UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ARTES, CIÊNCIAS E HUMANIDADES
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

CAIO DEUTSCH

Atribuição autoral com uso de classificadores de perfis

São Paulo

2020

CAIO DEUTSCH

Atribuição autoral com uso de classificadores de perfis

Dissertação apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação.

Área de concentração: Metodologia e Técnicas da Computação

Versão corrigida contendo as alterações solicitadas pela comissão julgadora em 05 de outubro de 2020. A versão original encontra-se em acervo reservado na Biblioteca da EACH-USP e na Biblioteca Digital de Teses e Dissertações da USP (BDTD), de acordo com a Resolução CoPGr 6018, de 13 de outubro de 2011.

Orientador: Prof. Dr. Ivandré Paraboni

São Paulo

2020

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

CATALOGAÇÃO-NA-PUBLICAÇÃO

(Universidade de São Paulo. Escola de Artes, Ciências e Humanidades. Biblioteca)
CRB 8 - 4936

Deutsch, Caio

Atribuição autoral com uso de classificadores de perfis / Caio
Deutsch ; orientador, Ivandré Paraboni. – 2020.
94 f : il.

Dissertação (Mestrado em Ciências) - Programa de Pós-
Graduação em Sistemas de Informação, Escola de Artes,
Ciências e Humanidades, Universidade de São Paulo.
Versão corrigida

1. Reconhecimento de padrões. 2. Inteligência artificial. 3.
Processamento de língua natural. 4. Reconhecimento de
texto. I. Paraboni, Ivandré, orient. II. Título

CDD 22.ed.– 006.4

Dissertação de autoria de Caio Deutsch, sob o título ”**Atribuição autoral com uso de classificadores de perfis**”, apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação, na área de concentração Metodologia e Técnicas da Computação, aprovada em 05 de outubro de 2020 pela comissão julgadora constituída pelos doutores:

Prof. Dr. Ivandré Paraboni

Universidade de São Paulo

Presidente

Prof. Dr. Ariadne Maria Brito Rizzoni Carvalho

Universidade Estadual de Campinas

Prof. Dr. José de Jesus Pérez Alcazár

Universidade de São Paulo

Agradecimentos

A realização deste mestrado me proporcionou conhecimento, resiliência e persistência, o qual estou imensamente grato.

Aos meus pais Noely Aparecida de Oliveira Deutsch e Hugo Deutsch, que me inspiraram desde pequeno com exemplo de muito trabalho e carinho.

A minha noiva Thaís Sêneda de Mecnas, que me apoiou em momentos difíceis com paciência, parceria e principalmente amor.

A minha amiga Clarissa Flor da Vitória Vieira, que me apoiou a fazer este mestrado e acreditou no meu potencial.

Ao meu amigo Arthur Marçal Flores, que participou da jornada de fazer o mestrado comigo, aprendendo e compartilhando experiências.

E finalmente ao meu orientador Ivandré Paraboni, pela dedicação e ideias importantes para a produção deste trabalho.

Resumo

DEUTSCH, Caio. **Atribuição autoral com uso de classificadores de perfis**. 2020. 94 f. Dissertação (Mestrado em Ciências) - Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2020.

A atribuição autoral (AA) busca identificar um autor de texto a partir de um conjunto de autores conhecidos. Autores deixam rastros em seus textos e é possível identificar características sociolinguísticas baseadas no estilos de escrita refletidos no texto destes autores. A atribuição autoral está cada vez mais demonstrando importância para diversas atividades sociais, em especial para a análise forense. Os trabalhos envolvendo AA demonstram resultados modestos e motivam a exploração de diferentes técnicas para melhorar a acurácia dos modelos atuais. A partir desses pontos, o presente trabalho apresenta uma proposta de pesquisa em nível de mestrado no campo de processamento de língua natural (PLN), com ênfase em AA, com o objetivo geral de melhorar o desempenho de classificadores de atribuição autoral utilizando técnicas de caracterização autoral (CA).

Palavras-chaves: Atribuição autoral. Caracterização autoral. Aprendizado de máquinas. Processamento de língua natural. Distorções textuais.

Abstract

DEUTSCH, Caio. **Author attribution using profile classifiers**. 2020. 94 p. Dissertation (Master of Science) - School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2020.

Author attribution (AA) seeks to identify a text author from a set of known authors. Authors leave traces in their texts and it is possible to identify sociolinguistic characteristics based on the writing styles reflected in the text of these authors. Author attribution is increasingly showing importance for various social activities, especially forensic analysis. Studies involving AA show modest results and motivate the exploration of different techniques to improve the accuracy of current models. From these perspective, this project presents a master's level research proposal in the field of natural language processing (NLP), with an emphasis in AA, with the general objective of improving the performance of AA classifiers using author profiling techniques.

Keywords: Author attribution. Author characterization. Machine learning. Natural language processing. Textual distortions.

Lista de figuras

Figura 1 – Arquitetura do modelo Word2vec	30
Figura 2 – Visualização do papel dos classificadores em AA	31
Figura 3 – Arquitetura do EACH-USP modelo <i>ensemble</i>	49
Figura 4 – Arquitetura do <i>ensemble</i> CA-Label	57
Figura 5 – Acurácia por quantidade de autores	59
Figura 6 – <i>Ensemble</i> CA-Prob ou CA-Label	61
Figura 7 – Curva de acurácia por quantidade de autores de AA com informação de gênero para o corpus b5-post	70
Figura 8 – Curva de acurácia por quantidade de autores de AA com informação de formação em TI para o corpus b5-post	71
Figura 9 – Curva de acurácia por quantidade de autores de AA com informação de idade para o corpus b5-post	71
Figura 10 – Curva de acurácia por quantidade de autores de AA com informação de gênero para o corpus BRmoral	75
Figura 11 – Curva de acurácia por quantidade de autores de AA com informação de formação em TI para o corpus BRmoral	75
Figura 12 – Curva de acurácia por quantidade de autores de AA com informação de idade para o corpus BRmoral	76
Figura 13 – Curva de acurácia por quantidade de autores de AA com informação de escolaridade para o corpus BRmoral	76
Figura 14 – Curva de acurácia por quantidade de autores de AA com informação de visão política para o corpus BRmoral	77
Figura 15 – Curva de acurácia por quantidade de autores de AA com informação de grau de religiosidade para o corpus BRmoral	77
Figura 16 – Curva de acurácia por quantidade de autores de AA com informação de gênero para o corpus <i>The Blog Authorship</i>	78
Figura 17 – Curva de acurácia por quantidade de autores de AA com informação de idade para o corpus <i>The Blog Authorship</i>	79
Figura 18 – Curva de acurácia por quantidade de autores de AA com informação de gênero para o subcorpus <i>Twisty</i> e idioma espanhol	82

Figura 19 – Curva de acurácia por quantidade de autores de AA com informação de gênero para o subcorpus <i>Twisty</i> e idioma francês	82
Figura 20 – Curva de acurácia por quantidade de autores de AA com informação de gênero para o subcorpus <i>Twisty</i> e idioma português	83
Figura 21 – Curva de acurácia por quantidade de autores de AA com informação de gênero para o subcorpus <i>Twisty</i> e idioma italiano	83
Figura 22 – Curva de acurácia por quantidade de autores de AA com informação de gênero para o subcorpus <i>Twisty</i> e idioma alemão	83
Figura 23 – Curva de acurácia por quantidade de autores de AA com informação de gênero para o subcorpus <i>Twisty</i> e idioma holandês	84

Lista de quadros

Quadro 1 – Tipos de recursos estilométricos	21
Quadro 2 – Histórico de PAN-CLEF para AA	22
Quadro 3 – Histórico de PAN-CLEF para CA	24
Quadro 4 – Conjuntos de dados avaliados	41
Quadro 5 – Exemplo da aplicação dos algoritmos DV-MA e DV-SA	48
Quadro 6 – Resumo dos estudos da área de AA considerados	55
Quadro 7 – Córpus considerados	61
Quadro 8 – Configuração das bases de AA e CA	64

Lista de tabelas

Tabela 1 – Resultados médios do modelo CA-Label com diferentes margens de acerto comparados ao <i>baseline</i> de Custódio e Paraboni (2018)	58
Tabela 2 – Exemplo de como combinar características preditas de CA e modelos de AA utilizando probabilidades	62
Tabela 3 – Exemplo de como combinar características preditas de CA e modelos de AA utilizando rótulos	63
Tabela 4 – Valores ótimos de k para seleção univariada nas tarefas de CA consideradas	65
Tabela 5 – Resultados de classificação de CA para o corpus b5-post	67
Tabela 6 – Resultados de classificação de CA para o corpus BRmoral	67
Tabela 7 – Resultados de classificação de CA para o corpus <i>The Blog Authorship</i> .	67
Tabela 8 – Resultados de classificação de CA para o corpus <i>Twisty</i>	68
Tabela 9 – Acurácia média de AA no corpus b5-post	68
Tabela 10 – Acurácia por quantidade de autores de AA com informação de gênero para o corpus b5-post	69
Tabela 11 – Acurácia por quantidade de autores de AA com informação de formação em TI para o corpus b5-post	69
Tabela 12 – Acurácia por quantidade de autores de AA com informação de idade para o corpus b5-post	70
Tabela 13 – Acurácia média de AA no corpus BRmoral	71
Tabela 14 – Acurácia por quantidade de autores de AA com informação de gênero para o corpus BRmoral	72
Tabela 15 – Acurácia por quantidade de autores de AA com informação de formação em TI para o corpus BRmoral	72
Tabela 16 – Acurácia por quantidade de autores de AA com informação de idade para o corpus BRmoral	73
Tabela 17 – Acurácia por quantidade de autores de AA com informação de escolaridade para o corpus BRmoral	73
Tabela 18 – Acurácia por quantidade de autores de AA com informação de grau de religiosidade para o corpus BRmoral	74

Tabela 19 – Acurácia por quantidade de autores de AA com informação de visão política para o corpus BRmoral	74
Tabela 20 – Acurácia média de AA no corpus <i>The Blog Authorship</i>	77
Tabela 21 – Acurácia por quantidade de autores de AA com informação de gênero para o corpus <i>The Blog Authorship</i>	78
Tabela 22 – Acurácia por quantidade de autores de AA com informação de idade para o corpus <i>The Blog Authorship</i>	78
Tabela 23 – Acurácia média de AA no corpus <i>Twisty</i>	79
Tabela 24 – Acurácia por quantidade de autores de AA com informação de gênero para o subcorpus <i>Twisty</i> e idioma espanhol	80
Tabela 25 – Acurácia por quantidade de autores de AA com informação de gênero para o subcorpus <i>Twisty</i> e idioma francês	80
Tabela 26 – Acurácia por quantidade de autores de AA com informação de gênero para o subcorpus <i>Twisty</i> e idioma português	80
Tabela 27 – Acurácia por quantidade de autores de AA com informação de gênero para o subcorpus <i>Twisty</i> e idioma italiano	81
Tabela 28 – Acurácia por quantidade de autores de AA com informação de gênero para o subcorpus <i>Twisty</i> e idioma alemão	81
Tabela 29 – Acurácia por quantidade de autores de AA com informação de gênero para o subcorpus <i>Twisty</i> e idioma holandês	81
Tabela 30 – Resultados de AA - Média de acurácia com tarefas combinadas de CA para o corpus b5-post	84

Lista de abreviaturas e siglas

PLN	Processamento de língua natural
AA	Atribuição autoral
CA	Caracterização autoral
BOW	<i>Bag of Word</i> ou modelo de espaço de vetores
EN	Língua inglesa
ES	Língua espanhola
DU	Língua holandesa
FR	Língua francesa
GE	Língua alemã
PL	Língua polonesa
PT	Língua portuguesa
IT	Língua italiana
AR	Língua árabe

Sumário

1	Introdução	16
1.1	<i>Hipótese</i>	17
1.2	<i>Objetivo</i>	18
1.3	<i>Método</i>	18
1.4	<i>Organização do documento</i>	18
2	Conceitos fundamentais	19
2.1	<i>Atribuição Autoral</i>	19
2.2	<i>Caracterização Autoral</i>	22
2.3	<i>Córpus para CA</i>	25
2.4	<i>Modelos de representação de texto</i>	27
2.4.1	Modelos tradicionais de n-gramas, bag of n-gramas e variações . . .	27
2.4.2	Modelos de <i>word embeddings</i>	29
2.5	<i>Métodos de aprendizado supervisionado de máquinas aplicados a AA .</i>	30
2.6	<i>Modelos baseados em distância específicos para AA</i>	33
2.7	<i>Considerações</i>	34
3	Revisão Bibliográfica	36
3.1	<i>Abordagens tradicionais de AA</i>	36
3.1.1	Uma nova abordagem para atribuição autoral	36
3.1.2	Segmentação de autores com técnicas de segmentação hierárquica .	37
3.1.3	Explorando <i>Word Embedding</i> e n-gramas de Caracteres para seg- mentação de autores	38
3.1.4	Atribuição Autoral em mídias sociais	38
3.1.5	Atribuição Autoral utilizando método de identificação de impostores	40
3.1.6	Explorando características importantes da atribuição autoral	41
3.2	<i>Uso de perfis autorais em AA</i>	42
3.2.1	Modelos baseados em perfil para resolução de atribuição autoral . .	43
3.2.2	Utilização de técnicas baseadas em perfil para identificação de gênero e idade	44
3.2.3	Utilizando modelagem de tópicos	45

3.3	<i>Outras abordagens</i>	46
3.3.1	Representado estilo de escrita em atribuição autoral	46
3.3.2	Atribuição Autoral utilizando distorção de texto	47
3.3.3	Aplicações entre domínios na AA	49
3.3.4	Melhorando a atribuição autoral de tópicos cruzados através do pré processamento dos dados	50
3.3.5	Atribuição Autoral por consenso entre características	51
3.3.6	A semântica de <i>frames</i> na atribuição autoral	51
3.4	<i>Considerações</i>	53
4	Estudos exploratórios	56
4.1	<i>Modelo proposto</i>	56
4.2	<i>Conjunto dos dados</i>	57
4.3	<i>Procedimento</i>	58
4.4	<i>Resultados</i>	58
4.5	<i>Discussão</i>	59
5	Metodologia	60
5.1	<i>Córpus</i>	60
5.2	<i>Modelos Desenvolvidos</i>	61
6	Avaliação	64
6.1	<i>Conjuntos de treinamento e teste</i>	64
6.2	<i>Procedimento</i>	65
6.2.1	Avaliação dos classificadores de CA individuais	65
6.2.2	Avaliação do <i>ensemble</i> de AA	66
7	Resultados	67
7.1	<i>Caracterização Autoral</i>	67
7.2	<i>Atribuição Autoral</i>	68
7.2.1	Domínio Facebook (córpus b5-post)	68
7.2.2	Domínio opinativo (córpus BRmoral)	71
7.2.3	Domínio de blogs (córpus <i>The Blog Authorship</i>)	77
7.2.4	Domínio Twitter (córpus <i>Twisty</i>)	79

7.2.5	Análise complementar: uso de múltiplos classificadores de CA . . .	84
7.3	<i>Considerações</i>	85
8	Conclusão	86
8.1	<i>Contribuições</i>	86
8.2	<i>Limitações e trabalhos futuros</i>	86
	Referências ¹	87

¹ De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

1 Introdução

A atribuição autoral (AA) (do inglês, *Authorship Attribution*) busca identificar o autor de texto a partir de um conjunto de autores conhecidos (POTTHAST *et al.*, 2017). Quando os textos destes autores pertencem a um dos candidatos a serem identificados, este conjunto é denominado fechado (KESTEMONT *et al.*, 2018). Já quando os textos pertencem a autores que não estão no conjunto de autores a serem identificados, o conjunto é denominado aberto. O escopo deste trabalho utiliza exclusivamente conjuntos fechados.

Do ponto de vista da pesquisa em Processamento de Línguas Naturais (PLN), a AA é um problema computacional atual e relevante, sendo inclusive tema de competições internacionais como a série PAN-CLEF (KESTEMONT *et al.*, 2018). A atribuição autoral é importante também para diversas aplicações do cotidiano da sociedade, destacam-se a detecção de plágio e investigação forense (SARI; STEVENSON; VLACHOS, 2018). Sua aplicação pode auxiliar em outras atividades, por exemplo, a desvendar a identidade de autores de notícias falsas (PENG; CHOO; ASHMAN, 2016), na atribuição autoral de código-fonte (ALSULAMI *et al.*, 2017), ou até mesmo na identificação de pseudônimos (JUOLA, 2012).

Desde o final da década de 1990, os estudos sobre AA sofreram uma mudança considerável. A vasta quantidade de textos eletrônicos disponíveis através da internet (e-mails, blogs, fóruns on-line etc) e mídias digitais aumentou a necessidade de lidar eficientemente com essa informação (POTHA; STAMATATOS, 2018). Este fato teve um impacto significativo em áreas de pesquisa como aprendizado de máquina e PLN.

Do ponto de vista computacional, a tarefa de AA pode ser tratada de diversas formas. No estudo de Jockers e Witten (2010), a tarefa é tratada como um problema de vizinhos próximos utilizando *k-nearest neighbors* (k-NN), e um problema de utilizando máquinas de vetores de suporte (SVM). Sua aplicação também se estende a redes neurais recorrentes (BAGNALL, 2016) e redes neurais de convolução (SOLORIO *et al.*, 2017).

De especial interesse para o presente trabalho, foi observado que o estudo em Custódio e Paraboni (2018) apresentou um modelo com distorção textual (STAMATATOS, 2009) que combina diversas fontes de conhecimentos em um *ensemble* de máquina para o problema de AA de conjunto fechado. O modelo final do trabalho obteve resultados

positivos para a tarefa de AA, sendo considerado o sistema de melhor desempenho global na competição PAN-CLEF de 2018 (KESTEMONT *et al.*, 2018).

Apesar destes resultados positivos, o problema de AA ainda possui espaço para melhorias. Avanços significativos têm sido observados, por exemplo, na área correlata de Caracterização Autoral (CA) ou Caracterização de Perfis Autorais, que visa identificar características de autores como gênero, faixa etária e outros (NGUYEN *et al.*, 2016). Sistemas de CA possuem aplicações como identificação de perfil do usuário com detecção automática de gênero (BURGER *et al.*, 2011), faixa etária (NGUYEN *et al.*, 2014), localização geográfica (EISENSTEIN *et al.*, 2010) ou afiliação de usuários (PIERGALLINI *et al.*, 2014) com base em uma análise de suas escolhas linguísticas.

