria.

Cabe mencionar que a média não é a única maneira de agregar as visões projetadas, sendo possível o uso de outras abordagens, muito embora não tenhamos tentado métodos diferentes nesta fase de combinação das visões. Mesmo assim, apesar de termos priorizado o emprego da média, é importante ter em mente o fato de que possivelmente há alternativas melhores na literatura.

Figura 1 – Um diagrama da abordagem de aprendizagem multivisão para atribuição de autoria proposta neste trabalho. Os diferentes conjuntos de características (visões) são extraídos a partir do *corpus* disponível. Em seguida, cada *visão* é projetada em um espaço de dissimilaridade que as torna diretamente comparáveis. Depois, uma matriz de dissimilaridades é obtida pela média das matrizes de dissimilaridades individuais. Finalmente, após a fusão dos dados, estes são dados como entrada para um classificador com o objetivo de obter um modelo de atribuição de autoria.



3.1 EXTRAÇÃO DE CARACTERÍSTICAS

Um passo importante na tarefa de atribuição de autoria é a caracterização do estilo de escrita dos autores candidatos. Neste trabalho, as características são extraídas a partir de 4 categorias distintas: léxica; caractere; sintática; e semântica. A lista de características em cada categoria é apresentada na Tabela 1.

Tabela 1 – Esta tabela contém o conjunto de características usadas na abordagem multivisão proposta para representar documentos. Cada coluna contem a lista de características para uma categoria específica.

Léxico	**Sintático**	**Caracter**	**Semântico**
Lemma (tf-idf)	POS-tags	tri-gramas de caractere	*Word embeddings*
Lemma (frequências)	palavras funcionais (frequências)		
Tamanho médio de palavras	Tamanho médio de sentenças		
Taxa de palavras únicas			
Diversidade léxica			
Tamanho do vocabulário			
Taxa de tipo de token			
Taxa de tipo de token (raiz)			

Cada característica textual categorizada na Tabela 1 foi tratada como uma *visão* diferente, exceto para *tamanho médio de palavras*, *tamanho médio de sentenças*, *taxa de palavras únicas*, *diversidade léxica*, *tamanho do vocabulário*, *taxa de tipo de token* e *palavras funcionais*, as quais foram combinadas para formar uma única *visão* que neste trabalho será denominada *características tradicionais*. Há duas razões por trás dessa escolha: (1) essas características costumam ser utilizadas em estudos tradicionais e, portanto, representam uma *visão* por si mesmas; (2) ao serem utilizadas individualmente, a influência destas características poderia ser ofuscada na presença das demais nas outras *visões*.

Além da *taxa de tipo de token*, que relaciona a quantidade de palavras únicas com o número total de palavras no texto — é conhecida como Type Token Ratio (TTR) na literatura em inglês, a *taxa de tipo de token (raiz)* também foi considerada, a qual relaciona o número de termos únicos com a raiz quadrada do número total de palavras contidas em um texto, conforme proposto por Guiraud (1960) e discutido posteriormente por Malvern et al. (2004).

É importante mencionar que a decisão de usar tri-gramas de caractere se deve a pesquisas anteriores apontando o sucesso obtido com este tipo de característica em textos escritos em inglês (FOURKIOTI; SYMEONIDIS; ARAMPATZIS, 2019). De qualquer modo, não se trata de uma restrição, já que outros tamanhos podem ser facilmente incorporados.

Por fim, vale dizer que as bibliotecas de Python Spacy[1], NLTK[2] e GENSIM[3] foram utilizadas para extrair características linguísticas dos documentos. Para características semânticas, *word embeddings* com 300 dimensões foram obtidos usando o método word2vec proposto por Mikolov et al. (2013).

As três fases restantes da presente proposta — projeção de visões, fusão e aprendizagem — lidam especificamente com a construção do modelo. Por tal motivo, são descritas nesta próxima seção.

3.2 MODELO MULTIVISÃO PARA ATRIBUIÇÃO DE AUTORIA

É importante definirmos formalmente um *corpus* multivisão, antes de apresentar o método em mais detalhes. Considere que $\mathbf{C} = \{(\mathcal{X}_1, y_1), (\mathcal{X}_2, y_2), \ldots, (\mathcal{X}_n, y_n)\}$ seja um *corpus* com n documentos. Cada par $(\mathcal{X}_i, y_i)$ é uma representação multivisão de um documento e seu autor y_i. De tal modo, cada $\mathcal{X}_i$ representa uma família de conjuntos de características, ou seja, $\mathcal{X}_i = \{\mathbf{X_i^{(1)}}, \mathbf{X_i^{(2)}}, \ldots, \mathbf{X_i^{(k)}}\}$, onde $\mathbf{X_i^{(j)}}$ indica o conjunto de características do documento i para a j-ésima visão.

Em nosso contexto, especificamente, cada $\mathbf{X_i^{(j)}}$ é um dos conjuntos de características listados na Tabela 1. A abordagem multivisão que é proposta nesta dissertação é baseada no método híbrido, agnóstico de modelo e multivisão de Cao et al. (2019). Cada visão

[1] <https://spacy.io/>
[2] <https://www.nltk.org/>
[3] <https://radimrehurek.com/gensim/>

dos dados é projetada no mesmo espaço de descrição usando a medida de dissimilaridade supervisionada proposta pelos autores. Esta medida, denominada Dissimilaridade Random Forest (DRF), é computada a partir do método *Random Forest* (BREIMAN, 2001) ajustado em cada visão de maneira individual. A razão pela escolha desse método é precisamente a sua robustez em altas dimensões, além da possibilidade de explorar os rótulos das instâncias para identificar dissimilaridades.

É válido mencionar que *Random Forest* é um método de comitê baseado em árvore de decisão e voltando tanto para problemas de classificação quanto de regressão. O algoritmo realiza uma amostragem aleatória com reposição do conjunto de dados original múltiplas vezes, ajustando árvores de decisão nesses subconjuntos e agregando seus resultados para fazer predições. Além disso, o ajuste das árvores de decisão também considera amostras das características em cada ponto de decisão e tal estrutura pode ser utilizada para computar as (dis)similaridades entre documentos.

Considere que uma Random Forest composta por m árvores de decisão seja ajustada em uma *visão* j específica de algum *corpus*. Seja $l_q^{(j)}(\mathbf{X_i^{(j)}})$ um nó folha (ou nó resposta) que é alcançado ao tentar classificar o documento $\mathbf{X_i^{(j)}}$ na q-ésima árvore. Nós dizemos que dois documentos $\mathbf{X_i^{(j)}}$ e $\mathbf{X_o^{(j)}}$ são similares quando um mesmo nó folha é alcançado por ambos ao tentar classificá-los em uma dada árvore. Ou seja, a medida de dissimilaridade para o conjunto de documentos respectivamente à uma árvore em particular e uma *visão* j é dada por:

$$d_q^{(j)}(\mathbf{X_i^{(j)}}, \mathbf{X_o^{(j)}}) = \begin{cases} 0, & \text{se} \quad l_q^{(j)}(\mathbf{X_i^{(j)}}) = l_q^{(j)}(\mathbf{X_o^{(j)}}) \\ 1, & \text{do contrário} \end{cases} \tag{3.1}$$

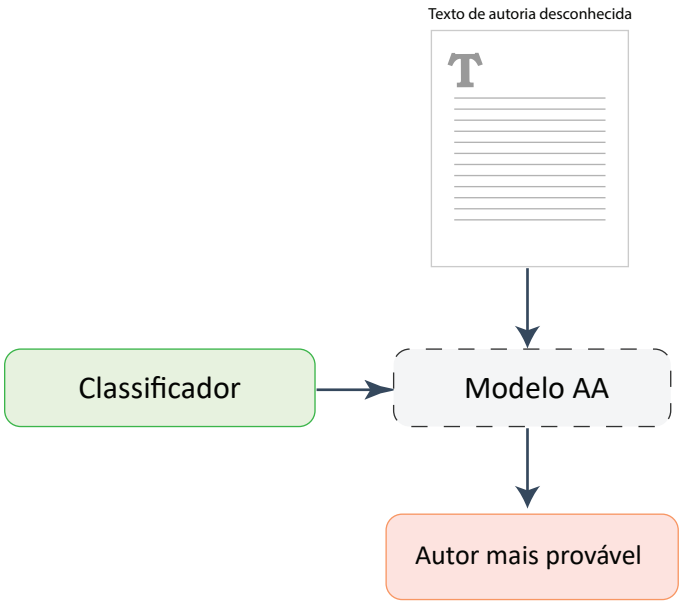
As medidas de dissimilaridade individuais de cada árvore são, então, agregadas para compor a medida de DRF. A função agregadora utilizada aqui é a média, conforme mostrado na Equação 3.2. Além disso, trata-se da mesma função agregadora utilizada por Cao et al. (2019). Mesmo assim, o uso da média, nesse caso, não é algo imposto pelo método, havendo a possibilidade de usar outras funções dependendo da aplicação.

$$d^{(j)}(\mathbf{X_i^{(j)}}, \mathbf{X_o^{(j)}}) = \frac{1}{M} \sum_{q=1}^{M} d_q^{(j)}(\mathbf{X_i^{(j)}}, \mathbf{X_o^{(j)}}) \tag{3.2}$$

As *visões* são, em seguida, projetadas em um mesmo espaço de dissimilaridade usando a Equação 3.2. Em cada *visão*, os documentos são representados como vetores de dissimilaridades, onde cada vetor contém os valores de dissimilaridade em relação a todos os outros documentos. Uma vez que os documentos do *corpus* tenham sido projetados num espaço de dissimilaridade, tem-se início a próxima fase do método: *fusão das visões*.

Todas as *visões* são fundidas por meio da média de seus valores, conforme Cao et al. (2019). Especificamente, a *visão* conjunta que representa o *corpus* é a matriz de dissimilaridade média de todas as *visões*, conforme descrito formalmente na Equação 3.3.

Figura 2 – Cada classificador treinado na fase de aprendizagem, usando as representações intermediárias obtidas por meio da fusão das visões projetadas, resulta em um modelo de atribuição de autoria. Esses modelos recebem como entrada as instâncias de teste e determinam sua autoria mais provável.



Novamente, é importante ressaltar que esta não é a única maneira de agregar as *visões*, havendo a possibilidade de substituir a média por outras funções de agregação mais adequadas nesta fase.

$$d(\mathcal{X}_i, \mathcal{X}_o) = \frac{1}{kM} \sum_{j=1}^{k} \sum_{q=1}^{M} d_q^{(j)}(\mathbf{X_i^{(j)}}, \mathbf{X_o^{(j)}}) \tag{3.3}$$

As *visões* fundidas a partir do *corpus* original formam um conjunto de dados *univisão* representado por uma matriz $n \times n$. Cada linha nesta matriz pode ser vista como um vetor n-dimensional representando um documento cujos valores são as dissimilaridades agregadas considerando todas as *visões* – dissimilaridades estas que são obtidas usando o algoritmo Random Forest. O conjunto de dados resultante é, em seguida, usado na quarta fase do método aqui proposto: *aprendizagem*. Nesta fase, os diferentes classificadores são treinados utilizando as representações intermediárias obtidas na etapa imediatamente anterior. Cada classificador gera um modelo que é usado para atribuir autoria nas instâncias de testes, conforme o fluxo mostrado na Figura 2.

Apesar de (CAO et al., 2019) usar Random Forest também na fase de aprendizagem, é conveniente destacar que uma das vantagens em utilizar uma abordagem de aprendizado multivisão que independe de modelo é a possibilidade de trabalhar com qualquer algoritmo de classificação. Por tal razão, vários algoritmos de aprendizagem são testados durante esta fase final do nosso método. Lembremos que uma das hipóteses levantadas neste trabalho é a de que os dados fundidos obtidos durante as etapas anteriores tornam os classificadores

mais consistentes, o que permitiria o uso de vários algoritmos. Em outras palavras, deseja-se investigar se o uso da abordagem multivisão torna a tarefa de atribuição de autoria menos dependente de classificadores. Para verificar essa e outras hipóteses previamente listadas nesse trabalho, foi conduzida uma série de experimentos, os quais são reportados do próximo capítulo.

4 RESULTADOS EXPERIMENTAIS

Neste capítulo, são reportados os experimentos conduzidos em 4 diferentes *corpus* para verificar as hipóteses levantadas na introdução deste trabalho (ver Capítulo 1). Especificamente, uma série de experimentos foi conduzida para verificar quando a estratégia multivisão proposta nesta dissertação, e descrita no Capítulo 3, resulta em melhor acurácia na tarefa de atribuição de autoria, bem como torna os classificadores mais consistentes. Isto pouparia os linguistas e especialistas em domínio do ônus de decidir qual classificador usar para a tarefa.

Primeiro, os *corpus* utilizados nos experimentos são brevemente descritos nas seções 4.1, 4.2, 4.3 e 4.4. Em seguida, os resultados experimentais obtidos em cada conjunto de dados são apresentados. Para cada conjunto de dados, a apresentação dos resultados está dividida em duas partes. Primeiro, a performance geral do método multivisão é comparada com a performance das abordagens univisão, as quais se baseiam em *visões* individuais e na abordagem multivisão simplista (fusão precoce), lembrando que esta se baseia na concatenação de todas as visões para formar um único conjunto univisão. Em seguida, a consistência de vários classificadores é avaliada na tarefa de atribuição de autoria baseada na abordagem multivisão, onde é feita uma comparação das abordagens univisão e multivisão.

Ao final deste capítulo, nós apresentamos uma análise de sensibilidade onde investigamos a complementariedade de alguns conjuntos de características extraídos a partir dos dados textuais. A discussão abrange tanto visões individuais quanto suas combinações. O objetivo é identificarmos visões que melhoram ou pioram o desempenho de outras.

Todos os experimentos foram conduzidos usando validação cruzada estratificada 5-folds com 5 repetições. A performance de classificação foi medida com a média dos escores F1. Cabe mencionar que a escolha da quantidade de árvores utilizada pelo algoritmo Random Forest é fundamental na fase do método em que o algoritmo computa as distâncias entre os documentos para permitir a projeção das *visões*. Tal número foi fixado em 100 árvores. Avaliações prévias com quantidades menores de árvores levaram a resultados marginais, enquanto que quantidades maiores resultaram em melhorias de performance pouco relevantes.

