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

Daniel Cirne Vilas-Boas dos Santos[1], **Cleber Zanchettin**[1]

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

{dcvs,cz}@cin.ufpe.br

Abstract. *The growth of digital documents, associated with their usage in several knowledge areas requires computational resources for its comprehension and analysis. The literature proposes distinguishing authors by their writing style and keywords. However, these studies do not fall under the educational context for English text. This research is unique because it explores authorship analysis within a dataset composed of school activities written by undergraduate students in Portuguese. Due to the low number of samples, we used robust journalistic datasets as references. Through the experiments, it was verified that in restricted domains, stylometric representations are superior to textual representations, which suffer influence from the topic in broader corpora. We found out that the Extremely Randomized Trees associated with the proposed stylometric features overcome every other model tested, in all the datasets, reaching an average of 0.70 accuracy and 0.81 AUC.*

Resumo. *O aumento no volume de documentos digitais associado ao seu uso no processo de verificação de aprendizagem demanda recursos computacionais para compreensão e análise de autoria. A literatura propõe distinguir os autores pelo estilo de escrita e palavras-chave. Entretanto, estes trabalhos não estão inseridos no contexto educacional e são majoritariamente em inglês. Este artigo se distingue por explorar a verificação de autoria numa base de atividades pedagógicas escritas na língua portuguesa. Devido ao baixo volume de exemplos, usamos bases jornalísticas robustas como referência. Por meio dos experimentos verificamos que em domínios específicos, representações baseadas em características de estilo são superiores à abordagens textuais, que sofrem influência do tópico em corpora mais abrangentes. Este trabalho revelou que o modelo Extremelly Randomized Trees associado as características de estilo propostas foi superior aos demais modelos em todas as bases utilizadas, alcançando uma média de 70% na taxa de acerto e AUC 0.81.*

1. Introdução

A utilização de recursos tecnológicos na educação, apesar de trazer diversos benefícios, também possui mazelas inerentes a essa inserção. O número de práticas prejudiciais, como a reprodução de informações na web, comércio de trabalhos escolares e divisão paralela de atividades entre alunos têm se tornado mais frequentes [Curtis and Tremayne 2019] [Singh and Remenyi 2016]. Os recursos na *web* são excelentes fontes para pesquisa, porém compreender as informações e retratá-las com seu

DOI: 10.5753/sbie.2021.217413

próprio entendimento é uma importante parte do processo de aprendizado. A facilidade de comprar ou reproduzir conteúdos, além de sedutora, é difícil de ser identificada durante a correção das atividades pelos avaliadores, prejudicando todos envolvidos no processo de ensino e aprendizagem.

Desta forma, a criação de recursos computacionais capazes de mitigar ou apoiar o processo de verificação da aprendizagem se fazem necessários. Por estarmos tratando da comunicação escrita, a classificação de documentos com foco na autoria (análise de autoria) tem se mostrado como uma estratégia eficiente em problemas similares [Tempestt et al. 2017]. Nossos principais desafios são, $i)$ o pequeno número de documentos por autor, pois é provável que só exista um volume significativo de documentos por estudante após algum tempo de curso; $ii)$ domínio restrito ao conteúdo do curso ou disciplina, que colabora com a presença de muitos documentos semelhantes; e $iii)$ uso indiscriminado de ferramentas de busca, que leva a construção de textos compostos por excertos de outros autores, dificultando a análise de estilo.

Este trabalho é baseado na dissertação de mestrado de um dos autores [dos Santos 2021], e tem como propósito explorar a atividade de atribuição de autoria em atividades pedagógicas escritas por estudantes para suportar os avanços da tecnologia na educação e desencorajar práticas de atrapalhem o processo de verificação da aprendizagem de maneira não punitiva [Botelho and da Silva Martins 2020]. Após a condução de um estudo de caso que avaliou diversas abordagens para resolução da atividade, a solução proposta se fundamentou no uso de PLN para pré-processamento e extração de características de estilo na língua portuguesa associados à comitês de árvores de decisão extremamente aleatórias [Geurts et al. 2006a].

O artigo foi estruturado em cinco seções. Após a introdução do trabalho, na segunda seção é apresentada a fundamentação teórica, na terceira seção temos a metodologia de pesquisa, incluindo o detalhamento das características construídas, a quarta seção detalha os experimentos executados e resultados obtidos, e na última fazemos as considerações finais.

2. Fundamentação

A análise de autoria contribui para computação forense, combate ao plágio e solução de casos com autoria contestada ou anônima [Varela 2017]. Seu avanço está diretamente relacionado ao desenvolvimento da Aprendizagem de Máquina (AM) e o Processamento de Linguagem Natural (PLN). Na análise de autoria, se destacam as tarefas de atribuição, verificação e *profiling* de autoria, que validam, identificam ou caracterizam os autores dos documentos respectivamente [Stamatatos 2009] [Juola 2008].

A estilometria defende o uso de características de estilo para quantificar e definir o estilo de escrita dos autores. Segundo [Stamatatos 2009] e [Varela 2017], cada autor possui um estilo único de escrita, que é composto por múltiplos fatores, como vícios de linguagem, uso e composição de pontuação, palavras, frases e parágrafos, legibilidade, concordância e riqueza de vocabulário. Porém, a produção de tais características exige esforço humano para extração, construção e avaliação. A seleção de características é apontada como um dos maiores desafios da estilometria, pois não há consenso sobre quais são mais relevantes, variando bastante de acordo com o problema [Neal et al. 2017].

