*DOI: 10.12957/palimpsesto.2023.65733*

# Atribuição de autoria por meio de ferramentas computacionais – um estudo de caso

Andre Luiz Siqueira Alencar (PUC-SP)[i]

**RESUMO**

O objetivo deste artigo é avaliar duas ferramentas computacionais gratuitas, que, usadas conjuntamente, podem trazer resultados significativos para uma primeira análise do estilo de um autor: o *Orange Canvas* e o etiquetador *TreeTagger*. Três *workflows* foram criados no *Orange Canvas*, por meio das *tagsets* do português disponíveis na plataforma do *TreeTagger*. Três classificadores foram usados: *Logistic Regression*, *Support Vector Machine* e *Random Forest,* juntamente com dois *corpora* de estudo. Os resultados mostram que o *Logistic Regression* e a *tagset UD_Portuguese-Bosque* forneceram as melhores combinações para a análise dos *corpora*. Há indícios de que, se usados corretamente, as ferramentas aqui avaliadas podem fornecer os primeiros subsídios sólidos para investigações posteriores mais aprofundadas por parte do perito.

**Palavras-chave**: atribuição de autoria; *Orange Canvas*; *TreeTagger*.

**ABSTRACT**

The aim of this paper is to evaluate two free computational tools, that, combined, can bring about significant results to a first analysis of an author's style: *Orange Canvas* and the *TreeTagger* labeler. Three workflows were created in *Orange Canvas*, through the Portuguese *tagsets* available on the *TreeTagger* platform. Three classifiers were used: *Logistic Regression*, *Support Vector Machine* and *Random Forest*, along with two study *corpora*. The results show that *Logistic Regression* and the *UD_Portuguese-Bosque tagset* provided the best combinations for *corpora* analysis. There are indications that, if used correctly, these tools may provide the first solid subsidies for further investigation by the expert.

**Keywords:** authorship attribution; *Orange Canvas*; *TreeTagger*.

[i] In memoriam. ORCID: https://orcid.org/0000-0003-4697-8269

# INTRODUÇÃO

A Linguística Forense (LF) refere-se a um campo interdisciplinar que aplica os conhecimentos da linguística a quaisquer textos (falados ou escritos) implicados, direta ou indiretamente, no âmbito jurídico (OLSSON, 2008). Segundo Sousa-Silva e Coulthard (2016), a LF subdivide-se em três grandes áreas: 1) o estudo de textos legais, isto é, o discurso jurídico tal como se encontra nos textos escritos da lei; 2) estudo das interações nos processos legais, tais como interrogatórios, leitura de sentenças, etc.; e 3) estudo da língua enquanto ciência forense, cujo domínio recai sobre uma ampla gama de crimes tipificados (ou não) na lei, como plágio, ameaça e injúria, etc.

No que concerne à terceira grande área da LF, nomeadamente a atribuição de autoria, foco deste artigo, o papel do perito é a atribuição de um ou mais textos a um ou mais suspeitos, por meio da procura e identificação dos seus marcadores de estilo na escrita, isto é, do seu idioleto. Segundo McMenamin (2002), idioleto refere-se ao modo único e particular de escrever de um determinado indivíduo e que não pode ser replicado na escrita de outro indivíduo.

No encalço do idioleto de um autor, existem, basicamente, dois tipos de marcadores de estilo: 1) os marcadores do tipo *bottom-up* e 2) do tipo *top-down* (MCMENAMIN, 2002). Marcadores *bottom-up* referem-se àqueles encontrados nos textos, por meio da leitura e análise minuciosas do perito. No nível sintático, por exemplo, a ausência de pontuação (ou sua presença, quando não solicitada), de concordância entre sujeito e verbo, etc., podem ser importantes na identificação do idioleto de um determinado autor. McMenamin (2002) cita em seu livro centenas de marcadores desse tipo, distribuídos nos mais diversos níveis de análise, desde o formato do texto – passando por números e símbolos, abreviações, pontuação – até a frequência de palavras e frases.

Já os marcadores *top-down*, embora sejam importantes para a discriminação do estilo de um autor (MACIEJ, 2011), são geralmente de difícil acesso, uma vez que requerem o uso de linguagens de programação, como *Python*, *R* e *Java*, por exemplo, nem sempre consolidadas no arsenal metodológico dos iniciantes na atribuição de autoria. Marcadores *top-down* implicam uma abordagem *a priori* dos textos, isto é, são pré-selecionados, geralmente derivados de estudos linguísticos consolidados

(MCMENAMIN, 2002). Frequência e tamanho médio das palavras, razão *type/token*, bigramas, trigramas, etc., são apenas alguns exemplos, para os quais algum treinamento em programação é necessário.

Ao considerarmos marcadores *top-down* como informados por algum tipo de teoria linguística estabelecida, é possível alocarmos os etiquetadores morfossintáticos nessa mesma definição. A Linguística de *Corpus*, nomeadamente a tradição probabilística, também conhecida como abordagem baseada em *corpus*, faz uso massivo de anotação de *corpora*, por meio de estruturas linguísticas pré-definidas (ou também denominadas de variáveis/características linguísticas) que serão inseridas no *corpus* de estudo, a fim de que os pesquisadores possam investigar os padrões ali encontrados (BERBER SARDINHA, 2004). Desse modo, etiquetadores morfossintáticos oferecem uma possibilidade única de acesso a marcadores *top-down*. Combinados corretamente com programas que não requerem habilidade em linhas de comando, podem tornar-se em uma primeira ferramenta útil para a abordagem do estilo de um autor e sua identificação. Neste sentido, o objetivo deste artigo é demonstrar o uso de duas ferramentas gratuitas – o *Orange Canvas* e o etiquetador morfossintático *TreeTagger* – e avaliar sua utilização em conjunto, de modo que o perito linguista possa lançar mão de programas facilmente acessíveis na Internet, a fim de fornecer-lhe impressões sólidas e confiáveis para análises posteriores mais aprofundadas de suas demandas judiciais. A importância deste artigo reside no fato de que, na literatura forense, não há, até onde pudemos pesquisar, indícios da utilização das ferramentas aqui discutidas no âmbito pericial linguístico, o que torna pertinente o ineditismo das abordagens adotadas neste estudo de caso.

O artigo está organizado da seguinte maneira: a segunda seção trata da revisão da literatura de estudos em linguística forense que fizeram uso de técnicas validadas na identificação de autoria, e do apontamento de ferramentas computacionais que ainda não foram utilizadas para esse fim.

Na terceira seção, apresento a ferramenta *Orange Canvas*, por meio de um breve histórico do seu desenvolvimento e dos passos iniciais de algumas de suas diversas funcionalidades. Ao fim, apresento diversos estudos – entre artigos dissertações e teses – que utilizaram o *Orange Canvas* como principal ferramenta metodológica para suas pesquisas. Ainda na terceira seção, introduzo o etiquetador morfossintático *TreeTagger*,

ferramenta fundamental para a etiquetagem de uma grande quantidade de textos, apresentando as três *tagsets* do português utilizadas nesta pesquisa por meio de exemplos, de modo a posteriormente avaliarmos a performance de cada uma delas nos *corpora* de estudo.

Na quarta seção, são apresentados os *corpora* de estudo propriamente ditos: o *corpus* formado pelas Cartas Chilenas (CC) e pelo *corpus* de poetas mineiros (CPM), formado pelos escritores Tomás Antônio Gonzaga e Cláudio Manuel da Costa, a fim de confrontarmos seus textos entre si, para, em seguida, atribuirmos a autoria das CC a um dos dois poetas do nosso arcadismo. Na sequência desta seção, apresento o segundo estudo, desta vez formado por *corpora* de textos jornalísticos, a dizer, o *corpus* jornalístico (CJ), constituído por textos das jornalistas Vera Magalhães e Eliane Cantanhêde, a fim de validarmos os resultados do primeiro estudo.

Na quinta seção, apresento todo percurso metodológico – desde a compilação e tratamento dos *corpora*, passando pela construção de planilhas com o auxílio de *scripts* para a contagem da frequência das variáveis linguísticas contidas no *TreeTagger* – até a efetiva análise dos dados no *Orange Canvas*.

Na sexta seção, discuto primeiramente os resultados do processo metodológico do primeiro estudo, isto é, dos *corpora* CPM e CC, por meio da comparação dos resultados dos três classificadores; em seguida, utilizo a *tagset* com melhor performance nos classificadores, a fim de a validarmos no segundo estudo, ou seja, no *corpus* CJ.

Por fim, na sétima seção, apresento as conclusões dos dois estudos realizados, levando em conta o contexto geral de análise dos dois estudos.