Conforme já mencionado, foi admitida a hipótese de que o método multivisão aqui proposto deveria ser consistente com uma ampla gama de classificadores. Em decorrência disso, vários classificadores foram selecionados para uso nos experimentos. O objetivo é que o estudo contemple os classificadores mais utilizados, ao mesmo tempo em que são considerados os representantes de diferentes tipos de métodos. No total, a seleção inclui 13 classificadores com a maioria destes fazendo parte do pacote de ferramentas de aprendizagem de máquina *Scikit-Learn* (PEDREGOSA et al., 2011). Adicionalmente, foram

considerados outros dois métodos no estado da arte que não fazem parte daquele pacote: KTBoost (SIGRIST, 2019), e XGBoost (CHEN; GUESTRIN, 2016).

A preferência pelos classificadores disponíveis no *Scikit-Learn* se deve a popularidade do pacote em diversas soluções em AM e ciência de dados. Conforme já aludido, a intenção com este estudo é poupar, na medida do possível, os linguistas e outros especialistas do encargo envolvido na escolha do método certo para sua aplicação. Visto por essa perspectiva, o uso do referido pacote torna-se uma escolha bastante conveniente.

Os classificadores considerados para a tarefa estão listados na Tabela 2. Todos foram utilizados com os valores padrão de seus hiperparâmetros. Ainda que fosse possível melhorar a performance geral desses algoritmos por meio do ajuste fino de seus hiperparâmetros, optou-se por não fazê-lo, já que a intenção seria avaliar o desempenho deles como uma solução pronta para uso.

Tabela 2 – Lista dos classificadores utilizados nos experimentos deste estudo. Os 13 classificadores abrangem diferentes tipos de métodos. Essa lista inclui muitos dos classificadores mais usados em atribuição de autoria e classificação de textos. A coluna intitulada *instâncias* refere-se aos métodos de aprendizagem baseados em instâncias.

Classificador	Árvore	Função	Instâncias	Rede Neural	Bayesiano	Comitê
Decision Tree	×					
Random Forest	×					×
Logistic Regression		×				
Stochastic Gradient Descent (SGD)		×				
Support Vector Machine (SVM)		×				
K-Nearest Neighbors (KNN)			×			
Perceptron				×		
Multilayer perceptron (MLP)				×		
Gaussian Naive Bayes (GNB)					×	
Extreme Gradient Boosting (XGBoost)						×
Kernel and Tree Boosting (KTBoost)						×
AdaBoost						×
Gradient Boosting						×

A seguir, são descritos os conjuntos de dados utilizados durante os experimentos para validar a performance do método proposto nesta dissertação. Dois deles são compostos por textos literários, um por textos jurídicos e há um outro *corpus* que é composto por textos de artigos científicos. Todos os textos são escritos em inglês.

4.1 CONJUNTO DE OBRAS DA RENASCENÇA INGLESA

Trata-se de um conjunto de trabalhos da época da renascença inglesa que já foi utilizado em outras pesquisas na área de AA (AREFIN et al., 2014). O conjunto de dados é composto

por obras coletadas a partir de dois repositórios: o *corpus Shakespeare e seus Conteporâneos* (FROEHLICH, 2020); e o *corpus Folger Shakespeare* (Folger Shakespeare Library, n.d.). Esses repositórios, é importante destacar, são organizados manualmente por renomados estudiosos em ciências humanas. Além disso, a linguagem presente nos textos das obras foi padronizada para o inglês moderno (muitas dessas obras foram escritas em inglês antigo), o que é particularmente benéfico para a aplicação de ferramentas de processamento de linguagem natural. O conjunto de dados utilizado em nosso estudo é uma combinação dos dois *corpus* acima mencionados.

Para os experimentos executados ao longo da presente pesquisa, foram selecionados 218 textos de autoria solo, dentre os 230 textos que fazem parte do *corpus* original (incluindo os textos disputados). Essa quantidade de textos está associada com 17 autores. A quantidade média de palavras por documento é de 3653, aproximadamente. O maior texto tem 5753 palavras; o menor, 1781 palavras. É perceptível, portanto, que esse conjunto de dados é composto predominantemente por textos longos.

Os autores considerados para esta dissertação estão listados na Tabela 3. Na lista, cada autor é acompanhado por seu respectivo número de obras, quantidade média de palavras que usa por texto, bem como a quantidade média de palavras únicas.

Tabela 3 – Lista de autores cujas obras remontam ao período da renascença inglesa. Cada autor listado nesta tabela também acompanha seu respectivo número de obras. Em suma, esse conjunto de dados é composto por 218 textos associados com 17 autores. A quantidade média de palavras por documento é de aproximadamente 3653. O maior texto contém 5753 palavras; o menor é composto por 1781 palavras.

Autor	**Nº de textos**	**Média de palavras por texto**	**Média de palavras únicas por texto**
Shirley, James	31	3252,3	1807,8
Shakespeare, William	29	3907,2	2227,8
Heywood, Thomas	19	3592,4	1982,5
Fletcher, John	14	3505,9	1892,0
Jonson, Ben	14	4529,9	2594,0
Chapman, George	14	3619,3	2057,6
Middleton, Thomas	14	3599,9	2018,9
Brome, Richard	13	4041,6	2285,8
Massinger, Philip	13	3992,5	2276,0
Dekker, Thomas	9	3870,1	2212,2
Ford, John	9	4039,4	2344,1
Lyly, John	8	2775,4	1582,9
Marston, John	8	3760,9	2206,6
Peele, George	6	2912,7	1577,8
Marlowe, Christopher	6	3249,3	1834,0
Greene, Robert	6	2939,5	1630,7
Rowley, William	5	3565,0	2004,4

4.2 CONJUNTO DE OBRAS DA ERA VITORIANA

Gungor (2018) selecionou obras de 50 autores conhecidos da era Vitoriana com o objetivo de formar um extenso conjunto de dados para atribuição de autoria[1]. Isso resultou em um *corpus* composto por 53678 textos rotulados, todos escritos em inglês.

Para diminuir o tempo de aprendizagem dos modelos em nosso estudo, a quantidade de textos foi reduzida. Foram selecionados 14 autores de maneira arbitrária e os primeiros 50 textos de cada autor foram escolhidos a partir do *corpus* original. Isso resultou em um conjunto de dados composto por 700 textos. O maior texto contém 483 palavras; o menor possui 292 palavras. Além disso, 75% dos textos possui cerca de 423 palavras ou menos e há uma média de 404 palavras por documento, aproximadamente. Isso significa que, em sua maioria, os textos presentes nesse conjunto de dados são mais breves em comparação com aqueles presentes no *corpus* anteriormente descrito.

Por fim, os autores estão listados na Tabela 4, bem como o correspondente número de textos para cada um deles; o número médio de palavras usadas por texto; e a média de palavras únicas.

Tabela 4 – Lista de autores cujas obras remontam à Era Vitoriana. Esta tabela também apresenta o correspondente número de textos associado com cada autor, bem como o número médio de palavras usadas por texto e a média de palavras únicas.

Autor	**Nº de textos**	**Média de palavras por texto**	**Média de palavras únicas por texto**
Arthur Conan Doyle	50	413,2	287,6
Charles Darwin	50	424,8	293,6
Charles Dickens	50	388,1	255,0
Edith Wharton	50	391,2	256,3
Horace Greeley	50	415,5	291,8
James Baldwin	50	416,6	287,4
Jane Austen	50	405,2	273,5
John Muir	50	400,8	265,0
Joseph Conrad	50	390,3	255,4
Mark Twain	50	387,2	256,7
Nathaniel Hawthorne	50	394,4	259,6
Ralph Emerson	50	428,8	301,9
Robert Louis Stevenson	50	383,8	245,1
Rudyard Kipling	50	412,9	286,0

[1] https://dataworks.iupui.edu/handle/11243/23

4.3 JULGAMENTOS DO SUPREMO TRIBUNAL DA AUSTRÁLIA

Esse conjunto de dados é composto por sentenças de três juízes que atuaram no Supremo Tribunal da Austrália de 1913 a 1975. Trata-se de um conjunto de textos no qual os autores discutiram tópicos muito similares, conforme apontado por Sari, Stevenson e Vlachos (2018). Esse mesmo *corpus* também foi utilizado por Seroussi, Zukerman e Bohnert (2014) em seu estudo, no qual foram considerados apenas os textos de autoria não disputada e escritos em períodos onde apenas um dos três juízes serviram na suprema corte. Em nosso estudo, esse mesmo critério foi seguido.

O *corpus* é composto por 1342 textos associados com 3 autores. Há uma quantidade mediana de 443 palavras por documento, aproximadamente. O maior texto contém 2914 palavras; o menor é composto por apenas 39 palavras. A maior parte dos documentos possui mais de 400 palavras.

Tabela 5 – Esse conjunto de dados é composto por sentenças de três juízes que atuaram no Supremo Tribunal da Austrália de 1913 a 1975. São 1342 textos com uma quantidade mediana de 443 palavras por texto. Nesta tabela, estão listados os autores das sentenças e a quantidade de textos correspondente a cada autor, bem como a quantidade mediana de palavras que cada autor usou por texto, além da quantidade mediana de palavras únicas.

Autor	**Nº de textos**	**Quantidade mediana de palavras por texto**	**Quantidade mediana de palavras únicas por texto**
Dixon	902	543,0	309
McTiernan1965-1975	253	275,0	159,0
Rich1913-1928	187	185,0	119,0

4.4 CONJUNTO DE ARTIGOS CIENTÍFICOS DA REVISTA PLOS ONE

Trata-se de um conjunto de artigos científicos disponibilizados publicamente[2] pela revista científica Plos One[3]. Os artigos são predominantemente escritos na área de biologia e, para este trabalho, uma amostra contendo 230 textos associados com um total de 11 autores foi selecionada.

Nesse corpus, a maior parte dos artigos é escrita por mais de um autor, havendo poucos textos de autoria solo. Considerando que nosso método não é adequado para classificação *multi-label* — casos em que mais de um rótulo (autor) é atribuído para cada texto —, o seguinte protocolo em duas etapas foi seguido durante a preparação deste *corpus*, para que pudesse ser usado nos experimentos: primeiro, foram selecionados apenas os textos de autoria solo; em seguida, foram selecionados aqueles artigos escritos por múltiplos autores

[2] https://plos.org/text-and-data-mining/
[3] https://journals.plos.org/plosone/

e que contaram com a colaboração de algum autor solo já considerado na primeira etapa. Neste último caso, a autoria desses textos escritos de maneira colaborativa foi atribuída ao autor solo previamente considerado, já que seu estilo de escrita também estaria presente nesses textos e poderia ser identificado. Tal solução permitiu aumentar a quantidade de textos por autor no *corpus*.

A quantidade média de palavras por documento nesse *corpus* final é 1038, aproximadamente. O maior texto contém 2492 palavras; já o menor é composto por 161 palavras. A maior parte dos textos possui mais de 1000 palavras, o que representa um *corpus* composto por textos relativamente longos. A Tabela 6 lista os autores considerados, bem como o tamanho médio dos textos em termos da quantidade de palavras usada.

Tabela 6 – Este *corpus* é predominantemente composto por textos escritos na área de biologia para a revista científica *Plos One*. Nesta tabela, estão listados os autores considerados para o presente estudo, junto com o número de textos associados com o autor e os tamanhos médios dos textos considerando o número de palavras.

Autor	**Nº de textos**	**Média de palavras por texto**	**Média de palavras únicas por texto**
John P. A. Ioannidis	52	1096,2	310,9
Peter J. Hotez	39	1158,9	325,2
Paul T. Williams	21	1140,8	262,9
Hemai Parthasarathy	20	416,9	299,0
Liza Gross	18	1083,4	585,1
Walter R. Tschinkel	18	1485,8	316,2
Richard Robinson	14	590,0	360,4
Jane Bradbury	12	899,0	485,8
Henry Nicholls	12	956,7	484,0
Chin-Hsiao Tseng	12	1156,5	554,3
Virginia Gewin	12	1123,8	633,2

A Tabela 7 apresenta a quantidade de características por visão dos dados considerando cada um dos *corpus* anteriormente descritos. Especificamente, o leitor atento imediatamente perceberá o quanto a concatenação dos dados aumenta a complexidade de cada *corpus*.

Tabela 7 – Quantidade de características por visão dos dados. Nas colunas, temos os quatro *corpus* anteriormente descritos. Nas linhas, as visões dos dados e a quantidade de características presentes em cada uma.

Visão	**Renascença**	**Era vitoriana**	**Julgamentos**	**Plos One**
POS-tags	22018	21700	41602	7360
embedding	65400	210000	402600	69000
palavras funcionais	121208	389200	746152	127880
tradicional	122734	394100	755546	129490
n-gramas	707628	730100	3185908	570630
lemma (tfidf)	2180000	5793900	13420000	2300000
lemma (frequência)	8297298	5807200	30372144	5524600
concatenado	11516286	13346200	48923952	8728960

4.5 ANÁLISE DOS RESULTADOS

A seguir, são discutidos os resultados obtidos em cada um dos experimentos realizados considerando cada *corpus* anteriormente descrito. O procedimento adotado na análise dos resultados foi o mesmo para cada conjunto de dados. Conforme já aludido, primeiro a performance geral da abordagem multivisão é comparada com as abordagens de base; em seguida, a consistência dos classificadores na atribuição de autoria é avaliada.

4.5.1 Obras da Renascença inglesa

4.5.1.1 Performance geral da abordagem multivisão

Nesta análise, nos comparamos a performance geral da abordagem multivisão com aquela obtida nas abordagens de base: *fusão precoce* (visões concatenadas) e *univisão*, a qual refere-se a usar as visões dos dados individualmente para treinar os algoritmos de classificação. Um resumo dos resultados é apresentado na Tabela 8[4].

De acordo as performances observadas na tabela, a abordagem multivisão superou quase todas as demais neste experimento, resultando em escores F1 mais altos em 11 classificadores. A exceção está no desempenho obtido com as visões *frequência do lemma* e *concatenado*, as quais resultaram em escores maiores.