Para extração dessas características, se faz necessária a compreensão da linguagem

a nível de *tokens*, frases e parágrafos. Isso pode ser alcançado por atividades do PLN, baseadas em corpus, sistemas de referência léxica, expressões regulares ou análise e geração de regras gramaticais [Chowdhury 2003]. Com isso, os papéis morfossintáticos das palavras, estrutura frasal e principais entidades das frases revelam as principais características dos documentos. Assim, os documentos passam a ser representados por vetores numéricos interpretáveis pelos modelos de AM. Alternativas comuns para representação de documentos são vetores de contagem de palavras por TF-IDF (*Term Frequency–Inverse Document Frequency*) ou *Bag of Words* [Goldberg 2017].

As árvores de decisão (AD) são um conjunto de algoritmos de AM representados numa estrutura de árvore. O comitê de árvores extremamente aleatórias (*Extra randomized Trees* - ET), é um *ensemble* de ADs que constrói árvores com pontos de corte aleatórios a partir de toda a base de dados [Geurts et al. 2006b], gerando um conjunto robusto de árvores profundas. A associação de modelos baseados em AD a características de estilo tem resultados comprovados na literatura para atividades de atribuição e verificação de autoria [Khonji et al. 2015] [Pacheco et al. 2015] [Maitra et al. 2016]. Nota-se também que a associação da estilometria com modelos do tipo máquina de vetores de suporte (SVM), rede neurais recorrentes e aprendizagem não supervisionada vêm sendo empregadas para atividades de atribuição e verificação de autoria [Bevendorff et al. 2020a] [Yang et al. 2018].

3. Metodologia

Apesar da existência de *datasets* relevantes voltados para atribuição de autoria [Bevendorff et al. 2020b] [Rangel et al. 2020] [Varela 2017], nenhum destes atingiu por completo os requisitos desta pesquisa. Desta forma, por meio de uma parceria por professores de ensino superior, construímos uma base composta por atividades pedagógicas escritas por estudantes. Ao remover trabalhos realizados em grupo ou com pelo menos 3 exemplos por autor, a base foi reduzida a 84 documentos distribuídos entre 16 autores. Dado o baixo volume de exemplos e grande disparidade no número de trabalhos e *tokens* por autor na base de **estudantes**, utilizamos outras duas bases para efeitos comparativos. i) **notícias**: composta por colunas de um renomado portal de notícias brasileiro, e ii) **Varela**: coleção com 3.000 textos jornalísticos, escritos por 100 autores e distribuídos por assunto em 10 categorias [Varela 2017].

Durante o pré-processamento das bases, garantimos o anonimato e eliminamos vieses por meio da remoção de termos e palavras-chave que pudessem estar vinculados à autoria. Utilizamos representações textuais e numéricas para retratar os documentos. Para a textual, utilizamos TF-IDF e *word-embeddings* e na numérica, extraímos 74 características de estilo. Para construção destas características, foi implementado um algoritmo[1] capaz de extrair, pré-processar, computar e exportar estes valores. O mesmo foi disponibilizado sob licença de código aberto.

Para extração de características, fizemos uso massivo de recursos de Processamento de Linguagem Natural e Aprendizagem de Máquina. As características foram separadas em grupos lógicos. i) **lexicais**: representam a estrutura da escrita dos autores, compreendendo frequência e tamanho de parágrafos, frases e sílabas por palavra. Combinamos o algoritmo de separação de sílabas proposto por [SILVA 2011] com a

[1] https://github.com/daanielvb/text-extractor

implementação da biblioteca *Pyphen*[2] para obter uma separação mais precisa. ii) **caracteres e palavras-chave**: representam a frequência de aparição de palavras ou pontuação ao longo do texto. No grupo de termos pré-definidos, mensuramos pontuações menos comuns, conectivos lógicos e palavras capitalizadas. Também calculamos a frequência de aparições dos *n-grams* ($n = [2, ..., 5]$) mais frequentes do *corpora* em cada documento (*top-grams*); iii) **sintáticas**: indicam o papel morfo-sintático dos *tokens* e frases no documento. Para anotação sintática, foi treinado um *POS-Tagger* na língua portuguesa sobre a base MACMORPHO [Aluísio et al. 2003]. Desta forma, obtivemos a frequência de classes gramaticais e pudemos identificar frases verbais e nominais por meio do *chunking* com expressões regulares. Seguindo o trabalho de [Scarton and Aluísio 2010], grupos de palavras de conteúdo e funcionais foram mensuradas. Características mais refinadas, como flexões de gênero, plural, discurso e tempo verbal foram capturadas com adaptações da biblioteca SpaCy [Honnibal and Montani 2017]. iv) **semânticas**: quantificam os diversos papéis semânticos das palavras de acordo com o contexto. Aqui utilizamos um classificador pré-treinado para reconhecimento de entidades nomeadas. O mesmo foi disponibilizado por [Pires 2017], e se baseia no *dataset* HAREM [Freitas et al. 2010], sendo capaz de rotular os termos de acordo com as categorias de entidade do HAREM. v) **riqueza de vocabulário**: contém índices que indicam a legibilidade e variedade lexical do documento, por meio do cálculo de *hapax legomena* (local e global), repetição, legibilidade (Flesh-Kincaid) [Martins et al. 1996] e tradicionais medidas de riqueza de vocabulário [Tweedie and Baayen 1998]. vi) **aplicação**: características relativas ao domínio e aplicação, neste caso incluímos a incidência de erros ortográficos, *stopwords* e *collocations* ($n = [2, ..., 4]$).