## 1 REVISÃO DA LITERATURA

Os estudos de atribuição de autoria possuem uma longa tradição histórica no Ocidente, mas só recentemente alçaram ao estatuto de ciência no âmbito dos estudos de linguística aplicada. Remontam à Antiguidade Clássica, nomeadamente à Grécia Antiga, em que estudiosos alexandrinos se debruçaram sobre a autoria dos poemas épicos de Homero (MACIEJ, 2011). Na modernidade, sua primeira fase como ciência linguística ligada à estatística ocorreu no século XIX, quando o matemático inglês Augustus de Morgan – em uma carta a um amigo no ano de 1851 – sugeriu que o estilo

dos autores poderia ser diferenciado por meio de estatísticas escondidas nos itens lexicais dispostos nos textos. Em suma, Morgan argumentava que a média de palavras poderia ser prova dos traços característicos do estilo de um determinado autor. A partir desse instante, deu-se início ao período de uma técnica conhecida como "estilometria" (WILLIAMS, 1970), isto é, a aplicação do estudo do estilo linguístico, geralmente direcionado à escrita. No entanto, segundo Koppel et al (2009), da antiguidade à modernidade, em que métodos estilométricos se desenvolveram a passos largos com a tecnologia vigente de seu tempo, os problemas de atribuição de autoria do passado muito frequentemente se resumiam a casos idealizados, ou seja, havia um conjunto muito claro de autores suspeitos e uma ampla gama de textos à disposição dos investigadores. No universo dos tribunais, a atribuição de autoria está muito aquém daquele passado idílico, uma vez que, como uma ciência do dia a dia, deve lidar com textos escassos, fragmentos de textos até, além de uma centena de suspeitos ou mesmo nenhum suspeito sequer. Neste sentido, atualmente, existe uma quantidade enorme de abordagens à disposição do perito linguista, as quais sem exagero, podem ultrapassar a casa dos mil (JUOLA; VESCVI, 2011). Embora haja uma pletora de abordagens à disposição do perito, o problema da atribuição de autoria pode ser resumido, segundo Koppel et al (2009), a três cenários a serem considerados pelo perito linguista: 1) o cenário do perfilamento linguístico, em que não há um suspeito em particular, e o analista, por meio do texto periciado, precisa fornecer informações a respeito do autor do texto no que diz respeito à sua origem regional, seu grau de escolaridade, etc; 2) da "agulha num palheiro", em que existem centenas de suspeitos, geralmente com um número muito limitado de textos à disposição para análise; e 3) do problema de verificação, em que não há um conjunto definido de suspeitos, mas dentre eles há um suspeito em potencial e a tarefa do perito é identificá-lo como o autor do texto. Ainda segundo os autores, é possível lançar mão de três métodos distintos de análise de atribuição de autoria, cada um com centenas de abordagens: 1) o método invariante unitário, cuja hipótese reside na relação entre o tamanho da palavra e sua frequência relativa no texto; 2) método de análise multivariada, que aplica a metodologia bayesiana a um conjunto de palavras que são independentes do tópico dos textos, isto é, funcionam como elementos universais e independentes do uso consciente dos autores; e 3) do aprendizado de máquina, método utilizado neste estudo, em que um conjunto de

textos (*corpus*) é etiquetado, de acordo com classes gramaticais ou mesmo com aspectos de sintaxe e semântica e transformado em vetores numéricos por algoritmos de classificação. Com relação às abordagens de aprendizado de máquina – que mais nos interessam neste artigo – centenas de estudos já foram realizados, o que não surpreende, dada a infinidade de abordagens mencionadas acima.

Na tabela 1, é possível observar alguns exemplos de estudos variados, que levam em consideração o aprendizado de máquina em suas metodologias.

**Quadro 1**: Exemplos de pesquisas em atribuição de autoria e aprendizado de máquina.

| **Autor(es)** | **Corpus** | **Quantidade de autores** | **Variáveis linguísticas** | **Método de Classificação** | **Ano de referência** |
|---|---|---|---|---|---|
| P. Jeevan Kumar et al | Avaliação de produtos da *Amazon* | 10 | Peso dos documentos | NBM, LR e RF | 2017 |
| Peng et al | Autores gregos | 10 | N-gramas | NB | 2004 |
| Shrestha et al | *Tweets* | 9000 | N-gramas | CNN | 2017 |
| Pavelec et al | Jornais brasileiros | 10 | Conjunções | SVM | 2007 |
| Hou e Huang | Romances chineses | 4 | Tons e bigramas de rimas | SVM e RF | 2020 |

Fonte: AUTOR.

P. Jeevan Kumar et al (2017) aplicam métodos de aprendizado de máquina, a fim de classificarem a autoria de um *corpus* de quatro mil avaliações de produtos da *Amazon*, escritos por dez autores diferentes. A inovação do estudo se refere ao fato de que, no lugar de variáveis linguísticas, como advérbios, substantivos, tamanho da palavra, etc, os pesquisadores utilizaram apenas o peso do documento, isto é, a

agregação do peso dos termos vetorizados dos documentos. Após sua vetorização, três algoritmos foram utilizados para a classificação: *NaiveBayes Multinomial* (NBM), *Logistic Regression* (LR) e *Random Forest* (RF). Cada um dos modelos de classificação tinha a tarefa de predizer os autores dos textos. A acurácia do modelo resultou num alto percentual de acerto (97,71%), por meio do algoritmo NBM; já o RF e LR obtiveram resultados acima de 90%, muito satisfatórios para um estudo em atribuição de autoria, uma vez que apenas o peso do documento foi usado.

Com um método também inovador, Peng et al (2004) propuseram a ampliação dos algoritmos de NB por meio de modelos linguísticos com n-gramas, isto é, sequências de itens linguísticos de um texto, de modo a contornar os problemas inerentes aos modelos de NB tradicionais. A partir de um *corpus* de duzentos textos oriundos de diversos gêneros textuais da língua grega e dez estilos de autores diferentes, o modelo com bigramas atingiu uma acurácia de 86% na atribuição de autoria, uma diferença de 4% em relação a outros estudos muito mais profundos de análise com algoritmos de NB tradicionais, segundo os pesquisadores.

Um outro estudo de atribuição de autoria com n-gramas, mas por meio de um algoritmo diferente – o *Convolutional Neural Networks* (CNN) – foi conduzido por Shrestha et al (2017), no intuito de atribuírem a autoria de novecentos mil tweets provenientes de nove mil autores. Apesar da obtenção de 76,1% de acurácia – ainda considerado satisfatório nos estudos de atribuição de autoria – o resultado obtido deu-se principalmente pelo fato de os pesquisadores constatarem que 30% dos autores do *corpus* consistiam de *bots*. Após sua remoção, os níveis de acurácia caíram para 68,35%. Ainda que o segundo resultado não tenha sido promissor, os pesquisadores afirmam que se trata de um processo válido, já que permitiu entender como a arquitetura dos algoritmos aprende em tais situações.

No âmbito da língua portuguesa, Pavelec et al (2007) conduziram um estudo por meio das conjunções da língua portuguesa. A premissa residiu no fato de que as conjunções podem ser consideradas marcadores estilísticos confiáveis na atribuição de autoria em detrimento a um conjunto muito maior de marcadores estilísticos, pois uma mesma conjunção ou locução conjuntiva pode ser escrita de maneiras muito distintas, o que as caracterizariam como próprias do estilo de cada autor. Com um *corpus* de dez jornalistas e 150 textos, os pesquisadores utilizaram diferentes parâmetros do algoritmo

*Support Vector Machine* (SVM) e obtiveram uma acurácia de 75,1%, valor que, segundo os pesquisadores, está de acordo com diversos outros estudos na literatura.

Por fim, Hou e Huang (2019) apresentam um estudo original, ao considerar marcadores de estilo não convencionais da literatura forense. Com um *corpus* de romances chineses, composto por quatro autores, os pesquisadores utilizaram os algoritmos RF e SVM no intuito de atribuírem a autoria por meio de dois aspectos fonológicos: o tom e bigramas de rima. Após diversos testes com diferentes aspectos fonológicos da língua chinesa, os pesquisadores obtiveram resultados promissores para ambos os algoritmos, numa média de acurácia entre 85% e 95%.

Como se nota, o uso de técnicas de aprendizado de máquina oferece técnicas estatísticas efetivas, uma vez que a modelagem estatística da linguagem se preocupa com os vários níveis de análise linguística, tais como a semântica, a sintaxe, a fonologia, etc, o que permite afirmar que algoritmos de aprendizado de máquina fornecem subsídios sólidos para a atribuição de autoria. Nem todos os algoritmos necessariamente são 100% precisos, mas a condução de diversos estudos, por meio de diferentes algoritmos, compensa as imprecisões. Nenhum dos estudos aqui apontados utilizaram ferramentas gratuitas, o que torna esse tipo de estudo complicado, sendo necessários conhecimentos avançados de programação. No entanto, diversas ferramentas disponíveis livremente na internet são capazes de realizar praticar os mesmos trabalhos aqui discutidos. Ferramentas como *Weka* (2016) e *KNIME* (2007) produzem resultados semelhantes e possuem uma curva de aprendizado média, o que permite aos iniciantes na linguística forense o aprofundamento de suas pesquisas e demandas judiciárias. Na próxima seção, abordaremos o uso do *Orange Canvas* e *TreeTagger* para fins de atribuição.

## 2 AS FERRAMENTAS

### 2.1 *Orange Canvas*

Desenvolvido na Eslovênia, no laboratório de bioinformática da Faculdade de Ciências e Informação da Universidade de Liubliana (JOVIĆ; BRKIĆ; BOGUNOVIĆ, 2014), em 1996, o *Orange Canvas* é um programa baseado na linguagem *Python*, que

oferece uma ampla gama de ferramentas estatístico-computacionais, como se pode observar na figura abaixo.