É pertinente ressaltar, com base nos resultados observados, que o uso das representações concatenadas (fusão precoce) não resultou em melhor performance em comparação com a abordagem multivisão neste *corpus*. Trata-se, portanto, de uma evidência em potencial de que é importante, por certo, atentar para a informação complementar contida em cada *visão* dos dados. A concatenação é um processo de fusão mais simples, onde a representação univisão resultante é menos apropriada para os classificadores explorarem

[4] o código utilizado nos experimentos será disponibilizado em <https://github.com/luisfredgs/authorship-attribution-multi-view-learning>

Tabela 8 – Esta tabela mostra a performance dos classificadores usando diferentes *visões* e a abordagem multivisão no *corpus* de obras da renascença inglesa. A performance é medida em termos da média dos escores F1. Valores entre parêntesis representam os desvios padrão. Valores em negrito representam o maior escore obtido por um classificador (linha). As *visões* Concatenada e Tradicional são respectivamente a fusão simplificada de todas as *visões* e as características utilizadas em estudos tradicionais conforme descrito no Capítulo 3. A linha rotulada como Stochastic GD refere-se aos resultados do classificador *Stochastic Gradient Descent*. Gaussian NB refere-se ao classificador *Gaussian Naive Bayes*. A coluna rotulada como *p. função* refere-se à *visão* composta por palavras funcionais. A coluna *frequência* representa a frequência do lemma.

classificador	multivisão	concatenado	frequência	tradicional	p. função	tf-idf	POS-tag	embedding	n-gramas
Decision Tree	**0,52**(0,01)	0,49(0,02)	0,48(0,04)	0,23(0,01)	0,27(0,01)	0,38(0,01)	0,42(0,02)	0,25(0,02)	0,39(0,02)
Random Forest	**0,61**(0,02)	0,42(0,03)	0,34(0,04)	0,27(0,01)	0,40(0,06)	0,39(0,03)	0,43(0,02)	0,36(0,04)	0,41(0,02)
Logistic Regression	**0,62**(0,00)	0,20(0,00)	0,20(0,00)	0,02(0,00)	0,04(0,00)	0,05(0,00)	0,05(0,00)	0,01(0,00)	0,21(0,00)
Stochastic GD	0,60(0,02)	0,07(0,01)	**0,68**(0,02)	0,03(0,01)	0,30(0,03)	0,59(0,13)	0,33(0,07)	0,19(0,11)	0,07(0,00)
SVM	**0,73**(0,00)	0,08(0,01)	0,70(0,00)	0,04(0,00)	0,33(0,00)	0,68(0,00)	0,27(0,00)	0,03(0,00)	0,08(0,01)
KNN	**0,71**(0,00)	0,17(0,00)	0,26(0,00)	0,14(0,00)	0,41(0,00)	0,40(0,00)	0,51(0,00)	0,51(0,00)	0,17(0,00)
Perceptron	**0,32**(0,00)	0,08(0,00)	0,28(0,00)	0,01(0,00)	0,17(0,00)	0,14(0,00)	0,18(0,00)	0,10(0,00)	0,08(0,00)
MLP	**0,69**(0,02)	0,55(0,02)	0,63(0,02)	0,03(0,00)	0,06(0,00)	0,65(0,02)	0,06(0,01)	0,03(0,00)	0,16(0,02)
Gaussian NB	**0,54**(0,00)	0,40(0,00)	0,23(0,00)	0,16(0,00)	0,33(0,00)	0,23(0,00)	0,38(0,00)	0,51(0,00)	0,30(0,00)
XGBoost	0,52(0,00)	**0,64**(0,00)	0,54(0,00)	0,27(0,00)	0,55(0,00)	0,52(0,00)	0,54(0,00)	0,44(0,00)	0,60(0,00)
KTBoost	**0,67**(0,01)	0,54(0,02)	0,52(0,01)	0,28(0,01)	0,34(0,01)	0,45(0,01)	0,39(0,01)	0,34(0,01)	0,42(0,02)
AdaBoost	**0,27**(0,00)	0,13(0,00)	0,11(0,00)	0,10(0,00)	0,08(0,00)	0,04(0,00)	0,10(0,00)	0,07(0,00)	0,10(0,00)
Gradient Boosting	**0,64**(0,01)	0,52(0,02)	0,48(0,01)	0,28(0,01)	0,39(0,02)	0,45(0,01)	0,42(0,01)	0,37(0,02)	0,45(0,01)

completamente a informação contida em cada visão dos dados. Ao negligenciar essa informação complementar, deixamos de ganhar em termos de capacidade de generalização. Isso não acontece quando os dados multivisão são utilizados no processo de aprendizagem.

Ademais, é interessante notar que não houve ganho relevante de performance ao combinar as várias características textuais que formam a visão *tradicional*. De fato, esta visão dos dados resultou na pior performance alcançada com o presente *corpus*. Para efeito de comparação, a visão *palavras funcionais* é uma das representações dos dados que foi utilizada para compor a já mencionada visão *tradicional*, conforme indicado no Capítulo 3. Todavia, ao ser utilizada separadamente, seu ganho foi ainda superior. Esse comportamento poderia nos induzir a uma reflexão interessante. É bem possível que essa combinação das visões tenha diminuído a influência das outras representações que foram utilizadas para compô-la. Não obstante, seria necessária a execução de mais experimentos para confirmar essa hipótese. Trata-se, portanto, de uma oportunidade para trabalhos futuros.

Relativamente à performance dos classificadores na abordagem multivisão, destacam-se os algoritmos SVM e KNN, ambos tendo alcançado os mais altos escores. É oportuno lembrar que o método SVM já é comprovadamente robusto, mesmo em espaços de alta dimensionalidade, além de ser frequentemente adotado na literatura em atribuição de autoria — especificamente nas abordagens baseadas em instâncias. Por seu lado, o método KNN é frequentemente usado em abordagens não supervisionadas e baseadas em perfis.

Há, portanto, uma conexão entre esse resultado e o que é frequentemente mostrado na literatura (KOPPEL; SCHLER; ARGAMON, 2009).

Um aspecto importante desse corpus é seu acentuado desbalanceamento em comparação com os demais conjuntos de dados utilizados nesse estudo, contendo autores com menos de 6 obras, enquanto há outros com 30 obras. Como resultado, foi observado que autores com um número muito reduzido de obras, especificamente aqueles com menos de 8 textos no corpus, não contribuíram para um bom desempenho de classificação de seus textos — algo que, diga-se de passagem, já era esperado. Os autores *Peele, George* e *Rowley, William*, que estão entre aqueles com menos textos no *corpus*, resultaram em performances mais baixas de classificação em comparação com os demais.

Os resultados listados na Tabela 8 foram comparados estatisticamente para verificarmos quando as diferenças observadas nos escores são significativas. Para tal, foi seguida a metodologia proposta por Demšar (2006). Primeiro, o teste de Friedman foi utilizado para verificar a hipótese nula de que não há diferenças nas performances dos métodos considerando as diferentes visões dos dados. A hipótese foi rejeitada com um p-valor igual a $1,07192 \times 10^{-52}$. Com isso, o teste *post-hoc* de Nemenyi para comparações emparelhadas foi conduzido com o objetivo de identificar onde estão essas diferenças. Os p-valores resultantes desse teste são reportados na Tabela 9. O gráfico de distâncias críticas, considerando um nível de significância de 0,05, é apresentado na Figura 3

Tabela 9 – P-valores do teste *post-hoc* de Nemenyi para comparações emparelhadas.

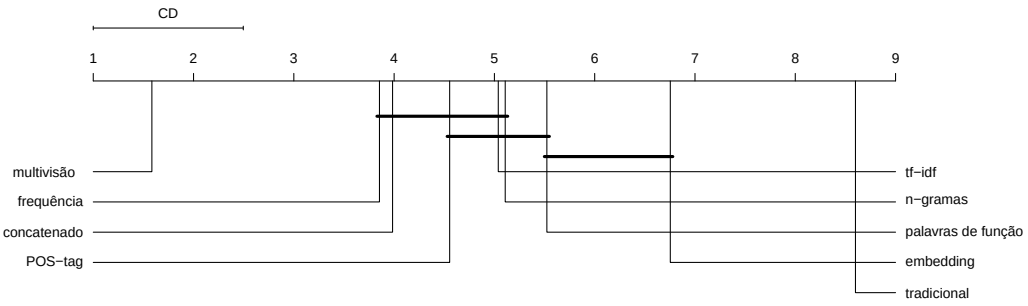
	multivisão	concatenado	frequência	tradicional	palavras funcionais	tf-idf	POS-tag	embedding
concatenado	$2,07 \times 10^{-5}$							
frequência	$8,08 \times 10^{-5}$	1						
tradicional	0	$6,71 \times 10^{-14}$	$8,38 \times 10^{-14}$					
palavras funcionais	$1,13 \times 10^{-13}$	0,0368	0,015	$5,4 \times 10^{-9}$				
tf-idf	$2,34 \times 10^{-11}$	0,41	0,249	$4,47 \times 10^{-12}$	0,985			
POS-tag	$2,29 \times 10^{-8}$	0,96	0,875	$9,2 \times 10^{-14}$	0,531	0,985		
embedding	$7,92 \times 10^{-14}$	$2,93 \times 10^{-7}$	$5,64 \times 10^{-8}$	0,00387	0,203	0,0107	0,000161	
n-gramas	$8,11 \times 10^{-12}$	0,319	0,182	$1,3 \times 10^{-11}$	0,995	1	0,966	0,0177

O teste estatístico confirma que, por certo, a abordagem multivisão supera as demais em comparação nesse experimento. Trata-se de um resultado bastante significativo, o qual demonstra que explorar a complementaridade das *visões* dos dados poderia trazer resultados positivos ao implementar soluções para problemas em atribuição de autoria.

De acordo com a Figura 3, resultante do teste *post-hoc*, as performances das *visões* concatenada, *frequência do lemma* e *Pos-tag* são equivalentes. Isso pode ser constatado pelo agrupamento formado entre as três. Isso indica que simplesmente concatenar as *visões* não levou a resultados estatisticamente diferentes daqueles obtidos ao utilizar essas outras representações dos dados individualmente.

Os resultados apresentados acima mostram que a abordagem multivisão foi capaz de trazer resultados positivos ao ser aplicada no problema de atribuição de autoria, superando a performance das abordagens comparadas, considerando o *corpus* de textos literários da

Figura 3 – Gráfico de distâncias críticas do teste *post-hoc* de Nemenyi com um nível de significância de 0,05



época da renascença inglesa. Tal fato responde a primeira pergunta de pesquisa levantada na introdução desta dissertação, visto que tanto o teste de Friedman quanto o teste *post-hoc* de Nemenyi demonstram melhoria em comparação com as outras abordagens.

A consistência dos classificadores, que também é uma medida importante para esta pesquisa, é tratada na seção seguinte.

4.5.1.2 Consistência em atribuição de autoria

A segunda pergunta de pesquisa enumerada na introdução indaga se o uso da abordagem multivisão permite que os classificadores concordem mais entre si sobre os autores atribuídos aos textos. Com isso, deseja-se saber se os dados multivisão tornam os algoritmos de classificação mais consistentes no que se refere à autoria.

Para tal, os classificadores foram tratados como avaliadores e a confiabilidade entre suas avaliações foi calculada. É oportuno mencionar que essa análise é baseada em uma metodologia geralmente usada em análise de conteúdo (KRIPPENDORFF, 2018). Neste estudo, a concordância de classificação (confiabilidade entre avaliadores) foi calculada usando a medida Fleiss' Kappa (FLEISS, 1971). Além disso, nossa análise foi orientada pela escala proposta por Landis e Koch (1977) para categorizar as concordâncias entre os avaliadores. A Tabela 10 lista essas categorias. Outrossim, a dispersão da medida de qualidade adotada nos experimentos (escore F1) foi analisada e tratada como um indicador de consistência.

A Figura 4 mostra as dispersões nos escores F1 (eixo y) dos classificadores com respeito às diferentes *visões* dos dados (eixo x). Cada *boxplot* presente na figura reflete as dispersões nos escores considerando todos os classificadores que foram treinados numa determinada

Tabela 10 – Categorização dos valores Kappa de concordância com (LANDIS; KOCH, 1977)

Kappa	**Categoria**
< 0	Baixa concordância
0,01 – 0,20	Ligeira concordância
0,21 – 0,40	Concordância razoável
0,41 – 0,60	Concordância moderada
0,61 – 0,80	Concordância substancial
0,81 – 1,00	Concordância quase perfeita

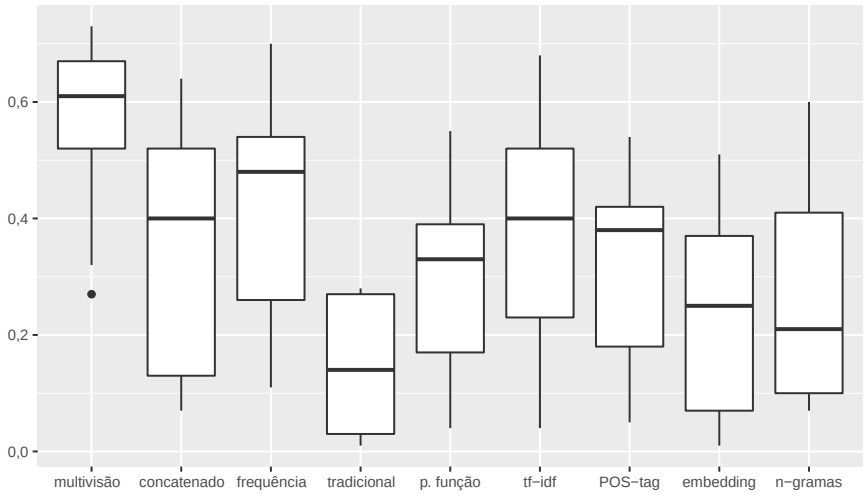


Figura 4 – A dispersão dos escores F1 com respeito à cada abordagem considerada. Como pode ser visto, o intervalo interquartil da abordagem multivisão é muito mais estreito do que nas outras abordagens. Isso mostra como os métodos treinados em dados multivisão concordam consistentemente com o verdadeiro autor dos textos no corpus, independentemente do classificador usado.

visão dos dados. De acordo com a referida figura, a abordagem multivisão não apenas resultou em um escore mediano superior, como também um IIQ mais estreito. Isso mostra que os diferentes classificadores mantiveram uma performance similar ao serem expostos a dados multivisão, posto que a dispersão nos escores F1 foi menor com esse conjunto de dados. Em outras palavras, vemos que a classificação permaneceu consistente ao longo da execução dos métodos.