Existe uma relação natural e pouco explorada na literatura entre os problemas computacionais de AA (que objetiva identificar um autor) e de CA (que objetiva estimar características autorais como gênero e idade). De forma mais específica, é possível usar conhecimento autoral (como gênero) proveniente de classificadores de CA, como de Santos e Paraboni (2019), para então reduzir o espaço de busca na tarefa de AA. Uma vez conhecido, por exemplo, o gênero do autor sob análise, esta informação efetivamente divide o conjunto de autores candidatos entre homens e mulheres, o que potencialmente aumenta acurácia da tarefa de identificação. Outras formas de classificação de CA como idade, visão política etc têm em princípio o mesmo potencial.

Com base nestas observações, o restante deste documento apresenta os resultados de pesquisa em nível de mestrado acadêmico tratando do problema computacional de AA com técnicas de CA.

1.1 Hipótese

A combinação de modelos de CA como os utilizados para predição de idade, gênero, religião, visão política, dentre outras características, pode ser um bom preditor para a tarefa de AA comparado a trabalhos já existentes e que não utilizam modelos de CA para a resolução de AA.

Esta hipótese será testada comparando-se o modelo de AA baseado em *ensemble* de Custódio e Paraboni (2018) com versões estendidas deste mesmo modelo às quais são acrescentados classificadores de CA. Espera-se que o uso de conhecimento de CA na tarefa

de AA apresente resultados superiores aos dos modelos que não dispõem deste tipo de conhecimento.

1.2 *Objetivo*

Propor modelos de atribuição autoral do tipo *ensemble* com auxílio de classificadores de caracterização autoral, de modo a obter resultados superiores ao de modelos que não utilizam este tipo de recurso.

1.3 *Método*

Este trabalho consiste em estender o modelo de AA denominado EACH-USP de Custódio e Paraboni (2018), acrescentando classificadores para diversas tarefas de CA. Serão utilizados conjuntos de dados em português disponibilizados pelos corpuses b5-post (RAMOS *et al.*, 2018) e BRmoral (SANTOS; PARABONI, 2019), *The Blog Authorship* (SCHLER *et al.*, 2006) no idioma inglês e *TwiSty* (VERHOEVEN; DAELEMANS; PLANK, 2016) nos idiomas espanhol, francês, português, italiano, alemão e holandês. Os resultados serão comparados a um *baseline* constituído do *ensemble* original de Custódio e Paraboni (2018) combinado com a classe majoritária de cada tarefa.

1.4 *Organização do documento*

O restante deste documento está organizado da seguinte forma. Os capítulos seguintes apresentam a revisão bibliográfica de conceitos fundamentais (Capítulo 2), discussões acerca de estudos correlatos encontrados na literatura (Capítulo 3), estudo exploratório (Capítulo 4), metodologia (Capítulo 5), avaliação (Capítulo 6), resultados (Capítulo 7) e conclusão (Capítulo 8).

2 Conceitos fundamentais

Este capítulo apresenta os conceitos fundamentais, relacionados ao escopo deste projeto de pesquisa, acerca das técnicas usadas na área de atribuição autoral.

2.1 Atribuição Autoral

A atribuição autoral é a tarefa de determinar o autor de um texto contestado, dado um conjunto de autores candidatos e amostras da sua escrita (JUOLA, 2008; STAMATATOS, 2009). A forma pela qual o autor escreve define uma subárea denominada estilometria, que compreende uma série de técnicas e medidas de escrita do autor, buscando extrair uma identidade de escrita padrão deste autor (NEAL *et al.*, 2018). Idealmente, a estilometria não deve ser afetada por mudanças nas variações de tópicos ou gêneros, e deve depender apenas do estilo de escrita pessoal dos autores (STAMATATOS, 2009).

Uma das principais aplicações da atribuição autoral através da identificação do estilo de escrita dos autores é a autoria forense. Esta tarefa procura identificar possíveis criminosos a partir da estilometria, prática que ganhou relevância nos últimos anos com as mídias sociais e a ampliação da internet em escala global. Também relacionado às evoluções da tecnologia, principalmente na área de armazenamento de dados não estruturados provenientes da internet como twitter, blogs, fóruns, foi possível avançar com diferentes técnicas de aprendizado de máquinas e medidas de captura do estilo de escrita dos autores no âmbito da atribuição autoral experimental (ROCHA *et al.*, 2017).

Do ponto de vista computacional, AA pode ser dividido em grupos (ROCHA *et al.*, 2017). Os primeiros estudos da estilometria são atribuídos à resolução da autoria de Augustus de Morgan através da frequência do comprimento das palavras em 1851 (NEAL *et al.*, 2018), estes estudos são denominados variantes unitários. A evolução destes estudos chegou em estatísticas multivariadas e mais recentemente a modelos envolvendo técnicas de similaridade.

Outro estudo envolvendo variantes unitárias é a lei de Zipf que relaciona a posição e a frequência das palavras. Essa relação foi descoberta por George Kingsley em 1932 (NEAL *et al.*, 2018), mas foi demonstrada ser limitada no trabalho de Savoy (2012). Estudos envolvendo estatística multivariada são iniciados com o trabalho de Mosteller e Wallace

(1963), onde uma versão do algoritmo Naïve Bayes é utilizado. Por fim, a atribuição autoral pode ser feita utilizando métodos baseados em vizinho mais próximo (STAMATATOS, 2009). Nesta abordagem, os documentos são organizados em vetores e assim comparados com alguma medida de distância ou dissimilaridade.

A comunidade científica caracterizou os tipos de conhecimento da estilometria em cinco níveis principais: léxico, caracteres, sintático, semântico e domínio (ou conteúdo) (NEAL *et al.*, 2018). Estas categorias são descritas brevemente a seguir.

A categoria léxica é uma maneira simples e intuitiva de visualizar um texto como uma sequência de *tokens* agrupados em sentenças, onde cada *token* representa uma palavra, número ou um sinal de pontuação. As primeiras tentativas de atribuição autoral foram baseadas em métricas simples, como contagens de comprimento das sentenças e contagens do comprimento das palavras (MENDENHALL, 1887). A vantagem de tais recursos simplificados é que eles podem ser aplicados a qualquer idioma e qualquer *corpus* sem premissas específicas (STAMATATOS, 2009).

A categoria de caracteres pode ser vista como uma sequência de caracteres e desta forma é possível definir várias medidas no nível de caractere como contagem de caracteres alfabéticos, dígitos de cada caractere, contagem de caracteres maiúsculos e minúsculos, frequências de letras, contagem de pontuação etc (VEL *et al.*, 2001). Esse tipo de informação é frequentemente disponível em qualquer idioma ou *corpus* e tem se provado bastante útil para quantificar o estilo de escrita do autor (GRIEVE, 2007).

A categoria sintática representa a tendência dos autores a usar padrões de sintaxe de forma similar inconscientemente, ou seja, a informação de sintaxe é considerada uma “impressão digital” mais confiável em comparação com as características léxicas. O sucesso das funções de palavras vem na representação do estilo de escrita, indicando uma utilidade da informação sintática, uma vez que são geralmente encontradas em certas estruturas. Por outro lado, este tipo de informação requer técnicas mais robustas de processamento de língua natural (PLN) para serem processadas (STAMATATOS, 2009).

Quanto mais detalhada a análise do texto necessária para extrair características estilométricas, menos precisas (e geralmente com mais ruído) as medidas são produzidas. Desta forma, existem técnicas de PLN capazes de capturar palavras e sentenças. São exemplos desta categoria a contagem das palavras, os analisadores semânticos (VARELA *et al.*, 2016), POS *tagging*, *word embeddings* (SARI; STEVENSON, 2016) etc. Estas técnicas são utilizadas para minimizar o ruído dos textos e extrair o significado semântico do autor.

Finalmente, a categoria de domínio, como mensagens de e-mail e mensagens de fóruns on-line, possibilita definir medidas estruturais para quantificar a estilometria. Estas medidas incluem o uso de saudações ou despedidas nas mensagens, tipos de assinaturas, uso de recuo, tamanho do parágrafo etc. Medidas relacionadas à distribuição de tags HTML (VEL *et al.*, 2001), contagem de cores de fonte e tamanho de fonte (ABBASI; CHEN, 2005) também podem ser utilizados para a escolha do autor em questão.

O Quadro 1 sumariza os tipos de recursos estilométricos e o conjunto de ferramentas e recursos computacionais necessários para suas medições.

Quadro 1 – Tipos de recursos estilométricos

Características	Ferramentas e recursos necessárias
Léxicas	Baseado em <i>tags</i> (comprimento de palavra, comprimento de sentenças, etc.)
	Riqueza de vocabulário
	Frequência de palavras
	Frequência de n-gramas
	Erros
Caracteres	Tipos de caracteres (letras, dígitos, etc)
	N-gramas de caracteres (tamanho fixo)
	N-gramas de caracteres (tamanho variável)
	Métodos de compressão
Sintáticas	Parte do discurso
	<i>Chunks</i>
	Frase e estrutura de frase
	Reescrita de regras de frequência
	Erros
Semânticas	Sinônimos
	Dependências semânticas
	Funcional
Domínios	Estrutural
	Conteúdo específico
	Linguagem específica

Fonte: Adaptado de Stamatatos (2009)

Tarefas de AA na competição PAN-CLEF

A competição PAN-CLEF (KESTEMONT *et al.*, 2018) reúne a comunidade científica para a resolução de tarefas compartilhadas, com o objetivo de avançar estudos de problemas relacionados a técnicas investigativas para textos digitais. Inicialmente, a

competição foi criada para explorar problemas relacionados a plágio. Sua primeira edição chamou *Plagiarism Analysis*.

AA tem sido tema de várias edições da competição PAN-CLEF. O quadro 2 sumariza o histórico das edições com suas respectivas características. A coluna *referência* traz a referência para os estudos que sumarizam as abordagens usadas pelos participantes. A coluna *idioma* especifica os tipos de idiomas abordados na competição do ano, os idiomas abordados foram o inglês (EN), espanhol (ES), italiano (IT), holandês (DU), português (PT), grego (GR) e polonês (PL). A coluna *Conjuntos de Dados* especifica os conjuntos de dados utilizados e por fim a coluna *Tarefas de Predição* especifica qual tarefa de AA foi predita. Os possíveis tipos são plágio (P), atribuição (A), verificação (V), clusterização (C) e quebra de estilo (Q).

Quadro 2 – Histórico de PAN-CLEF para AA

Referência	Idioma	Conjunto de Dados	Tarefas de Predição
Argamon e Juola (2011)	EN	<i>Enron</i> e-mails	A, V
Juola (2012)	EN	Facebook	A, C
Juola e Stamatatos (2013)	EN, ES e GR	textos publicados sobre ciências da computação	V
Stamatatos et al. (2014)	EN, ES, DU e GR	artigos de jornal e romances	V
Stamatatos et al. (2015)	EN, ES, DU e GR	textos de fóruns com gêneros e tópicos distintos	V
Rosso et al. (2016)	EN, DU e GR	artigos de jornais e revistas	C, P
Tschuggnall et al. (2017)	EN, ES, DU e GR	artigos de jornais e revistas	C, Q
Kestemont et al. (2018)	EN, ES, FR, IT e PL	textos de literatura de fantasia	A, Q

Fonte: Caio Deutsch 2020

Todas as edições utilizam o inglês como um dos idiomas abordados. A ultima edição em 2018 apresentou idiomas como o francês, italiano e polonês, nunca abordados em nenhuma edição do PAN-CLEF para AA anterior. Os textos com a temática de artigos de jornal são comuns, principalmente para a tarefa de verificação autoral (identificação de sub grupos antes da resolução da tarefa de AA).

2.2 Caracterização Autoral

A caracterização autoral (CA) (do inglês, *Author Profiling*) é uma subárea de PLN que procura identificar características de autores, como gênero ou idade (NGUYEN *et al.*, 2016). Estes padrões provêm de dialetos sociais representados por grupos de pessoas que compartilham características sociais semelhantes.

Embora o foco principal do presente estudo seja o problema de AA, modelos de CA serão um componente fundamental da solução a ser proposta. De forma mais específica,

modelos de CA serão empregados de modo a melhorar o desempenho de classificadores de atribuição autoral, como será discutido no Capítulo 4.

Do ponto de vista computacional, CA trata na maior parte das vezes da resolução de problemas de classificação multi-classe. A tarefa faz uso de métodos supervisionados a partir de corpus rotulados com classes autorais de interesse, como gênero, idade, nível educacional etc.

Alguns exemplos de estudos importantes na área são ferramentas que identificam o perfil do usuário mais facilmente com detecção automática de gênero (BURGER *et al.*, 2011), idade (NGUYEN *et al.*, 2014), localização geográfica (EISENSTEIN *et al.*, 2010) ou afiliação de usuários (PIERGALLINI *et al.*, 2014) com base em uma análise de suas escolhas linguísticas.

Na competição PAN-CLEF, além de trabalhos relacionados a AA, CA tem um papel fundamental para o avanço de abordagens relevantes para problemas que impactam a vida de pessoas em ambientes digitais. Mais detalhes da competição serão explorados na próxima seção.

Tarefas de CA na competição PAN-CLEF

CA vem sendo abordada pela competição PAN-CLEF desde 2013. O Quadro 3 sumariza o histórico das edições com suas respectivas características. A coluna *referência* traz a referência para os estudos que sumarizam as abordagens usadas pelos participantes. A coluna *idioma* especifica os tipos de idiomas abordados na competição do ano. Os idiomas abordados foram o inglês (EN), espanhol (ES), italiano (IT), holandês (DU), português (PT) e árabe (AR). A coluna *Conjuntos de Dados* especifica os conjuntos de dados utilizados. Por fim, a coluna *Tarefas de Predição* especifica qual tarefa de caracterização foi predita. Os tipos foram idade (I), gênero (G) e personalidade (P).

A grande maioria dos conjuntos de dados abordados são provenientes de redes sociais, como Twitter, Facebook e blogs. As tarefas de CA mais comuns são de gênero e idade. Alguns estudos como Burger *et al.* (2011), Nguyen *et al.* (2014) mostram que modelos computacionais de CA conseguem ser mais precisos do que um humano na identificação de autores de *tweets*. Este é um exemplo do avanço científico que a competição provê para a sociedade à prática forense.

Quadro 3 – Histórico de PAN-CLEF para CA

Referência	Idioma	Conjunto de Dados	Tarefas de Predição
RANGEL et al., 2013	EN, ES	Redes Sociais	I, G
RANGEL et al., 2014	EN, ES	Blogs, Twitter, Facebook e avaliações de hotéis	I, G
RANGEL et al., 2015	IT, DU	Blogs, Twitter, Facebook e avaliações de hotéis	I, P
RANGEL et al., 2016	EN, ES, DU	Twitter, blogs e redes sociais	I, G
RANGEL et al., 2017	EN, ES, PT, AR	Redes Sociais	I, G
RANGEL et al., 2018	EN, ES, AR	fotos e tweets do Twitter	G

Fonte: Caio Deutsch 2020

Exemplo de modelo de CA

O estudo de Hsieh, Dias e Paraboni (2018) apresenta resultados iniciais de algumas tarefas de caracterização autoral sobre um corpus do Facebook de português do Brasil. O trabalho foca nas tarefas mais comuns de CA como a previsão de gênero e idade, mas também aborda duas tarefas menos comuns de CA, como grau de religiosidade de um autor e formação em TI. As tarefas são modeladas utilizando diferentes fontes de conhecimento e resultados de abordagens alternativas são discutidos.

Os autores consideram cinco modelos sobre criação de perfil. O primeiro utiliza a contagem de palavras e o segundo a contagem de TF-IDF. A premissa por trás da utilização destes modelos é que o perfil do autor pode ser visto como uma tarefa de classificação de documentos. O terceiro modelo utiliza n-gramas de caracteres, populares na tarefa relacionada a previsão do autor (TSCHUGGNALL *et al.*, 2017). O quarto modelo utiliza informações lexicais psicolinguísticas, do tipo comumente utilizado em modelos de caracterização de personalidade utilizada em Mairesse *et al.* (2007). O quinto e último utiliza representações distribuídas de palavras (MIKOLOV *et al.*, 2013), comum em tarefas recentes da PLN. O corpus utilizado foi o b5-post (RAMOS *et al.*, 2018), descrito na seção 2.3.

Os resultados gerais variam de acordo com o tamanho do conjunto de dados e a definição de classe. Os resultados que mostram melhor desempenho foram observados no caso das classes binárias como gênero, com melhor F1 de 0,9 para classificação do gênero feminino e 0,86 para o gênero masculino, e formação de TI com melhor F1 de 0,8 para autores que não possuem formação e 0,64 para quem possui. A classificação por faixa etária foi moderadamente precisa, e graus de classificação de religiosidade tiveram a menor precisão de todas. Com relação aos métodos computacionais considerados, o

modelo TF-IDF geralmente supera os demais, com o Word2vec como o segundo melhor (particularmente no caso de identificação de gênero).

2.3 *Córpus para CA*

Há muitos *córpus* textuais rotulados com características autorais, nesta seção serão discutidos brevemente quatro exemplos destes *córpus*, e que serão utilizados ao longo do trabalho. O *córpus* b5-post é utilizado na seção de revisão bibliográfica 3 e no experimentos 4. Todos os quatro *córpus* são utilizados no capítulo 7 de resultados.

Córpus b5-post

O *córpus* b5-post (RAMOS *et al.*, 2018) é composto de textos e demografia dos autores. Este *córpus* foi utilizado em trabalhos de reconhecimento de personalidade (SANTOS; PARABONI; SILVA, 2017), escolhas léxicas (LAN; PARABONI, 2018) e será visto na seção de experimento deste projeto.

No total, o *córpus* possui 2,2 milhões de palavras, 193.758 frases de usuários com uma média de 189 frases por usuário baseadas em atualizações de status do Facebook. Este conjunto de dados possui até 1.000 atualizações de status de 1.019 usuários do Facebook em português do Brasil. Os dados são parcialmente rotulados com a idade, gênero, grau de religiosidade e formação de TI.

Córpus BRmoral

O córpus BRmoral (SANTOS; PARABONI, 2019) é composto de oito respostas sobre questões morais da sociedade atual, casamento, posse de armas, legalização do aborto, pena de morte, legislação de drogas, redução da idade criminal, cotas raciais e isenções de impostos da igreja. A representação é feita de forma textual e com pontuação de sentimentos, sendo 0 = 'totalmente contra' a 5 = 'Totalmente a favor'.

Os textos são rotulados com seis características do autor (idade, sexo, nível educacional, orientação política, grau de religiosidade e formação em TI) e cinco pontuações de bases morais obtidos a partir de questionários preenchidos por cada participante. Todo o texto e dados anexos foram fornecidos anonimamente pelos autores buscando apoiar a pesquisa em reconhecimento da postura moral, perfil do autor e classificação dos fundamentos morais do autor a partir do texto.