**Figura 1:** Tela inicial do *Orange Canvas*


Fonte: *Orange Canvas.*

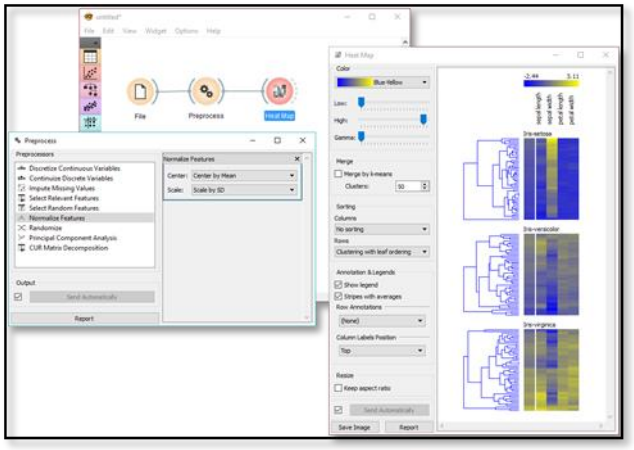
Na figura acima, as abas à esquerda correspondem às ferramentas à disposição do pesquisador. Por meio da opção *options* > *add-ons*, é possível expandir o número de abas e, consequentemente, o número de ferramentas possíveis de serem utilizadas. Por meio delas, pode-se construir modelos de aprendizado de máquina, minerar, analisar e visualizar dados, etc. Uma das grandes vantagens do *Orange Canvas* é a sua interface: para cada *workflow*, isto é, para cada conjunto de ações direcionadas a um determinado fim, o pesquisador conta com diversos componentes, chamados de *widgets* (figura 2). Cada aba contém diversos *widgets*, nos quais o pesquisador pode clicar ou arrastar para o espaço em branco (*canvas*).

**Figura 2:** Exemplo de *widget.*


Fonte: *Orange Canvas*.

Já no *canvas*, uma vez clicado, o *widget* se abre e diversas funcionalidades podem ser sintonizadas, de acordo com as necessidades dos dados a serem analisados. Por exemplo, na figura abaixo, três *widgets* foram usados para o *workflow*, a fim de 1) carregar os dados; 2) pré-processar os dados; e 3) visualizar os dados em um *heat-map*.

**Figura 3**: Exemplo de um *workflow* simples.



Fonte: *Orange Canvas*.

Observe que os *widgets Preprocess* e *Heat Map* foram clicados e suas respectivas janelas surgem com diversas funcionalidades, todas customizáveis. Também é interessante observar que cada *widget* se une ao seguinte, por meio de um traço ou linha cinza. Ao final da manipulação de cada *widget*, de acordo com as necessidades dos dados, o usuário o conecta a outro *widget*, manipula-o, conecta-o a outro *widget* e assim procede sucessivamente. No final, o pesquisador terá construído um *workflow* para o

seu projeto. Ainda na figura acima, nota-se que alguns *widgets* possuem entradas e saídas, representadas pelas linhas de conexão cinzas. Outros, como o *File*, possuem apenas a saída, o que é natural, pois as bases de dados são carregadas no programa, por meio da janela do próprio *widget* (que se abre, como vimos, quando nele clicamos).

O *Orange* ainda conta com uma grande base de dados embutidas, em diversas categorias disponíveis para o usuário, seja para aprendizado seja para estudos. Além do mais, há *widgets* que oferecem a possibilidade de uso de bases de dados *online*, como *PubMed*, *The Guardian*, *NY Times*, *Twitter*, dentre outros, bastando que o usuário forneça, no caso do *Twitter*, a sua API. Caso o usuário possua sua própria base de dados, o *Orange* permite a importação de diversas extensões de arquivos, como *.txt*, *.csv*, *.tsv*, *.xlsx*, etc. Vale ainda lembrar que, se por algum motivo, o usuário precisar criar seus próprios *scripts*, o *Orange* oferece um *widget* chamado *Python Scripts*, ampliando ainda mais as possibilidades de análise, pré-processamento e visualização dos dados.

Por fim, o *Orange* conta com um *site* e um canal no *Youtube*, onde o usuário pode aprender os fundamentos da ferramenta. Além do mais, diversos *sites*, como S*tackoverflow*, S*tackexchange*, dentre outros, contêm muitas perguntas e respostas de usuários que, por algum motivo, não conseguiram solucionar algum problema de sua análise no *Orange*. A consulta a essas plataformas é de grande valia no aprendizado do usuário iniciante.

Desde o seu lançamento, o *Orange* tem sido usado em diversas publicações, tais como em bioespectroscopia (TOPLAK et al, 2021), predição de cristais (KLUNNIKOVA et al, 2020), crédito financeiro (APAMPA, 2016), meteorologia (HARO RIVERA et al, 2018), biomedicina (SHI; ZHAO; WEI, 2018), audição humana (ALJABERY; KURNAZ, 2020), genética (STRAZAR et al, 2019), imageamento clínico (ALVES et al, 2021), etc. Além disso, diversos artigos de avaliação e comparação da ferramenta também foram publicados (DEMŠAR; ZUPAN, 2013; H et al, 2011; JOVIĆ; BRKIĆ; BOGUNOVIĆ, 2014; ZUPAN; DEMSAR, 2008), com revisões positivas, o que o torna uma ferramenta respeitada nos mais variados campos da atuação humana. No que diz respeito à atribuição de autoria, nenhum artigo, até onde pôde ser verificado, foi encontrado.

Na seção Metodologia, teremos a oportunidade de entender o funcionamento de um *workflow* no *Orange* com mais detalhes.

## 2.2 *TreeTagger*

O *TreeTagger* é um etiquetador morfossintático, desenvolvido por Helmut Schmid (1994), cuja função é etiquetar, isto é, anotar cada palavra de um texto, em relação aos aspectos morfossintáticos de uma determinada língua. Em seu núcleo, o *TreeTagger* utiliza um algoritmo do tipo árvore de decisão e, para cada item lexical do texto, atribui uma probabilidade de classificação, segundo os parâmetros morfológicos e sintáticos nele contidos. Além disso, o etiquetador gera uma coluna com os lemas de cada palavra, como mostra o quadro abaixo:

**Quadro 2**: Exemplo de anotação de texto no *Treetagger*.

| **ITEM LEXICAL** | ***POS*** | **LEMA**[1] |
|---|---|---|
| Que | CS | que |
| triste | AQ0CS0 | triste |
| , | Fc | , |
| Doroteu | NP00000 | doroteu |
| , | Fc | , |
| se | PP3CN00 | se |
| pôs | VMIS3S0 | pôr |
| a | DA0FS0 | o |
| tarde | NCFS000 | tarde |
| ! | Fat | ! |

Fonte: AUTOR.

Na primeira coluna, o *TreeTagger* devolve os itens lexicais do texto, tal como fornecido na entrada; a segunda coluna refere-se à etiquetagem, à anotação propriamente dita, executada pelo programa e se refere às partes do discurso (*part of speech*), isto é, às classes gramaticais presentes na língua. Essa coluna é dependente do tipo de *tagset*, isto é, do conjunto de etiquetas fornecidas ao programa. Nesse caso, trata-se de uma das três *tagsets* do português usadas nesta pesquisa, especificamente a *EAGLES FINE-GRAINED*, disponível na página oficial do *TreeTagger*. Por exemplo, para o item lexical *Que*, o etiquetador o anotou como sendo uma conjunção subordinada (CS); para *Doroteu*, nome próprio (NP00000); *tarde*, nome comum feminino singular

(NCFS000), e assim por diante. No entanto, etiquetadores não são 100% precisos em suas anotações. Observa-se, por exemplo, que o termo *Que* não foi corretamente classificado (conjunção subordinada, no lugar de advérbio). Esse tipo de problema é conhecido e se deve a vários fatores, dentre os quais, o tamanho do *corpus* de treino usado no etiquetador (AIRES et al, 2000). Para o português, Gamallo (2013) reportou uma acurácia de 91,39% e 96,03%, para os *corpora* de teste *Bosque_CF* e *Miscelâneo*, respectivamente.

As *tagsets* do português para o *TreeTagger* contêm dezenas de etiquetas diferentes. Exemplos de cada uma delas podem ser observados no quadro abaixo.

**Quadro 3**: Exemplos de *tagset* com suas respectivas *tags* (etiquetas) e significados.

| **EAGLES REDUZIDA** | | **UD_PORTUGUESE-BOSQUE** | | **EAGLES FINE-GRAINED** | |
|---|---|---|---|---|---|
| **TAG** | **SIGNIFICADO** | **TAG** | **SIGNIFICADO** | **TAG** | **SIGNIFICADO** |
| NCMS | NOME COMUM MASCULINO SINGULAR | ADJ.Masc.Sing | ADJETIVO MASCULINO SINGULAR | DI0MS0 | DETERMINANTE INDEFINIDO MASCULINO SINGULAR |
| SPS | PREPOSIÇÃO SIMPLES | NOUN.Fem.Sing | SUBSTANTIVO FEMININO SINGULAR | VMIS3S0 | VERBO PRINCIPAL INDICATIVO PASSADO 3ª. PESSOA SINGULAR |
| VMI | VERBO PRINCIPAL INDICATIVO | VERB.Inf | VERBO INFINITIVO | SP+DA | PREPOSIÇÃO + DETERMINANTE ARTIGO |
| DI0 | DETERMINANTE INDEFINIDO | DET.Fem.Sing | DETERMINANTE FEMININO SINGULAR | VMN0000 | VERBO PRINCIPAL INFINITIVO |
| RG | ADVÉRBIO | PUNCT.Comma | PONTUAÇÃO: VÍRGULA | AQ0MS0 | ADJETIVO QUALIFICATIVO MASCULINO SINGULAR |

Fonte: *TreeTagger*.