Também observamos valores medianos relativamente altos para dados léxicos (frequência do lemma, tf-idf do lemma), sintáticos (POS-tag) e concatenados (fusão precoce), mas a dispersão nos valores dos escores, neste caso, foi bem maior. É importante destacar que tal dispersão nos escores reflete uma abordagem que é mais dependente do algoritmo. Portanto, apesar de vermos um agrupamento de abordagens univisão alcançando escores relativamente altos, eles são muito mais dependentes de classificadores do que a abordagem multivisão. Essas observações indicam que o método multivisão foi, de fato, o menos

dependente de classificador nesse experimento.

Uma análise da concordância entre os diferentes classificadores, comparando os contextos univisão e multivisão, nos permitiria avaliar melhor esta afirmação. É pertinente ressaltar que concordância, neste âmbito, representa uma medida de consistência entre os classificadores, não de acurácia. Isso significa que nós não estamos avaliando quando os classificadores fazem predições corretas, mas, sim, quando todos eles predizem um mesmo autor para um dado texto ou não. Quanto mais predições para um mesmo autor, mais os classificadores concordam entre si.

Trata-se de uma importante medida, pois, um alto grau de concordância acompanhado de alta acurácia poderia indicar que a abordagem multivisão é capaz de alcançar alta performance, com consistência, independente da escolha do classificador.

Tabela 11 – Concordância dos classificadores por autor medido usando Fleiss' kappa. Na coluna *Multivisão*, é mostrado o grau de concordância para a abordagem multivisão; já na última coluna, temos a mesma medida para univisão considerando a frequência do lemma. A estatística Kappa varia de -1 a +1 e a interpretação dos valores é dada pela Tabela 10

Autor	**Multivisão**	**Univisão**
Brome, Richard	0,63	0,39
Chapman, George	0,34	0,24
Dekker, Thomas	0,45	0,26
Fletcher, John	0,71	0,38
Ford, John	0,49	0,17
Greene, Robert	0,11	0,09
Heywood, Thomas	0,47	0,31
Jonson, Ben	0,63	0,38
Lyly, John	0,43	0,26
Marlowe, Christopher	0,09	0,14
Marston, John	0,54	0,23
Massinger, Philip	0,87	0,61
Middleton, Thomas	0,62	0,27
Peele, George	0,16	0,05
Rowley, William	0,04	0,02
Shakespeare, William	0,95	0,41
Shirley, James	0,85	0,51

As variações na medida de concordância estão resumidas na Tabela 11. Para esse comparativo, foi considerada a *visão* individual com melhor escore mediano: *frequência do lemma*. Em outras palavras, selecionamos aquela que proporciona classificadores mais acurados e, ao mesmo tempo, mais consistentes entre as abordagens univisão.

Os resultados indicam que houve um aumento expressivo no valor dessa medida quando

consideramos o método multivisão dos dados. Apenas um autor não representou aumento na concordância (Marlowe). Para facilitar a visualização dos maiores ganhos no valor da medida de concordância, os autores relacionados com aumentos mais significativos estão destacados em negrito.

Outro resultado interessante é o número de autores com grau de concordância aumentando de *moderada* e *substancial* para *quase perfeita* (ver Tabela 10). Há 7 autores com concordância substancial ou quase perfeita no cenário multivisão, contra somente um autor com concordância em nível substancial (Massinger) no univisão.

Se combinarmos esse aumento proeminente no grau de concordância entre os classificadores com a superioridade na acurácia de classificação, podemos concluir que todos os classificadores consistentemente fizeram predições corretas quando treinados com dados multivisão. Isto posto, podemos afirmar, com base nos resultados anteriormente apresentados, que a procura por um classificador em particular para a tarefa de atribuição de autoria poderia ser um aspecto menos relevante do problema se o método multivisão for considerado, uma vez que todos os classificadores têm um desempenho similarmente bom em tal contexto. Em outros termos, verificamos neste *corpus* que é válida a hipótese de que a abordagem multivisão é capaz de isentar os linguistas e outros especialistas do encargo envolvido na escolha de um classificador ideal ao criar soluções para a atribuição de autoria.

4.5.2 Conjunto de Obras da Era Vitoriana

4.5.2.1 Performance geral da abordagem multivisão

O procedimento adotado nas análises previamente conduzidas é, também, seguido nas próximas análises que discutiremos. Portanto, primeiramente será avaliada a performance geral da abordagem multivisão em comparação com aquela obtida nas abordagens de base, as quais são: visão única e concatenada. Isto posto, apresentamos na Tabela 12 os desempenhos alcançados, em termos dos escores F1 médios, pelos métodos de classificação considerados neste estudo nas diferentes visões dos dados.

Os resultados apresentados na tabela sugerem que o método multivisão supera boa parte das outras visões com expressiva margem, ao mesmo tempo em que diferenças menores são observadas entre essa abordagem e a visão *concatenado*, bem como *frequência do lemma* e *tf-idf do lemma*.

Os resultados mostram que a visão *concatenado* resultou em um bom desempenho com alguns classificadores. Conquanto, sua performance é bastante instável. Isso poderia indicar que, com a alta dimensionalidade resultante da concatenação dos dados, surgem fronteiras linearmente separáveis e isso é aproveitado por alguns classificadores. Não obstante, tal prática é menos frutífera em termos de potencial de generalização, já que uma parte significativa dos classificadores não se saiu tão bem com essa visão dos dados.

Tabela 12 – Performance dos classificadores com diferentes *visões* e a abordagem multivisão no conjunto de obras da era vitoriana. A performance é medida em termos da média dos escores F1. Valores entre parêntesis representam os desvios padrão. Valores em negrito representam o maior escore obtido por um classificador (linha). As *visões* Concatenada e Tradicional são respectivamente a fusão simplificada de todas as *visões* e as características utilizadas em estudos tradicionais conforme descrito no Capítulo 3. A linha rotulada como Stochastic GD refere-se aos resultados do classificador *Stochastic Gradient Descent*. Gaussian NB refere-se ao classificador *Gaussian Naive Bayes*. A coluna rotulada como *p. função* refere-se à *visão* composta por palavras funcionais. A coluna *frequência* representa a frequência do lemma.

classificador	multivisão	concatenado	frequência	tradicional	p. função	tf-idf	POS-tag	embedding	n-gramas
Decision Tree	0,72(0,01)	**0,76**(0,01)	0,75(0,01)	0,13(0,01)	0,27(0,01)	0,68(0,01)	0,28(0,01)	0,31(0,01)	0,39(0,01)
Random Forest	**0,87**(0,01)	0,59(0,02)	0,63(0,01)	0,14(0,01)	0,37(0,02)	0,45(0,02)	0,35(0,01)	0,40(0,02)	0,42(0,02)
Logistic Regression	0,90(0,00)	0,18(0,00)	0,88(0,00)	0,10(0,00)	0,43(0,00)	**0,93**(0,00)	0,29(0,00)	0,43(0,00)	0,19(0,00)
Stochastic GD	0,83(0,02)	0,03(0,01)	**0,90**(0,01)	0,05(0,01)	0,49(0,03)	0,84(0,09)	0,26(0,02)	0,29(0,03)	0,04(0,02)
SVM	0,93(0,00)	0,02(0,02)	0,95(0,00)	0,10(0,00)	0,59(0,00)	**0,96**(0,00)	0,42(0,00)	0,53(0,00)	0,04(0,02)
KNN	**0,92**(0,00)	0,13(0,00)	0,69(0,00)	0,12(0,00)	0,41(0,00)	0,72(0,00)	0,35(0,00)	0,59(0,00)	0,13(0,00)
Perceptron	0,76(0,00)	0,03(0,00)	**0,82**(0,00)	0,02(0,00)	0,34(0,00)	0,49(0,00)	0,14(0,00)	0,23(0,00)	0,03(0,00)
MLP	0,91(0,00)	0,92(0,01)	**0,94**(0,00)	0,10(0,01)	0,65(0,00)	**0,94**(0,00)	0,35(0,01)	0,48(0,00)	0,16(0,01)
Gaussian NB	**0,84**(0,00)	0,80(0,00)	0,74(0,00)	0,07(0,00)	0,42(0,00)	0,73(0,00)	0,36(0,00)	0,58(0,00)	0,62(0,00)
XGBoost	0,70(0,00)	**0,91**(0,00)	0,88(0,00)	0,18(0,00)	0,60(0,00)	0,85(0,00)	0,49(0,00)	0,61(0,00)	0,73(0,00)
KTBoost	0,82(0,00)	**0,86**(0,01)	0,82(0,00)	0,16(0,00)	0,44(0,01)	0,74(0,01)	0,43(0,00)	0,46(0,00)	0,56(0,01)
AdaBoost	**0,26**(0,00)	0,09(0,00)	0,05(0,00)	0,08(0,00)	0,09(0,00)	0,10(0,00)	0,18(0,00)	0,08(0,00)	0,12(0,00)
Gradient Boosting	0,83(0,00)	**0,90**(0,00)	0,86(0,00)	0,16(0,00)	0,54(0,00)	0,79(0,01)	0,47(0,00)	0,55(0,01)	0,63(0,01)

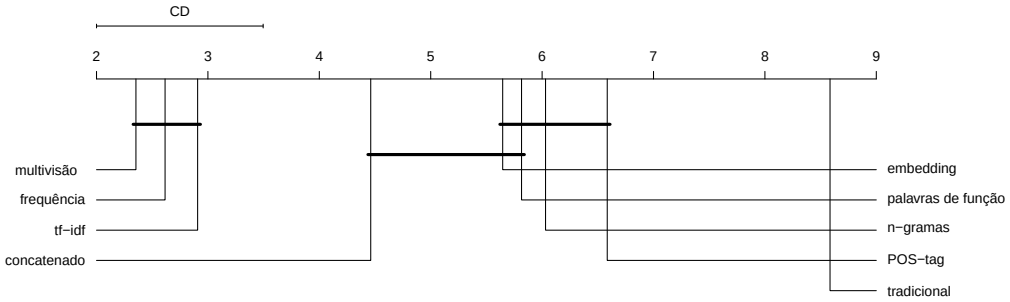
Ademais, vemos que parte dos dados léxicos (frequência do lemma e tf-idf do lemma) permitiram ganhos de performance significativos com alguns métodos. Concomitante a isso, essas visões mantiveram uma certa regularidade nos escores resultantes com outros classificadores.

Os dados também indicam que há uma relativa superioridade da abordagem multivisão em comparação com a visão concatenado. Contudo, um teste estatístico é necessário para verificar quando essa e outras diferenças observadas nos dados listados na Tabela 12 são estatisticamente significativas. Para tal, o teste de Friedman foi conduzido.

A hipótese nula de que não há diferenças nas performances dos métodos listados na Tabela 12, considerando diferentes visões dos dados, foi rejeitada com um p-valor igual a $1,4474 \times 10^{-60}$. Dessarte, o teste *post-hoc* de Nemenyi para comparações emparelhadas foi conduzido. Os p-valores correspondentes são reportados na Tabela 13 e o gráfico de distâncias críticas na Figura 5. Com um nível de significância de 0,05, o teste confirma que a abordagem multivisão é superior a todas as outras abordagens, com exceção de apenas duas que alcançaram performances análogas a essa. É o caso da *frequência do lemma* e *tf-idf do lemma*.

Esses resultados também indicam que explorar a complementariedade dos dados é mais benéfico para atribuição de autoria do que meramente concatenar os dados, prática esta que aumenta consideravelmente a dimensionalidade dos dados. Isso afeta demasiado boa parte dos classificadores, posto que não lidam satisfatoriamente com dados nessas

Figura 5 – Gráfico de distâncias críticas do teste *post-hoc* de Nemenyi com um nível de significância de 0,05



condições. O resultado é o baixo poder de generalização alcançado por eles. Todavia, o método multivisão é particularmente robusto em tais contextos, o que explicaria sua performance superior.

Tabela 13 – P-valores do teste post-hoc de Nemenyi para comparações emparelhadas.

	multivisão	concatenado	frequência	tradicional	palavras funcionais	tf-idf	POS-tag	embedding
concatenado	0,000391							
frequência	1	0,00387						
tradicional	0	$8,12 \times 10^{-14}$	0					
palavras funcionais	$2,08 \times 10^{-11}$	0,11	$9,76 \times 10^{-10}$	$2,93 \times 10^{-7}$				
tf-idf	0,966	0,0333	1	0	$5,11 \times 10^{-8}$			
POS-tag	$6,92 \times 10^{-14}$	0,000338	$1,05 \times 10^{-13}$	0,00105	0,805	$7,56 \times 10^{-13}$		
embedding	$2,59 \times 10^{-10}$	0,249	$1,01 \times 10^{-8}$	$3,43 \times 10^{-8}$	1	$4,28 \times 10^{-7}$	0,576	
n-gramas	$7,56 \times 10^{-13}$	0,0301	$4,19 \times 10^{-11}$	$3,77 \times 10^{-6}$	1	$2,87 \times 10^{-9}$	0,966	0,997

4.5.2.2 Consistência em atribuição de autoria

Passemos agora para a análise da concordância entre os classificadores, o que corresponde precisamente à análise da consistência desses métodos em atribuição de autora usando tanto os dados univisão quanto multivisão. É oportuno recordar que aqui estamos tratando da nossa segunda pergunta de pesquisa. A concordância entre as predições (confiabilidade) é avaliada seguindo a mesma metodologia já adotada nesta dissertação, ou seja, usando a medida Fleiss' Kappa.