No total são 510 participantes voluntários, sendo 65.8% do gênero masculino e média de 29.4 anos. A maioria dos participantes possui formação em TI (66.7%). São 4.080 textos (510 * 8) com uma média de 54 palavras ou 566.6 caracteres por texto.

Córpus *The Blog Authorship*

O córpus *The Blog Authorship* (SCHLER *et al.*, 2006) é composto de postagens de diversos tópicos no idioma de língua inglesa. Os textos são rotulados com duas características do autor (idade e sexo). Também existe a informação do tópico de cada documento e signo de cada autor.

No total são 19.320 autores distintos e 681.288 postagens coletadas do site blogger.com em agosto de 2004. O total de palavras foi de 140 milhões ou aproximadamente 35 postagens e 7.250 palavras por pessoa. O córpus foi balanceado com um número igual de entradas para os gêneros feminino e masculino. São 8.240 autores para idade entre 13-17 anos, 8.086 entre 23-27 anos e 2.994 entre 33-47 anos.

Córpus *Twisty*

O corpus *TwisTy* (VERHOEVEN; DAELEMANS; PLANK, 2016) é composto de *tweets* da rede social Twitter. Não houve critério de escolha de tópico e os idiomas extraídos foram o espanhol, português, francês, alemão, holandês e italiano. Cada idioma possui um corpus diferente e a única tarefa de CA é o gênero.

Cada corpus possui uma quantidade diferente de *tweets* e autores. O corpus francês é o maior com 1.044.604 *tweets* (43% masculino e 57% feminino) e 786 autores, o segundo maior corpus é o de idioma espanhol com 745.184 *tweets* (45% masculino e 55% feminino) e 615 autores, na sequência vem o idioma alemão com 38.9578 (48% masculino e 52% feminino) e 228 autores, seguido do idioma holandês com 149.276 *tweets* (46% masculino e 54% feminino) e 97 autores, seguido do idioma italiano com 122.133 *tweets* (27% masculino e 73% feminino) e 93 autores, e por fim o menor idioma é o português com 86.154 *tweets* (39% masculino e 61% feminino) e 59 autores.

2.4 Modelos de representação de texto

Nesta seção serão conceituados os modelos de representação textual mais utilizados na literatura, os modelos tradicionais e *word embeddings*. Os modelos tradicionais, como os n-gramas, *bag of n-grams* e variações de n-gramas, são mais utilizados em baselines (GOLDBERG; HIRST, 2017). Já os modelos de *word embedding* são menos comuns e exigem um alto poder computacional para treinamento.

2.4.1 Modelos tradicionais de n-gramas, bag of n-gramas e variações

O modelo *bag of words* (BOW, ou saco de palavras) é representado pela vetorização de um documento de texto, ou seja, um conjunto desordenado de palavras com a sua posição ignorada, mantendo apenas a frequência de cada palavra do documento (JURAFSKY; MARTIN, 2000).

O modelo BOW pode ser simbolicamente representado por um vetor de números naturais $d \in N$ para contagem e frequência de palavras. Um documento pode ser representado por um vetor w com dimensão d e seu total de palavras deve ser igual ao total de palavras únicas dentre as palavras do vocabulário do texto original. Este vetor w pode

conter a frequência de palavras, desconsiderando a estrutura gramatical e não leva em conta a ordem das palavras.

O modelo BOW é utilizado com sucesso em diversas tarefas relacionadas ao PLN (TURNERY; PANTEL, 2010). Uma das suas desvantagens é a sua representação por matrizes esparsas, o que gera modelos com alta dimensionalidade. Por exemplo, ao considerar um vocabulário de 30.000 palavras para representar um documento de 15 palavras, este modelo terá apenas 15 dimensões com valores não zeros e muitos valores com apenas zeros (GOLDBERG; HIRST, 2017).

O modelo de n-gramas, similar ao BOW, representa vetores de sequências de n termos consecutivos. Estes termos podem ser divididos em diferentes tipos, como n-gramas de palavras, n-gramas de caracteres ou n-gramas de POS, onde n é a quantidade de termos sequenciais. Por exemplo, quando temos $n = 1$ os termos são denominados unigramas, quando $n = 2$ bigramas, $n = 3$ trigramas e assim por diante.

Trabalhar com contagem de termos como o modelo BOW auxilia na explicação do tipo de corpus que esta sendo estudado. No entanto, o poder preditivo dos modelos BOW e n-gramas é menor quando estes dados estão em frequências absolutas. O trabalho de Salton, Wong e Yang (1975) resolve este problema relativizando os termos e criando o modelo TF-IDF, onde a contagem dos termos é substituída pela frequência dos termos do documento (TF) (*term frequency*) ponderada pela frequência no corpus (IDF) (*inverse document frequency*). A normalização TF consiste em dividir cada termo pela frequência total dos termos, conforme 1.

$$TF(t, d) = \left(\frac{Contagem(t, d)}{\sum_{j=1}^N Contagem(j, d)} \right) \quad (1)$$

Já a normalização IDF penaliza o termo t e considera a quantidade de documentos onde ele ocorre, dado por $DF(t, D)$, considerando o número total de documento D , conforme 2.

$$IDF(t, D) = \log \left(\frac{D}{1 + DF(t, D)} \right) \quad (2)$$

2.4.2 Modelos de *word embeddings*

A forma de pensar em entidades como representações de símbolos discretos, e interações entre entidades como conjuntos de relações discretas entre símbolos formando um grafo, predominou como forma de pensar na comunidade de redes neurais (ROOKER, 1989).

Em contraste com este contexto tradicional, representações distribuídas formam entidades com um vetor de valor (“ou padrão de ativação”), e o significado da entidade, assim como sua relação com outras entidades, são capturados pelas ativações destes vetores e as semelhanças entre diferentes vetores. Para o contexto de PLN, isso significa que palavras (ou sentenças) não devem ser mapeadas como dimensões discretas, mas sim mapeadas para um espaço de baixa dimensão compartilhada, onde cada palavra é associada a um vetor d-dimensional, e o significado da palavra é capturado por sua relação com outras palavras e pelos valores de ativação em seu vetor (GOLDBERG; HIRST, 2017).

Word2vec - Representação distribuída de palavras

O algoritmo conhecido popularmente por Word2vec foi desenvolvido em Mikolov *et al.* (2013). Este algoritmo utiliza um modelo de linguagem neural e é composto de duas implementações de contexto diferentes, o CBOW e Skip-Gram. Ambas implementações possuem objetivos diferentes de otimização, o *Negative-Sampling* e o *Hierarchical Softmax* (GOLDBERG; HIRST, 2017).

A implementação CBOW busca calcular a probabilidade de uma palavra central, considerando palavras anteriores e posteriores. Os vetores são agrupados pela média e o neurônio central utiliza ativação linear, conforme (3).

$$P(D = 1|w, c_{1:k}) = \frac{1}{1 + e^{(w \cdot c_1 + w \cdot c_2 + \dots + w \cdot c_k)}} \quad (3)$$

A implementação *skip-gram* desacopla a dependência entre as palavras. Para um contexto de k palavras $c_{i:k}$, o *skip-gram* assume que as palavras c_i no contexto são independentes umas das outras, essencialmente tratando de como k contextos diferentes (ou seja, um par de contexto de palavras $(w, c_{i:k})$) será representado em D como k diferentes contextos: $(w, c_1), \dots, (w, c_n)$. A função $s(w, c)$ é definida como na implementação do

CBOW, mas cada contexto é um vetor de incorporação única (GOLDBERG; HIRST, 2017). As equações que definem o *skip-gram* podem ser visualizadas em (6).

$$P(D = 1|w, c_i) = \frac{1}{1 + e^{-w \cdot c_i}} \quad (4)$$

$$P(D = 1|w, c_{1:k}) = \prod_{i=1}^k P(D = 1|w, c_i) = \prod_{i=1}^k \frac{1}{1 + e^{-w \cdot c_i}} \quad (5)$$

$$\log P(D = 1|w, c_{1:k}) = \sum_{i=1}^k \log \frac{1}{1 + e^{-w \cdot c_i}} \quad (6)$$

A arquitetura do Word2vec com as implementações CBOW e Skip-Gram são representadas na figura 1. A primeira parte da ilustração, representada por CBOW, faz a predição da palavra com base no contexto da própria palavra. Já a segunda parte, representada por Skip-gram, faz a predição do contexto atual o qual a palavra se encontra.

Figura 1 – Arquitetura do modelo Word2vec

Fonte: Mikolov et al. (2013)

2.5 Métodos de aprendizado supervisionado de máquinas aplicados a AA

Nesta seção serão discutidos alguns dos principais métodos de aprendizado de máquinas utilizados nas tarefas de AA. Normalmente o papel do classificador de aprendizado de máquinas (AM) vem *a posteriori* ao tratamento dos dados com representação textual

(ROCHA *et al.*, 2017). A Figura 2 mostra o papel do classificador em um contexto de resolução clássica de uma tarefa de AA. Na figura uma mensagem suspeita de um determinado autor é selecionada para treinamento de um modelo para identificar indivíduos suspeitos em redes sociais. A mensagem passa por técnicas de n-gramas de caracteres ou n-gramas de palavras e é escolhida uma forma de representação textual para seleção de *features*, no caso o *bag of words*. Por fim, os modelos desenvolvidos são utilizados para classificar novas entradas de mensagens, identificando autores suspeitos das redes sociais.

Figura 2 – Visualização do papel dos classificadores em AA

Fonte: (ROCHA *et al.*, 2017)

Naïve Bayes

O algoritmo Naïve Bayes é um caso especial de uma Rede Bayesiana, na qual a topologia é conhecida. Esta topologia pressupõe que todas as variáveis pertencentes ao modelo devem estar conectadas unicamente à variável independente (MARON, 1961), que é o documento do autor em questionamento.

Assim como nas redes bayesianas, o algoritmo Naïve Bayes utiliza o teorema de Bayes. A principal diferença das redes bayesianas é a forte independência entre as variáveis

explicativas, ou variáveis dependentes. Pode-se definir o Naïve Bayes como sendo um subproblema de redes bayesianas no que diz respeito à topologia (RENNIE *et al.*, 2003).

O efeito de independência das variáveis explicativas torna o processo de aprendizado de máquinas muito mais rápido do ponto de vista computacional. Este comportamento torna o Naïve Bayes um ótimo classificador para *baselines* na tarefa de AA e PLN em geral, pois estas tarefas normalmente requerem grande poder computacional para serem desenvolvidas.

A equação do Naïve Bayes, supondo independência entre as variáveis explicativas, pode ser visualizadas em (7).

$$P(C|X_1, X_2, \dots, X_n) = \frac{P(C) \prod_{i=1}^n P(X_i|C)}{P(X_1, \dots, X_n)} \quad (7)$$

onde o conjunto de dados $X_1, X_2, \dots, X_n$ são as variáveis independentes e C a variável dependente.

O classificador escolhe a classe c_i ($c_i \in C$) que possui maior probabilidade segundo a avaliação em (7), dada uma determinada instância de entrada. Esse método é conhecido como regra MAP e pode ser definido como uma aplicação da seguinte equação de maximização:

$$\operatorname{argmax}_{c_j} P(C = c_j) \prod_{i=1}^n P(X_i = x_i | C = c_j) \quad (8)$$

Logística Regularizada

O método conhecido na estatística como “Regressão Logística Multinomial” é um classificador multi-classes com interpretação probabilística. A regressão logística possui vantagens para problemas de AA em comparação com demais métodos clássicos de aprendizado de máquina. Por exemplo, sua vantagem em relação ao Naïve Bayes é a não independência das variáveis explicativas. Em relação a Máquina de Vetores de Suporte (SVM, do inglês: Support Vector Machine), sua vantagem é ser um classificador multi-classes por definição. Além da aplicação como classificador, esse método é utilizado como normalizador da camada de saída em redes neurais artificiais. A logística *softmax* foi

utilizada em Custódio e Paraboni (2018) e escolhida como classificador nos experimentos apresentados no Capítulo 4.

A equação logística definida em Christensen (2006) pode ser vista em (11).

$$f_1(X) - \ln Z = \ln P(Y = 1|X) \quad (9)$$

$$\dots \quad (10)$$

$$f_n(X) - \ln Z = \ln P(Y = n|X) \quad (11)$$

onde $f_n(X)$ é uma função que transforma a matriz X em um vetor $P(Y = n|X)$, resultando na probabilidade de X ser da classe n . A normalização é introduzido por um fator $\ln Z$ que produz um efeito de balanceamento das probabilidades.

Aplicando-se a função exponencial em ambos os lados, temos que:

$$P(Y = n|X) = \frac{1}{Z} e^{f_n(X)} \quad (12)$$

Regressão logística *softmax*

A regressão logística *softmax* é uma generalização da regressão logística que pode ser usado para a classificação de várias classes sob a suposição de que as classes são mutuamente exclusivas.

Substituindo Z na equação (13), obtêm-se a função *softmax* que define a probabilidade de Y ser da classe n dado X .

$$P(Y = n|X) = \text{softmax}(X, n) = \frac{e^{f_n(X)}}{\sum_{w=1}^W e^{f_w(X)}} \quad (13)$$

2.6 Modelos baseados em distância específicos para AA

Para a tarefa de atribuição autoral algumas distâncias merecem destaque; são elas: Distância de Kešelj (KESELJ *et al.*, 2003), Distância de Stamatatos (POTHA; STAMATATOS, 2014) e SCAP (FRANTZESKOU *et al.*, 2006).

Distância de Kešelj e Stamatatos

O estudo em Keselj *et al.* (2003) define a função de dissimilaridade conhecida na literatura como o método original *Common N-Grams* (CNG), no qual, dado dois perfis $P(x)$ e $P(y)$, a medida simétrica é descrita como

$$d_0(P(x), P(y)) = \sum_{g \in P(x) \cup P(y)} \left(\frac{2(f_x(g) - f_y(g))}{f_x(g) + f_y(g)} \right)^2 \quad (14)$$

onde $f_x(g)$ e $f_y(g)$ são frequências normalizadas da quantidade de n-gramas g nos documentos x e y respectivamente. Esta função é muito eficaz quando os tamanhos de perfil são de similares. No entanto, foi no trabalho de Potha e Stamatatos (2014) que, a partir deste método, foi elaborada uma variação da equação (14):

$$d_1(P(x), P(y)) = \sum_{g \in P(x)} \left(\frac{2(f_x(g) - f_y(g))}{f_x(g) + f_y(g)} \right)^2 \quad (15)$$

Esta alternativa do método original CNG é mais estável para perfis que são menores do que outros e funcionam melhor para conjuntos de dados fechados.

SCAP - Source Code Authorship Profiles

A equação (15) não é simétrica, pois pressupõe que o primeiro documento é o questionado e possivelmente mais curto ou muito mais curto que o segundo documento de autoria conhecida. Outra medida alternativa utilizada é o SCAP, desenvolvido em Frantzeskou *et al.* (2006) e com resultados muito bons para problemas de interseção de perfil simplificada (SPI).

$$SPI(P(x), P(y)) = |P(x) \cap P(y)| \quad (16)$$

2.7 Considerações

Neste capítulo foram apresentados os principais conceitos e técnicas relacionadas ao problema computacional de AA. A ampla abordagem do problema leva a utilização

de técnicas básicas de aprendizado de máquinas, como Naïve Bayes, até técnicas mais complexas envolvendo redes neurais com *word embeddings*, ou a junção de ambos.

3 Revisão Bibliográfica

Este capítulo apresenta uma revisão bibliográfica exploratória sobre estudos de AA. Os estudos considerados foram selecionados a partir dos repositórios Scopus, ACL Anthology, IEEE e PAN-CLEF. Foram priorizados artigos mais recentes e de maior fator de impacto. Para facilitar a discussão, esta revisão foi dividida em estudos que utilizam técnicas tradicionais (seção 3.1), baseadas em perfil (seção 3.2), e outras abordagens (seção 3.3). Por fim, este capítulo é finalizado com uma seção de discussão apresentando as principais conclusões desta revisão (seção 3.4).

3.1 Abordagens tradicionais de AA

Esta seção detalha estudos que utilizam *bag of words*. Esta representação de texto é composta basicamente por modelos de n-gramas de palavras, modelos de n-gramas de caracteres e n-gramas de POS. Estes modelos normalmente servem como *baseline* nos estudos de AA. Também serviram como inspiração para o experimento da seção 4.1.

3.1.1 Uma nova abordagem para atribuição autoral

O estudo em Reddy *et al.* (2018) propõe uma nova abordagem de atribuição autoral na qual uma medida de peso do documento é usado para representar o vetor de documento ao invés de usar características ou termos. Os resultados foram satisfatórios quando comparados aos da atual literatura para conjuntos de dados de opiniões de usuários.

O conjunto de dados utilizado foi o de opiniões de usuários do site www.amazon.com, sendo 10 autores com 400 opiniões cada um. O paradigma utilizado pelo autor do artigo foi o baseado em instância e, diferentemente do paradigma baseado em perfil, não concatena os documentos dos autores *a priori*. O pré-processamento dos dados compõe a retirada de *stop words*, *stemming* do termos e retirada de termos com frequência menor do que cinco. Por fim, a medida de peso dos termos é aplicada seguindo o modelo *Discriminative Feature Selection Term Weight* (DFSTW). Este modelo procura dar pesos maiores para termos que estão na classe do autor alvo e pesos menores para termos que todos os autores utilizam.

Os experimentos foram feitos com *baselines* utilizando *Naïve Bayes*, Regressão Logística e *Random Forest* no conjunto de dados sem a utilização dos pesos. Quebras de 1.000 até 8.000 termos mais frequentes foram feitas. Na sequência, os testes foram realizados com os mesmos classificadores com a utilização dos pesos. Os resultados mostram um ganho de 15,73 pontos percentuais do *baseline* Naïve Bayes sem pesos para o com peso, sendo 82,16% de acurácia para sem peso e 97,89% de acurácia para com peso.

3.1.2 Segmentação de autores com técnicas de segmentação hierárquica

O estudo em Gómez-Adorno *et al.* (2017) apresenta uma segmentação hierárquica de diferentes características extraídas de documentos, como a digitação ou emissão de n-gramas de caracteres e n-gramas de palavras. Esta abordagem foi a vencedora da competição PAN-CLEF 2017.