4. Experimentos e discussões

Os experimentos foram divididos nas seguintes fases: 1) análise exploratória, 2) avaliação de modelos, 3) otimização das soluções, e 4) compreensão da solução. Para construir visualizações interpretáveis, utilizamos PCA, TSNE [Van der Maaten and Hinton 2008] e LSA [Deerwester et al. 1990] para redução de dimensionalidade, observou-se que as fronteiras de separabilidade dos exemplos por autoria na base de estudantes era muito pequena, tanto nas representações textuais como de estilo. Por outro lado, nas bases jornalísticas a separabilidade e segregação são mais evidentes (Figura 1).

Ao aplicar algoritmos de *clustering soft* (Fuzzy c-médias) e *hard* (K-médias) a partir de conhecimento prévio (K = número de autores) ou na distribuição, através da silhueta [Thinsungnoena et al. 2015] ou índice de partição difusa (FPC) [Bezdek 2013], constatamos que não foi possível agrupar os estudantes por autoria em nenhum dos cenários.

Para as bases jornalísticas, observamos agrupamentos capazes de segregar a maioria das obras de alguns autores. Nos experimentos a partir da representação textual, destacamos que vários agrupamentos demonstraram ser ocasionados em razão do tema do documento (Figura 2). O fato dos autores na base Varela terem escrito sobre apenas um tema pode enviesar os classificadores. Ao analisar as palavras mais frequentes dos agrupamentos, percebemos haver coesão por disciplina ou assunto (Figura 3).

Os exemplos foram separados para treinamento e teste, respeitando a proporção 70/30% de maneira estratificada. Os exemplos foram imputados numa série de classifi-

[2]https://pyphen.org

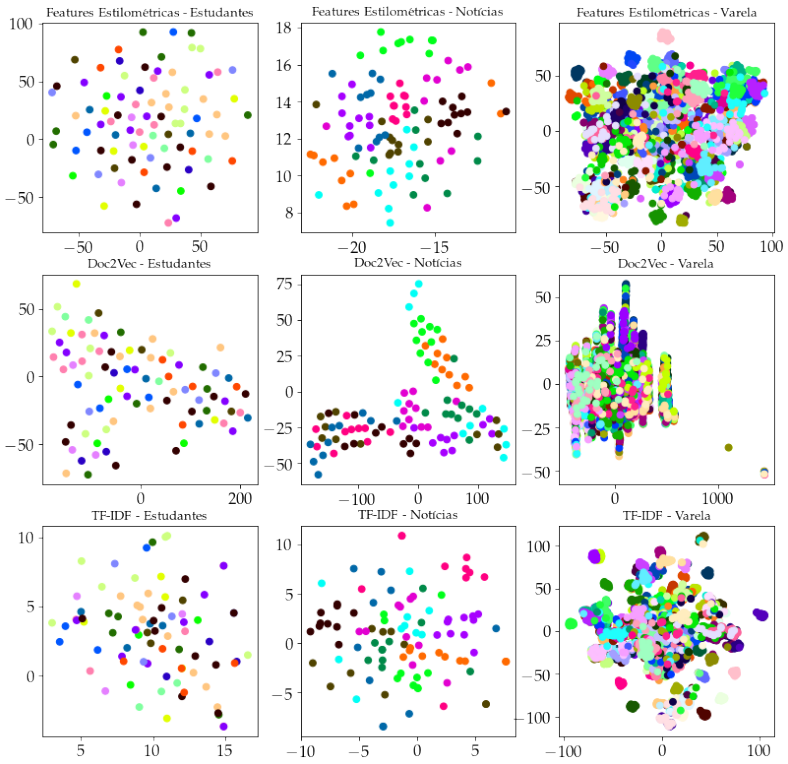

Figura 1. Apresentação das bases de dados em 2 dimensões após redução de dimensionalidade com PCA + TSNE ou LSA (TF-IDF).

cadores reconhecidos na literatura (SVM, Regressão Linear, Naive Bayes, *Random Forest*, Árvores Extra, Perceptron Multicamada (MLP), Redes Neurais customizadas, Redes Neurais Convolucionais e LSTM), utilizando os parâmetros *default* da implementações do *Sklearn*, *Tensorflow* e *Keras*. Os classificadores foram avaliados por meio da acurácia e área abaixo da curva (AUC), dado os desbalanceamento; aqueles que não superaram os modelos de base de referência em ao menos uma das representações ou *dataset* foi removido.

Apesar dos resultados positivos na literatura [Chowdhury et al. 2018] [Shrestha et al. 2017], o uso de *Deep Learning* combinado a *embeddings* pré-treinados (GloVe, FastText e *Word2Vec*) [Jang et al. 2019] [Bojanowski et al. 2017] não alcançou resultados satisfatórios, especialmente na base de estudantes. Isso pode estar relacionado ao baixo volume de exemplos e a perda de palavras importantes, ausentes nos *embeddings*.