A Figura 6 exibe as dispersões nos escores F1 dos classificadores com respeito às diferentes *visões* dos dados. Iniciando pela comparação entre as abordagens multivisão e concatenado, é patente que a primeira resultou em escores mais estáveis, com performance consistente, o que é evidenciado pelo estreito IIQ correspondente. Soma-se a isso o escore mediano superior obtido frente àquele alcançado pela abordagem *concatenado*. Por seu

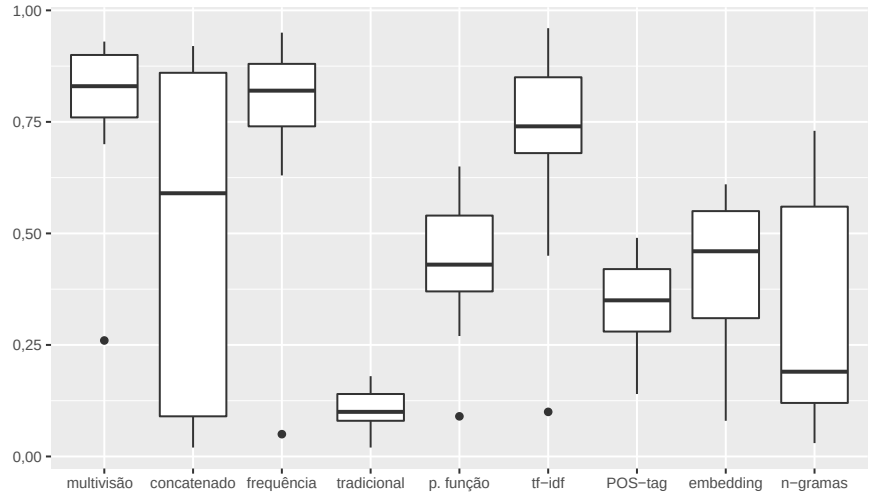


Figura 6 – A dispersão dos escores F1 com respeito à cada abordagem considerada. Como pode ser visto, a amplitude interquartil da abordagem multivisão é compatível com a que pode ser observada para a *frequência do lemma*. São, portanto, performances equivalentes. Por outro lado, o método multivisão é mais estável do que a abordagem *concatenado*.

lado, esta resultou em escores muito mais dispersos, revelando-se bem mais dependente do classificador. Ao mesmo tempo, resultados similarmente bons foram alcançados com as *visões* frequência do lemma e tf-idf do lemma, equiparáveis àqueles obtidos com a abordagem multivisão.

Também é possível perceber uma consistência na classificação com a visão *tradicional*, a qual foi a menos dependente de classificador. Em compensação, essa *visão* resultou em escores substancialmente inferiores àqueles obtidos com as outras. A Figura 6 também mostra que os n-gramas de caractere, a exemplo dos dados concatenados, também foram mais dependentes do classificador nesse experimento, o que é evidenciado pelo amplo IIQ associado. É importante recordar, baseado na revisão de literatura, que os n-gramas em nível de caractere costumam captar mais as sutilizas do estilo autoral em conjuntos de dados mais ruidosos, tais como aqueles onde são mais frequentes os erros de ortografia e usos anormais de pontuação. Todavia, esse tipo de ruído é menos observado em textos formais ou literários. Isso poderia explicar o baixo escore mediano resultante dessa visão dos dados nesse experimento, bem como a instabilidade evidente no desempenho associado.

Novamente, medimos o grau de concordância entre os classificadores nas abordagens univisão e multivisão. Para a tarefa, a abordagem univisão com o melhor desempenho foi considerada: *frequência do lemma*. É pertinente ressaltar que quanto mais os classificadores predizem um mesmo autor, maior a concordância entre eles. Isso se traduz em maior consistência ao atribuir autoria. Além disso, lembremos que um alto grau de concordância acompanhado de alta acurácia se traduz na possibilidade de alcançar melhor performance independente da escolha do algoritmo.

Tabela 14 – Concordância dos classificadores por autor medido usando Fleiss' kappa. Na coluna *Multivisão*, é mostrado o grau de concordância para a abordagem multivisão; já na última coluna, temos a mesma medida para univisão considerando a frequência do lemma. A estatística Kappa varia de -1 a +1 e a interpretação dos valores é dada pela Tabela 10

Autor	**Multivisão**	**Univisão**
Arthur Conan Doyle	0,84	0,28
Charles Darwin	0,55	0,23
Charles Dickens	0,42	0,20
Edith Wharton	0,57	0,29
Horace Greeley	0,69	0,31
James Baldwin	0,65	0,30
Jane Austen	0,50	0,24
John Muir	0,65	0,30
Joseph Conrad	0,72	0,28
Mark Twain	0,76	0,36
Nathaniel Hawthorne	0,60	0,43
Ralph Emerson	0,55	0,25
Robert Louis Stevenson	0,68	0,24
Rudyard Kipling	0,77	0,40

As variações na medida de concordância estão resumidas na Tabela 14. Os dados apresentados indicam que os classificadores passaram a concordar bem mais sobre a autoria dos textos ao serem expostos a dados multivisão. Não houve decréscimos nos valores. Os dois autores com aumentos mais expressivos no nível de concordância estão destacados em negrito. Em um desses casos, o valor da medida triplicou. Ademais, há 8 autores com grau de concordância variando entre substancial e quase perfeita no cenário multivisão, algo que não foi observado no cenário univisão.

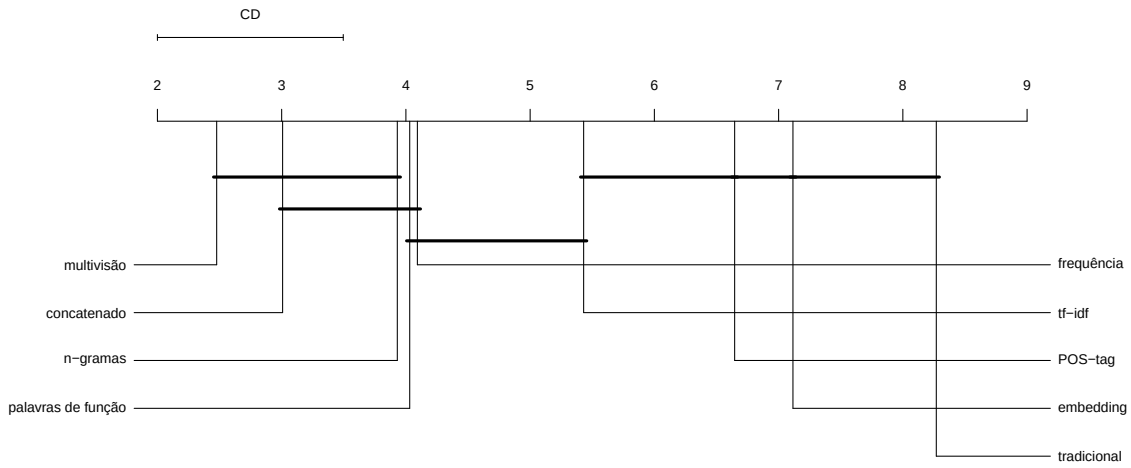
Trata-se de mais uma evidência de que todos os classificadores consistentemente concordam sobre a autoria dos textos e com desempenho similarmente bom, quando utilizam dados multivisão. Tal resultado reforça a ideia de que a busca por um classificador em particular para a tarefa de atribuição de autoria é menos relevante diante da abordagem multivisão. De fato, isso poderia indicar que a abordagem multivisão alivia o ônus da escolha de classificadores na tarefa de atribuição de autoria.

4.5.3 Corpus de sentenças do Supremo Tribunal da Austrália

4.5.3.1 Performance geral da abordagem multivisão

A Tabela 15 apresenta as performances obtidas com os classificadores executados nas diferentes visões desse *corpus*. Comparando os desempenhos das abordagens *multivisão* e *concatenado*, observamos que ambas foram próximas. A diferença mais evidente é que, no

Figura 7 – Gráfico de distâncias críticas do teste *post-hoc* de Nemenyi com um nível de significância de 0,05



caso desta última, os escores são bem mais dispersos, indicando que não houve consistência na classificação. Em contrapartida, os dados *multivisão* resultaram em escores muito mais regulares, o que indica menos dependência do classificador.

Aqui, vemos um resultado interessante no que se refere às visões *frequência do lemma*, *palavras funcionais* e *n-gramas* de caractere. Especificamente, as duas últimas resultaram em melhor desempenho nesse *corpus* do que nos anteriores. Por certo, há uma conexão entre esse resultado e o estudo de Sari, Stevenson e Vlachos (2018), o qual destaca que palavras funcionais, bem como características léxicas, são mais eficazes em textos onde os autores discutem tópicos similares. Essencialmente, é o que acontece nesse *corpus* de julgamentos australianos. No caso de *n-gramas*, o fato de essas características textuais em nível de caractere capturarem tanto o estilo de escrita quanto as preferências dos autores por tópicos específicos (SARI; STEVENSON; VLACHOS, 2018), seria suficiente para explicar tal resultado.

O teste de Friedman foi conduzido para verificar quando há ou não diferenças estatísticas entre os resultados observados na Tabela 15. A hipótese nula é a de que não há diferenças entre as performances, a qual foi rejeitada com um p-valor de $4,46727 \times 10^{-54}$. Isto posto, foi aplicado o teste *post-hoc* de Nemenyi para comparações emparelhadas, com o objetivo de identificarmos onde estão essas diferenças. Os p-valores desse teste são reportados na Tabela 16, a qual mostra as diferenças estatísticas entre cada visão dos dados. O gráfico de distâncias críticas resultante desse teste é exibido na Figura 7, considerando um nível de significância de 0,05.

O teste confirma, com um nível de significância de 0,05, que as performances das abordagens *concatenado* e *n-grama* de caractere são equivalentes àquela obtida com os dados *multivisão*. Isso poderia indicar que, em *corpus* de tópicos específicos, o método

Tabela 15 – Performance dos classificadores com diferentes *visões* e a abordagem multivisão, considerando o *corpus* de Julgamentos do Supremo Tribunal da Austrália. A performance é medida em termos da média dos escores F1. Valores entre parêntesis representam os desvios padrão. Valores em negrito representam o maior escore obtido por um classificador (linha). As *visões* Concatenada e Tradicional são respectivamente a fusão simplificada de todas as *visões* e as características utilizadas em estudos tradicionais conforme descrito no Capítulo 3. A linha rotulada como Stochastic GD refere-se aos resultados do classificador *Stochastic Gradient Descent*. Gaussian NB refere-se ao classificador *Gaussian Naive Bayes*. A coluna rotulada como *p. função* refere-se à *visão* composta por palavras funcionais. A coluna *frequência* representa a frequência do lemma.

classifier	multi-view	concatenated	frequency	traditional	function-words	tf-idf	POS-tag	embedding	n-grams
Decision Tree	**0,68**(0,01)	0,67(0,00)	0,60(0,01)	0,46(0,01)	0,63(0,00)	0,57(0,01)	0,56(0,01)	0,51(0,01)	0,65(0,00)
Random Forest	**0,74**(0,01)	0,63(0,02)	0,59(0,02)	0,47(0,01)	0,67(0,01)	0,56(0,02)	0,61(0,01)	0,55(0,02)	0,65(0,02)
Logistic Regression	**0,80**(0,00)	0,60(0,00)	0,73(0,00)	0,27(0,00)	0,43(0,00)	0,50(0,00)	0,36(0,00)	0,27(0,00)	0,55(0,00)
Stochastic GD	0,79(0,01)	0,43(0,02)	**0,83**(0,01)	0,32(0,01)	0,71(0,04)	0,80(0,01)	0,52(0,01)	0,28(0,01)	0,41(0,03)
SVM	0,82(0,00)	0,52(0,05)	**0,83**(0,00)	0,31(0,00)	0,68(0,00)	0,78(0,00)	0,47(0,00)	0,27(0,00)	0,42(0,07)
KNN	**0,81**(0,00)	0,55(0,00)	0,42(0,00)	0,44(0,00)	0,50(0,00)	0,41(0,00)	0,48(0,00)	0,34(0,00)	0,55(0,00)
Perceptron	0,78(0,00)	0,43(0,00)	**0,82**(0,00)	0,38(0,00)	0,64(0,00)	0,74(0,00)	0,51(0,00)	0,51(0,00)	0,41(0,00)
MLP	0,81(0,00)	**0,84**(0,00)	0,78(0,00)	0,37(0,01)	0,83(0,00)	0,73(0,01)	0,62(0,01)	0,74(0,01)	0,69(0,01)
Gaussian NB	**0,76**(0,00)	0,48(0,00)	0,35(0,00)	0,43(0,00)	0,37(0,00)	0,38(0,00)	0,38(0,00)	0,69(0,00)	0,61(0,00)
XGBoost	0,78(0,00)	**0,87**(0,00)	0,82(0,00)	0,47(0,00)	0,79(0,00)	0,76(0,00)	0,65(0,00)	0,72(0,00)	0,85(0,00)
KTBoost	0,74(0,00)	**0,85**(0,01)	0,78(0,01)	0,47(0,00)	0,79(0,00)	0,74(0,00)	0,66(0,01)	0,68(0,01)	0,83(0,00)
AdaBoost	0,75(0,00)	0,75(0,00)	0,74(0,00)	0,44(0,00)	0,74(0,00)	0,67(0,00)	0,64(0,00)	0,64(0,00)	**0,77**(0,00)
Gradient Boosting	0,79(0,00)	**0,85**(0,00)	0,77(0,00)	0,48(0,00)	0,80(0,00)	0,71(0,00)	0,67(0,00)	0,71(0,00)	**0,85**(0,00)

Tabela 16 – P-valores do teste post-hoc de Nemenyi para comparações

	multivisão	concatenado	frequência	tradicional	palavras funcionais	tf-idf	POS-tag	embedding
concatenado	0,974							
frequência	0,022	0,368						
tradicional	0	$8,39 \times 10^{-14}$	$7,44 \times 10^{-14}$					
palavras funcionais	0,0333	0,453	1	$6,77 \times 10^{-14}$				
tf-idf	$2,8 \times 10^{-8}$	$1,61 \times 10^{-5}$	0,119	$1,24 \times 10^{-7}$	0,0852			
POS-tag	$7,48 \times 10^{-14}$	$1,36 \times 10^{-12}$	$3,77 \times 10^{-6}$	0,0208	$1,86 \times 10^{-6}$	0,217		
embedding	$6,46 \times 10^{-14}$	$8,28 \times 10^{-14}$	$1,12 \times 10^{-8}$	0,283	$4,86 \times 10^{-9}$	0,0135	0,988	
n-gramas	0,0623	0,599	1	$5,93 \times 10^{-14}$	1	0,047	$5,66 \times 10^{-7}$	$1,21 \times 10^{-9}$

multivisão não é tão robusto comparado a essas outras visões. Quiçá, o uso individual de representações léxicas, palavras funcionais ou n-gramas de caractere seja mais conveniente em tais contextos — até mesmo um conjunto multivisão dessas representações.