O método utilizado para a segmentação foi o *Hierarchical Cluster Analysis* (HCA) com abordagem aglomerativa *bottom-up*. Esta abordagem consiste em iniciar um segmento próprio de cada autor e, a cada iteração, incluir novos pares no mesmo. Esta junção de segmentos foi feita utilizando-se o algoritmo de ligações médias, onde a distância média de cossenos entre todos os documentos nos dois segmentos são consideradas para decidir se a junção é feita ou não. Para avaliar o desempenho do algoritmo, o autor utilizou o *score* Calinski Harabaz (CALINSKI; HARABASZ, 1974). Quanto maior este *score*, maior a chance do segmento estar mais bem definido.

O desempenho de cada característica foi avaliado independentemente com uma otimização utilizando *grid search* e *F-Bcubed* como métrica de ajuste. Esta métrica consiste em uma média harmônica entre precisão e revocação. Os parâmetros avaliados no *grid search* foram $N \in \{0, 50, 100, \dots, 60.000\}$, onde N é a frequência termos que mais aparecem no conjunto de dados, dígitos de n-gramas de caracteres $\in \{3\}$, não dígitos de n-gramas de caracteres $\in \{2, 3, 4, \dots, 8\}$ e n-gramas de palavras $\in \{1, 2, 3\}$.

Os resultados aplicados ao conjunto de dados PAN-CLEF 2017 apresentam os melhores resultados de *F-Bcubed score* para a competição do ano. A métrica superou em 3 pontos percentuais o *baseline* da competição, sendo a média para as três línguas avaliadas (inglês, grego e holandês) de 57,33% para a métrica *F-Bcubed score*.

3.1.3 Explorando *Word Embedding* e n-gramas de Caracteres para segmentação de autores

O estudo de Sari e Stevenson (2016) apresenta um modelo que utiliza k-médias com otimização do coeficiente *Silhouette* (ROUSSEEUW, 1987) para segmentação de autores. O trabalho também apresenta uma comparação entre n-gramas de caracteres e *word embeddings*.

A ideia por trás do modelo do autor é unir duas técnicas já utilizadas na literatura sobre atribuição autoral: n-gramas de caracteres e *word embeddings*. Embora, segundo o autor, a última não seja tão popular na tarefa de atribuição autoral, *word embeddings* pode capturar características semânticas do autor (WANG *et al.*, 2015). A outra técnica citada é o uso de n-gramas. Esta técnica demonstrou ser capaz de capturar o efeito de escrita do autor em Sapkota *et al.* (2015).

Um fator de suma importância para a tarefa de segmentação de autores é a quantidade correta do número de segmentos de classificação dos documentos, pois este número representa a quantidade de autores treinados e posteriormente segmentados. Para encontrar a quantidade certa de segmentos, o autor do artigo propõe utilizar k-médias com a otimização do coeficiente de Silhouette.

O conjunto de dados da competição PAN-CLEF 2016 inclui 18 problemas em 3 línguas distintas (inglês, holandês e grego) e 2 gêneros (artigos de notícias e revisões). Para cada problema, a linguagem e o gênero são uniformes, mas os tópicos são diferentes. Os tamanhos dos documentos variam de algumas centenas a alguns milhares de palavras.

Os resultados comprovaram a hipótese de que modelos de *word embeddings* possuem um desempenho superior ao de n-gramas de caracteres. O resultado da medida MAP, respectivamente para os idiomas inglês e holandês, foram 0,037 e 0,046 para *word embeddings* e 0,021 e 0,059 para n-gramas. O desempenho geral do sistema foi de 0,795 para a medida F-BCubed, sendo a terceira melhor da competição no ano de 2016.

3.1.4 Atribuição Autoral em mídias sociais

O estudo em Rocha *et al.* (2017) fez uma revisão dos principais métodos de atribuição autoral para problemas aplicados a mídias sociais, especificamente para a área forense.

Também foi explorada a utilização de métodos supervisionados específicos para conjuntos de dados pequenos, como no *Twitter*.

O autor introduz o artigo explicando a dificuldade de fazer tarefas de atribuição autoral em textos pequenos. Segundo o autor, quanto maior o tamanho do texto, mais fácil de identificar o estilo de escrita.

O conjunto de dados utilizado foi o do *Twitter*, onde o máximo de caracteres por *tweet* é de 140 caracteres, considerado pouco para atribuição autoral. Para o experimento, foram considerados 10 milhões de *tweets* provenientes de 10.000 autores distintos. O período de extração dos dados contemplou seis meses em 2014.

Para os experimentos realizados pelo autor, foram utilizados quatro classificadores: SVM *power mean*, florestas aleatórias, a medida de distância SCAP e o método de compressão de textos *prediction by partial matching* (PPM-5)(TEAHAN; HARPER, 2003). A preparação dos dados incluiu remoção de palavras que não pertencem a língua inglesa, *tweets* com menos de quatro palavras, *tweets* marcados como *retweets* e *tweets* que contém a meta etiqueta RT.

O primeiro experimento foi realizado com o classificador florestas aleatórias. Também foram combinadas diferentes combinações de características, como palavras $n\text{-gram} \in \{1, \dots, 5\}$, caracteres $n\text{-gram}$ e POS $n\text{-gram} \in \{1, \dots, 5\}$. A quantidade de autores foi fixada em 50 e a quantidade de *tweets* por usuário foi variando de 50 a 1.000 mensagens. A conclusão com o experimento foi uma relação linear de crescimento da acurácia de classificação dos autores quando a quantidade de *tweets* era aumentada e o classificador que considerou todas as características apresentou a melhor acurácia.

O segundo experimento buscou comparar os classificadores. As mesmas variações de combinações do experimento um foram consideradas mas com os demais classificadores citados. A conclusão do autor foi que o classificador PPM-5 obteve os melhores resultados, chegando a uma acurácia de 56%. Os classificadores que apresentaram resultados inferiores ao PPM-5 foram, respectivamente, PMSVM, árvores aleatórias e SCAP.

Por fim, o último experimento avaliou o desempenho dos experimentos anteriores em um conjunto de dados que possui autores que não estavam na base treino/teste. Para o *baseline* foi estabelecido o número de 50 autores aleatórios conhecidos com 600 mensagens destes autores mais 600 mensagens de autores não conhecidos. Também foram feitas variações da quantidade de autores de 100 e 150. Os classificadores utilizados foram o W-SVM, RBF SVM e Regressão Logística. Os resultados mostram que ao aumentar o

número de autores, a acurácia cai de forma linear até sua conversão. A acurácia foi de 72% com 50 autores para 68% com 150 autores. O classificador que demonstrou melhores resultados foi o RBF SVM.

3.1.5 Atribuição Autoral utilizando método de identificação de impostores

O estudo em Potha e Stamatatos (2017) analisou o método de identificação de documentos *outliers* dentro de um conjunto de documentos de um determinado autor. Este método também é conhecido como o método do impostor, bastante utilizado nas competições do PAN-CLEF. O modelo apresentado no artigo superou a ultima versão do método do impostor utilizado na competição PAN-CLEF de 2015.

O autor começa introduzindo conceitos teóricos do método do impostor, inicialmente desenvolvidos por Koppel e Winter (2014). O método tenta transformar o problema de atribuição do autor em uma classificação binária, uma vez que calcula se os textos são do autor candidato ou são dos impostores, verificando se estão mais próximos dos textos do autor original. O método original do impostor utiliza textos aleatórios extraídos dos autores antes de transformar o problema de multi-classe para binário. A modificação proposta pelo autor do artigo utiliza o máximo e o mínimo da medida de similaridade dos textos investigados. Dessa forma, é possível aumentar a probabilidade de considerar documentos de impostores que tenham uma semelhança temática ou no estilo de escrita com os documentos conhecidos.

O método GI original possui uma série de parâmetros que necessitam ser otimizados. Os estudos anteriores focaram em tentar otimizar estes parâmetros separadamente (KHONJI; IRAQI, 2014; SEIDMAN, 2013). O experimento do artigo procura simplificar este processo de *fine-tuning* variando o parâmetro do número de repetições.

O experimento foi conduzido utilizando n-gramas de caracteres igual a cinco e uma medida de mínimo e máximo de dissimilaridade. Os conjuntos de dados utilizados foram os da competição PAN-CLEF 2014 e 2015. Para incluir documentos fora dos conjuntos de dados originais, foram feitas *queries* no buscador Bing e inclusos mais de 1.000 documentos de autores desconhecidos.

Os resultados demonstraram que a mudança proposta pelo autor para o método do impostor melhorou significativamente o desempenho da curva ROC nos conjuntos de dados mencionados quando comparados com os ganhadores da competição em seus

respectivos anos de atuação e o método original do impostor. O método proposto teve seu desempenho superior para todas as línguas do PAN-CLEF 2014 menos para revisões em holandês e dissertações em espanhol. Para o conjunto de dados PAN-CLEF 2015, os resultados superaram todas as línguas menos o idioma de língua inglesa.

3.1.6 Explorando características importantes da atribuição autoral

O estudo em Sari, Stevenson e Vlachos (2018) fez uma análise de quatro conjuntos de dados variando diferentes características importantes na literatura de atribuição autoral, como o estilo de escrita e tópicos. Por fim, comparou a acurácia dos resultados entre os diferentes cenários.

O artigo começa introduzindo a falta de trabalhos existentes na literatura que falam sobre *benchmarks* de conjuntos de dados, especificamente para a área de atribuição autoral. Os autores também exploram técnicas apontadas na literatura que desempenham bem para a tarefa de atribuição autoral, como a utilização de n-gramas (PENG *et al.*, 2003; KOPPEL; SCHLER; ARGAMON, 2011; SAPKOTA *et al.*, 2015; STEVENSON; VLACHOS; SARI, 2017; SOLORIO *et al.*, 2017).

Os conjuntos de dados utilizados foram: Judgment (SEROUSSI; SMYTH; ZUKERMAN, 2011), CCAT10, CCAT50 (STAMATATOS, 2008) e IMDb62 (SEROUSSI; ZUKERMAN; BOHNERT, 2010). A escolha destes conjuntos de dados foi motivada pela grande presença na literatura para a área de atribuição autoral. A quantidade de autores, documentos e médias de palavras e caracteres pode ser visualizado no Quadro 4.

Quadro 4 – Conjuntos de dados avaliados

	Judgment	CCAT10	CCAT50	IMDb62
Gênero	julgamentos legais	notícias	notícias	revisões de filmes
Quantidade de autores	3	10	50	62
Quantidade total de documentos	1.342	1.000	5.000	79.550
Média de caracteres por documento	11.957	3.089	3.058	1.401
Média de palavras por documento	2.367	580	584	288

Fonte: Adaptado de Sari, Stevenson e Vlachos (2018)

O autor utilizou a técnica *Latent Dirichlet Allocation* (LDA) para criar um modelo probabilístico sobre os textos dos autores, para cada conjunto de dados. Este modelo foi utilizado como parâmetro no modelo de *Jensen-Shannon Divergence* (JSD) para calcular uma medida de dissimilaridade entre tópicos de autores de cada conjunto de dados (PAHLAVAN; LEVESQUE, 2005). O autor do artigo mostrou quais conjuntos

de dados possuem maior dissimilaridade entre os tópicos. Também concluiu que gênero influencia significativamente na similaridade dos autores, independente do conjunto de dados avaliado.

O estudo explorou três características dentre os conjuntos de dados citados.

1. **Estilo:** O estilo de escrita que o autor possui, como utilização de pontuação exagerada, palavras específicas, uso de muitas palavras para contextualizar o texto etc;
2. **Conteúdo:** A preferência de tópico do autor. Todas as palavras de função são removidas;
3. **Híbrida:** Uma combinação de Estilo e Conteúdo.

O experimento foi feito com as técnicas Regressão Logística (LR) e Redes Neurais (RN) com uma camada oculta. Os testes foram realizados para cada um dos quatro conjuntos de dados, e foram feitos inicialmente com todas as características, com remoção de estilo, remoção de conteúdo e remoção de ambos.

Os resultados mostram que, ao remover o conteúdo, a acurácia diminuiu para conjuntos de dados de notícias por exemplo. Foi de 96% para 80% para um autor específico do conjunto de dados CCAT10, e de 84% para 64% para outro autor do mesmo conjunto de dados. Estilo e Híbrido mostram pouca mudança nos resultados. Para o conjunto de dados *Judgment*, a remoção de conteúdo não apresentou tanto impacto quanto nos conjuntos de dados CCAT10 e CCAT50. A remoção do estilo de escrita foi mais relevante, sendo que a acurácia geral para LR foi de 89,43% para 85,56%. Com a remoção de conteúdo, a acurácia foi de 89,43% para 88%.

3.2 *Uso de perfis autorais em AA*

Esta seção detalha trabalhos que utilizam o paradigma baseado em perfil. Este paradigma assume que cada documento de cada autor é independente dos demais. Também foi analisado o estudo de Ucelay *et al.* (2016), que utiliza este tipo de paradigma para a resolução de CA, e que serviu de inspiração para o experimento do Capítulo 4.

3.2.1 Modelos baseados em perfil para resolução de atribuição autoral

O estudo em Potha e Stamatatos (2014) utilizou o paradigma baseado em perfil para a resolução do problema de atribuição autoral. Além disso, também mostrou que essa resolução foi superior aos resultados do ganhador da competição PAN-CLEF 2013. Por fim, o estudo fez um comparativo com o paradigma baseado em instância.

O artigo introduziu três tipos de abordagens para atribuição autoral: tópico, sentimento e estilo, afirmando que o último é de suma importância para identificação de gênero e informações sobre o autor. Os autores definem termos relacionados a diferentes tipos de problemas de atribuição autoral, como atribuição por conjunto fechado (documentos dos candidatos que são elegíveis para a identificação do autor e que de fato contém o autor do texto), atribuição por conjunto aberto (documentos dos candidatos que são elegíveis para a identificação do autor e que não necessariamente possuem o autor do texto) e verificação de autoria (identificação de sub grupos a priori, ou seja, antes de resolver a tarefa de atribuição autoral). Também definiram os paradigmas baseados em perfil e em instância. Por fim, os autores definiram funções de dissimilaridade conhecidas na literatura como o método original *Common N-Grams* (CNG) no qual, dado dois perfis $P(x)$ e $P(y)$, a medida simétrica é descrita como

$$d_1(P(x), P(y)) = \left(\frac{2(f_x(g) - f_y(g))}{f_x(g) + f_y(g)} \right)^2 \quad (17)$$

onde $f_x(g)$ e $f_y(g)$ são as frequências normalizadas de ocorrência de g n -gramas nos documentos x e y respectivamente. Esta função é eficaz para perfis de tamanhos similares.

Outra medida introduzida por Potha e Stamatatos (2014) é o método *Simplified Profile Intersection* (SPI), que apresentaria bons resultados em problemas envolvendo atribuição autoral de código fonte, conforme a equação:

$$SPI(P(x), P(y)) = P(x) \cap P(y) \quad (18)$$

onde SPI é definida como a contagem dos n -gramas de ambos os perfis comparados. Esta medida, diferentemente do CNG, é uma medida de similaridade. Para comparação de ambas, utiliza-se o inverso das medidas, ou seja, $1 - SPI(P(x), P(y))$.

Apresentar uma alta pontuação de dissimilaridade não é suficiente para a tarefa de atribuição autoral. É necessário uma resposta binária positiva, caso o documento

questionado seja estimado pelo mesmo autor, ou negativa, caso seja estimado o contrário. A equação utilizada em Potha e Stamatatos (2014) foi a seguinte:

$$p_+(x) = 1 - \text{scale}(\text{score}(x, \text{dissimFunction}), Y) \quad (19)$$

O conjunto de dados utilizado foi da competição PAN-CLEF de 2013. Sobre o experimento, o autor utilizou a medida ROC-AUC como métrica de ajuste do modelo e variou os parâmetros $L_k \in 1,000, 2,000, \dots, 20,000$, $L_u \in 1,000, 2,000, 10,000$, $n \in 3,4,5$, e $d \in d_0, d_1, \text{SPI}$

Os melhores resultados foram com os parâmetros L_u igual a 1.000, 10.000 e 10.000, L_k igual a 1.000, 2.000 e 2.000, n igual a 5 e d igual a d_1 , respectivamente para as línguas, inglesa, grega e espanhola. As medidas ROC-AUC para estes parâmetros foram de 0,845 e superaram o primeiro lugar da competição, que obteve 0,777.

3.2.2 Utilização de técnicas baseadas em perfil para identificação de gênero e idade

O estudo em Ucelay *et al.* (2016) propôs uma variação da abordagem baseada em perfil, aplicada com sucesso para a resolução do problema de atribuição autoral em Potha e Stamatatos (2014), para a resolução de um problema de caracterização autoral. Este estudo teve como objetivo obter resultados superiores aos ganhadores da competição PAN 2016 para a identificação de gênero e idade dos autores.

O artigo introduz dois paradigmas de atribuição autoral. A abordagem baseada em instância e a abordagem baseada em perfil, focando o estudo no último. Para a construção dos perfis, o autor utiliza técnicas de estilometria. Estas técnicas envolvem frequência de pronomes, artigos, preposições, quantidade de *hyperlinks* e média de palavras. Também utiliza n-gramas, e enfatiza que a utilização de trigramas é efetivo para língua inglesa. Cada perfil apresenta características distintas da combinação de estilometria. Por exemplo, indivíduos do gênero feminino que publicam textos em *blogs* tendem a utilizar mais pronomes do que indivíduos do gênero masculino. Também relacionado à construção de perfis, os autores utilizaram técnicas baseadas em conteúdo, ou seja, palavras que possuem uma relação com os tópicos abordados. Por exemplo, mulheres tendem a falar mais sobre assuntos pessoais, como shopping e suas respectivas mães. Por outro lado, homens tendem a falar mais sobre política e tecnologia.

Os autores apresentam conceitos teóricos de medidas de similaridade. As técnicas apresentadas foram a distância relativa de Keseljic (KRD) para dois perfis distintos e a intersecção simplificada de perfis (SPI) que calcula a quantidade de características para ambos os perfis, sendo o último utilizado pelo autor no estudo.

Foram criados 10 perfis distintos: *female* 18-24, *male* 18-24, *female* 25-34, *male* 25-34, *female* 35-49, *male* 35-49, *female* 50-64, *male* 50-64, *female* 65-XX e *male* 65-XX. Segue o algoritmo abaixo com o passo a passo para a criação dos perfis.

1. Concatenação de arquivos de extensão .xml em um único arquivo de extensão do tipo .txt;
2. Remoção de tags e imagens;
3. Criação de n-gramas e ordenação dos mais frequentes para cada perfil;
4. Salva o perfil com os L n-gramas mais frequentes;
5. Verifica similaridade com SPI e prediz quais são os perfis mais parecidos entre si.