Diante de uma quantidade menor de modelos, experimentamos usar técnicas de mudança de escala (escalonamento padrão, normalização mínimo-máximo e *power transformer* [Weisberg 2001]) e ajuste de hiperparâmetros, usando a busca em *grid* para otimização. Nesta etapa, avaliamos os classificadores por meio da validação cruzada estratificada com 3 *folds* por 10 iterações (dado o número mínimo de documentos por autor). Os modelos baseados na representação de estilo apresentaram maiores ganhos do que os textuais, assim como classificadores pautados em Árvores de Decisão e Redes Neurais se comparados aos probabilísticos. O escalonamento padrão alcançou maiores ganhos na representação textual e a normalização mínimo-máximo na de estilo.

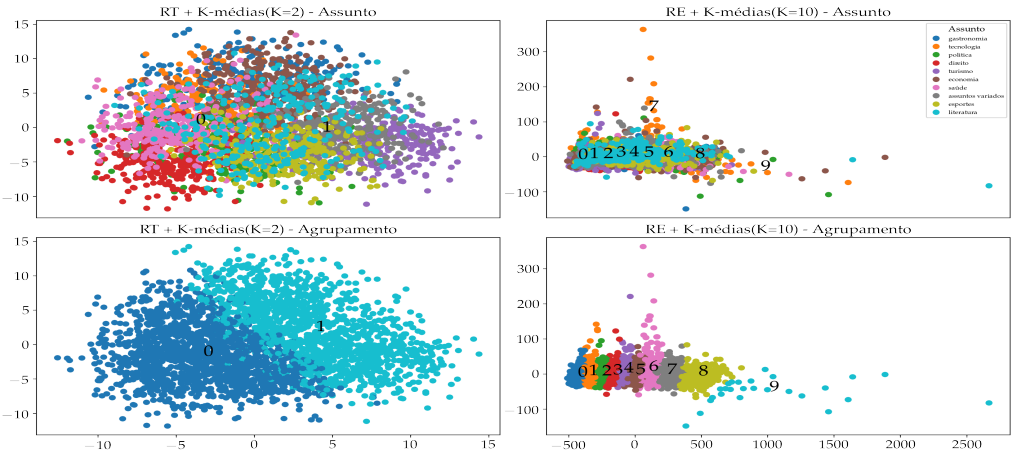

Figura 2. Distribuição dos exemplos da base Varela por assunto usando K relativo à silhueta

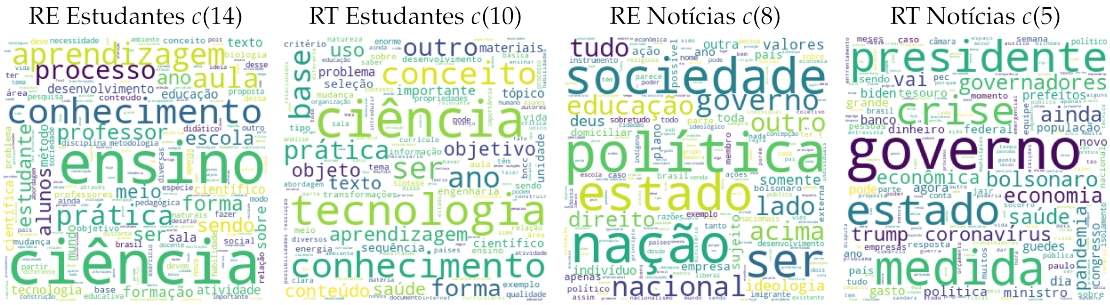

Figura 3. Nuvens de palavras de agrupamentos das bases de estudantes e notícias após aplicação do K-médias com K relativo a autoria

As dificuldades previstas para a base de estudantes se confirmaram por meio dos resultados finais (Figura 4). O melhor classificador para esta base foi o comitê de Árvores Extra (*Extra Trees* - ET) na representação de estilo com 39% de acurácia e 0.61 AUC. Na base de notícias os resultados chegaram a 94% de acurácia e 0.98 AUC com o mesmo modelo. Para Varela, a MLP textual alcançou o maior resultado, com 88% de acurácia de 0.90 AUC, seguida pelo ET, também na representação de estilo. No critério geral, considerando as três bases de dados, o *ranking* (Figura 5) confirma que o ET na representação de estilo foi superior aos demais classificadores, alcançando 71% de acurácia média.

Para ratificar os resultados encontrados, avaliamos os três melhores classificadores no critério geral por meio do teste estatístico de análise da variância simples (*One-way* ANOVA), que é o mais adequado para este cenário [Demšar 2006]. Foram coletadas as métricas dos três classificadores por 30 iterações, usando partições aleatórias. Dado que m é o número de grupos sob análise e n a quantidade de observações, realizamos o cálculo de F e p, usando α=0.05, grau de liberdade no numerador $df1 = m - 1$ e grau de liberdade no denominador $df2 = m - n$. A hipótese nula que defende não haver diferença significativa entre os grupos observados foi rejeitada (Tabela 1).