Os comentários sobre a análise de consistência dos classificadores em atribuição de autoria estão disponíveis a seguir.

4.5.3.2 Consistência em atribuição de autoria

De acordo com a Figura 8, a qual mostra as dispersões nos escores F1, a abordagem multivisão foi a mais consistente e menos dependente de classificador. Isso é evidenciado por seu IIQ mais estreito que o de todas as outras. Se a compararmos com a performance da visão *concatenado*, vemos que esta foi bem mais instável, posto que seu IIQ é muito

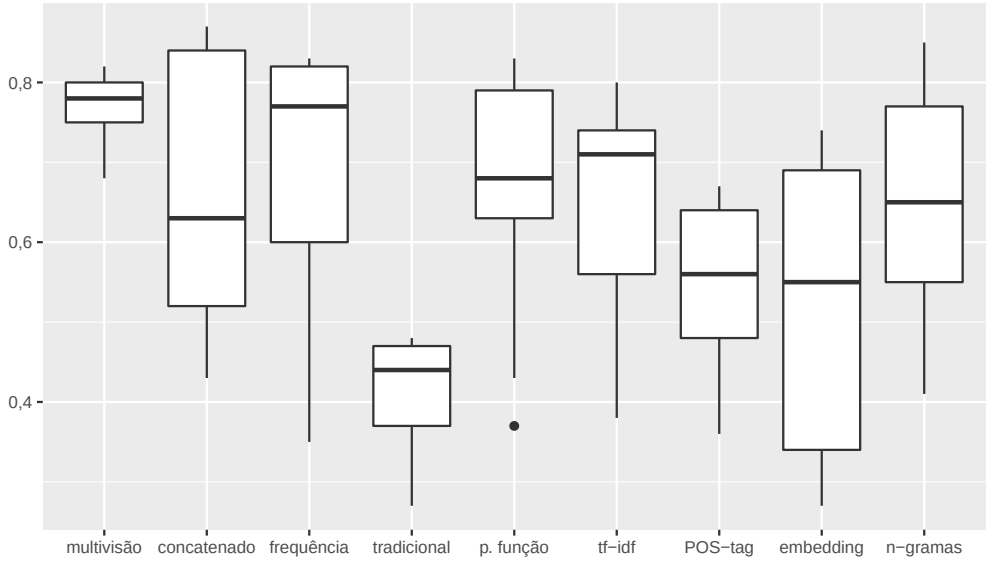


Figura 8 – A dispersão dos escores F1 com respeito à cada abordagem considerada. O estreito IIQ da abordagem multivisão indica que esta foi a mais consistente, com menos dependência do classificador. Por outro lado, não houve diferença estatisticamente significativa entre usar dados multivisão e concatenados nesse experimento, já que ambos os métodos de fusão levaram a resultados equivalentes.

mais amplo. De fato, *concatenado*, junto com *embedding*, foram as abordagens menos consistentes. A visão a ter uma performance pior foi *tradicional*. Tal resultado já era esperado, considerando seu desempenho similar nas análises anteriores.

Uma avaliação da medida de concordância entre os classificadores nas abordagens multivisão e univisão é discutida a seguir. Novamente, tratamos os classificadores como avaliadores e calculamos a confiabilidade entre suas avaliações. A abordagem univisão com melhor escore mediano foi considerada: frequência do lemma. A Tabela 17 exibe os valores correspondentes e a variação entre os dois cenários considerados.

Tabela 17 – Concordância dos classificadores por autor medido usando Fleiss' kappa. Na coluna *Multivisão*, é mostrado o grau de concordância para a abordagem multivisão; já na última coluna, temos a mesma medida para univisão considerando a frequência do lemma. A estatística Kappa varia de -1 a +1 e a interpretação dos valores é dada pela Tabela 10

Autor	**Multivisão**	**Univisão**
McTiernan1965-1975	0,85	0,53
Rich1913-1928	0,85	0,49
Dixon	0,75	0,51

Conforme evidenciado na Tabela 17, a abordagem multivisão resultou no aumento de

concordância entre os classificadores com respeito aos três autores presentes no *corpus*, todos com grau de concordância variando entre substancial e quase perfeita.

Ainda que o método *multivisão* tenha se mostrado menos dependente de classificador neste *corpus*, concomitante ao aumento da concordância de classificação, seu desempenho foi inferior neste experimento se compararmos aos resultados anteriores. É possível, assim, que o método seja menos adequado para problemas de atribuição de autoria onde os autores, especificamente, discutem tópicos específicos. Não obstante, isso também pode ser apenas um efeito do número reduzido de autores neste conjunto de dados em particular.

4.5.4 Corpus de artigos científicos da revista Plos One

4.5.4.1 Performance geral da abordagem multivisão

As performances de cada abordagem nos diferentes algoritmos são apresentadas na Tabela 18. Esses resultados mostram que os conjuntos de características que resultaram nas melhores performances foram *multivisão*, *frequência do lemma* e *tf-idf do lemma*.

Algo que podemos notar facilmente é a baixa dispersão nos escores alcançados pelos classificadores quando expostos a dados multivisão. De fato, não apenas há consistência nos valores dos escores, como também eles foram substanciais. Trata-se de um resultado que não é observado na maior parte das outras abordagens. Performance similar pode ser notada com respeito às abordagens *frequência do lemma* e *tf-idf do lemma*. Exceto essas, todas as outras apresentam desempenhos menos estáveis.

Comparando as performances dos métodos *multivisão* e *concatenado*, vemos que o segundo foi muito mais dependente do classificador. Isso poderia indicar, mais uma vez, que a fusão baseada na mera concatenação dos dados não gera resultados melhores do que um método de fusão que permita aos classificadores explorarem melhor as informações contidas nas múltiplas visões dos dados.

Mesmo que esses resultados indiquem uma aparente superioridade do método multivisão em relação à fusão precoce (concatenado) e à outras abordagens, uma validação estatística foi conduzida para verificar quando as diferenças observadas nas performances listadas na Tabela 18 são estatisticamente significantes.

Começamos por testar a hipótese nula de que não há diferenças entre as performances dos algoritmos usando diferentes visões dos dados. Para tal, o teste de Friedman foi conduzido. A hipótese acima foi rejeitada com um p-valor de $6,49154 \times 10^{-70}$. Por conseguinte, o teste post-hoc de Nemenyi para comparações emparelhadas foi realizado para localizarmos as diferenças entre as performances. Os p-valores correspondentes estão listados na Tabela 19. O correspondente gráfico de distâncias críticas, considerando um nível de significância de 0,05, é exibido na Figura 9

Os testes estatísticos mostram, com um nível de significância de 0,05, que não há diferenças significativas entre as performances resultantes do método multivisão e algumas

Tabela 18 – Performance dos classificadores com diferentes *visões* do corpus de textos da Plos One e a abordagem multivisão. A performance é medida em termos da média dos escores F1. Valores entre parêntesis representam os desvios padrão. Valores em negrito representam o maior escore obtido por um classificador (linha). As *visões* Concatenada e Tradicional são respectivamente a fusão simplificada de todas as *visões* e as características utilizadas em estudos tradicionais conforme descrito no Capítulo 3. A linha rotulada como Stochastic GD refere-se aos resultados do classificador *Stochastic Gradient Descent*. Gaussian NB refere-se ao classificador *Gaussian Naive Bayes*. A coluna rotulada como *p. função* refere-se à *visão* composta por palavras funcionais. A coluna *frequência* representa a frequência do lemma.

classificador	multivisão	concatenado	frequência	tradicional	p. função	tf-idf	POS-tag	embedding	n-gramas
Decision Tree	**0,93**(0,01)	0,90(0,01)	**0,93**(0,01)	0,79(0,01)	0,85(0,01)	0,91(0,01)	0,84(0,01)	0,86(0,00)	0,85(0,01)
Random Forest	**0,96**(0,01)	0,93(0,00)	0,95(0,01)	0,82(0,02)	0,91(0,01)	0,91(0,00)	0,88(0,01)	0,89(0,02)	0,91(0,01)
Logistic Regression	**0,97**(0,00)	0,41(0,00)	0,95(0,00)	0,03(0,00)	0,10(0,00)	0,53(0,00)	0,10(0,00)	0,03(0,00)	0,41(0,00)
Stochastic GD	0,94(0,01)	0,09(0,03)	**0,97**(0,00)	0,13(0,03)	0,60(0,07)	0,96(0,03)	0,31(0,04)	0,34(0,14)	0,10(0,03)
SVM	0,97(0,00)	0,16(0,04)	**0,98**(0,00)	0,06(0,00)	0,62(0,00)	**0,98**(0,00)	0,37(0,00)	0,03(0,00)	0,16(0,04)
KNN	**0,95**(0,00)	0,42(0,00)	0,60(0,00)	0,41(0,00)	0,63(0,00)	0,69(0,00)	0,56(0,00)	0,72(0,00)	0,42(0,00)
Perceptron	0,92(0,00)	0,05(0,00)	**0,96**(0,00)	0,06(0,00)	0,39(0,00)	0,94(0,00)	0,16(0,00)	0,17(0,00)	0,04(0,00)
MLP	**0,97**(0,00)	**0,97**(0,00)	**0,97**(0,00)	0,03(0,00)	0,44(0,03)	**0,97**(0,00)	0,15(0,01)	0,12(0,01)	0,55(0,03)
Gaussian NB	0,95(0,00)	0,96(0,00)	**0,97**(0,00)	0,44(0,00)	0,87(0,00)	**0,97**(0,00)	0,81(0,00)	0,95(0,00)	0,86(0,00)
XGBoost	0,93(0,00)	**0,96**(0,00)	**0,96**(0,00)	0,83(0,00)	0,94(0,00)	0,96(0,00)	0,90(0,00)	0,94(0,00)	0,94(0,00)
KTBoost	**0,97**(0,00)	0,95(0,00)	0,94(0,00)	0,83(0,00)	0,90(0,00)	0,95(0,01)	0,86(0,00)	0,88(0,00)	0,90(0,00)
AdaBoost	**0,72**(0,00)	0,37(0,00)	0,31(0,00)	0,16(0,00)	0,17(0,00)	0,27(0,00)	0,19(0,00)	0,21(0,00)	0,14(0,00)
Gradient Boosting	**0,97**(0,00)	0,95(0,00)	0,96(0,01)	0,84(0,00)	0,90(0,00)	0,95(0,00)	0,88(0,01)	0,89(0,01)	0,92(0,01)

Tabela 19 – P-valores do teste post-hoc de Nemenyi para comparações emparelhadas.

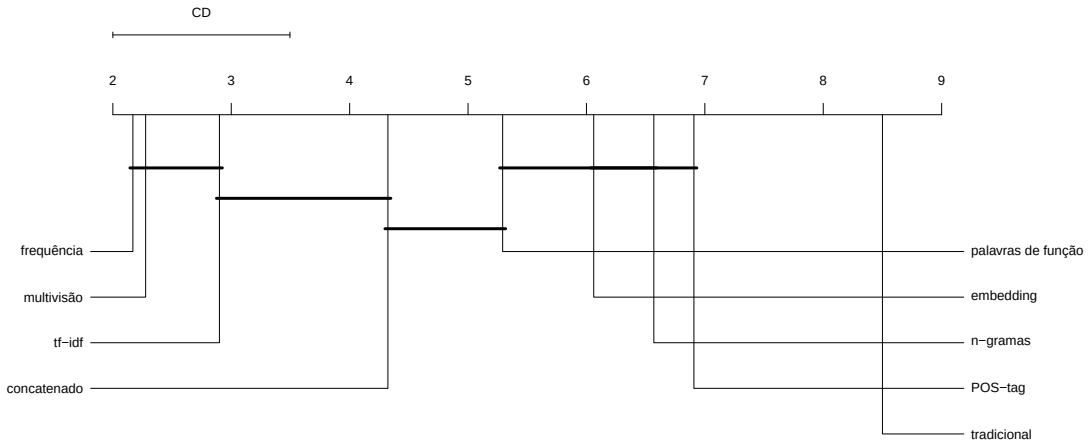
	multivisão	concatenado	frequência	tradicional	palavras funcionais	tf-idf	POS-tag	embedding
concatenado	0,000691							
frequência	1	0,000252						
tradicional	0	$7,44 \times 10^{-14}$	0					
palavras funcionais	$1,24 \times 10^{-8}$	0,531	$2,87 \times 10^{-9}$	$8,75 \times 10^{-10}$				
tf-idf	0,933	0,0747	0,846	0	$2,24 \times 10^{-5}$			
POS-tag	$6,52 \times 10^{-14}$	$2,65 \times 10^{-6}$	$8,56 \times 10^{-14}$	0,0257	0,022	$9,7 \times 10^{-14}$		
embedding	$2,5 \times 10^{-13}$	0,00902	$1,31 \times 10^{-13}$	$1,36 \times 10^{-5}$	0,805	$1,68 \times 10^{-9}$	0,708	
n-gramas	$6,27 \times 10^{-14}$	0,000102	$1,08 \times 10^{-13}$	0,00191	0,163	$8,47 \times 10^{-13}$	0,999	0,98

abordagens léxicas, as quais são: *tf-idf do lemma* e *frequência do lemma*. Por outro lado, há diferenças estatisticamente significativas entre aquela e o método *concatenado*. A avaliação da consistência dos classificadores na atribuição de autoria nos permitiria obter melhores conclusões. Tal avaliação é discutida a seguir.