No quadro acima, estão dispostos alguns exemplos das três *tagsets* usadas nesta pesquisa. A *tagset EAGLES Reduzida* (ERE), proposta pelo projeto EAGLES (desenvolvido em 1996), é um conjunto de etiquetas reduzidas, isto é, contam apenas com informações básicas da estrutura do português. Contém onze divisões (dez

correspondentes às classes gramaticais, e uma, aos sinais de pontuação). A UD_Portuguese-Bosque (UPB) é baseada na Gramática de Restrição, aplicada ao *corpus Bosque*, que faz parte de uma *corpus* maior, denominado *Floresta Sintá(c)tica* (RADEMAKER et al, 2017). Contém basicamente as mesmas etiquetas que o ERE, exceto pelo fato de terem sido treinadas em diferentes *corpora*, com algoritmos diferentes. Já a *EAGLES FINE-GRAINED* (EFG), diferente da ERE, contém uma granularidade maior, isto é, contém informações mais específicas a respeito do léxico. Por exemplo, nota-se, no quadro acima, que, para a etiqueta *determinante* (DI0) na *tagset* ERE, há apenas uma informação adicional: indefinido. Para a EFG, a mesma etiqueta seria, além de indefinido, masculino e singular (DI0MS0).

# 3 OS CORPORA

## 3.1 As *Cartas Chilenas*

O primeiro *corpus* de estudo contempla As *Cartas Chilenas* (CC). Refere-se a um conjunto de treze cartas escritas por Tomás Antônio Gonzaga, poeta mineiro do século XVIII. Seu conteúdo gira em torno da crítica direcionada a Cunha Meneses, então governador de Minas Gerais, devido aos desserviços prestados ao Estado (MENDES, 2010). Com um tom satírico e mordaz, o poema é narrado por Critilo, morador da capital do Chile (Santiago do Chile), que conta ao seu amigo Doroteu os desserviços e despotismo do governador chileno Fanfarrão Minésio. Obviamente, Critilo, Santiago do Chile e Minésio Fanfarrão são apenas subterfúgios literários para se referirem a Tomás Antônio Gonzaga, Vila Rica e Cunha Meneses, respectivamente. Uma vez palco de debates fervorosos acerca de sua autoria, foram definitivamente atribuídas a Tomás Antônio Gonzaga, por Rodrigo Lapa e Manuel Bandeira (BOSI, 2017), e não a Cláudio Manuel da Costa (como se supunha), poeta conterrâneo de Gonzaga, importantíssimo no cenário de Minas Gerais do século XVIII.

A fim de avaliarmos os potenciais de atribuição de autoria das duas ferramentas aqui utilizadas, consideraremos as CC como de autoria desconhecida. Para o confronto, foram coletadas todas as obras de Tomás Antônio Gonzaga e Cláudio Manuel da Costa, denominado *corpus* de poetas mineiros (CPM), disponíveis no *site* Domínio Público[2].

Todo o procedimento de coleta, pré-processamento e anotação dos *corpora* aqui mencionados serão discutidos na *Metodologia*.

### 3.2 O *corpus* jornalístico

O segundo estudo contém dois *subcorpora* compostos de textos de duas jornalistas distintas: Vera Magalhães e Eliane Cantanhêde, denominado *corpus* jornalístico (CJ) (MARCONDES; BERBER SARDINHA, 2021). Vera Magalhães é jornalista reconhecida nos meios midiáticos, tendo passagem pela F. de São Paulo, Estado de São Paulo, Veja, dentre outros. É também apresentadora do programa Roda Viva, da TV Cultura. Eliane Cantanhêde é jornalista do Estado de São Paulo. Cobriu diversos momentos importantes da história do Brasil, como as "Diretas Já", a Assembleia Nacional Constituinte de 1987-88, dentre outros. Os textos do CJ retratam os cenários da política brasileira no contexto do mandato do presidente Jair Bolsonaro e foram compilados do jornal Estado de São Paulo, a partir do mesmo registro: coluna de política. A motivação de uso desse *corpus* reside no fato de que, para além de já estarem compilados, ambas as jornalistas versam sobre temas semelhantes e recobrem um mesmo período[3]. A essa homogeneidade temática, somam-se outros fatores, como escolaridade e emprego semelhantes e mesmo sexo, o que torna a identificação de autoria num processo menos artificial possível. No conjunto, o *corpus* contém um total de 360 textos, dentre os quais 144 pertencem a Vera Magalhães e 216, a Vera Cantanhêde. Para cada jornalista, foram retirados dez textos, que serão usados como textos desconhecidos, denominado *corpus* de jornalistas desconhecidos (CJD) e base para a atribuição de autoria, após o treino e teste no restante dos *subcorpora*.

Na próxima seção, discutiremos todo o percurso metodológico de processamento dos *corpora*, por meio do *TreeTagger* e *Orange Canvas*.

## 4 METODOLOGIA

### 4.1 Pré-processamento

Antes que possamos avaliar o *TreeTagger* e o *Orange Canvas*, é preciso seguirmos algumas etapas preliminares: 1) coleta dos *corpora*; 2) pré-processamento dos *corpora*; 3) etiquetagem automática dos *corpora* com as três *tagsets* do português; e 4) contagem e normalização da frequência das etiquetas de cada texto e produção de planilhas em arquivos de extensão *.csv*.

Com relação ao primeiro estudo – CC e CPM – todos os textos, incluindo Tomás Antônio Gonzaga (TAG) e Cláudio Manoel da Costa (CMC), está resumido no quadro abaixo.

**Quadro** 4: *Corpus* do primeiro estudo.

| Título | Autor | Fonte | Site |
|---|---|---|---|
| Cartas Chilenas | Tomáz Antônio Gonzaga | Universidade da Amazônia | http://www.dominiopublico.gov.br/ |
| Marília de Dirceu | Tomáz Antônio Gonzaga | Universidade da Amazônia | http://www.dominiopublico.gov.br/ |
| Culto Métrico | Cláudio Manuel da Costa | Universidade Federal de Santa Catarina | http://www.dominiopublico.gov.br/ |
| Epicédio | Cláudio Manuel da Costa | Universidade Federal de Santa Catarina | http://www.dominiopublico.gov.br/ |
| Munúsculo Métrico | Cláudio Manuel da Costa | Universidade Federal de Santa Catarina | http://www.dominiopublico.gov.br/ |
| O Parnaso Obsequiso e Obras Poéticas | Cláudio Manuel da Costa | Universidade Federal de Santa Catarina | http://www.dominiopublico.gov.br/ |
| Sonetos Inéditos | Cláudio Manuel da Costa | Universidade Federal de Santa Catarina | http://www.dominiopublico.gov.br/ |
| Vila Rica | Cláudio Manuel da Costa | Universidade Federal de Santa Catarina | http://www.dominiopublico.gov.br/ |

Fonte: AUTOR.

A segunda etapa consistiu no pré-processamento dos *corpora*, uma vez que, ao serem coletados diretamente do *site*, os arquivos se encontram no formato *.pdf*, extensão de arquivo incompatível com a etiquetagem automática. Desse modo, os textos foram convertidos para o único formato aceito pelo *TreeTagger*, a extensão *.txt*. Qualquer conversor de extensão para *.txt* pode ser usado, desde que o arquivo final contenha apenas os textos dos escritores, sem cabeçalhos, informações sobre o livro, notas de rodapé, etc. Optamos por um conversor online – *Online Convert* – pela agilidade e

precisão na conversão. Após esse passo, os arquivos foram conferidos e todos os elementos estranhos, como os citados há pouco, foram retirados. Neste passo, é possível usar editores de textos, como *Notepad++*, por exemplo, ou mesmo *scripts* de limpeza de textos, caso o usuário tenha familiaridade com programação. Para a limpeza dos textos, utilizamos o *Text-Cleaner*[4], um pequeno programa com interface gráfica, que permite o uso de funções pré-programadas, bem como a utilização de *regex* – expressões regulares para alteração de padrões de caracteres nos textos.

Após a limpeza, alguns procedimentos adicionais foram aplicados. Como os textos contêm muitas palavras grafadas no português antigo, resolvemos modificá-las para a ortografia moderna, a fim de melhorar a performance do reconhecimento das *tagsets.* Termos como "Co'a" (Com a), "d'ouro" (de ouro), "d'alta" (da alta), etc, foram modificados. Uma segunda modificação foi necessária, tendo em vista o fato de que, até onde pudemos averiguar, as *tagsets* têm muita dificuldade em classificar verbos e seus objetos quando se apresentam unidos, como "vê-lo", disse-lhe", "dar-lhe-ia", etc. Assim, o hífen de cada um desses verbos foi afastado de seus termos, de modo que as *tagsets* reconheçam com mais precisão os elementos constituintes. Essas modificações são também possíveis em editores de texto e, no estudo deste artigo, optamos pela ferramenta *TextCleaner*.