O conjunto de dados utilizado para criação dos perfis foi o da competição PAN-CLEF 2016, sendo que a quebra entre treino e teste foi de 80% e 20% respectivamente. Os testes foram feitos variando-se o parâmetro L, descrito acima, com os valores de 2.000, 4.000, 6.000, 8.000 e 10.000. Também foram variados o número de n-gramas de 2 a 5. O melhor resultado utilizou os parâmetros L igual a 4.000 e trigramas, sendo que a acurácia chegou a 77% para o idioma holandês, 74% para o inglês e 74% para o espanhol.

3.2.3 Utilizando modelagem de tópicos

O estudo em Potha e Stamatatos (2018) utilizou um método baseado em perfil com a abordagem *latent semantic indexing* (LSI) para resolução de verificação autoral. Este método é independente de língua e demonstrou resultados superiores quando comparados com *Latent Dirichlet Allocation* (LDA) no conjunto de dados PAN-CLEF 2015.

O autor introduz o artigo falando sobre a diferença entre atribuição autoral em conjuntos fechados e abertos. Atribui uma dificuldade maior em trabalhar com conjuntos abertos do que fechados e afirma como a literatura foca na resolução de problemas de conjunto aberto. Também explica como é possível transformar problemas de atribuição autoral em problemas de verificação autoral.

A principal ideia do artigo é utilizar unigramas de palavras e n-gramas de caracteres como pré processamento dos dados, e aplicar técnicas de modelagem de tópicos para reduzir a dimensionalidade e extrair representações mais compactas e menos esparsas de textos.

Os conjuntos de dados utilizados foram os do PAN-CLEF 2014 e 2015. As linguagens desses conjuntos de dados são o holandês, inglês, grego e espanhol. Os tópicos são diversos, como artigos de notícias, revisões, textos literários etc. Cada problema de verificação autoral inclui até 10 textos de um autor e um texto de um autor desconhecido. Foram incluídos 500 documentos externos provenientes de buscas no Bing dos termos mais frequentes das bases de treino dos conjuntos de dados do PAN-CLEF. O primeiro resultado da busca no Bing era incorporado no conjunto de dados do experimento. Os documentos foram concatenados utilizando o paradigma baseado em instância.

O experimento foi feito com as variações dos seguintes parâmetros: n -gramas de palavras $\in \{1, 3, 4 \text{ e } 5\}$, k documentos por autor $\in \{20, 50, 70, 100, 150, \dots, 500\}$ e a medida de similaridade $\in \{\text{cossenos, cossenos ponderados, distância euclidiana, distância euclidiana ponderada, mínimo-máximo, mínimo-máximo ponderado}\}$. Também foram considerados pelo menos 5 termos frequentes nos textos. Os classificadores utilizados foram o LDA e LSI.

Os resultados demonstraram que a curva ROC desempenhou melhor para o LSI do que LDA para todos os cenários propostos. Este desempenho é explicado pois o LSI é mais estável quando o número de tópicos varia. Os modelos LSI funcionam melhor com unigramas de palavras e semelhança de cosseno.

3.3 Outras abordagens

Esta seção detalha outros estudos relacionados com AA e CA. Trabalhos entre domínios, por exemplo, são menos comuns na literatura de AA e exigem um foco maior na extração do estilo de escrita dos autores (SUNDARARAJAN; WOODARD, 2018).

3.3.1 Representado estilo de escrita em atribuição autoral

O estudo em Sundararajan e Woodard (2018) analisou cenários linguísticos que caracterizam um estilo de escrita, especificamente para a sintaxe e palavras léxicas como substantivos, verbos, adjetivos e advérbios. Também são avaliados cenários variando os estilos citados para entre domínios, de domínio único, entre gêneros e de gênero único.

O desenvolvimento dos modelos sintáticos de língua, utilizados na classificação da atribuição de autoria, consiste na construção de modelos de gramática probabilística livres de contexto (PCFG). Estes modelos são feitos para cada autor, utilizando árvores de sentenças de suas postagens de treinamento. Enquanto algumas abordagens (RAGHAVAN; KOVASHKA;

MOONEY, 2010) usam as regras de reescrita diretamente para construir PCFG's, os autores utilizaram cadeias de Markov verticais e horizontais (KLEIN; MANNING, 2003) nas árvores de análise antes de construir PCFG's.

O cópua IMDB1M (SEROUSSI; SMYTH; ZUKERMAN, 2011) foi utilizado nos experimentos. Este conjunto de dados é de domínio único e consiste em 204.809 postagens, 66.816 revisões e 22.116 usuários. Para o experimento, foram utilizados 674 usuários, com uma média de 238 postagens por pessoa. Também foi utilizado o cópua entre domínios *The Guardian* (STAMATATOS, 2013). Este conjunto de dados é composto de opiniões sobre artigos e revisões de livros escritos por 13 autores, com uma média de 34 postagens por tópico e gênero. Os tópicos testados envolvem temas como política, dia a dia do Reino Unido, o mundo e a sociedade.

A discussão final do artigo fala sobre os resultados obtidos ao realizar os experimentos com as seguintes combinações de modelos: sintático, sintático + léxicos e de características, utilizado como *baseline* de comparação o resultado do experimento em Teahan e Harper (2003). Estes três modelos foram utilizados para os dois conjuntos de dados citados, variando entre domínios, domínio único, entre gênero e gênero único. Os resultados demonstraram que a utilização de léxicos para domínio único não melhoraram a acurácia da classificação. No cenário entre domínios, o modelo sintático demonstrou melhores resultados entre gêneros do que entre cruzados.

3.3.2 Atribuição Autoral utilizando distorção de texto

O artigo em Stamatatos (2017) mostrou que a utilização de distorção textual antes da extração do estilo de escrita do autor pode ajudar a melhorar modelos de atribuição autoral entre domínios. O método proposto tenta mascarar as informações específicas dos tópicos que não estão relacionadas ao estilo de escrita dos autores.

A ideia principal do método é fornecer uma nova versão de textos que seja mais neutra em relação aos textos originais, mantendo a maioria das informações relacionadas ao estilo de escrita do autor. Este método é inspirado na abordagem de distorção de texto introduzida por Granados *et al.* (2011), Granados, Camacho e Rodríguez (2012). A diferença do método proposto para sua inspiração é a adequação para a atribuição autoral. A aplicação do método original foi uma generalização de agrupamentos temáticos.

Um dos algoritmos do método consiste em substituir cada dígito por # e cada letra por * para cada conjunto de k palavras mais frequentes do texto. Este algoritmo é nomeado *Distorted View with Multiple Asterisks* (DV-MA). Outra variação deste algoritmo consiste em substituir cada sequência de dígitos por # e cada sequência de letras por *, e também para

cada conjunto de k sentenças mais frequentes. Este algoritmo é nomeado *Distorted View with Single Asterisks* (DV-SA). Um exemplo da aplicação destes algoritmos pode ser visualizada no quadro 5. A primeira linha representa a frase original. As demais linhas representam a frase após a aplicação destes algoritmos com a variação do parâmetro k .

Quadro 5 – Exemplo da aplicação dos algoritmos DV-MA e DV-SA

Texto Original	The cars, slightly smaller than the Ford Taurus and expected to be priced in the 15,000–17,000 range, could help GM regain a sizeable piece of the mid-size car market, a segment it once dominated.
DV-MA, k=0	***** , *****\$## , ## - \$## , ## ***** , ***** ** ***** * ***** ***** ** ** _ **** ** ***** , * ***** ** *****
DV-MA, k=100	The **** , ***** ***** than the **** ***** and ***** to be ***** in the \$## , ## - \$##,## ***** , could ***** a ***** of the *** - ***** , a ***** it *****
DV-MA, k=1.000	The **** , ***** ***** than the **** ***** and expected to be ***** in the \$## , ## - \$## , ## range , could help ** ***** a ***** ***** of the *** - size car market , a ***** it once *****
DV-SA, k=1.000	The * , * * than the * * and expected to be * in the \$# , # - \$# , # range , could help * * a * * of the * - size car market , a * it once * .
Granados et al. (2012), k=1.000	*** cars slightly smaller **** * ford taurus *** ***** ** * priced ** * ***** ***** gm regain * sizeable piece ** * mid **** * ***** * segment ** **** dominated .

Fonte: Adaptado de Stamatatos (2017)

Um dos conjuntos de dados utilizados foi o CCAT-10, com 10 autores distintos e 100 textos por autor. Os temas abordados neste corpus são indústria e corporações. O outro conjunto de dados utilizado foi o *The Guardian UK newspaper*, com 13 autores e 10 textos por autor para cada um dos 4 temas comentados: política, dia a dia do Reino Unido, o mundo e a sociedade.

Os experimentos foram conduzidos em ambos os conjuntos de dados e consistem em um *grid search* dos seguintes parâmetros: $n \in \{3, 4, 5, 6, 7, 8\}$ sendo n o número de n-gramas, $n \in \{1, 2, 3\}$ sendo n o número de *token* n-gramas, $f_t \in \{5, 10, 15, \dots, 50\}$ sendo f_t as palavras mais frequentes do conjunto de treino e $k \in \{0, 100, 200, \dots, 500, 1.000, 2.000, \dots, 5.000\}$ sendo k o parâmetro de distorção apresentado pelo método. Os testes foram divididos em três, sendo um *baseline* sem distorção textual, a aplicação do algoritmo 1 (DV-MA) e a aplicação do algoritmo 2 (DV-SA).

Os resultados se demonstraram significativos para o conjunto de dados entre domínios, o *The Guardian*. A acurácia geral do experimento *baseline* foi de 66,2%. Para o algoritmo DV-MA foi de 75,1%, e para o algoritmo DV-SA foi de 74%. Já para conjunto de dados de domínio único, o CCAT-10, os resultados demonstraram que o *baseline* possui uma acurácia maior que os algoritmos testados.

3.3.3 Aplicações entre domínios na AA

O estudo em Custódio e Paraboni (2018) apresentou o método de melhor desempenho na competição PAN-CLEF 2018. Este método combinou a predição de três classificadores independentes, n-gramas de caracteres padrões, n-gramas de caracteres com distorção textual e n-gramas de palavras. Essa abordagem foi aplicada a um conjunto de dados entre domínios.

A ideia por trás dessa abordagem está em combinar diferentes fontes de conhecimento, tais como preferências lexicais, inflexão morfológica, uso de letras maiúsculas e estruturas de texto. Por exemplo, modelos baseados em palavras ou conteúdo podem indicar preferências de uso de palavras e, conseqüentemente, ajudar a distinguir um autor de outro. Já modelos baseados em caracteres são conhecidos por sua capacidade de captar outros efeitos, como tempo, gênero, pontuação ou espaçamento (CUSTÓDIO; PARABONI, 2018).

O método proposto consiste em substituir o comprimento fixo de n-gramas e o classificador linear originais do *baseline* da competição por uma variação do comprimento de n-gramas e um classificador logístico multinomial. O chamado EACH-USP *ensemble model* combina três tipos de classificadores: o *Std.charN* que é um modelo de comprimento variável de n-gramas de caracteres, o *Dist.charN* que é um modelo de n-gramas de caracteres com comprimento variável mas com distorções não diacríticas e por último o *Std.wordN* que é um modelo de palavras de n-gramas com comprimento variável. O experimento foi conduzido com um *grid search* variando os parâmetros $n\text{-gram} \in \{\text{Start}=(1 \text{ to } 3), \text{End}=(1 \text{ to } 5)\}$, frequência mínima de documentos $\in \{0,01, 0,05, 0,1, 0,5\}$, frequência máxima de documentos $\in \{0,25, 0,5, 0,9, 1,0\}$, TF $\in \{\text{normal}, \text{sublinear}\}$, IDF $\in \{\text{normal}, \text{smoothed}\}$, normalização de documentos $\in \{L1, L2\}$ e percentual de variância explicada pela PCA $\in \{0,10, 0,25, 0,50, 0,75, 0,90, 0,99\}$. A figura 3 representa o *ensemble* de modelos descrito.

Figura 3 – Arquitetura do EACH-USP modelo *ensemble*

Fonte: Custódio e Paraboni (2018)

Os resultados superaram os do *baseline* para todas as línguas do conjunto de dados (Português, Inglês, Espanhol, Italiano e Polonês), sendo a média da medida F de 71,14% para a combinação dos três modelos e 61,2% para o *baseline*.

3.3.4 Melhorando a atribuição autoral de tópicos cruzados através do pré processamento dos dados

O estudo em Markov, Stamatatos e Sidorov (2017) mostrou que o pré processamento dos dados antes de organizar o conjunto de dados com n-gramas pode melhorar o desempenho da atribuição autoral para domínios únicos e entre domínios. O artigo citou o método de pré processamento dos dados feito em Sapkota *et al.* (2015) e propõe uma variação deste método. Esta variação consiste em:

1. **Dígitos:** Substituir todos os números por 0. Exemplo, 12.345 passam a ser 00.000. Esta substituição é motivada por números não determinarem o estilo de escrita do autor. No entanto, as informações abreviadas como 1k para 1.000 podem caracterizar a forma de escrita do autor;
2. **Pontuação:** As pontuações são separadas procurando capturar suas frequências individualmente. Também é adicionado um espaço entre as pontuações extraídas;
3. **Nomes:** Os nomes estão altamente relacionados com os tópicos dos textos, segundo os autores. Portanto, os nomes são substituídos por pelo símbolo # procurando não perder a informação de que aquele termo simboliza um nome;
4. **Termos mais frequentes:** Os termos mais frequentes são substituídos pelos símbolos % e \$. Esta substituição é motivada pela ampliação do poder de captura dos termos quando trigramas são utilizados.

O conjunto de dados utilizado foi o CCAT_10 da *Reuters Corpus Volume 1*. Este cópys consiste em 10 autores e 100 histórias por autor de sobre o único tópico, notícias corporativas. O outro conjunto de dados, este contemplando tópicos cruzados, foi o jornal *The Guardian*. Este possui 13 autores com 10 documentos por autor de áreas como política, sociedade, mundo e o Reino Unido.

O experimento foi conduzido utilizando-se o algoritmo em Sapkota *et al.* (2015) como *baseline* e o algoritmo citado pelo autor para pré processamento dos dados, seguindo da extração de trigramas com a exclusão de trigramas com frequência inferior a cinco. Os classificadores utilizados foram o *Support Vector Machine* (SVM) e *Naïve Bayes*. A métrica de comparação foi a acurácia.

O resultado do experimento demonstrou que entre domínios, a acurácia foi 10 pontos percentuais acima do *baseline* para os classificadores citados e o tamanho das características foi 80% menor, reduzindo o tempo de processamento dos testes.

3.3.5 Atribuição Autoral por consenso entre características

O estudo em Patchala e Bhatnagar (2018) utilizou uma abordagem baseada em *templates* de árvores de características pré-estabelecidas para atribuição de autoria. Esta abordagem é independente de tópicos dos autores. Também se demonstra que as sub-árvores dos *templates*, combinados com características sintáticas do autor, melhoram a atribuição autoral.

Os *templates* são criados utilizando-se as frequências de sub-árvores de várias alturas, derivadas das árvores de sentenças dos documentos. A ideia por trás destes *templates* é que o estilo gramatical é capturado neste processo. Em adição, também são incluídos três outros tipos de recursos: nomes, n-gramas de caracteres e palavras de função.

Assim que os *templates* dos autores são criados, o autor utiliza a abordagem de *Dempster-Shafer* (DS) (YAGER; LIU, 2008) para medir a similaridade dos autores. Foram utilizados três conjuntos de dados: *News Articles*, com 10 colunistas e 50 artigos de cada colunista, *Reuter_50_50* (HOUVARDAS; STAMATATOS, 2006; STAMATATOS,) com 50 autores e 5.000 documentos no total (50 para cada autor), e *Blogs* (SCHLER *et al.*, 2006; KOPPEL; SCHLER; ARGAMON, 2011) com 100 autores com 10 postagens para cada autor.

O experimento foi iniciado com *baselines* utilizando *Naïve Bayes*, SVM e o método proposto por (KOPPEL; SCHLER; ARGAMON, 2011), que serve de comparação para os testes com DS. Os resultados demonstram a robustez da abordagem proposta sob várias perspectivas, e os melhores resultados foram obtidos utilizando-se a combinação baseada em regras de *Dempster* de todas as seis características (trigramas, funções de palavras, trigramas POS, árvores com sub-árvores de 2, 3 e 4 tamanhos). Para os conjuntos de dados *News Article*, *Reuter_50_50* e *Blogs*, respectivamente, a acurácia foi de 80%, 64% e 54% no *baseline* de (KOPPEL; SCHLER; ARGAMON, 2011). Já para o modelo de DS com as seis características citadas, a acurácia passou para 88%, 73% e 68%, respectivamente, para os mesmos conjuntos de dados citados.

3.3.6 A semântica de *frames* na atribuição autoral

O estudo em Hinh, Shin e Taylor (2016) apresentou uma técnica que incorpora a utilização de *frames* semânticos como um método de atribuição autoral. A ideia por trás deste método é a

de que os *frames* podem trazer uma visão dos níveis semânticos do texto dos autores, fator de suma importância para o estilo de escrita. O resultado final é o chamado “*bag of frames*” que representa com sucesso o estilo de escrita do autor.

Os autores do artigo utilizaram a implementação *FrameNet* de Baker, Fillmore e Lowe (1998). Esta implementação consiste em *frames* que representam um sentido único de uma palavra (e seus sinônimos) e que contém *frames* de elementos. Estes *frames* de elementos são objetos que representam informações adicionais sobre o sentido da sentença. Os elementos do quadro (FE) são classificados como essenciais (de presença obrigatória) e não essenciais (geralmente informações adicionadas, como localização, horário etc). Cada um dos *frames* é mapeado para uma lista de unidades lexicais e vice-versa. As unidades lexicais são mapeadas para vários *frames*.

O conjunto de dados utilizado pelo autor foi o *The Extended Brennan-Greenstadt Corpus*, com 45 autores e 15-22 textos por autor. Este conjunto de dados foi extraído com a ferramenta *Amazon’s Mechanical Turk* (AMT).

Para o experimento, foram consideradas onze características semânticas de *frames* (HINH; SHIN; TAYLOR, 2016). As principais foram o número total de contagens de tipos FE, número médio de FE usado por quadro, número médio de FE duplicado por sentença e contagem de *frames* com relação FE usada.

Para o *baseline* foram consideradas nove características básicas de estilo de escrita: número de palavras únicas, densidade lexical, contagem de caracteres (com espaço em branco), contagem de caracteres (sem espaço), sílabas médias por palavra, comprimento médio de sentenças, contagem de sentenças, índice de legibilidade *Gunning-Fog* e índice de legibilidade de *Flesch-Kincaid* Brennan, Afroz e Greenstadt (2012).