4.5.4.2 Consistência em atribuição de autoria

A Figura 10 mostra as dispersões nos escores F1 obtidos pelos classificadores durante a execução nas diferentes representações dos dados. De acordo com os dados presentes na figura, vemos que, de fato, as três melhores abordagens foram *multivisão*, *frequência do lemma* e *tf-idf do lemma*. Temos uma visão clara de que essas representações dos dados resultaram em desempenhos muito menos dependentes do classificador, o que é visto pelo seu IIQ muito menor. Esse resultado também é uma evidência de que não houve diferença

Figura 9 – Gráfico de distâncias críticas do teste *post-hoc* de Nemenyi com um nível de significância de 0,05



entre usar dados multivisão e usar características léxicas individualmente.

Todavia, o uso do método de fusão híbrida ainda mostrou-se mais apropriado à situação do que a simples concatenação dos dados, ao ser menos dependente do método de classificação adotado. Ao resultar em performances similarmente boas em diferentes classificadores, o uso de dados multivisão, por certo, pouparia os linguistas e outros especialistas dos encargos envolvidos na escolha do classificador ideal para uma solução em atribuição de autoria.

A avaliação da medida de concordância entre os classificadores ao atribuir autoria com base nas abordagens multivisão e univisão é discutida em seguida. Os mesmos passos adotados nas análises anteriores foram seguidos nesta ocasião também. A visão *frequência do lemma* foi utilizada como abordagem univisão de base, posto que se saiu melhor em comparação com as outras. A Tabela 20 exibe as variações nas medidas de concordância entre os classificadores, considerando tanto o cenário univisão quanto o multivisão. Os resultados indicam que, de fato, o método multivisão proporcionou um aumento no grau de concordância entre os classificadores sobre a autoria dos textos. *Walter R. Tschinkel* foi o único a não representar aumento no grau de concordância no contexto multivisão comparado ao univisão. Outrossim, com exceção deste, todos os autores tiveram grau de concordância quase perfeita no cenário multivisão. Apesar disso, é conveniente salientar que o grau de concordância já era bom no cenário univisão. Mesmo assim, no geral, tais resultados reforçam que o uso de dados multivisão é verdadeiramente vantajoso na implementação de soluções para problemas em atribuição de autoria. Diante dessa conclusão, podemos afirmar que a procura por um classificador ideal para a tarefa tem menos relevância em tais contextos. Isso facilita o trabalho dos linguistas ao propor soluções para problemas em atribuição de autoria, conforme já aludido nas análises anteriores.

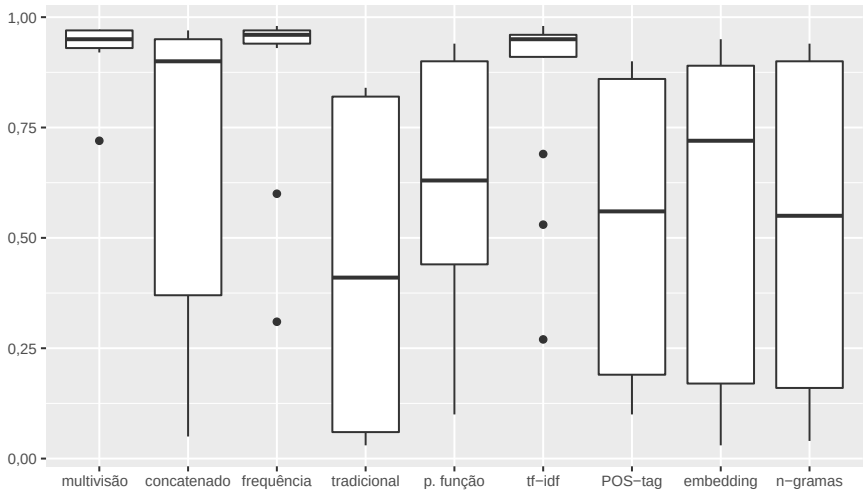


Figura 10 – A dispersão dos escores F1 com respeito à cada abordagem considerada.

Tabela 20 – Concordância dos classificadores por autor medido usando Fleiss' kappa. Na coluna *Multivisão*, é mostrado o grau de concordância para a abordagem multivisão; já na última coluna, temos a mesma medida para univisão considerando a frequência do lemma. A estatística Kappa varia de -1 a +1 e a interpretação dos valores é dada pela Tabela 10

Autor	**Multivisão**	**Univisão**
Richard Robinson	0,95	0,94
Jane Bradbury	0,90	0,78
Paul T. Williams	0,91	0,79
Henry Nicholls	0,94	0,77
Liza Gross	0,90	0,71
Chin-Hsiao Tseng	0,91	0,83
Virginia Gewin	0,94	0,81
Peter J. Hotez	0,85	0,81
John P. A. Ioannidis	0,99	0,79
Hemai Parthasarathy	0,92	0,80
Walter R. Tschinkel	0,94	0,97

4.5.5 Análise de sensibilidade

Os resultados discutidos anteriormente partem de um cenário onde o desempenho resultante da combinação de todas as visões em um conjunto multivisão é comparado com um número pequeno de visões únicas, sem entrar em detalhes a respeito da complementariedade das visões. Não foi discutido o impacto que determinadas visões únicas poderiam ter no conjunto multivisão, ou, ainda, se há algum conjunto multivisão com poucas visões dos dados que poderia levar a modelos com capacidade preditiva competitiva.

Nossa discussão, a partir de agora, tratará justamente desse aspecto do problema. Para tanto, nós medimos os desempenhos de 120 combinações multivisão, de diferentes tamanhos, envolvendo as 7 visões únicas anteriormente mencionadas. Nos próximos parágrafos, discutiremos a respeito do comportamento das combinações e visões únicas cujos resultados despertaram mais atenção nesse experimento.

Nós usamos o teste unilateral de Wilcoxon emparelhado para múltiplas comparações visando comparar os desempenhos entre pares de visões (GARCíA; HERRERA, 2008). Os p-valores foram ajustados usando o critério de FDR (False Discovery Rate) proposto por Benjamini e Hochberg (1995). Tais pares abrangem não apenas visões únicas como também as diferentes combinações que formam conjuntos multivisão. Para esta análise, nós selecionamos o *corpus* de textos da era vitoriana anteriormente descrito. A escolha é justificada pelo fato de este ser o conjunto de dados com o melhor balanceamento entre classes dentre aqueles utilizados nesse estudo, além de conter o maior número de autores.

O problema foi analisado sob duas perspectivas: 1) quantas representações (combinações ou visões únicas) uma determinada visão ou combinação supera em desempenho e 2) quais visões melhoram, ou pioram, a performance resultante de uma combinação. O termo *performance* é empregado aqui para denotar a capacidade de uma representação em gerar modelos mais acurados.

A Figura 11 resume alguns resultados que discutiremos. Nesta figura, cada círculo representa uma visão dentre aquelas que selecionamos para análise. O número contido em cada círculo indica quantas representações aquela visão superou em termos de performance. As interseções entre os círculos representam os conjuntos multivisão resultantes das combinação das visões com base no método proposto nesta dissertação.

A observação mais importante a ser feita corresponde ao desempenho da visão *n-gramas de caractere*. Especificamente, o desempenho desta visão supera a performance de outras 118 representações dos dados — considerando visões únicas e suas combinações. Posto em outros termos, se considerarmos todas as 120 combinações somadas às 7 visões únicas, 93% dessas representações tem um desempenho inferior ao da visão única n-gramas. É oportuno mencionar que, entre as visões que compõem os 7% das representações com desempenho que não foi superado por *n-gramas de caractere*, estão as visões individuais *tf-idf* e *frequência* do lemma, além do conjunto multivisão composto por todas as visões. Isso é consoante com os resultados apresentados na Tabela 12.

(a) *n-gramas*, *tradicional* e *palavras funcionais* (b) *n-gramas*, *tf-idf* e *frequência* do lemma

Figura 11 – Análise da complementariedade das visões. Cada círculo representa uma das visões que selecionamos para esta análise. O número contido nos círculos mostra quantos conjuntos multivisão, bem como visões individuais, foram superados pela respectiva visão. Os números contidos nas interseções indicam quantas visões foram superadas pelo conjunto multivisão composto pelas visões combinadas. Cada interseção entre os círculos representa uma combinação.

Outro importante resultado observado diz respeito ao conjunto multivisão resultante da combinação entre *n-gramas*, *tf-idf* e *frequência* do lemma, o qual teve sua performance superada pela maior parte das visões. Trata-se de uma observação importante, pois, isso indica que as duas últimas diluem o desempenho da primeira. De fato, as performances de *tf-idf* e *frequência* do lemma são pouco significantes, conforme indicado na Figura 11. Portanto, ambas não complementam n-gramas e a performance do conjunto multivisão composto por esses elementos é pouco expressiva. Esses resultados apontam, portanto, para o uso individual dessas representações como sendo uma estratégia mais conveniente. Outra importante constatação a respeito da visão n-gramas é que nossos resultados corroboram o que é frequentemente mencionado na literatura a respeito da robustez desse recurso linguístico para representar estilos de escrita.

As visões *palavras funcionais* e *tradicional* também mostraram um comportamento interessante. Ambas superaram os desempenhos 75% e 88% das representações, respectivamente. Todavia, uma combinação entre as duas não gerou ganhos, fazendo esse número cair, conforme mostrado na Figura 11. Esse resultado já era esperado, posto que *palavras funcionais* já é um subconjunto da visão *tradicional*. Portanto, ambas as visões não se complementam.

A combinação de *tradicional* com *n-gramas* gerou o melhor conjunto multivisão obtido nesse experimento com os textos da era vitoriana, sendo complementar a várias outras combinações. Isso sugere que, muito provavelmente, esse conjunto tenha influenciado em grande medida os resultados anteriores. Outros conjuntos multivisão a mostrarem desempenhos relevantes foram *palavras funcionais-tradicional* e *tf-idf do lemma-tradicional*.

Tabela 21 – Na coluna da direita desta tabela, estão algumas combinações multivisão que resultam em modelos melhores após a inclusão dos pares indicados na coluna da esquerda. Em outras palavras, a tabela aponta algumas combinações que se complementam.

Combinação	**Melhora o desempenho de**
{*tradicional*, *n-gramas*}	{*POS-tags*, *Embedding*, *Palavras funcionais*} {*POS-tags*, *Embedding*} {*POS-tags*, *Palavras funcionais*} {*POS-tags*, *tf-idf*} {*POS-tags*, *tf-idf*, *frequência*} {*palavras funcionais*, *tf-idf*, *frequência*}
{*palavras funcionais*, *tradicional*}	{*POS-tags*, *n-gramas*} {*POS-tags*, *n-gramas*, *tf-idf*} {*POS-tags*, *n-gramas*, *tf-idf*, *frequência*} {*POS-tags*, *n-gramas*, *frequência*} {*POS-tags*, *tf-idf*} {*POS-tags*, *tf-idf*, *frequência*} {*POS-tags*, *frequência*} {*Embedding*, *n-gramas*} {*Embedding*, *n-gramas*, *tf-idf*} {*Embedding*, *n-gramas*, *tf-idf*, *frequência*} {*Embedding*, *n-gramas*, *frequência*} {*Embedding*, *tf-idf*} {*Embedding*, *tf-idf*, *frequência*} {*Embedding*, *frequência*}
{*tradicional*, *tf-idf do lemma*}	{*POS-tags*, *Embedding*, *Palavras funcionais*} {*POS-tags*, *Embedding*, *Palavras funcionais*, *frequência*} {*POS-tags*, *Embedding*} {*POS-tags*, *Embedding*, *frequência*} {*POS-tags*, *Palavras funcionais*} {*POS-tags*, *Palavras funcionais*, *frequência*} {*POS-tags*, *frequência*} {*Embedding*, *Palavras funcionais*} {*Embedding*, *Palavras funcionais*, *frequência*} {*Embedding*, *frequência*} {*Palavras funcionais*, *frequência*}

A Tabela 21 mostra algumas combinações que passam a gerar melhores resultados quando essas representações são incluídas. Em outras palavras, isso indica que os conjuntos multivisão resultantes da combinação entre os elementos presentes na primeira e na segunda coluna da referida tabela possivelmente ajudariam a produzir modelos melhores para atribuição de autoria.

5 CONCLUSÃO E FUTURAS DIREÇÕES

Este trabalho propõe uma nova abordagem multivisão baseada em instâncias para atribuição de autoria. O método faz uso de um framework de fusão híbrido que é independente de modelo e baseia-se no aprendizado supervisionado de representações paralelas dos dados, as quais foram chamadas de visões nesse estudo. Tal estrutura permite explorar a complementaridade dos diferentes tipos de características dos dados que representam o estilo de escrita de um autor. A abordagem proposta não requer que um único tipo de recurso linguístico seja o mais adequado para atribuir autoria com base no estilo de escrita. Pelo contrário, os resultados experimentais indicam que a combinação de características léxicas, de caractere, sintáticas e semânticas é muito eficaz para atribuição de autoria.

No estudo, foram levantadas duas perguntas de pesquisa: (1) que a combinação de diferentes características dos dados em um conjunto multivisão melhora a acurácia dos classificadores em atribuição de autoria; e (2) que isso também contribui para aliviar o ônus da escolha do classificador ideal para lidar com o problema, uma vez que os classificadores se tornam mais consistentes quando utilizam dados multivisão. Uma série de experimentos foi conduzida para validar essas hipóteses. Tais experimentos foram realizados em 4 diferentes conjuntos de dados com diferentes números de autores e de textos para verificar se a abordagem proposta: (1) melhora a acurácia geral da classificação; e (2) gera modelos mais consistentes, que realmente percebem o estilo de escrita de um autor e, assim, tendem a concordar mais entre si sobre a autoria dos textos.

De fato, a Literatura em aprendizagem multivisão assume que modelos com maior capacidade preditiva podem ser obtidos ao considerarmos as múltiplas representações dos dados. A mesma literatura não menciona, todavia, que modelos mais consistentes podem ser obtidos. Conforme mostrado em nosso estudo, ao treinar os classificadores com múltiplas representações dos mesmos dados, fomos capazes de obter modelos de atribuição de autoria mais consistentes do que aqueles que adotam apenas uma única representação dos dados. Essa é, portanto, uma importante contribuição do nosso método.