O fato de um algoritmo baseado na representação textual ser superior somente na base Varela despertou curiosidade por causa das observações preliminares durante o

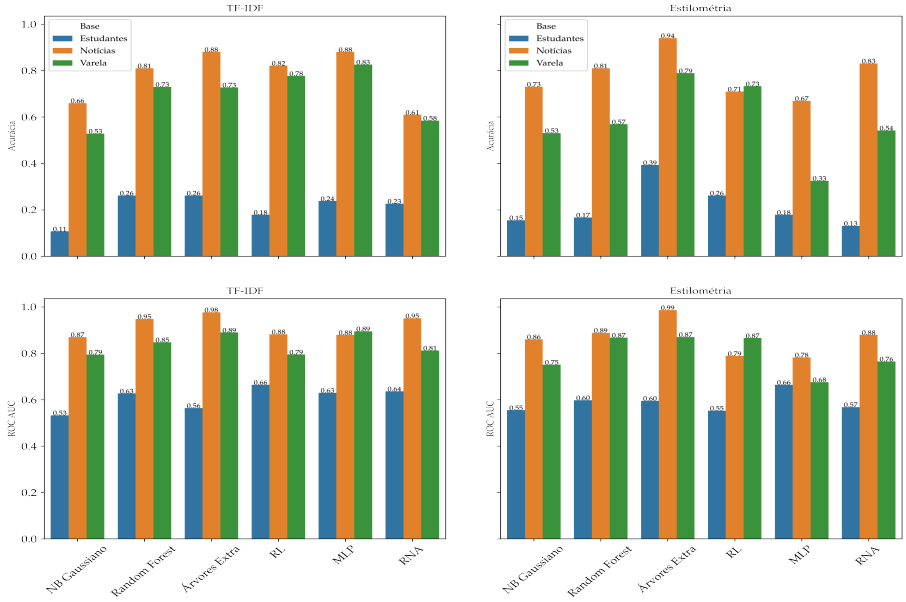

Figura 4. Resultados experimentais pós otimização agrupados por métrica, modelo e base

Table 1. Análise de variância simples a partir da acurácia dos classificadores com ANOVA *one-way*

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

agrupamento e sua composição: mais numerosa e distribuída entre 10 assuntos diferentes, passível de sofrer maior influência do tópico do que o estilo de escrita [Gamon 2004]. Assim, um experimento considerando um subconjunto de documentos de um único assunto nesta base foi realizado. O ET alcançou 92% de acurácia e 0.96 de AUC, apresentando apenas 1% de diferença para a MLP textual. Esse experimento provê indícios de que as características de estilo são mais eficazes em domínios específicos do que abrangentes, como foi visto neste experimento e nos anteriores para as bases de estudantes e notícias.

Para aumentar o entendimento sobre os fatores críticos na construção da nossa solução, melhorar a compreensão acerca das predições e suportar o entendimento da relevância das características de estilo para verificação de autoria [Goebel et al. 2018], realizamos um estudo de interpretabilidade dos modelos por meio dos valores Shapley [Shapley 1953]. A técnica foi escolhida por ser capaz de satisfazer diversos axiomas, como eficiência, linearidade, simetria, monotonicidade e proporcionalidade [Sundararajan and Najmi 2020]. Interpretamos os melhores classificadores de estilo e construímos gráficos que demonstram as características de maior influência sobre os modelos por meio da biblioteca SHAP[3].

Como critério de seleção, analisamos as características mais importantes (primeiro

[3]https://github.com/slundberg/shap

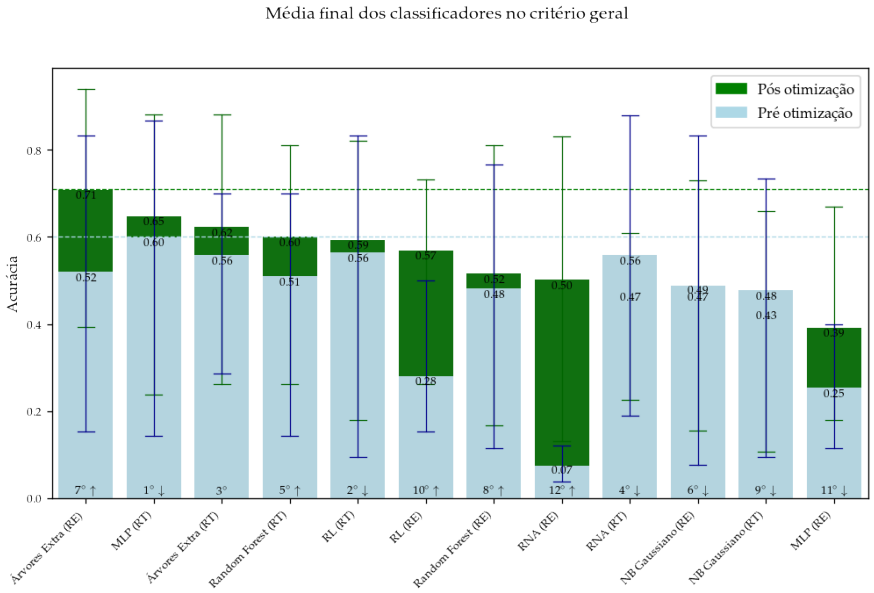

Figura 5. Acurácia média pré e pós otimização agrupados por classificador e representação

quartil) para cada classificador de acordo com os valores Shapley. Na base de estudantes, as características baseadas em caracteres, lexicais e de riqueza de vocabulário predominam igualmente entre as de maior importância. Na base de notícias também há dominância das características baseadas em caracteres, seguidas pelas de riqueza de vocabulário, lexicais e sintáticas. Para a base Varela, se destacam as características lexicais e sintáticas, seguidas por aplicação e baseada em caracteres.

O ordenamento dos grupos de características mais importantes entre as duas primeiras bases e a última são um contraponto interessante, visto que as primeiras possuem domínio restrito, e a outra possui maior diversidade de autores, temas e vocabulário, comprovando a influência do domínio sobre a relevância dos grupos de características.