O experimento foi conduzido com a escolha de dois autores aleatórios, calculando-se sua taxa de erro com um classificador SVM para ambos os casos do *baseline*, e considerando-se todas as características de quadro descritas anteriormente. Na sequência, mais dois autores aleatórios eram incluídos e o processo de cálculo de erro feito novamente. Este processo seguiu até que todos os autores fossem incluídos no experimento. Os resultados demonstraram que a taxa de erro do *baseline* sempre se mostrou acima do experimento com as características de *frames* descritas. A conclusão do autor é a de que a utilização de *frames* aumenta a acurácia da classificação dos autores comparada com as nove características básicas do *baseline*.

3.4 Considerações

Esse capítulo apresentou uma revisão bibliográfica acerca dos estudos de AA de destaque na literatura em PLN e áreas afins. O quadro 6 apresenta uma visão geral desta revisão, contendo as colunas com a citação de cada artigo, o conjunto de dados, o idioma, a tarefa, o tipo de conhecimento e o método de aprendizado de máquina utilizado em cada estudo.

Os conjuntos de dados são, em sua maioria, da competição PAN-CLEF, contemplando as edições de 2013 a 2018. Outros conjuntos de dados também foram utilizados como o jornal *The Guardian*, os *blogs* IMDB e conversas de Twitter.

Os idiomas considerados pelos estudos foram o inglês (EN), espanhol (ES), holandês (DU), italiano (IT), polonês (PL) e grego (GR). Os estudos são, quase em sua totalidade, dedicados ao idioma inglês, e foi encontrado apenas um estudo utilizando o idioma português (PT).

As tarefas consideradas pelos estudos foram atribuição (A), verificação (V), agrupamento (C). A grande maioria dos estudos analisados foi de atribuição e foi analisado um estudo relacionado a CA que motivou a utilização desta abordagem para este projeto.

Os tipos de conhecimento considerados foram, em sua maioria, n-gramas de caracteres (C), palavras (W) e POS (P). Também foram utilizadas técnicas menos comuns na literatura, como PCFG e CNG. Ambas tentam capturar o estilo de escrita do autor. Estas técnicas são utilizadas no pré processamento dos dados antes da utilização de técnicas de aprendizado de máquinas.

Apenas um estudo apresentou modelo de *embedding*. O treinamento foi feito com o próprio conjunto de dados, sem utilizar um modelo de representação distribuída já treinada como o Word2vec da google por exemplo.

Os métodos de aprendizado de máquinas utilizados incluíram SVM, Florestas Aleatórias (RF), Naïve Bayes, Regressão Logística (LR). Outras técnicas de aproximação baseadas em distância foram utilizadas como a Intersecção Simplificada de Perfis (SPI), *Latent Dirichlet Allocation* (LDA).

Por fim, podem-se chegar às seguintes observações ao analisar os estudos de AA. Trabalhos envolvendo *word embedding* são custosos computacionalmente, mas demonstram um grande poder de captura do estilo de escrita dos autores (SARI; STEVENSON, 2016). Os estudos envolvendo *frames* semânticos demonstraram resultados satisfatórios na captura do estilo de escrita dos autores e seu poder computacional é menor que a utilização de *word embedding*. Trabalhos envolvendo modelos de *bag of words* são muito utilizados em *baselines*. Trabalhos entre domínios apresentaram diferentes combinações de capturas distintas do estilo de escrita do autor. Este é o caso do estudo (CUSTÓDIO; PARABONI, 2018) que utiliza um *ensemble* de três diferentes

formas de captura do estilo de escrita e o estudo Stamatatos (2017) que combina distorção textual com modelos de *bag of words*.

Quadro 6 – Resumo dos estudos da área de AA considerados

Estudo	Conjunto de Dados	Idioma	Tarefa	Conhecimento	Método
Potha e Stamatatos (2014)	PAN2013	EN	A, V, C	CNG	SPI
Sari e Stevenson (2016)	PAN2016	EN, DU, GR	C	C, W, D2V	K-médias
Hinh, Shin e Taylor (2016)	<i>The Extended Brennan-Greenstadt Corpus</i>	EN	A	C, W	SVM
Rocha <i>et al.</i> (2017)	Twitter	EN	A, V	C, W, P	SVM, RF e SCAP
Gómez-Adorno <i>et al.</i> (2017)	PAN2017	EN, ES, DU e GR	C	C, W	HCA
Markov, Stamatatos e Sidorov (2017)	CCAT10 e <i>The Guardian</i>	EN	A	C, W	SVM, Naïve Bayes
Stamatatos (2017)	CCAT10 e <i>The Guardian</i>	EN	A	C, W	SVM
Potha e Stamatatos (2017)	PAN2015	EN, DU, ES, GR	A	C, W	Impostor
Reddy <i>et al.</i> (2018)	Opiniões www.amazon.com	EN	A	C	LR, Naïve Bayes
Patchala e Bhatnagar (2018)	<i>News Articles</i> , <i>Reuter .50.50</i> e <i>Blogs</i>	EN	A	P	SVM, Naïve Bayes
Sari, Stevenson e Vlachos (2018)	Judgment, CCAT10, CCAT50 e IMDB62	EN	A	C	LDA, LR, Naïve Bayes
Custódio e Paraboni (2018)	PAN2018	EN, ES, FR, IT e PL	A	C, W, P	<i>ensemble model</i>
Sundararajan e Woodard (2018)	IMDB1M e <i>The Guardian</i>	EN	A	PCFG	Cadeias de Markov
Potha e Stamatatos (2018)	PAN2014 e PAN2013	EN, DU, ES, GR	V, C	C, W	LDA, LSI

Fonte: Caio Deutsch 2020

4 Estudos exploratórios

Conforme discutido no capítulo 3 de revisão bibliográfica, diversos estudos demonstram esforços para a resolução da tarefa de AA. Entre estes estudos, vários tipos de algoritmos e diferentes tipos de conhecimento são utilizados para a resolução da tarefa de AA, inclusive de forma combinada como na estratégia de *ensemble*, como em Custódio e Paraboni (2018). Apesar disso, os resultados dos experimentos envolvendo AA demonstram resultados relativamente modestos e ainda possuem grande espaço para melhorar o desempenho de seus classificadores.

Este cenário contrasta os estudos revisados na área de CA, em que resultados expressivos (principalmente na classificação de gênero) como em Hsieh, Dias e Paraboni (2018) foi observado. Motivado por estas premissas, sugere-se assim a possibilidade de utilizar classificadores de CA em auxílio à tarefa de AA. De forma mais específica, foi realizado um experimento de AA em que o modelo de Custódio e Paraboni (2018) foi expandido com o acréscimo de informação de gênero autoral. Este experimento é descrito nas próximas seções, e seguido de uma análise sobre a combinação dos temas descritos.

4.1 Modelo proposto

Para a resolução da tarefa de AA foi utilizado o modelo EACH-USP de Custódio e Paraboni (2018), com o acréscimo de informação de gênero autoral como forma de auxiliar à tarefa de atribuição autoral. O modelo de Custódio e Paraboni (2018) é composto de um *ensemble* de três outros classificadores: n-gramas de caracteres padrões, n-gramas de caracteres com distorção textual e n-gramas de palavras. Este modelo foi discutido em detalhes na seção 3.3.3.

Por questões de simplicidade, este experimento não utiliza classificadores de gênero propriamente ditos, mas apenas a informação de gênero autoral proveniente do próprio conjunto de dados a ser classificado, e que pode assim ser visto como um classificador ótimo.

O resultado do modelo EACH-USP retorna uma probabilidade para cada modelo individual. A estas três probabilidades, foi acrescentada uma informação proveniente do próprio corpus utilizado. Esta é a informação de gênero autoral, representada por uma variável binária. A motivação para esta estratégia de acréscimo de informação de gênero autoral ao modelo de AA é proveniente da possibilidade de que ela possa reduzir o espaço de busca. Neste caso específico foi explorada a divisão dos autores candidatos entre homens e mulheres, mas esta estratégia pode

em princípio ser generalizada para outros tipos de características como idade, visão política, grau de religiosidade etc.

No final do processo, estas três probabilidades são combinadas com a variável binária de gênero e uma regressão logística multinomial é executada com as quatro variáveis. O acréscimo de informação foi introduzido no *ensemble* original de Custódio e Paraboni (2018) conforme a Figura 4. Este modelo foi chamado de CA-Label.

Figura 4 – Arquitetura do *ensemble* CA-Label

Fonte: Adaptado de Custódio e Paraboni (2018)

É importante destacar que a utilização de informação de gênero real obtido do corpus é uma simplificação da realidade, realizada unicamente com o intuito de verificar a viabilidade da estratégia geral de uso de conhecimento de gênero em auxílio à tarefa de AA. No desenvolvimento da futura pesquisa, este acréscimo de informação de gênero será fornecido diretamente por meio de modelos computacionais de CA, e não utilizando rótulos pré-existentes no corpus.

Buscando simular este comportamento predito por modelos computacionais de CA, foram geradas outras variáveis binárias de gênero com um ruído, a partir da rotulação original do corpus. Este ruído altera a distribuição original de gêneros da variável, simulando a acurácia predita por modelos computacionais de CA. As acurácias assumidas foram de 100% (rotulação original do corpus) 90%, 80%, 70%, 60% e 50%.

4.2 Conjunto dos dados

O conjunto de dados utilizado foi o corpus b5-post (RAMOS *et al.*, 2018), detalhado na seção 2.3. Para o experimento foi utilizada uma parte deste corpus, na qual 50 autores que

possuem mais frases foram filtrados. Uma frase foi definida por um conjunto de palavras até uma quebra de linha.

4.3 Procedimento

O experimento foi dividido em um *baseline* e a aplicação do modelo CA-Label para cada variação de acurácia gerada. No *baseline* foi aplicado o *ensemble* EACH-USP de Custódio e Paraboni (2018) sobre o conjunto de dados detalhado da seção anterior. Na aplicação dos modelos CA-Label, a informação de gênero com sua respectiva acurácia foi introduzido no *ensemble* original, conforme explicado na seção 4.1.

Tanto no *baseline* quanto na aplicação dos modelos CA-Label, foram selecionados 2 autores aleatoriamente para a aplicação do *ensemble*. As métricas F1, acurácia, precisão e revocação foram medidas. Foi acrescentado mais 1 autor aleatório e aplicado o *ensemble* novamente. Este processo foi feito até chegar nos 50 autores, e repetido 20 vezes procurando minimizar o efeito de aleatoriedade dos autores incrementados. Por fim, foi calculada uma média dos resultados obtidos.

4.4 Resultados

A Tabela 1 representa a média dos resultados de cada simulação variando os autores para as métricas avaliadas.

Tabela 1 – Resultados médios do modelo CA-Label com diferentes margens de acerto comparados ao *baseline* de Custódio e Paraboni (2018)

Métrica	<i>baseline</i>	100%	90%	80%	70%	60%	50%
F1	0,32	0,42	0,38	0,35	0,33	0,32	0,32
Precisão	0,33	0,42	0,39	0,36	0,34	0,33	0,32
Revocação	0,32	0,42	0,38	0,35	0,33	0,32	0,32
Acurácia	0,33	0,42	0,38	0,36	0,34	0,33	0,32

Fonte: Caio Deutsch 2020

Com base nestes resultados, é possível constatar que as métricas apresentadas na Tabela 1, são superiores ao *baseline* para os classificadores CA-Label com mais do que 70% de margem de acerto. Em complemento a estes resultados, é possível visualizar a perda de acurácia na predição com o acréscimo de autores na figura 5.

Figura 5 – Acurácia por quantidade de autores

Fonte: Caio Deutsch 2020

As curvas apresentadas também mostram que os modelos com acurácia acima de 70% superam a curva do modelo *baseline*. Ou seja, independentemente da quantidade de autores, o acréscimo de informação de gênero ajuda na classificação do autor desde que o classificador de CA obtenha uma predição com acurácia de 70% ou mais na classificação de gênero.

4.5 Discussão

O experimento descrito na seção anterior, ainda que de forma idealizada, sugere que o uso de conhecimento de CA pode auxiliar à tarefa de AA, desde que o classificador a ser utilizado possua uma margem mínima de acurácia.

Conforme visto no artigo de Hsieh, Dias e Paraboni (2018) da seção 3, classificadores de CA para gênero chegaram a uma acurácia de até 0,9. Isso significa que incorporar um classificador de CA ao *ensemble* de AA teria melhorado os resultados da tarefa neste cenário específico. Uma proposta que faz uso de classificadores de CA em auxílio à tarefa de AA será discutida nos próximos capítulos.

5 Metodologia

O estudo exploratório do Capítulo 4 sugeriu que a utilização de modelos de AA em combinação com rótulos de CA pode auxiliar na predição do autor de um determinado documento. Por um lado, modelos de AA como o *ensemble* de Custódio e Paraboni (2018) buscam capturar o estilo de escrita do autor como pontuação, elementos morfológicos, palavras e expressões de função. Já a utilização de características do autor como gênero e idade procura reduzir o espaço de busca dos autores. Os resultados obtidos sugerem o potencial de uso de métodos de CA em auxílio à AA para o conjunto de dados proposto e motivam métodos oriundos da combinação destes dois temas.

O foco do presente estudo é a substituição da informação real, utilizada no experimento anterior, por classificadores de CA (como o gênero e idade extraídos diretamente do *corpus*). Ao combinar CA e AA desta forma, espera-se obter uma solução para o problema de AA que apresente resultados superiores a modelos tradicionais, como o próprio modelo EACH-USP que não fazem uso de classificadores de CA.

Em linhas gerais, o método proposto consiste na implementação de classificadores individuais para diferentes tarefas de CA, e da integração destes classificadores ao EACH-USP. Nas seções a seguir são descritos os dados a serem utilizados na construção destes modelos e os componentes de CA e AA.

5.1 *Cópus*

O presente estudo utiliza quatro *cópus*: b5-post (RAMOS *et al.*, 2018), BRmoral (SANTOS; PARABONI, 2019), *The Blog Authorship* (SCHLER *et al.*, 2006) e *Twisty* (VERHOEVEN; DAELEMANS; PLANK, 2016). O Quadro 7 apresenta, para cada *cópus*, a quantidade de autores, as tarefas de CA consideradas (gênero (G), faixa etária (F), grau de escolaridade (E), grau de religiosidade (R), formação em TI (T) e posição política (P)), e os idiomas tratados (português (PT), inglês (EN), espanhol (ES), francês (FR), italiano (IT), alemão (GE), holandês (DU)).

As tarefas gênero e formação em TI são classes binárias. Já as tarefas faixa etária, grau de escolaridade, grau de religiosidade e posição política são classes ternárias.

Quadro 7 – Crpus considerados

C��rpus	Dom��nio	Autores	Tarefas	Idioma
b5-post	Facebook	1.019	G, F, T	PT
BRmoral	Opini��es	510	G, F, E, R, T, P	PT
<i>The Blog Authorship</i>	Blogs	19.320	G, F	EN
<i>Twisty</i>	Twitter	1.878	G	ES, FR, PT, IT, GE, DU

Fonte: Caio Deutsch 2020

5.2 Modelos Desenvolvidos

Dado que o papel de CA no presente estudo apenas de apoio tarefa principal de AA, foi adotado o uso de classificadores de CA simples. A tcnica de representao textual utilizada foi baseada em contagens de palavras ponderadas por TF-IDF com unigramas de palavras e o classificador foi a regresso logstica nos moldes de Santos e Paraboni (2019) e Hsieh, Dias e Paraboni (2018), que utilizaram os mesmos crpus deste estudo.

No que diz respeito tarefa de AA propriamente dita, o presente trabalho prope um modelo de AA estendido que, alm dos classificadores de AA propostos no modelo EACH-USP, acrescenta um classificador de CA conforme ilustrado na Figura 6.

Figura 6 – *Ensemble* CA-Prob ou CA-Label

Fonte: Caio Deutsch 2020

Cada mdulo do *ensemble* de AA produz probabilidades de um documento de entrada pertencer a um determinado autor, e a estas predies foram concatenadas as predies (na forma de probabilidades ou rtulos de classificao, como ser discutido a seguir) produzidas pelos classificadores de CA desenvolvidos. Finalmente, ambos os tipos de predio (dos classificadores de AA e CA) so concatenados e tratados de forma conjunta como os atributos de aprendizado do *ensemble* de AA.

Foram consideradas duas estratégias para a combinação do modelo EACH-USP e os modelos de CA desenvolvidos, aqui denominadas de CA-Prob e CA-Label. Ambas as estratégias estão relacionadas à forma pela qual os modelos de CA são incorporados ao *ensemble* e diferem no momento em que os modelos de AA iniciam o processo de treinamento.

A estratégia chamada de CA-Prob utiliza como atributos de aprendizado para a tarefa de AA as probabilidades produzidas pelo classificador de CA (parte inferior da Figura 6) concatenadas às probabilidades dos classificadores de AA (parte superior da Figura 6). A motivação para utilizar a estratégia de resultados probabilísticos vem da forma como a arquitetura do *ensemble* combina os resultados das características do autor com sua forma de escrita. Ou seja, uma vez que a saída dos classificadores de AA são probabilidades, acrescentar probabilidades obtidas por classificadores de CA é uma forma simples de estender o modelo de *ensemble*. De acordo com a estratégia adotada, todas as probabilidades de uma classe são incorporadas ao modelo. Por exemplo, um modelo de CA de gênero pode resultar em 0,43 de chance de um determinado indivíduo ser do gênero masculino e 0,57 de chance de ser feminino. Neste caso, ambas as probabilidades são incorporadas ao modelo. Se a tarefa de CA possuir uma saída ternária, por exemplo escolaridade, as três probabilidades são incorporadas ao modelo.

A tabela 2 mostra um exemplo de como os dados de AA e CA foram combinados para CA-Prob. Neste caso, as probabilidades de cada tarefa (em negrito) foram incluídas em conjunto com as probabilidades dos três modelos do EACH-USP.

Tabela 2 – Exemplo de como combinar características preditas de CA e modelos de AA utilizando probabilidades

Probabilidade 0	Probabilidade 1	...	Probabilidade n	Gênero M	Gênero F
0,12	0,01	...	0,04	0,43	0,57
0,47	0,02	...	0,01	0,78	0,22
0,67	0,05	...	0,24	0,88	0,11

Fonte: Caio Deutsch 2020

Como alternativa ao uso de probabilidades de CA, foi considerado também o uso direto dos rótulos de classificação produzidos. Esta estratégia, denominada CA-Label, é motivada pelos pesos da regressão logística para resultados binários ou ternários que é maior do que o peso individual de suas probabilidades, podendo ajudar mais na escolha do autor em questão. Por exemplo, o mesmo modelo de gênero descrito no exemplo anterior possui uma chance de 0,57 de o indivíduo ser feminino. Neste caso, é considerado um valor binário de 0 para indivíduos femininos e 1 para masculinos. A mesma lógica é aplicada para as características ternárias.