Embora esses procedimentos sejam uma decisão controversa, já que em LF os textos devem ser analisados do modo como são recebidos, independentemente de erros ortográficos, grafia, etc., ela foi necessária já que as *tagsets* usadas não reconhecem esses termos. Terminado todo o pré-processamento, os arquivos foram salvos com o respectivo título e nome dos autores.

Como mostra o Quadro 4, TAG é o autor com menor quantidade de textos. Em aprendizado de máquina, a quantidade importa e, portanto, para contornamos esse problema, optamos por dividir a obra *Marília de Dirceu* (1792) em partes iguais, de modo a obtermos um número significativo de observações. A obra contém milhares de versos, e optamos por dividi-la em 150 partes iguais. Trata-se de uma divisão puramente subjetiva, uma vez que não há um parâmetro para esse tipo de questão, apenas bom senso. Somente as *Cartas Chilenas* (1845) não foram divididas, uma vez que se trata dos textos questionados da pesquisa. Contudo, aplicamos nelas as mesmas modificações descritas para os *corpora* dos autores. O mesmo procedimento foi feito com as obras de

CMC, porém, com uma pequena diferença. Por exemplo, para *Obras Poéticas*, cada registro (epicédios, sonetos, romances, etc.) foi agrupado individualmente e, logo em seguida, divididos em partes iguais. Ao final, obtivemos 150 partes, a exemplo de TAG. O procedimento foi feito por meio de um pequeno *script* em *Python*[5], de modo a dividir os textos corretamente e acelerar o processo.

### 4.2 Processamento no *Treetagger*

Após todo o pré-processamento, os arquivos foram pós-processados pelo *TreeTagger*. Cada um dos c*orpora* foi etiquetado com as três *tagsets* do português, como mostra o quadro abaixo:

**Quadro 5**: Etiquetagem dos corpora de estudo.

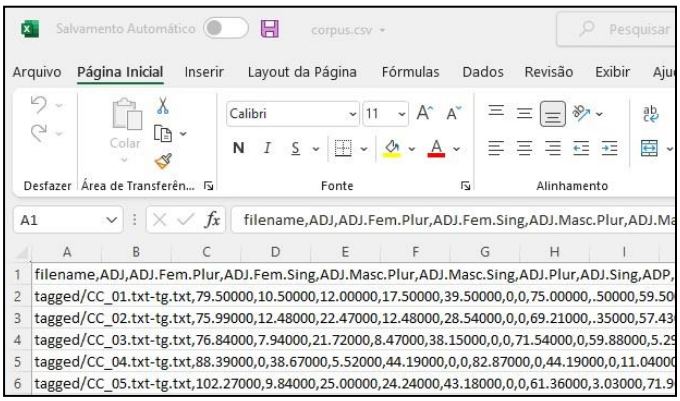
| ***CORPUS*** | **ESTUDO** | **ETIQUETAS** |
|---|---|---|
| *CARTAS CHILENAS* | 01 | **REDUZIDAS**<br>**UD_PORTUGUESE-BOSQUE**<br>**EAGLES FINE-GRAINED** |
| TOMÁZ ANTÔNIO GONZAGA | 01 | |
| CLÁUDIO MANUEL DA COSTA | 01 | |
| *CORPUS* JORNALÍSTICO | 02 | |

Fonte: AUTOR.

Finalizadas as três etiquetagens – todas salvas em pastas diferentes – os arquivos estão prontos para receber o último *script*[6].

Nesta última etapa, todos os *corpora* foram processados pelo *script*, cuja função é 1) construir uma *wordlist* baseada na coluna dos lemas, produzida pelo *TreeTagger*; 2) contar as etiquetas referentes à *wordlist* criada; 3) normalizar a frequência da contagem das etiquetas; e 4) retornar uma planilha em *.csv*, com todas as etiquetas e observações processadas. Ao final, o *script* retorna, na pasta *spreadsheet* (criada automaticamente pelo *script*), uma planilha, denominada *corpus*, como mostra a figura abaixo.

**Figura 4**: planilha dos *corpora* CC e CPM.




Fonte: AUTOR.

A figura acima mostra a planilha final da *tagset* UD_PORTUGUESE-BOSQUE, para o CC e CPM. Ao final, há três planilhas – uma para cada *tagset* do estudo 1. O mesmo procedimento foi realizado para o estudo 2. Como mencionado anteriormente, o CJ contém 316 textos de duas jornalistas brasileiras – Eliane Cantanhêde e Vera Magalhães – conhecidas jornalistas no cenário brasileiro. Um resumo do *corpus* pode ser visto no quadro abaixo.

**Quadro 6**: Resumo do CJ.

| | |
|---|---|
| **CORPUS** | CORPUS JORNALÍSTICO |
| **AUTORAS** | VERA MAGALHÃES/ELIANE CANTANHÊDE |
| **VEÍCULO** | ESTADO DE SÃO PAULO |
| **ÉPOCA** | 2019 A 2020 |
| **TÓPICO** | POLÍTICA BRASILEIRA CONTEMPORÂNEA |
| **REGISTRO** | COLUNA POLÍTICA DE JORNAL |
| **ORIENTAÇÃO POLÍTICA** | CENTRO-DIREITA |
| **TOTAL DE TEXTOS** | X ELIANE; Y VERA |

Fonte: adaptado de Marcondes e Berber-Sardinha (2021).

Três planilhas também foram produzidas para cada *tagset*. Feito isso, cria-se manualmente uma coluna, em cada planilha, chamada "autor", e insere-se, no caso do CPM, as respectivas iniciais dos nomes dos poetas. Por exemplo, TAG, para Tomáz Antônio Gonzaga e CMC, para Cláudio Manuel da Costa. Na planilha do *corpus* CC, insere-se uma coluna com o nome "autor" e as iniciais CC. Isso é extremamente importante, uma vez que o *Orange Canvas* espera uma coluna categórica, por meio da

qual possa treinar os algoritmos de classificação. O mesmo procedimento foi realizado no CJ, havendo apenas a mudança das iniciais para as jornalistas correspondentes.

Com as planilhas prontas, pode-se iniciar a etapa de análise dos dados no *Orange*. Para efeitos didáticos, apresentaremos um *workflow* completo, por meio de apenas uma *tagset* usada no estudo 1.

### 4.3 Processamento no *Orange Canvas*

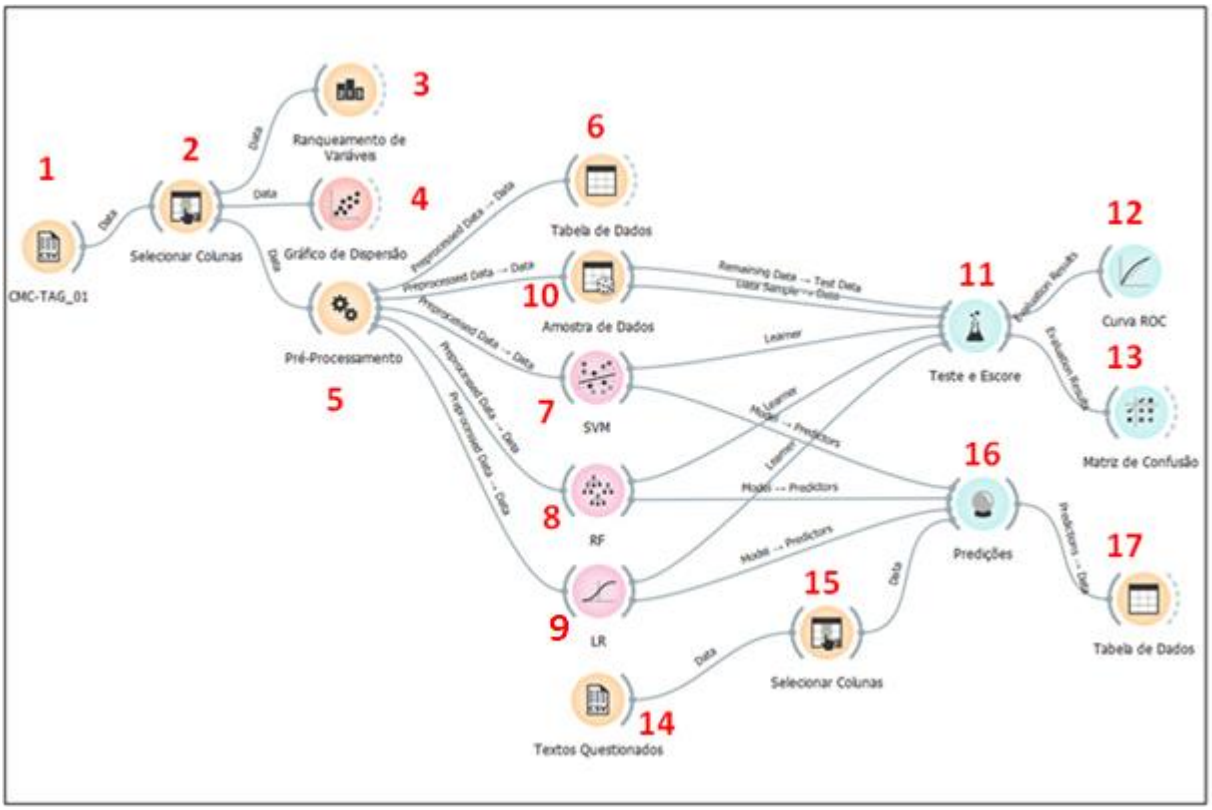



**Figura 5**: Exemplo de *workflow* em *Orange*.

Fonte: AUTOR.