No lado esquerdo da Figura 6, observamos grande homogeneidade na importância das características entre os autores, influenciando negativamente para as baixas taxas de acerto observadas para esta base. Enxergamos algumas peculiaridades nesta imagem - i) A importância das *stopwords* na maioria dos autores é um indicativo de artigos e preposições no texto, que pode estar associado a esse grupo de estudantes, ii) Autores 0 e 10 aparentam possuir um estilo de estruturação de texto através de parágrafos que destoam dos demais, e iii) A incidência de termos na primeira pessoa e termos no infinitivo (futuro) para os autores 15 e 11 respectivamente, são indicativos de estilos de escrita que os diferenciam. Do lado direito, há uma maior heterogeneidade na importância das características, que explica a maior facilidade de diferenciação dos autores durante os experimentos. Pontuamos também que: i) O impacto causado pela incidência de vírgulas indica que o autor 5 pode ter um estilo de escrita mais pausado ou sem pausas, ii) Alta frequência de *top 4-grams* e *5-grams* para o autor 8 são indicativos da presença de sequências de palavras incomuns, iii) O autor 7 se destaca pelos termos não etiquetados, que tem relação direta com o *global hapax* e pode ser indicativo de uso de palavras únicas ou estrangeirismos, e iv) Importância das exclamações em textos jornalísticos, que a pri-

ori consideramos não ser uma prática muito comum, mas que pode ter sido introduzida por influência do autor 8.

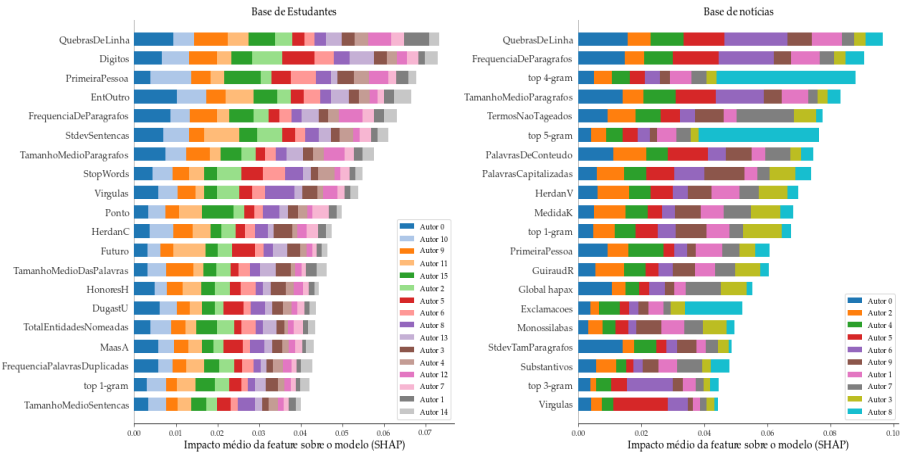

Figura 6. Valores Shapley para base de estudantes e notícias

Relativo a base Varela, devido ao amplo número de autores, foi difícil distinguir características individualmente importantes. A visão macro nos permite confirmar os grupos de características que se destacam globalmente. Para compreender individualmente como ocorre a verificação de autoria, selecionamos apenas dez autores de maneira aleatória e fizemos uma nova análise (Figura 7). Os resultados do subconjunto ratificam observações globais. Individualmente, destacamos: i) O autor 58 se distingue dos demais pelo seu índice de legibilidade (*BR-Flesch*), número de vírgulas e tamanho médio das frases. Analisando o conteúdo textual dos documentos escritos por esse autor, verificamos que todos estão dentro do assunto saúde e há uma série de termos médicos, tais como: ”arritmia”, ”hipertireoidismo” e ”amiloidose”. Concluímos que este autor se distinguiu pelo cunho técnico-científico de seus textos, ii) Os autores 51 e 82 se sobressaem pelo número de erros ortográficos, contudo ao verificar que o tema principal destes documentos é turismo, constatamos uma quantidade de palavras estrangeiras superior aos demais documentos, e iii) Um valor destoante no número de citações do autor 95 nos levou a pensar que seus textos poderiam conter citações literárias ou depoimentos. Entretanto, verificamos que os textos continham entrevistas com autoridades da área da economia.

Como resultado desta análise, inferimos que, independente da base, para atividades de atribuição de autoria, o ET se beneficia principalmente de características lexicais, baseadas em caracteres, sintáticas e de riqueza de vocabulário. Do ponto de vista individual, foi possível enxergar padrões de estilo de escrita relacionados à autoria. Para a base de textos jornalísticos, onde temos maior segurança de que os textos são propriedade intelectual de seus autores, há uma maior heterogeneidade na relação entre autoria e características de estilo mais influentes. Já na base de estudantes, onde há uma probabilidade maior de utilização de fontes de informação na internet e colaboração entre os autores [Curtis and Tremayne 2019], as características são mais homogêneas.