A tabela 3 mostra um exemplo de como os dados de AA e CA foram combinados para CA-Label. As n probabilidades da tabela são o resultado dos três modelos do EACH-USP, sendo

n o resultado da quantidade de modelos adicionados ao EACH-USP (no caso são 3 modelos) vezes a quantidade de autores. A estas probabilidades, foram acrescentadas as tarefas de CA (em negrito). No exemplo, 1 representa o gênero masculino e 0 feminino.

Tabela 3 – Exemplo de como combinar características preditas de CA e modelos de AA utilizando rótulos

Probabilidade 0	Probabilidade 1	...	Probabilidade n	Gênero Predito
0,12	0,01	...	0,04	1
0,47	0,02	...	0,01	1
0,67	0,05	...	0,24	0

Fonte: Caio Deutsch 2020

6 Avaliação

Este capítulo descreve a forma de avaliação dos modelos propostos. Estes modelos compreendem a construção e divisão de treinamento e teste para as tarefas de CA e AA em diferentes corpú.

6.1 Conjuntos de treinamento e teste

O Quadro 8 apresenta estatísticas descritivas dos conjuntos de dados utilizados nesta avaliação, com informações do idioma, quantidade de autores, quantidade de unidades textuais em cada corpú, e o tipo unidade textual considerada na tarefa de AA.

Quadro 8 – Configuração das bases de AA e CA

Córpus	Idioma	Autores	Unidade	Tipo de Unidade
b5-post	PT	1.019	128.310	Post
BRmoral	PT	510	10.236	Frase
<i>The Blog Authorship</i>	EN	19.320	918.298	Post
<i>TwisTy</i>	ES	615	745.184	Tweet
<i>TwisTy</i>	FR	786	1.044.604	Tweet
<i>TwisTy</i>	PT	59	86.154	Tweet
<i>TwisTy</i>	IT	93	122.133	Tweet
<i>TwisTy</i>	GE	228	389.578	Tweet
<i>TwisTy</i>	DU	97	149.276	Tweet

Fonte: Caio Deutsch 2020

A organização dos dados em treinamento e teste leva em conta a diferença de granularidade das tarefas de CA e AA. Enquanto a tarefa de CA utiliza como entrada a concatenação de todos os textos de um autor em um único documento, a tarefa de AA utiliza unidades textuais menores, já que a concatenação resultaria em um conjunto de instâncias de aprendizado igual ao número de autores a classificar.

Para a avaliação da tarefa de AA, selecionou-se os 20 autores com maior volume de textos em cada corpú, reservando-se o restante para construção dos classificadores de CA conforme será discutido a seguir. Esta seleção foi motivada por experimentos preliminares nos quais constatou-se que o classificador *ensemble* de AA apresentava baixo desempenho em problemas com pouco volume de dados. Conforme já discutido na seção 5.1, os textos destes 20 autores considerados na tarefa de AA foram quebrados em unidades individuais (tweets, posts etc.).

A porção restante dos dados (isso é, os textos de todos os demais autores exceto os 20 autores selecionados para a avaliação da tarefa de AA) foram utilizados para treinar classificadores de CA em cada corpú, conforme o Quadro 7. Cada classificador de CA recebe como entrada o

conjunto de textos concatenados de cada autor de treinamento, e é então utilizado para prever as características autorais de cada um dos 20 autores selecionados para a avaliação de AA discutida acima. As previsões fornecidas são representadas tanto na forma de probabilidades como de rótulos numéricos para uso das estratégias CA-Prob e CA-Label discutidas na seção 5.2.

6.2 Procedimento

A avaliação dos modelos propostos foi dividida em duas etapas. A primeira buscou avaliar os classificadores de CA e a segunda avaliar a tarefa de AA propriamente dita. Cada uma dessas avaliações será discutida individualmente nas seções a seguir.

6.2.1 Avaliação dos classificadores de CA individuais

Embora o foco principal do trabalho seja a tarefa de AA, em uma primeira etapa a avaliação contemplou também o desempenho dos classificadores de CA que compõem a solução de AA proposta. Cada classificador de CA foi treinado de forma independente dos demais. Em todos os casos, a representação textual foi um modelo de contagem de unigramas de palavras ponderado por TF-IDF.

A seleção univariada de atributos foi feita a partir do método de busca em malha (do inglês, *GridSearch*) com validação cruzada de 10 partições. A função utilizada foi a ANOVA com a métrica F1. A seleção das melhores características foi feita na base de treinamento de CA, na qual 3.000 delas foram inicialmente selecionadas e um incremento de 1.000 em 1.000 características foi realizado até um limite de 20.000. Os valores ótimos de k obtidos para cada tarefa de cada corpus estão na Tabela 4.

Tabela 4 – Valores ótimos de k para seleção univariada nas tarefas de CA consideradas

Córpus	Gênero	TI	Idade	Escolaridade	Religiosidade	Política
b5-post	1.300	6.000	1.600	-	-	-
BRmoral	5.000	3.000	9.000	10.000	9.000	7.000
<i>The Blog Authorship</i>	17.000	-	19.000	-	-	-
<i>Twisty-ES</i>	8.000	-	-	-	-	-
<i>Twisty-FR</i>	19.000	-	-	-	-	-
<i>Twisty-PT</i>	6.000	-	-	-	-	-
<i>Twisty-IT</i>	3.000	-	-	-	-	-
<i>Twisty-GE</i>	8.000	-	-	-	-	-
<i>Twisty-DU</i>	6.000	-	-	-	-	-

6.2.2 Avaliação do *ensemble* de AA

Para a avaliação de AA propriamente dita, o conjunto de 20 autores selecionados foi permutado em múltiplos cenários variando-se a constituição dos subconjuntos de autores de treinamento e teste. Mais especificamente, um conjunto de autores inicial foi aleatoriamente estabelecido e, para cada experimento, a quantidade de autores foi incrementada, chegando a 20 autores no final de cada simulação. Este processo foi feito 20 vezes permutando-se o conjunto inicial, e foi adotado para reduzir efeitos do desbalanceamento das classes de CA.

Para todos os *corpus*, com exceção do *The Blog Authorship*, as simulações iniciaram com 2 autores e finalizaram com 20 autores incrementando 2 autores a cada simulação (2, 4, 6 ..., 20). Apenas para o *corpus The Blog Authorship*, as simulações iniciaram com 5 autores e finalizaram com 20 autores (5, 10, 15 e 20). Esta diferenciação foi necessária em virtude do volume de dados deste *corpus* específico, e tempos de treinamento proibitivamente elevados.

Como sistema de *baseline*, foi utilizado o modelo EACH-USP. Por fim, cada característica foi combinada ao modelo EACH-USP de forma individual e comparado ao *baseline*.

7 Resultados

Este Capítulo apresenta os resultados dos experimentos realizados nos quatro corpos descritos no capítulo 5. Os resultados foram divididos em avaliação dos classificadores de CA (seção 7.1) e avaliação dos classificadores de AA propriamente ditos (seção 7.2).

7.1 Caracterização Autoral

As Tabelas 5, 6, 7 e 8 apresentam os resultados as tarefas de classificação para os corpos b5-post, BRmoral, *The Blog Authorship* e *Twisty*, respectivamente. Estes resultados mostram as métricas acurácia, precisão, revocação e F1 dos resultados da classificação de cada categoria para 20 autores selecionados.

Tabela 5 – Resultados de classificação de CA para o corpo b5-post

Tarefa	Acurácia	Precisão	Revocação	F1
Gênero	0,80	0,60	0,61	0,60
Formação em TI	0,90	0,72	0,72	0,72
Idade	0,45	0,57	0,44	0,42

Fonte: Caio Deutsch 2020

Tabela 6 – Resultados de classificação de CA para o corpo BRmoral

Tarefa	Acurácia	Precisão	Revocação	F1
Gênero	0,87	0,87	0,87	0,90
Formação em TI	0,78	0,81	0,77	0,85
Idade	0,37	0,52	0,42	0,55
Escolaridade	0,42	0,44	0,48	0,60
Grau Religiosidade	0,54	0,63	0,56	0,55
Visão Política	0,32	0,31	0,38	0,45

Fonte: Caio Deutsch 2020

Tabela 7 – Resultados de classificação de CA para o corpo *The Blog Authorship*

Tarefa	Acurácia	Precisão	Revocação	F1
Gênero	0,65	0,67	0,66	0,65
Idade	0,76	0,73	0,81	0,76

Fonte: Caio Deutsch 2020

Os resultados sugerem que diferentes tarefas possuem resultados variados, em alguns casos elevados (como gênero no corpo BRmoral) e em outros relativamente baixos (como idade no corpo b5-post). Estas diferenças se explicam por diversos fatores, como a dificuldade inerente

Tabela 8 – Resultados de classificação de CA para o corpus *Twisty*

Idioma	Tarefa	Acurácia	Precisão	Revocação	F1
ES	Gênero	0,85	0,85	0,86	0,85
FR	Gênero	0,80	0,80	0,75	0,76
PT	Gênero	0,70	0,67	0,63	0,64
IT	Gênero	0,75	0,58	0,56	0,57
GE	Gênero	0,80	0,81	0,80	0,80
DU	Gênero	0,55	0,46	0,48	0,43

Fonte: Caio Deutsch 2020

a certas tarefas de CA (HSIEH; DIAS; PARABONI, 2018), a definição de classes (binárias, ternárias etc.), volume de dados e outros.

7.2 Atribuição Autoral

Esta seção apresenta resultados dos cenários de AA para os quatro corpus apresentados, que incorporam os classificadores de CA discutidos nas seções anteriores. Os resultados da avaliação de cada corpus são discutidos individualmente.

Todos os resultados das duas abordagens propostas (CA-Label e CA-Prob) e do sistema de *baseline* tal serão apresentados em três formatos: primeiramente com uma visão geral da acurácia média obtida, seguida de uma visão detalhada por cenário de AA (ou seja, para cada tarefa de CA incorporada e variando-se o número de autores candidatos), e finalmente em forma gráfica ilustrando o desempenho dos modelos à medida que o conjunto de candidatos aumenta. Em todos os casos, os melhores resultados para cada tarefa considerada aparecem em destaque.

7.2.1 Domínio Facebook (corpus b5-post)

A Tabela 9 apresenta uma versão resumida dos resultados de acurácia média para os 20 cenários executados no corpus b5-post.

Tabela 9 – Acurácia média de AA no corpus b5-post

Tarefa de CA acrescentada	<i>Baseline</i>	CA-Prob	CA-Label
Gênero	0,48	0,56	0,56
Formação em TI	0,48	0,54	0,53
Idade	0,48	0,57	0,62

Fonte: Caio Deutsch 2020

Este resultado global sugere que as abordagens que fazem uso de conhecimento de CA são superiores ao modelo de *baseline* original que não tem acesso a este tipo de conhecimento.

Mais especificamente, o uso do conhecimento idade para o modelo CA-Label apresenta resultados consistentemente superiores às alternativas. Ao comparar os cenários CA-Label e CA-Prob para cada categoria nota-se que não houve ganho para as categorias binárias gênero e formação em TI. No entanto, para a categoria ternária idade o ganho se mostrou superior (em negrito).

As Tabelas 10, 11 e 12 mostram, para cada tarefa de CA individual, o desempenho dos modelos em cenários de AA com números crescentes de autores.

Tabela 10 – Acurácia por quantidade de autores de AA com informação de gênero para o córpus b5-post

Autores	<i>Baseline</i>	CA-Prob	CA-Label
2	0,80	0,82	0,84
4	0,61	0,68	0,71
6	0,52	0,60	0,62
8	0,48	0,54	0,56
10	0,44	0,53	0,54
12	0,42	0,50	0,50
14	0,39	0,48	0,47
16	0,38	0,46	0,46
18	0,37	0,45	0,44
20	0,36	0,43	0,42

Fonte: Caio Deutsch 2020

Tabela 11 – Acurácia por quantidade de autores de AA com informação de formação em TI para o córpus b5-post

Autores	<i>Baseline</i>	CA-Prob	CA-Label
2	0,80	0,82	0,83
4	0,61	0,67	0,68
6	0,52	0,59	0,60
8	0,48	0,52	0,52
10	0,44	0,52	0,51
12	0,42	0,48	0,48
14	0,39	0,46	0,46
16	0,38	0,44	0,43
18	0,37	0,43	0,42
20	0,36	0,42	0,41

Fonte: Caio Deutsch 2020

Tabela 12 – Acurácia por quantidade de autores de AA com informação de idade para o córpus b5-post

Autores	<i>Baseline</i>	CA-Prob	CA-Label
2	0,80	0,83	0,83
4	0,61	0,68	0,76
6	0,52	0,60	0,70
8	0,48	0,56	0,65
10	0,44	0,54	0,61
12	0,42	0,53	0,57
14	0,39	0,51	0,55
16	0,38	0,49	0,54
18	0,37	0,48	0,51
20	0,36	0,47	0,50

Fonte: Caio Deutsch 2020

Os resultados sugerem que usar conhecimento de CA melhora a tarefa de AA em todos os cenários considerados. Além disso, conforme já observado em Custódio e Paraboni (2018), o desempenho da tarefa cai à medida que o número de candidatos aumenta para todas as configurações consideradas. Especialmente para os resultados da Tabela 12, o modelo CA-Label superou CA-Prob em todos os cenários de aumento de autores.

Finalmente, os resultados detalhados são ilustrados para cada tarefa de CA individual nas figuras 7, 8 e 9.

Figura 7 – Curva de acurácia por quantidade de autores de AA com informação de gênero para o córpus b5-post

Fonte: Caio Deutsch 2020

Figura 8 – Curva de acurácia por quantidade de autores de AA com informação de formação em TI para o cópús b5-post

Fonte: Caio Deutsch 2020

Figura 9 – Curva de acurácia por quantidade de autores de AA com informação de idade para o cópús b5-post

Fonte: Caio Deutsch 2020

7.2.2 Domínio opinativo (cópús BRmoral)

A Tabela 13 apresenta uma versão resumida dos resultados de acurácia média para os 20 cenários executados no cópús BRmoral.

Tabela 13 – Acurácia média de AA no cópús BRmoral

Tarefa de CA acrescentada	<i>Baseline</i>	CA-Prob	CA-Label
Gênero	0,40	0,42	0,47
Formação em TI	0,40	0,41	0,47
Idade	0,40	0,42	0,47
Escolaridade	0,40	0,41	0,48
Grau de Religiosidade	0,40	0,42	0,53
Visão Política	0,40	0,41	0,49

Fonte: Caio Deutsch 2020

Este resultado global também sugere que o uso de conhecimento de CA é superior ao modelo de *baseline* original que não tem acesso a este tipo de conhecimento. Também é possível observar que algumas características obtiveram resultados superiores a outras. Este é o caso da característica grau de religiosidade que obteve o maior ganho dentre as alternativas (em negrito). O ganho de acurácia para gênero, formação em TI, escolaridade e visão política foram similares. Já o menor ganho ao utilizar os rótulos foi a característica idade. Os cenários que utilizaram probabilidades resultantes dos modelos de caracterização autoral não mostrou ganhos significativos quando comparados ao *baseline*.

As Tabelas 14, 15, 16, 17, 18 e 19 mostram, para cada tarefa de CA individual, o desempenho dos modelos em cenários de AA com números crescentes de autores.

Tabela 14 – Acurácia por quantidade de autores de AA com informação de gênero para o córpus BRmoral

Autores	<i>Baseline</i>	CA-Prob	CA-Label
2	0,78	0,75	0,76
4	0,56	0,59	0,67
6	0,48	0,48	0,55
8	0,38	0,39	0,49
10	0,38	0,39	0,45
12	0,33	0,36	0,41
14	0,30	0,33	0,37
16	0,26	0,30	0,35
18	0,26	0,29	0,32
20	0,26	0,28	0,31

Fonte: Caio Deutsch 2020

Tabela 15 – Acurácia por quantidade de autores de AA com informação de formação em TI para o córpus BRmoral

Autores	<i>Baseline</i>	CA-Prob	CA-Label
2	0,78	0,75	0,77
4	0,56	0,60	0,65
6	0,48	0,51	0,56
8	0,38	0,37	0,47
10	0,38	0,36	0,47
12	0,33	0,35	0,39
14	0,30	0,31	0,38
16	0,26	0,28	0,35
18	0,26	0,27	0,35
20	0,26	0,29	0,31

Fonte: Caio Deutsch 2020

Tabela 16 – Acurácia por quantidade de autores de AA com informação de idade para o córpus BRmoral

Autores	<i>Baseline</i>	CA-Prob	CA-Label
2	0,78	0,82	0,85
4	0,56	0,59	0,62
6	0,48	0,47	0,53
8	0,38	0,37	0,43
10	0,38	0,40	0,48
12	0,33	0,34	0,41
14	0,30	0,34	0,38
16	0,26	0,29	0,38
18	0,26	0,28	0,33
20	0,26	0,28	0,32

Fonte: Caio Deutsch 2020

Tabela 17 – Acurácia por quantidade de autores de AA com informação de escolaridade para o córpus BRmoral

Autores	<i>Baseline</i>	CA-Prob	CA-Label
2	0,78	0,85	0,83
4	0,56	0,60	0,68
6	0,48	0,45	0,58
8	0,38	0,38	0,47
10	0,38	0,39	0,43
12	0,33	0,34	0,41
14	0,30	0,31	0,38
16	0,26	0,28	0,37
18	0,26	0,26	0,32
20	0,26	0,28	0,32

Fonte: Caio Deutsch 2020

Tabela 18 – Acurácia por quantidade de autores de AA com informação de grau de religiosidade para o corpus BRmoral

Autores	<i>Baseline</i>	CA-Prob	CA-Label
2	0,78	0,78	0,89
4	0,56	0,61	0,72
6	0,48	0,51	0,64
8	0,38	0,42	0,53
10	0,38	0,37	0,48
12	0,33	0,37	0,43
14	0,30	0,32	0,45
16	0,26	0,30	0,40
18	0,26	0,29	0,40
20	0,26	0,28	0,36

Fonte: Caio Deutsch 2020

Tabela 19 – Acurácia por quantidade de autores de AA com informação de visão política para o corpus BRmoral

Autores	<i>Baseline</i>	CA-Prob	CA-Label
2	0,78	0,76	0,88
4	0,56	0,56	0,71
6	0,48	0,46	0,58
8	0,38	0,41	0,46
10	0,38	0,36	0,43
12	0,33	0,33	0,39
14	0,30	0,34	0,39
16	0,26	0,28	0,33
18	0,26	0,29	0,34
20	0,26	0,27	0,34

Fonte: Caio Deutsch 2020

Os resultados também sugerem que usar conhecimento de CA melhora a tarefa de AA em todos os cenários considerados e o desempenho do modelo é inversamente proporcional ao número de autores (classes). O modelo CA-Label superou CA-Prob em todos os cenários de incremento de autores, com exceção apenas da característica escolaridade para dois autores. Para algumas tarefas específicas como grau de religiosidade, visão política e escolaridade, a diferença de desempenho entre os modelos CA-Label e CA-Prob foi maior que as alternativas.