Foi utilizada uma variedade de algoritmos no estado da arte frequentemente usados em classificação de textos e em estudos de atribuição de autoria para avaliar o desempenho da abordagem proposta. Os experimentos foram repetidos 5 vezes e, para cada turno, a performance dos modelos foi avaliada usando o esquema de validação cruzada com 5-folds. Os resultados experimentais revelaram que a abordagem multivisão proposta nesta dissertação alcançou melhor performance do que a simples concatenação dos dados em 3 dos 4 conjuntos de dados utilizados. Além disso, os testes estatísticos de hipótese confirmaram que a fusão das visões em representações multivisão gera resultados superiores à maior parte das abordagens univisão consideradas em todos os conjuntos de dados. Ao utilizarmos textos de tópico específico, o uso de dados multivisão revelou-se menos

vantajoso, ainda que tenha resultado em desempenhos mais estáveis. Tais cenários parecem beneficiar em maior medida características mais específicas dos dados. É o exemplo de n-gramas de caractere e palavras funcionais, além de certos tipos de dados léxicos, de modo que combiná-las com outras características textuais menos relevantes em tais contextos não pareceu proporcionar ganhos significativos.

Em nossos resultados, foi observado que as visões *frequência do lemma* e *tf-idf do lemma* ocasionalmente resultaram em modelos com desempenhos tão bons quanto o do método multivisão. Trata-se de um resultado consoante com o que é frequentemente disposto na literatura a respeito das características léxicas. Por serem suficientemente robustas, sobretudo em textos de tópico geral, esse tipo de recurso linguístico já têm uma longa história na área de atribuição de autoria. Em textos de tópico específico, vale ressaltar, essa influência das características léxicas é mais diluída entre as outras visões dos dados, conforme evidenciado com nosso experimento utilizando o *corpus* de textos jurídicos.

No que se refere à consistência dos métodos em atribuição de autoria, observou-se um expressivo aumento no grau de acordo entre os classificadores em torno da autoria dos textos quando usamos dados multivisão. Foi observado, ainda, que na maior parte das ocasiões o grau de acordo variou entre acordo substancial e quase perfeito. Outra constatação importante é que o método multivisão frequentemente resultou em desempenhos de classificação mais estáveis em comparação com o método de fusão precoce e outras abordagens de visão única dos dados. Isso se traduz em uma abordagem que é menos dependente de classificadores. Trata-se de um resultado notável, uma vez que isso significa diminuir o esforço para os linguistas e outros especialistas em escolher um classificador para criar soluções em atribuição de autoria.

Diante desses resultados, concluímos que a abordagem multivisão foi, de fato, mais eficaz do que algumas abordagens univisão frequentemente empregadas na literatura, sobretudo aquelas que se baseiam na concatenação dos dados. O método é: (1) estatisticamente significativamente mais acurado do que boa parte das abordagens comparadas; e (2) é mais consistente ao longo dos diferentes tipos de classificadores, tonando-se, assim, menos dependente de algoritmos.

O método é limitado a trabalhar com conjuntos de dados contendo quantidades mais brandas de amostras, posto que o custo computacional da abordagem tornar-se demasiadamente elevado com grandes massas de dados. É pertinente lembrar, com base nos resultados experimentais, que o método funciona bem com poucas amostras, além de ser particularmente resistente aos efeitos da alta dimensionalidade.

Há uma outra possível limitação especificamente relacionada ao tamanho dos textos utilizados. É possível que a performance resultante do conjunto multivisão seja prejudicada pelo uso de textos muito curtos, por exemplo, com 140 caracteres ou menos. Isto porque algumas características textuais distribuídas ao longo das diferentes visões passam a exercer menos influência na percepção do estilo do autor como um efeito da quantidade

reduzida de textos. Conforme mencionado no Capítulo 2, as características sintáticas, bem como aquelas baseadas em caracteres, exercem mais influência na percepção do estilo de escrita do que outras ao lidar com textos curtos. Tal fato sugere que, possivelmente, o uso de uma abordagem multivisão baseada na combinação destes dois recursos linguísticos seja mais indicado para textos curtos do que um conjunto composto por mais visões dos dados, já que estaríamos eliminando características que pouco contribuem para uma boa performance nesse contexto. Todavia, este problema não foi abordado em nosso estudo. Portanto, uma possível futura direção seria estender esse trabalho para investigar a performance do método multivisão em cenários onde os textos sejam curtos, tais como aqueles presentes em plataformas de *microblogging* e similares.

REFERÊNCIAS

AREFIN, A. S.; VIMIEIRO, R.; RIVEROS, C.; CRAIG, H.; MOSCATO, P. An information theoretic clustering approach for unveiling authorship affinities in shakespearean era plays and poems. *PLOS ONE*, Public Library of Science, v. 9, n. 10, p. 1–12, 10 2014. Disponível em: <https://doi.org/10.1371/journal.pone.0111445>.

BALTRUŠAITIS, T.; AHUJA, C.; MORENCY, L. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 41, n. 2, p. 423–443, 2019. Disponível em: <https://doi.org/10.1109/TPAMI.2018.2798607>.

BALTRUšAITIS, T.; AHUJA, C.; MORENCY, L. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 41, n. 2, p. 423–443, 2019. Disponível em: <https://doi.org/10.1109/TPAMI.2018.2798607>.

BENJAMINI, Y.; HOCHBERG, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, [Royal Statistical Society, Wiley], v. 57, n. 1, p. 289–300, 1995. ISSN 00359246. Disponível em: <http://www.jstor.org/stable/2346101>.

BREIMAN, L. Random forests. *Machine Learning*, v. 45, n. 1, p. 5–32, 2001. Disponível em: <https://doi.org/10.1023/A:1010933404324>.

CAO, H.; BERNARD, S.; SABOURIN, R.; HEUTTE, L. Random forest dissimilarity based multi-view learning for Radiomics application. *Pattern Recognition*, v. 88, p. 185 – 197, 2019. ISSN 0031-3203. Disponível em: <http://www.sciencedirect.com/science/article/pii/S003132031830400X>.

CAO, H.; BERNARD, S.; SABOURIN, R.; HEUTTE, L. *A Novel Random Forest Dissimilarity Measure for Multi-View Learning*. 2020.

CHAO, G.; SUN, S. Consensus and complementarity based maximum entropy discrimination for multi-view classification. *Information Sciences*, v. 367-368, p. 296 – 310, 2016. ISSN 0020-0255. Disponível em: <http://www.sciencedirect.com/science/article/pii/S002002551630411X>.

CHEN, T.; GUESTRIN, C. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016. (KDD '16), p. 785–794. ISBN 978-1-4503-4232-2. Disponível em: <http://doi.acm.org/10.1145/2939672.2939785>.

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.

CRUZ, R. M.; CAVALCANTI, G. D.; TSANG, I. R.; SABOURIN, R. Feature representation selection based on classifier projection space and oracle analysis. *Expert Systems with Applications*, v. 40, n. 9, p. 3813 – 3827, 2013. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417412013383>.

CURTIS, C.; SHAH, S. P.; CHIN, S. F.; TURASHVILI, G.; RUEDA, O. M.; DUNNING, M. J.; SPEED, D.; LYNCH, A. G.; SAMARAJIWA, S.; YUAN, Y.; GRAF, S.; HA, G.; HAFFARI, G.; BASHASHATI, A.; RUSSELL, R.; MCKINNEY, S.; GROUP, M.; LANGEROD, A.; GREEN, A.; PROVENZANO, E.; WISHART, G.; PINDER, S.; WATSON, P.; MARKOWETZ, F.; MURPHY, L.; ELLIS, I.; PURUSHOTHAM, A.; BORRESEN-DALE, A. L.; BRENTON, J. D.; TAVARE, S.; CALDAS, C.; APARICIO, S. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, v. 486, n. 7403, p. 346–352, 2012.

DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, JMLR.org, v. 7, p. 1–30, 2006. ISSN 1532-4435.

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

DUQUE, A. B.; CARVALHO, F. d. A. T. de; VIMIEIRO, R. A multiview clustering approach for mining authorial affinities in literary texts. In: *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*. [S.l.: s.n.], 2019. p. 818–823.

EBRAHIMPOUR, M.; PUTNIņš, T. J.; BERRYMAN, M. J.; ALLISON, A.; NG, B. W.-H.; ABBOTT, D. Automated authorship attribution using advanced signal classification techniques. *PLOS ONE*, Public Library of Science, v. 8, n. 2, p. 1–12, 02 2013. Disponível em: <https://doi.org/10.1371/journal.pone.0054998>.

FLEISS, J. L. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, v. 76, p. 378–382, 1971.

Folger Shakespeare Library. *Shakespeare's Plays, Sonnets and Poems*. n.d. Accessed on 2020-04-28. Disponível em: <https://shakespeare.folger.edu/>.

FOURKIOTI, O.; SYMEONIDIS, S.; ARAMPATZIS, A. Language models and fusion for authorship attribution. *Information Processing & Management*, v. 56, n. 6, p. 102061, 2019. ISSN 0306-4573. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0306457318306769>.

FROEHLICH, H. *Replication Data for: "Dramatic Structure and Social Status in Shakespeare's Plays"*. Harvard Dataverse, 2020. Disponível em: <https://doi.org/10.7910/DVN/PPD3OO>.

FU, Y.; CAO, L.; GUO, G.; HUANG, T. S. Multiple feature fusion by subspace learning. In: *Proceedings of the 2008 International Conference on Content-Based Image and Video Retrieval*. New York, NY, USA: Association for Computing Machinery, 2008. (CIVR '08), p. 127–134. ISBN 9781605580708. Disponível em: <https://doi.org/10.1145/1386352.1386373>.

GARCíA, S.; HERRERA, F. An extension on "statistical comparisons of classifiers over multiple data sets"for all pairwise comparisons. *Journal of Machine Learning Research*, v. 9, 12 2008.

GIANCINTO, G.; ROLI, F. Design of effective neural network ensembles for image classification purposes. *Image Vision And Computing Journal*, v. 19, p. 699–707, 2001.

GÖNEN, M.; ALPAYDIN, E. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, v. 12, n. 64, p. 2211–2268, 2011. Disponível em: <http://jmlr.org/papers/v12/gonen11a.html>.

GRIEVE, J. Quantitative Authorship Attribution: An Evaluation of Techniques. *Literary and Linguistic Computing*, v. 22, n. 3, p. 251–270, 07 2007.

GUIRAUD, P. Probl èmes et méthodes de la statistique linguistique. In: . [S.l.: s.n.], 1960.

GUNGOR, A. *Benchmarking authorship attribution techniques using over a thousand books by fifty Victorian era novelists*. Tese (Doutorado) — Purdue University, 2018.

JOCKERS, M.; WITTEN, D. A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, v. 25, p. 215–223, 05 2010.

JUOLA, P.; SOFKO, J.; BRENNAN, P. A Prototype for Authorship Attribution Studies. *Literary and Linguistic Computing*, v. 21, n. 2, p. 169–178, 04 2006.

KOCHER, M.; SAVOY, J. Distributed language representation for authorship attribution. *Digital Scholarship in the Humanities*, v. 33, n. 2, p. 425–441, 08 2017. ISSN 2055-7671. Disponível em: <https://doi.org/10.1093/llc/fqx046>.

KOPPEL, M.; SCHLER, J.; ARGAMON, S. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, v. 60, n. 1, p. 9–26, 2009. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20961>.

KOPPEL, M.; SCHLER, J.; BONCHEK-DOKOW, E. Measuring differentiability: Unmasking pseudonymous authors. *Journal of Machine Learning Research*, v. 8, p. 1261–1276, 2007. ISSN 1532-4435.

KRIPPENDORFF, K. *Content analysis: An introduction to its methodology*. 4th. ed. [S.l.]: Sage publications, 2018.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. *Biometrics*, v. 33, n. 1, p. 159–174, 1977.

LAYTON, R.; WATTERS, P.; DAZELEY, R. Automated unsupervised authorship analysis using evidence accumulation clustering. *Natural Language Engineering*, Cambridge University Press, v. 19, n. 1, p. 95–120, 2013.

MALVERN, D.; RICHARDS, B.; CHIPERE, N.; DURÁN, P. *Lexical Diversity and Language Development: Quantification and Assessment*. Palgrave Macmillan UK, 2004. ISBN 9780230511804. Disponível em: <https://books.google.com.br/books?id=R78WDAAAQBAJ>.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; Dean, J. *Efficient Estimation of Word Representations in Vector Space*. 2013. arXiv:1301.3781 p.

MOSTELLER, F.; WALLACE, D. *Inference and Disputed Authorship: The Federalist*. [S.l.]: Addison-Wesley Publishing Company, 1964.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

SARI, Y.; STEVENSON, M.; VLACHOS, A. Topic or style? exploring the most useful features for authorship attribution. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, 2018. p. 343–353. Disponível em: <https://www.aclweb.org/anthology/C18-1029>.

SEROUSSI, Y.; ZUKERMAN, I.; BOHNERT, F. Authorship attribution with topic models. *Computational Linguistics*, MIT Press, Cambridge, MA, USA, v. 40, n. 2, p. 269–310, jun. 2014. ISSN 0891-2017. Disponível em: <https://doi.org/10.1162/COLI_a_00173>.

SIGRIST, F. *KTBoost: Combined Kernel and Tree Boosting*. 2019. arXiv:1902.03999 p.

STAMATATOS, E. Authorship attribution based on feature set subspacing ensembles. *International Journal on Artificial Intelligence Tools*, v. 15, n. 05, p. 823–838, 2006.

STAMATATOS, E. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, v. 60, n. 3, p. 538–556, 2009. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21001>.

SUN, S. A survey of multi-view machine learning. *Neural computing and applications*, Springer, v. 23, n. 7-8, p. 2031–2038, 2013.

XU, C.; TAO, D.; XU, C. A Survey on Multi-view Learning. *arXiv e-prints*, p. arXiv:1304.5634, 04 2013.

ZHAO, J.; XIE, X.; XU, X.; SUN, S. Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, v. 38, p. 43 – 54, 2017.

ZHENG, R.; LI, J.; CHEN, H.; HUANG, Z. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, v. 57, n. 3, p. 378–393, 2006. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20316>.