Na base de estudantes, o *corpora* é composto por vários documentos tratando do mesmo tema e questionamentos, o que acarreta numa restrição de vocabulário, ideologia e formato vinculado às normas institucionais. Por outro lado, as bases jornalísticas, apesar de tratarem de temas relacionados (direito, economia, política), são compostas por docu-

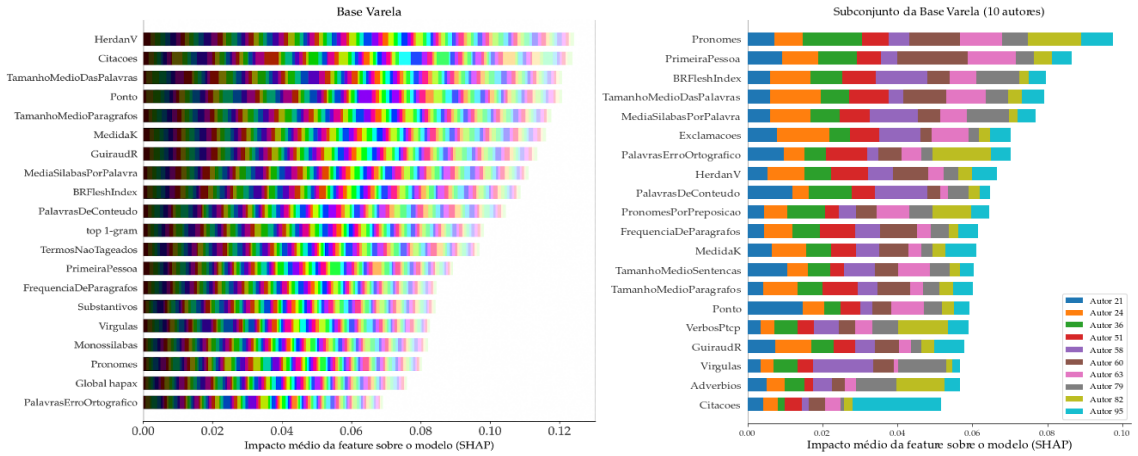

Figura 7. Valores Shapley para base Varela

mentos escritos por jornalistas e tratam de temas mais abrangentes. As características de estilo reforçam as observações: $i)$ menor influência de *top-grams* e *hapax* na base de estudantes. Ou seja, não há uma incidência de palavras únicas nos *corpora* suficiente para distinguir autores, devido às limitações já mencionadas. $ii)$ alta influência de palavras repetidas, relacionado ao perfil dos autores, que ainda são estudantes e possivelmente possuem habilidades linguísticas menos desenvolvidas que os jornalistas.

5. Conclusões e Trabalhos Futuros

Neste artigo apresentamos uma investigação inovadora ao explorar o uso de características de estilo para identificação de autoria em trabalhos escolares na língua portuguesa. A solução proposta provê um conjunto de características de estilo com efetividade comprovada em nosso idioma e pode servir de fundamento para trabalhos futuros. As análises, experimentos e observações trazem avanços sobre o uso da estilometria para análise de autoria em atividades pedagógicas.

Neste estudo constatamos a forte influência exercida pelo tópico dos documentos durante a atribuição de autoria, enaltecendo a importância da utilização de características agnósticas ao conteúdo [Halvani et al. 2020]. Verificou-se também que em bases com alta sobreposição de palavras e limitação de assuntos, as abordagens estilométricas foram superiores às textuais.

Mesmo que a solução proposta não tenha obtido elevadas taxas de acerto na distinção dos estudantes, nenhum outro classificador utilizado durante o estudo foi capaz de superá-la. Para as bases de referência, os resultados alcançados podem não superar o estado da arte, mas solucionam a atividade proposta como uma alternativa eficiente e interpretável. O trabalho se limita por não explorar comitês híbridos com classificadores textuais e estilométricos e pelo tamanho das bases, que limitaram os experimentos. Em trabalhos futuros, pretendemos incrementar a base de estudantes e avaliar o desempenho da solução diante de conjuntos mais expressivos, aplicar a solução em atividades compartilhadas de análise de autoria, como o PAN[4], e realizar um estudo de campo para avaliar a inserção da solução em ambientes virtuais de aprendizagem.

[4]https://pan.webis.de/clef21/pan21-web/author-identification.html

References

Aluísio, S., Pelizzoni, J., Marchi, A. R., de Oliveira, L., Manenti, R., and Marquiafável, V. (2003). An account of the challenge of tagging a reference corpus for brazilian portuguese. In *International Workshop on Computational Processing of the Portuguese Language*, pages 110–117. Springer.

Bevendorff, J., Ghanem, B., Giachanou, A., Kestemont, M., Manjavacas, E., Potthast, M., Rangel, F., Rosso, P., Specht, G., Stamatatos, E., et al. (2020a). Shared tasks on authorship analysis at pan 2020. In *European Conference on Information Retrieval*, pages 508–516. Springer.

Bevendorff, J., Ghanem, B., Giachanou, A., Kestemont, M., Manjavacas, E., Potthast, M., Rangel, F., Rosso, P., Specht, G., Stamatatos, E., et al. (2020b). Shared tasks on authorship analysis at pan 2020. In *European Conference on Information Retrieval*, pages 508–516. Springer.

Bezdek, J. C. (2013). *Pattern recognition with fuzzy objective function algorithms*. Springer Science & Business Media.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Botelho, J. C. and da Silva Martins, M. R. A. (2020). Avaliação da aprendizagem: novas perspectivas para velhos problemas. *Revista Encantar-Educação, Cultura e Sociedade*, 2.

Chowdhury, G. G. (2003). Natural language processing. *Annual review of information science and technology*, 37(1):51–89.

Chowdhury, H. A., Imon, M. A. H., and Islam, M. S. (2018). A comparative analysis of word embedding representations in authorship attribution of bengali literature. In *2018 21st International Conference of Computer and Information Technology (ICCIT)*, pages 1–6. IEEE.