A figura acima contém o *workflow* completo da *tagset* ER. Todos os *widgets* necessários foram importados ao *canvas*, sequencialmente, dependendo do processo a ser realizado. Após cada etapa, foram renomeados, para melhor compreensão dos procedimentos adotados[7]. Para melhor o compreendermos, dividimos as quatorze etapas

em cinco blocos, dispostos a partir dos tópicos contidos em cada um, como se pode notar na figura abaixo.

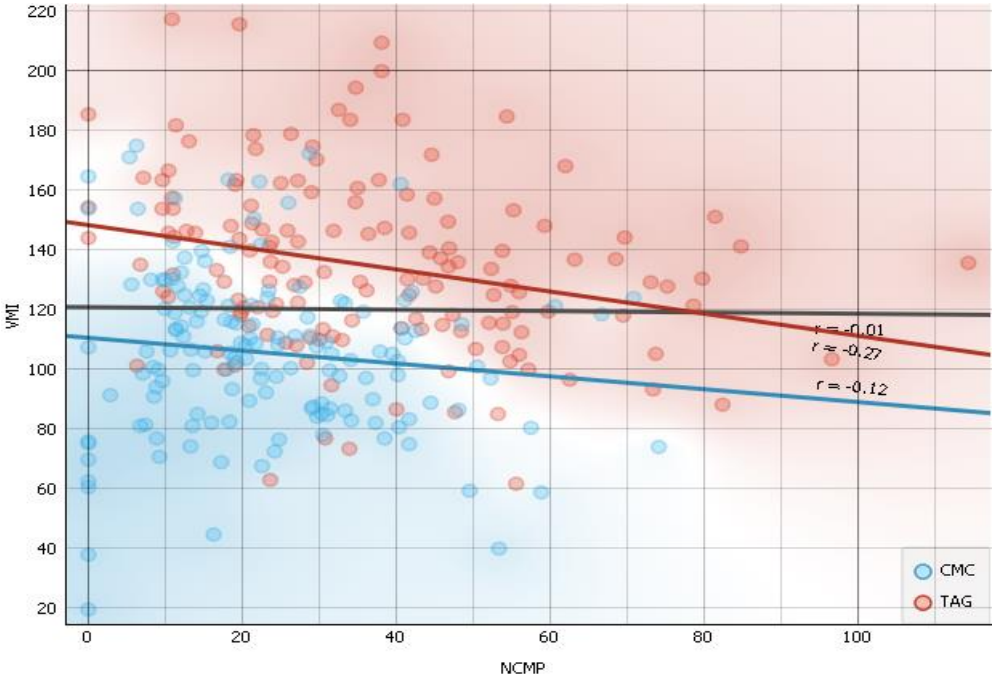
**Quadro 7**: Resumo em tópicos dos blocos de análise do *workflow*.

| **BLOCO 1** | **BLOCO 2** | **BLOCO 3** | **BLOCO 4** | **BLOCO 5** |
|---|---|---|---|---|
| PRÉ-PROCESSAMENTO DOS DADOS | SEPARAÇÃO DOS DADOS EM TREINO E TESTE | TREINO E TESTE DOS CLASSIFICADORES POR *CROSS VALIDATION*: AVALIAÇÃO DOS RESULTADOS | PREDIÇÃO DE DADOS NOVOS: ATRIBUIÇÃO DE AUTORIA DAS CARTAS CHILENAS | VERIFICAÇÃO DOS ACERTOS NA ATRIBUIÇÃO DE AUTORIA |
| **ITENS 1 A 6** | **ITEM 10** | **ITENS 7, 8, 9, 11,12 E 13** | **ITENS 14, 15, 16** | **ITEM 17** |

Fonte: AUTOR.

No primeiro bloco, o item 1 corresponde ao **carregamento do arquivo**, por meio do widget *CSV File Import* e, ao clicarmos no *widget*, as informações do arquivo são mostradas[8]. Com o item 2 – **seleção de colunas** – é possível ajustar corretamente as variáveis categóricas e contínuas para os seus respectivos papéis ou mesmo excluí-las. Como o *Orange* reconheceu corretamente as variáveis, nenhum ajuste foi necessário. Os itens 3 e 4 são variáveis, isto é, podem conter outros *widgets*. Em nosso caso, optamos, no item 3, apenas pela **verificação das melhores variáveis preditoras de classe** e a **dispersão das variáveis** em relação aos autores com o item 4. Um exemplo de gráfico de dispersão pode ser visto abaixo.

**Gráfico 1**: gráfico de dispersão de duas variáveis.




Fonte: AUTOR.

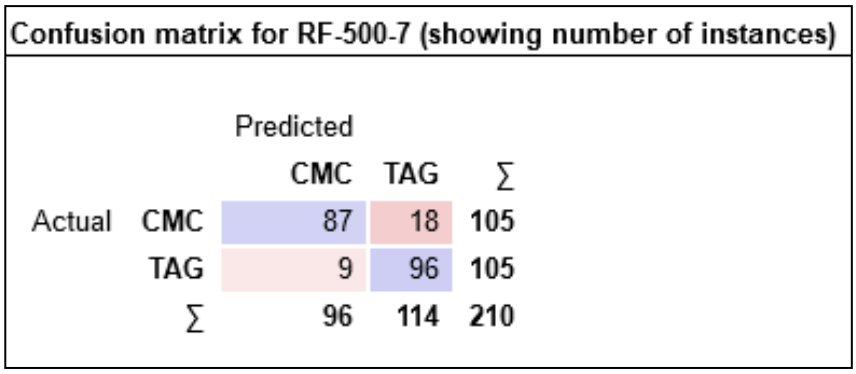
No gráfico acima, encontra-se a dispersão entre as variáveis VMI (verbo principal no indicativo) e NCMP (substantivo comum masculino plural), representadas pelos traços em vermelho e azul. Como se nota, Cláudio Manoel está mais bem definido do que Tomáz Antônio, com pouca sobreposição das variáveis. Após essa inspeção prévia, executa-se o **pré-processamento dos dados** com o item 5. Diversos ajustes podem ser feitos com esse ícone. Optamos apenas por centralizar e escalar as variáveis, a fim de obtermos o desvio padrão em uma mesma escala para todas as variáveis. Também optamos por utilizar a função remover valores perdidos, para variáveis que eventualmente não tenham algum valor declarado. A verificação dos dados obtidos do pré-processamento ocorre no item 6. Nessa etapa, é possível construir novas variáveis, isto é, somá-las, dividi-las em uma outra série de combinações, de acordo com as necessidades da pesquisa.

No segundo bloco, o perito recorre à **divisão dos dados em treino e teste** a serem utilizadas em ambas as partições. A partição realizada neste estudo corresponde a

uma proporção de .7, isto é, 70% de dados para o treino e 30% para o teste e corresponde ao item 10 da figura 5.

Quanto ao terceiro bloco, as amostras de dados particionadas anteriormente são enviadas para o *Test & Score*, item 11 da figura. Neste bloco, executam-se o **treinamento e teste dos dados** e a **avaliação dos classificadores**, elencados na figura pelos itens 7, 8 e 9. No *Teste e Escore*, clica-se em *Cross Validation*, a fim de que o teste seja feito em diversos blocos da base de treinamento para cada algoritmo; o *Cross Validation* ou Validação Cruzada é uma técnica que avalia a performance do algoritmo, "quebrando" os dados de treinamento em partes iguais. Neste estudo, optamos por uma "quebra" de dez partes para os dados de treinamento. Após o ajuste da Validação Cruzada e o envio dos dados de teste ao *Test & Score*, carregamos os três classificadores elencados para esta pesquisa. Optamos pelo uso dos algoritmos *Support Vector Machine* (SVM), *Random Forest* (RF) e *Logistic Regression* (LR), cujo detalhamento de informações pode ser encontrado clicando-se com o botão direito no *widget* do algoritmo. Cada um dos classificadores foi ajustado, a fim de obtermos as melhores combinações de parâmetros possíveis. Por exemplo, no caso do RF, para esta *tagset*, a melhor combinação deu-se entre 500 árvores de decisão, com sete variáveis para cada divisão das árvores. A verificação dessas combinações é feita em tempo real pelo *widget Test & Score* (11), mantendo-se as janelas de ambos os *widgets* (classificador e *Test & Score*) abertos e ir ajustando os parâmetros do algoritmo. Ainda neste terceiro bloco, os itens 12 e 13 correspondem à *curva ROC* e à matriz de confusão, respectivamente. Trata-se da **avaliação dos classificadores** utilizados, propriamente dita. A matriz de confusão mostra diversos resultados, como acurácia, precisão, *recall* etc.; já a *curva ROC* serve como diagnóstico dos classificadores, medindo a razão entre os verdadeiros positivos e falsos positivos. Para a matriz de confusão, usamos a acurácia, no intuito de medir a eficiência de acerto dos algoritmos. Um exemplo de matriz de confusão pode ser visto abaixo.