Curtis, G. J. and Tremayne, K. (2019). Is plagiarism really on the rise? results from four 5-yearly surveys. *Studies in Higher Education*, pages 1–11.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30.

dos Santos, D. C. V.-B. (2021). Estudo comparativo entre abordagens estilométricas e textuais para atribuição de autoria em trabalhos escolares. Master's thesis, Centro de Informática – Universidade Federal de Pernambuco (UFPE).

Freitas, C., Carvalho, P., Gonçalo Oliveira, H., Mota, C., and Santos, D. (2010). Second harem: advancing the state of the art of named entity recognition in portuguese. In *quot; In Nicoletta Calzolari; Khalid Choukri; Bente Maegaard; Joseph Mariani; Jan Odijk; Stelios Piperidis; Mike Rosner; Daniel Tapias (ed) Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)(Valletta*

17-23 May de 2010) European Language Resources Association. European Language Resources Association.

Gamon, M. (2004). Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 611–617.

Geurts, P., Ernst, D., and Wehenkel, L. (2006a). Extremely randomized trees. *Machine learning*, 63(1):3–42.

Geurts, P., Ernst, D., and Wehenkel, L. (2006b). Extremely randomized trees. *Machine learning*, 63(1):3–42.

Goebel, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., Stumpf, S., Kieseberg, P., and Holzinger, A. (2018). Explainable ai: the new 42? In *International cross-domain conference for machine learning and knowledge extraction*, pages 295–303. Springer.

Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis lectures on human language technologies*, 10(1):1–309.

Halvani, O., Graner, L., and Regev, R. (2020). A step towards interpretable authorship verification. *arXiv preprint arXiv:2006.12418*.

Honnibal, M. and Montani, I. (2017). spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420.

Jang, B., Kim, I., and Kim, J. W. (2019). Word2vec convolutional neural networks for classification of news articles and tweets. *PloS one*, 14(8):e0220976.

Juola, P. (2008). *Authorship attribution*, volume 3. Now Publishers Inc.

Khonji, M., Iraqi, Y., and Jones, A. (2015). An evaluation of authorship attribution using random forests. In *2015 International Conference on Information and Communication Technology Research (ICTRC)*, pages 68–71. IEEE.

Maitra, P., Ghosh, S., and Das, D. (2016). Authorship verification-an approach based on random forest. *arXiv preprint arXiv:1607.08885*.

Martins, T. B., Ghiraldelo, C. M., Nunes, M. d. G. V., and de Oliveira Junior, O. N. (1996). *Readability formulas applied to textbooks in brazilian portuguese*. Icmsc-Usp.

Neal, T., Sundararajan, K., Fatima, A., Yan, Y., Xiang, Y., and Woodard, D. (2017). Surveying stylometry techniques and applications. *ACM Computing Surveys (CSUR)*, 50(6):1–36.

Pacheco, M. L., Fernandes, K., and Porco, A. (2015). Random forest with increased generalization: A universal background approach for authorship verification. In *CLEF (Working Notes)*.

Pires, A. R. O. (2017). Named entity extraction from portuguese web text. Master's thesis, Faculdade de Engenharia da Universidade Do Porto.

Rangel, F., Giachanou, A., Ghanem, B., and Rosso, P. (2020). Overview of the 8th author profiling task at pan 2020: Profiling fake news spreaders on twitter. In *CLEF*.

Scarton, C. E. and Aluísio, S. M. (2010). Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh-metrix para o português. *Linguamática*, 2(1):45–61.

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317.

Shrestha, P., Sierra, S., González, F. A., Montes, M., Rosso, P., and Solorio, T. (2017). Convolutional neural networks for authorship attribution of short texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 669–674.

SILVA, D. d. C. (2011). Algoritmos de processamento da linguagem e síntese de voz com emoções aplicados a um conversor texto-fala baseado em hmm. *Doutorado, Programa de Engenharia Elétrica, Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia (COPPE/UFRJ), Rio de Janeiro*.

Singh, S. and Remenyi, D. (2016). Plagiarism and ghostwriting: The rise in academic misconduct. *South African Journal of Science*, 112(5-6):1–7.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.

Sundararajan, M. and Najmi, A. (2020). The many shapley values for model explanation. In *International Conference on Machine Learning*, pages 9269–9278. PMLR.

Tempestt, N., Kalaivani Sundararajan, A. F., Yan, Y., Xiang, Y., and Woodard, D. (2017). Surveying stylometry techniques and applications. *ACM Computing Surveys*, 50(6).

Thinsungnoena, T., Kaoungkub, N., Durongdumronchaib, P., Kerdprasopb, K., and Kerdprasopb, N. (2015). The clustering validity with silhouette and sum of squared errors. *learning*, 3(7).

Tweedie, F. J. and Baayen, R. H. (1998). How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities*, 32(5):323–352.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Varela, P. J. (2017). *Uma abordagem computacional baseada em análise sintática multílingue na atribuição da autoria de documentos digitais*. PhD thesis, Pontifícia Universidade Católica do Paraná.

Weisberg, S. (2001). Yeo-johnson power transformations. *Department of Applied Statistics, University of Minnesota. Retrieved June*, 1:2003.

Yang, M., Chen, X., Tu, W., Lu, Z., Zhu, J., and Qu, Q. (2018). A topic drift model for authorship attribution. *Neurocomputing*, 273:133–140.