**Figura 6**: matriz de confusão para o *Random Forest*.



Confusion matrix for RF-500-7 (showing number of instances)

| | | Predicted CMC | TAG | ∑ |
|---|---|---|---|---|
| Actual | CMC | 87 | 18 | 105 |
| | TAG | 9 | 96 | 105 |
| | ∑ | 96 | 114 | 210 |

Fonte: AUTOR.

A matriz de confusão para o RF – com 500 árvores e sete variáveis em cada divisão – previu corretamente 87 textos para CMC e 96 para TAG, isto é, uma porcentagem de 90,6% para CMC e 84,2% para TAG, em uma base de treino de 210 observações, como se nota na figura 6. A seguir, estendemos uma segunda linha do *widget Data Sample* para o *Test & Score*, a fim de enviarmos os dados restantes, ou seja, os dados de teste. É possível ver o desempenho dos dados de teste, por meio dos mesmos *widgets curva ROC* e Matriz de Confusão.

Após as avaliações, o bloco 4 consiste na **predição dos dados novos**, isto é, na atribuição de autoria do corpus CC, principal objeto deste estudo. Para isso, os dados são carregados por meio do *widget File Import*, item 14 da figura 5. Em seguida, o item 15 permite o ajuste correto das variáveis a serem enviadas ao item 16, que corresponde ao *widget de* predições. Os três classificadores (itens 7, 8 e 9) são finalmente ligados ao item 16 de predições, de modo que possam analisar os dados desconhecidos.

Finalmente, no bloco 5, os resultados são enviados para o item 17, em que o perito linguista procede à **verificação dos acertos e erros de atribuição de autoria** do *corpus* CC.

Todo o procedimento é repetido para as duas *tagsets* restantes do português, obviamente, salvando cada *workflow* em arquivos separados. Os três *workflows* foram então comparados e apenas o resultado com a melhor *tagset* foi usada para o segundo estudo, de modo a confirmamos ou não sua eficácia.

Discutiremos os resultados de todo esse processo na seção a seguir.

# 5 RESULTADOS

## 5.1 Estudo 1 – As *Cartas Chilenas*

Matrizes de confusão são o meio mais fácil de identificar a eficiência dos classificadores. Resumimos apenas os resultados da acurácia dos três classificadores e suas *tagsets* no quadro 6[9]. A acurácia se refere à taxa de predições corretas, isto é, à divisão das predições corretas pelo total de observações:

$$AC = \frac{VP + VN}{VP + VN + FP + FN}$$

onde AC corresponde à acurácia, VP (verdadeiro positivo), VN (verdadeiro negativo), FP (falso positivo) e FN (falso negativo).

**Quadro 8**: Resultado das matrizes de confusão dos três *workflows* nos dados de teste para o estudo 1.

| **WORKFLOW 1** | | | |
|---|---|---|---|
| **CLASSIFICADOR** | **TAGSET** | **RESULTADOS** | |
| SVM (C :0.50, g: auto, c=1.05) | EAGLES REDUZIDA | TAG: 92.7% | CMC: 85.7% |
| RF (trees: 500, atributes: 7) | EAGLES REDUZIDA | TAG: 88.4% | CMC: 85.1% |
| LR (Lasso - C = 0.12) | EAGLES REDUZIDA | TAG: 86.8% | CMC: 76.9% |
| **WORKFLOW 2** | | | |
| **CLASSIFICADOR** | **TAGSET** | **RESULTADOS** | |
| SVM (SIGM – C: 0.5, g: 0.01, c: 1.09) | UD_PORTUGUESE-BOSQUE | TAG: 91.7% | CMC: 77.8% |
| RF (trees: 500, atributes: 7) | UD_PORTUGUESE-BOSQUE | TAG: 88.9% | CMC: 88.9% |
| LR (Lasso - C =4) | UD_PORTUGUESE-BOSQUE | TAG: 89.7% | CMC: 80.4% |
| **WORKFLOW 3** | | | |
| **CLASSIFICADOR** | **TAGSET** | **RESULTADOS** | |
| SVM (POL – C: 0.5, g: 0.03, c: 1.13, d: 1,5) | EAGLES FINE-GRAINED | TAG: 87.5% | CMC: 80% |
| RF (trees: 500, atributes: 9) | EAGLES FINE-GRAINED | TAG: 85.7% | CMC: 92.7% |
| LR (Ridge – C: 0.3) | EAGLES FINE-GRAINED | TAG: 95.1% | CMC: 87.8% |

Fonte: AUTOR.

Os resultados do quadro acima referem-se à porcentagem de acerto de cada autor nos dados de teste, isto é, após o ajuste dos parâmetros e treinamento dos classificadores na base de treino. Por exemplo, no *workflow 1*, com a *tagset Eagles Reduzida*, o SVM[10]

classificou corretamente 92,7% dos textos de TAG e 85,7% dos textos de CMC, e assim por diante. Embora sejam valores altos, com um média de 85% de sucesso para os três *workflows* – média muito significativa – a classificação de dados novos (CC) não corresponde, necessariamente, à taxa de sucesso dos testes, como se nota no quadro abaixo.

**Quadro 9**: Resultados de classificação dos três *workflows* nas *Cartas Chilenas*.

| **WORKFLOW 1** | |
|---|---|
| **CLASSIFICADOR** | **RESULTADO** |
| SVM | 69.2% |
| RF | 38.4% |
| LR | 76.9% |
| **WORKFLOW 2** | |
| **CLASSIFICADOR** | **RESULTADO** |
| SVM | 38.4% |
| RF | 23% |
| LR | 84.6% |
| **WORKFLOW 3** | |
| **CLASSIFICADOR** | **RESULTADO** |
| SVM | 46.1% |
| RF | 7.7% |
| LR | 38.4% |

Fonte: AUTOR.

Claramente, o *workflow* 2 obteve o melhor desempenho em relação ao restante, por meio do classificador LR e da *tagset* UPB, classificando corretamente como de TAG onze das treze CC. O RF obteve o pior desempenho (*workflow* 3), com uma taxa de acerto de 7,7% (2 de 13 cartas). Já o SVM obteve resultados medianos, com uma média de acerto de 51%, o que o coloca como o segundo melhor classificador.

## 5.2 Estudo 2 – CJ

A fim de confirmarmos o desempenho do *workflow* 2, conduzimos um segundo estudo, dessa vez, com o CJ. Mantivemos os três classificadores e apenas a *tagset* UPB. Os quadros 9 e 10 representam, respectivamente, a matriz de confusão e o resultado no CJD.

**Quadro 10**: Resultados das matrizes de confusão do *workflow* 2 nos dados de teste para o estudo 2.

| ***WORKFLOW* 2** | | | |
|---|---|---|---|
| **CLASSIFICADOR** | **TAGSET** | **RESULTADOS** | |
| SVM | UD_PORTUGUESE-BOSQUE | EL: 91% | VR: 95.1% |
| RF | UD_PORTUGUESE-BOSQUE | EL: 86.6% | VR: 87.8% |
| LR | UD_PORTUGUESE-BOSQUE | EL: 96.8% | VR: 95.6% |

Fonte: AUTOR.

De acordo com o quadro acima, nos dados de teste, o LR obteve a melhor performance: 96,8% de acerto para Elaine e 95,6% para Vera. No CJD (quadro 10), mais uma vez, tal como no estudo 1, observa-se que o LR obteve o melhor desempenho possível, acertando todos os textos das jornalistas, o que aponta para o LR – juntamente com a *tagset* UPB – como um bom candidato na condução inicial de análise de autoria de textos.

**Quadro 11**: Resultados dos três *workflows* nos dados novos do CJ.

| ***WORKFLOW* 3 – TAGSET: UD_PORTUGUESE-BOSQUE** | |
|---|---|
| **CLASSIFICADOR** | **RESULTADO** |
| SVM | ELAINE: 80% VERA: 80% |
| RF | ELAINE: 80% VERA: 80% |
| LR | ELAINE: 100% VERA: 100% |

Fonte: AUTOR.

## CONSIDERAÇÕES FINAIS

Em Linguística Forense, nomeadamente na área de atribuição de autoria, marcadores de estilo são um dos elementos mais importantes na busca do estilo de um determinado autor ou suspeito de um crime que envolva dados linguísticos. Para marcadores do tipo *top-down*, isto é, marcadores informados por alguma teoria, sua seleção e manipulação podem se tornar difíceis, devido à necessidade de manuseio de ferramentas tecnológicas nem sempre acessíveis aos iniciantes, como *Python*, *Java*, *R*, etc. Este artigo procurou fornecer indícios de que ferramentas gratuitas podem ser um substituto inicial para a abordagem de atribuição de autoria de textos.

Como vimos, lançamos mão de duas ferramentas gratuitas: o *TreeTagger*, etiquetador de palavras automático, e o *Orange Canvas*, programa *open-source* para mineração de dados. Etiquetamos os dois *corpora* de estudo com as três *tagsets* disponíveis para o português no site oficial do *TreeTagger* e, logo em seguida, realizamos três *workflows* no *Orange* para os *corpora* CPM e CC. Obtivemos resultados muito significativos nos testes com o *workflow* 2 (*tagset* UPB) em que o classificador SVM obteve uma acurácia média de 84,7%; o RF, de 88,9% e o LR, de 85%. Nas CC, isto é, nos dados novos, o LR obteve a melhor performance, com uma média de acerto de 84,6%.

Realizamos um segundo estudo, a fim de confirmarmos a eficiência do *workflow* 2, por meio do corpus CJ. Novamente, o classificador LR mostrou-se superior aos demais, com uma performance de acerto no CJD de 100%. Portanto, de um modo geral, os resultados sugerem que o uso do LR, juntamente com a *tagset* UPB, fornecem a melhor combinação nos estágios iniciais da atribuição de autoria. Da mesma forma, também é possível inferir que o RF não é melhor opção para casos de autoria, uma vez que, em todos os *workflows*, seus resultados foram muito abaixo da expectativa, com uma média de 23%. No entanto, é difícil afirmar que o LR e a *tagset* UPB serão sempre a melhor solução para todos os tipos de casos.

Essa limitação advém em parte por conta da acurácia do próprio etiquetador, cuja eficiência não atinge o limite possível. Embora o tamanho do *corpus* de treinamento tenha um peso significativo na sua acurácia, Tian e Lo (2015) afirmam a necessidade de se treinar os etiquetadores em função dos erros reportados, mais do que em *corpora* maiores. Assim, estudos futuros e mais aprofundados ainda são necessários, não só da combinação dos classificadores, mas também da acurácia do etiquetador.

No entanto, há indícios óbvios do potencial da combinação dessas duas ferramentas para uma primeira abordagem dos textos por parte do perito. Por exemplo, pode-se usar a classificação dos melhores marcadores de estilo, por meio do *widget Rank*, para a condução de outros testes mais aprofundados, não só qualitativos, mas também quantitativos, como o *t-score*, *z-score*, erro amostral, etc., além dos próprios resultados de classificação. Em suma, acreditamos no potencial dessas duas ferramentas, e que esse tipo de primeira abordagem dos textos possa trazer ganhos significativos para a comunidade da linguística forense.

## Referências

AIRES, R. V. X. et al. Combining Multiple Classifiers to Improve Part of Speech Tagging: A Case Study for Brazilian Portuguese. *Proceedings of the 15th Brazilian Symposium on Artificial Intelligence*, 2000.

ALJABERY, M. A.; KURNAZ, S. Applying datamining techniques to predict hearing aid type for audiology patients. *Journal of Information Science and Engineering*, v. 36, n. 2, p. 205-215, 2020.

ALVES, A. F. F. et al. Combining machine learning and texture analysis to differentiate mediastinal lymph nodes in lung cancer patients. *Physical and Engineering Sciences in Medicine*, v. 44, n. 2, p. 387-394, 2021.

APAMPA, O. Evaluation of Classification and Ensemble Algorithms for Bank Customer Marketing Response Prediction. *Journal of International Technology & Information Management*, v. 25, n. 4, p. 85-100, 2016.

BERBER SARDINHA, T. *Linguística de Corpus.* São Paulo: Manole, 2004.

BOSI, A. *História Concisa da Literatura Brasileira.* São Paulo: Cultrix, 2017.

DEMŠAR, J.; ZUPAN, B. Orange: Data mining fruitful and fun - A historical perspective. *Informatica (Slovenia)*, v. 37, n. 1, p. 55-60, 2013.

GAMALLO, P.; GARCIA, M. *FreeLing e TreeTagger: Um Estudo Comparativo no Âmbito do Português.* [s.l: s.n.], 2013. (no prelo)

HARO RIVERA, S. et al. MÉTODOS DE CLASIFICACIÓN EN MINERÍA DE DATOS METEOROLÓGICOS. *Perfiles*, v. 2, n. 20, p. 107-113, 2018.

HOU, R; HUANG, R. Robust stylometric analysis and author attribution based on tones and rimes. *Natural Language Engineering*, 26(1), Cambridge University Press: p. 49–71, 2020.

JOVIĆ, A.; BRKIĆ, K.; BOGUNOVIĆ, N. An overview of free software tools for general data mining. *37th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2014 - Proceedings.* Anais.2014.

JUOLA, Patrick; VESCOVI, Darren. Analyzing Stylometric Approaches to Author Obfuscation. In: *Advances in Digital Forensics VII – International Conference on Digital Forensics, 2011, Orlando*. Orlando: Springer, pp.115-125, 2011.

KLUNNIKOVA, Y. V. et al. Machine learning application for prediction of sapphire crystals defects. *Journal of Electronic Science and Technology*, v. 18, n. 1, p. 1-8, 2020.

KOPPEL, M; SCHLER, J; ARGAMON, S. Computational Methods in Authorship Attribution. *Journal of the American Society for Information Science and Technology (JASIST)*. Vol. 60, pp 9-26, 2009.
MACIEJ, E. Style-markers in authorship attribution: a cross-language study of authorial fingerprint. *Studies in Polish Linguistics*, v. 6, p. 99–114, p. 99-114, 2011.

MARCONDES, M. V.; BERBER SARDINHA, T. *Temos uma digital linguística? Detecção de autoria com Linguística Forense e Linguística de Corpus.* São Paulo: [s.n.], 2021. (não publicado)

MCMENAMIN, G. R. *Forensic linguistics : advances in forensic stylistics.* Florida: CRC Press, 2002.

MENDES, M. DE F. *Cartas Chilenas, de Tomás Antônio de Gonzaga: um estudo historiográfico dos recursos linguísticos e argumentativos.* [s.l.] Pontifícia Universidade Católica, São Paulo, 2010.

OLSSON, J. *Forensic Linguistics.* London: Continuum, 2008.

P. JEEVAN KUMAR, G; SRIKANTH REDDY, T; AGHUNADHA, R. Document Weighted Approach for Authorship Attribution. *International Journal of Computational Intelligence Research.* Volume 13, Número 7, pp. 1653-1661, 2017.

PAVELEC, D; JUSTINO, E; OLIVEIRA, LUIZ S. Author Identification using Stylometric Features. *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial*, Vol. 11, núm.36, pp. 59-65, 2007.

PENG, F; SCHUURMANS, D; WANG, S. Augmenting Naive Bayes Classifiers with Statistical Language Models. *Information Retrieval* **7**, pp. 317–345, 2004.

RADEMAKER, A. et al. *Universal Dependencies for Portuguese.* Disponível em: <https://github.com/UniversalDependencies/UD_Portuguese-Bosque>. Acesso em: 10 jan. 2022.

SCHMID, H. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing. Anais.* Manchester, UK.: 1994

SHI, J.; ZHAO, G.; WEI, Y. Computational QSAR model combined molecular descriptors and fingerprints to predict HDAC1 inhibitors. *Medecine/Sciences,* v. 34, p. 52-58, 2018.

SHRESTHA, P; SIERRA, S; GONZÁLEZ, F; MONTES, M; ROSSO, P; SOLORIO, T. CONVOLUTIONAL. *Neural Networks for Authorship Attribution of Short Texts.* pp. 669-674, 2017.

SOUSA-SILVA, R.; COULTHARD, M. Linguística Forense. In: DINIS-OLIVEIRA, R. J.; MAGALHÃES, T. (org.). *O que são as Ciências Forenses?* Conceitos, Abrangência e Perspectivas Futuras. Lisboa: Pactor, 2016, p. 137-144.

STRAZAR, M. et al. ScOrange - A tool for hands-on training of concepts from single-cell data analytics. *Bioinformatics. Anais.*2019.

TIAN, Y.; LO, D. A comparative study on the effectiveness of part-of-speech tagging techniques on bug reports. *IEEE 22nd International Conference on Software Analysis, Evolution, and Reengineering, SANER 2015 - Proceedings. Anais.*2015.

TOPLAK, M. et al. Quasar: Easy machine learning for biospectroscopy. *Cell***s**, v. 10, n. 9, p. 1-10, 2021.

WILLIANS, B. *Style and vocabular: Numerical studies*. Londres: Griffin, 1970.
ZUPAN, B.; DEMSAR, J. Open-Source Tools for Data Mining. *Clinics in Laboratory Medicine*, v. 28, n. 1, p. 37-54, 2008.

Recebido em: 03/03/2022
Aceito em: 30/11/2022

[1] O *Treetagger* não gera três colunas nomeadas. Trata-se apenas de uma apresentação didática dos resultados.

[2] www.dominiopublico.gov.br/

[3] Período de 01/01/2019 a 31/07/2020.

[4] O *Text-Cleaner* contém scripts criados pelo autor do artigo e uma interface gráfica criada pelo programador *Python* Roger Amaro.

[5] Criado pelo autor desta pesquisa.

[6] Criado pelo Professor Doutor Tony Berber Sardinha, da PUC-SP.

[7] Mantivemos, no entanto, os nomes originais dos *widgets* nas explicações, uma vez que são facilmente identificados no Orange pelos seus desenhos característicos.

[8] O Orange identificou corretamente 300 observações, 66 variáveis linguísticas contínuas (produzidas pelo *TreeTagger*) e uma variável categórica (a coluna autor, produzida manualmente).

[9] A matriz de confusão também mostra outros valores, como os de Tipo I e Tipo II, além das classificações incorretas.

[10] Em parênteses, para cada classificador, estão relacionados os parâmetros ajustados, por meio dos *widgets*.