Os resultados detalhados nas tabelas anteriores são ilustrados para cada tarefa de CA individual nas figuras 10, 11, 12, 13, 14, 15.

Figura 10 – Curva de acurácia por quantidade de autores de AA com informação de gênero para o corpus BRmoral

Fonte: Caio Deutsch 2020

Figura 11 – Curva de acurácia por quantidade de autores de AA com informação de formação em TI para o corpus BRmoral

Fonte: Caio Deutsch 2020

Figura 12 – Curva de acurácia por quantidade de autores de AA com informação de idade para o corpus BRmoral

Fonte: Caio Deutsch 2020

Figura 13 – Curva de acurácia por quantidade de autores de AA com informação de escolaridade para o corpus BRmoral

Fonte: Caio Deutsch 2020

Figura 14 – Curva de acurácia por quantidade de autores de AA com informação de visão política para o corpus BRmoral

Fonte: Caio Deutsch 2020

Figura 15 – Curva de acurácia por quantidade de autores de AA com informação de grau de religiosidade para o corpus BRmoral

Fonte: Caio Deutsch 2020

7.2.3 Domínio de blogs (corpus *The Blog Authorship*)

A Tabela 20 apresenta uma versão resumida dos resultados de acurácia média para os 20 cenários executados no corpus *The Blog Authorship*.

Tabela 20 – Acurácia média de AA no corpus *The Blog Authorship*

Tarefa de CA acrescentada	<i>Baseline</i>	CA-Prob	CA-Label
Gênero	0,61	0,63	0,63
Idade	0,61	0,63	0,65

Fonte: Caio Deutsch 2020

Este resultado global também sugere o uso de conhecimento de CA na tarefa de AA, porém de forma menos expressiva do que nos outros domínios considerados. A característica ternária idade apresentou o maior ganho quando comparado ao *baseline*.

As Tabelas 21 e 22 mostram, para cada tarefa de CA individual, o desempenho dos modelos em cenários de AA com números crescentes de autores.

Tabela 21 – Acurácia por quantidade de autores de AA com informação de gênero para o corpus *The Blog Authorship*

Autores	Baseline	CA-Prob	CA-Label
5	0,77	0,81	0,82
10	0,64	0,63	0,63
15	0,54	0,55	0,56
20	0,49	0,52	0,52

Fonte: Caio Deutsch 2020

Tabela 22 – Acurácia por quantidade de autores de AA com informação de idade para o corpus *The Blog Authorship*

Autores	Baseline	CA-Prob	CA-Label
5	0,77	0,81	0,84
10	0,64	0,62	0,65
15	0,54	0,55	0,57
20	0,49	0,52	0,53

Fonte: Caio Deutsch 2020

Os resultados também sugerem que usar conhecimento de CA melhora a tarefa de AA em todos os cenários considerados e que o desempenho do modelo é inversamente proporcional ao número de autores (classes). O modelo CA-Label superou CA-Prob em todos os cenários de aumento de autores para a característica idade. Para a característica gênero houve pouca ou nenhuma diferença entre os dois modelos.

Os resultados detalhados nas tabelas anteriores são ilustrados para cada tarefa de CA individual nas figuras 16 e 17.

Figura 16 – Curva de acurácia por quantidade de autores de AA com informação de gênero para o corpus *The Blog Authorship*

Fonte: Caio Deutsch 2020

Figura 17 – Curva de acurácia por quantidade de autores de AA com informação de idade para o córpus *The Blog Authorship*

Fonte: Caio Deutsch 2020

7.2.4 Domínio Twitter (córpus *TwisTy*)

A Tabela 23 apresenta uma versão resumida dos resultados de acurácia média para os subconjuntos do córpus *TwisTy* nos idiomas considerados.

Tabela 23 – Acurácia média de AA no córpus *TwisTy*

Idioma	Tarefa de CA acrescentada	<i>Baseline</i>	CA-Prob	CA-Label
ES	Gênero	0,69	0,81	0,75
FR	Gênero	0,69	0,83	0,78
PT	Gênero	0,69	0,77	0,76
IT	Gênero	0,63	0,67	0,65
GE	Gênero	0,66	0,78	0,74
DU	Gênero	0,65	0,70	0,70

Fonte: Caio Deutsch 2020

Este resultado global também sugere o uso de conhecimento de CA. Nestes resultados, o modelo CA-Prob superou o CA-Label, diferentemente dos demais córpis avaliados. A maior diferença entre os dois modelos foi para o idioma espanhol. Este idioma também obteve o melhor resultado dos classificadores de CA (Tabela 8) dentre os demais idiomas avaliados. O único idioma que não obteve resultados superiores em comparação aos modelos CA-Prob e CA-Label foi o holandês. Este idioma obteve o pior resultado dentre os classificadores de CA.

As Tabelas 24, 25, 26, 27, 28 e 29 mostram, para cada a tarefa gênero, o desempenho dos modelos em cenários de AA com números crescentes de autores e variação de idioma.

Este resultado global também sugere o uso de conhecimento de CA. Para este subcórpis, os resultados de CA-Prob superaram o CA-Label para a maioria dos cenários. Apenas para o idioma Holandês, que obteve os piores resultados dos classificadores de CA, os modelos CA-Label e CA-Prob desempenharam de forma similar quando a quantidade de autores foi aumentada.

Tabela 24 – Acurácia por quantidade de autores de AA com informação de gênero para o subcórpus *Twisty* e idioma espanhol

Autores	<i>Baseline</i>	CA-Prob	CA-Label
2	0,91	0,96	0,97
4	0,82	0,91	0,91
6	0,76	0,88	0,85
8	0,71	0,85	0,81
10	0,67	0,83	0,77
12	0,65	0,81	0,75
14	0,63	0,79	0,72
16	0,61	0,77	0,70
18	0,59	0,76	0,68
20	0,57	0,75	0,67

Fonte: Caio Deutsch 2020

Tabela 25 – Acurácia por quantidade de autores de AA com informação de gênero para o subcórpus *Twisty* e idioma francês

Autores	<i>Baseline</i>	CA-Prob	CA-Label
2	0,91	0,95	0,95
4	0,81	0,89	0,86
6	0,75	0,86	0,80
8	0,71	0,83	0,77
10	0,68	0,80	0,73
12	0,65	0,78	0,71
14	0,63	0,76	0,69
16	0,61	0,74	0,66
18	0,59	0,73	0,65
20	0,58	0,72	0,64

Fonte: Caio Deutsch 2020

Tabela 26 – Acurácia por quantidade de autores de AA com informação de gênero para o subcórpus *Twisty* e idioma português

Autores	<i>Baseline</i>	CA-Prob	CA-Label
2	0,92	0,95	0,95
4	0,82	0,88	0,87
6	0,77	0,83	0,83
8	0,72	0,80	0,78
10	0,68	0,76	0,75
12	0,65	0,74	0,72
14	0,63	0,72	0,70
16	0,60	0,70	0,67
18	0,58	0,68	0,65
20	0,56	0,66	0,63

Fonte: Caio Deutsch 2020

Tabela 27 – Acurácia por quantidade de autores de AA com informação de gênero para o subcórpus *Twisty* e idioma italiano

Autores	<i>Baseline</i>	CA-Prob	CA-Label
2	0,87	0,88	0,88
4	0,76	0,79	0,78
6	0,69	0,73	0,72
8	0,65	0,70	0,68
10	0,61	0,67	0,64
12	0,59	0,64	0,62
14	0,56	0,61	0,59
16	0,53	0,59	0,56
18	0,51	0,57	0,54
20	0,50	0,56	0,53

Fonte: Caio Deutsch 2020

Tabela 28 – Acurácia por quantidade de autores de AA com informação de gênero para o subcórpus *Twisty* e idioma alemão

Autores	<i>Baseline</i>	CA-Prob	CA-Label
2	0,88	0,94	0,93
4	0,78	0,88	0,84
6	0,71	0,82	0,79
8	0,68	0,80	0,76
10	0,63	0,77	0,72
12	0,61	0,75	0,70
14	0,59	0,73	0,68
16	0,58	0,72	0,67
18	0,56	0,70	0,65
20	0,54	0,68	0,63

Fonte: Caio Deutsch 2020

Tabela 29 – Acurácia por quantidade de autores de AA com informação de gênero para o subcórpus *Twisty* e idioma holandês

Autores	<i>Baseline</i>	CA-Prob	CA-Label
2	0,89	0,90	0,91
4	0,78	0,81	0,83
6	0,71	0,75	0,75
8	0,67	0,72	0,72
10	0,64	0,68	0,68
12	0,61	0,66	0,65
14	0,59	0,64	0,64
16	0,57	0,62	0,62
18	0,55	0,61	0,61
20	0,53	0,60	0,59

Fonte: Caio Deutsch 2020

Também é possível observar que para os idiomas espanhol, francês e alemão, a diferença de desempenho entre os modelos CA-Prob e CA-Label foi mais acentuada do que para os demais idiomas avaliados.

Os resultados detalhados nas tabelas anteriores são ilustrados para cada tarefa de CA individual nas figuras 18, 19, 20, 21, 22 e 23.

Figura 18 – Curva de acurácia por quantidade de autores de AA com informação de gênero para o subcórpus *TwisTy* e idioma espanhol

Fonte: Caio Deutsch 2020

Figura 19 – Curva de acurácia por quantidade de autores de AA com informação de gênero para o subcórpus *TwisTy* e idioma francês

Fonte: Caio Deutsch 2020

Figura 20 – Curva de acurácia por quantidade de autores de AA com informação de gênero para o subcórpus *TwisTy* e idioma português

Fonte: Caio Deutsch 2020

Figura 21 – Curva de acurácia por quantidade de autores de AA com informação de gênero para o subcórpus *TwisTy* e idioma italiano

Fonte: Caio Deutsch 2020

Figura 22 – Curva de acurácia por quantidade de autores de AA com informação de gênero para o subcórpus *TwisTy* e idioma alemão

Fonte: Caio Deutsch 2020

Figura 23 – Curva de acurácia por quantidade de autores de AA com informação de gênero para o subcórpus *TwisTy* e idioma holandês

Fonte: Caio Deutsch 2020

7.2.5 Análise complementar: uso de múltiplos classificadores de CA

Dado que o uso de classificadores individuais de CA auxilia a tarefa de AA em todos os cenários discutidos, é natural questionar se o uso simultâneo de mais de um classificador teria ganhos similares ou superiores. Para ilustrar esta questão, foi conduzido um experimento complementar de AA no cópua b5-post considerando os classificadores de CA (gênero, formação em TI e idade) em todas suas combinações possíveis. A Tabela 30 mostra o *ensemble* de mais de uma tarefa para o cópua b5-post.

Tabela 30 – Resultados de AA - Média de acurácia com tarefas combinadas de CA para o cópua b5-post

Tarefas de CA acrescentadas	<i>Baseline</i>	CA-Prob	CA-Label
Gênero	0,48	0,56	0,56
Formação em TI	0,48	0,54	0,53
Idade	0,48	0,57	0,62
Gênero + Formação em TI	0,48	0,59	0,59
Idade + Formação em TI	0,48	0,63	0,68
Gênero + Idade	0,48	0,64	0,70
Gênero + Idade + Formação em TI	0,48	0,67	0,72

Fonte: Caio Deutsch 2020

Os resultados sugerem que a utilização de mais tarefas combinadas com o *ensemble* de Custódio e Paraboni (2018) apresenta melhores resultados do que o uso de classificadores de CA individuais. Este é o caso de gênero com formação em TI, que desempenhou melhor que gênero ou formação em TI individualmente. A mesma lógica pode ser aplicada para as demais combinações de tarefas e seus desempenhos individuais. O modelo que desempenhou melhor foi o que utilizou as três tarefas combinadas.

7.3 Considerações

Os resultados descritos na seção anterior sugerem que o uso da informação proveniente de modelos de CA combinado com classificadores de AA pode auxiliar a tarefa de predição de autores. Estes resultados oferecem suporte à hipótese de pesquisa central do presente estudo, discutida no Capítulo 1.

Por fim, é possível observar que a inclusão de mais de uma tarefa de CA ao EACH-USP sugere melhores resultados na predição autoral. Os resultados da combinação de tarefas de CA para o corpus b5-post mostram que, ao incluir as três tarefas disponíveis no corpus (gênero, idade e formação em TI), temos um resultado superior a inclusão destas tarefas combinadas e a utilização destas tarefas individualmente. Esta conclusão só reforça a ideia de redução do espaço de busca do problema de AA com a utilização de modelos de CA.

8 Conclusão

O trabalho apresentou uma pesquisa em nível de mestrado dedicado à resolução da tarefa de AA com técnicas de CA. A seguir são relacionadas as principais contribuições deste trabalho, suas limitações e possibilidades de trabalhos futuros.

8.1 Contribuições

Este trabalho apresentou duas contribuições principais para as áreas de PLN, Atribuição Autoral e Caracterização Autoral.

- Um novo método de AA baseado em classificadores de CA, com resultados superiores ao do modelo EACH-USP para diferentes domínios textuais, idiomas e tarefas de CA.
- Evidências de que a combinação de múltiplos classificadores de CA pode melhorar ainda mais o desempenho da tarefa de AA (conforme seção 7.2.5).

8.2 Limitações e trabalhos futuros

O presente estudo consiste de uma extensão do *ensemble* de Custódio e Paraboni (2018) e, como tal, só considera o problema de AA em conjunto fechado, ou seja, quando um dos autores candidatos é garantidamente o autor do documento a ser verificado. Uma extensão do presente trabalho para conjuntos abertos é deixada como trabalho futuro.

Dado que o modelo de Custódio e Paraboni (2018) obteve o melhor resultado na competição PAN-CLEF 2018, este foi o único *baseline* considerado para fins de avaliação. É possível entretanto que, atualmente, já haja outros modelos de AA com os quais o presente estudo possa ou deva ser comparado. Uma avaliação desta natureza é também deixada como trabalho futuro.

Finalmente, cabe destacar ainda que os resultados obtidos são específicos para os *cópus* b5-post, BRmoral, *The Blog Authorship* e *Twisty*, cada qual associado a um domínio, idioma e tarefas de CA específicas. Além disso, cabe destacar que foram considerados apenas conjuntos de até 20 autores candidatos. Para conjuntos de candidatos maiores é possível que o modelo utilizado exija modificações que não foram abordadas no presente estudo.

Os modelos propostos e seus resultados estão disponíveis no repositório do autor no *github.com*¹. Os *cópus* foram organizados e separados entre conjuntos de treinamento e teste, respeitando a divisão de AA e CA, e podem ser obtidos mediante solicitação ao autor.

¹ https://github.com/caiodeutsch/authorship_attribution

Referências²

- ABBASI, A.; CHEN, H. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, v. 20, n. 5, p. 67–75, 2005. Citado na página 21.
- ALSULAMI, B.; DAUBER, E.; HARANG, R. E.; MANCORIDIS, S.; GREENSTADT, R. Source code authorship attribution using long short-term memory based networks. In: *Computer Security - ESORICS 2017 - 22nd European Symposium on Research in Computer Security, September 11-15, Proceedings, Part I*. Oslo, Norway: [s.n.], 2017. p. 65–82. Citado na página 16.
- BAGNALL, D. Authorship clustering using multi-headed recurrent neural networks. In: *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, 5-8 September*. Évora, Portugal: [s.n.], 2016. p. 791–804. Citado na página 16.
- BAKER, C. F.; FILLMORE, C. J.; LOWE, J. B. The berkeley framenet project. In: *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, COLING 1998, August 10-14, Université de Montréal. Proceedings of the Conference*. Montréal, Quebec, Canada: [s.n.], 1998. p. 86–90. Citado na página 52.
- BRENNAN, M.; AFROZ, S.; GREENSTADT, R. Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Trans. Inf. Syst. Secur.*, v. 15, n. 3, p. 12:1–12:22, 2012. Citado na página 52.
- BURGER, J. D.; HENDERSON, J. C.; KIM, G.; ZARRELLA, G. Discriminating gender on twitter. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, A meeting of SIGDAT, a Special Interest Group of the ACL*. Edinburgh, UK: [s.n.], 2011. p. 1301–1309. Citado 2 vezes nas páginas 17 e 23.
- CALINSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. *Communications in Statistics*, p. 1–27, 1974. Citado na página 37.
- CHRISTENSEN, C. M. The ongoing process of building a theory of disruption. *Journal of Product Innovation Management*, v. 23, n. 1, p. 39–55, 2006. Citado na página 33.
- CUSTÓDIO, J. E.; PARABONI, I. EACH-USP ensemble cross-domain authorship attribution: Notebook for PAN at CLEF 2018. In: *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, September 10-14, 2018*. Avignon, France: [s.n.], 2018. Citado 15 vezes nas páginas 10, 16, 17, 18, 33, 49, 53, 55, 56, 57, 58, 60, 70, 84 e 86.
- EISENSTEIN, J.; O'CONNOR, B.; SMITH, N. A.; XING, E. P. A latent variable model for geographic lexical variation. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, 9-11 October 2010, MIT Stata Center, A meeting of SIGDAT, a Special Interest Group of the ACL*. Massachusetts, USA: [s.n.], 2010. p. 1277–1287. Citado 2 vezes nas páginas 17 e 23.
- FRANTZESKOU, G.; STAMATATOS, E.; GRITZALIS, S.; KATSIKAS, S. K. Source code author identification based on n-gram author profiles. In: *Artificial Intelligence Applications and Innovations, 3rd IFIP Conference on Artificial Intelligence Applications and Innovations (AIAI) 2006, June 7-9, 2006*. Athens, Greece: [s.n.], 2006. p. 508–515. Citado 2 vezes nas páginas 33 e 34.
- GOLDBERG, Y.; HIRST, G. *Neural Network Methods in Natural Language Processing*. [S.l.]: Morgan & Claypool, 2017. Citado 4 vezes nas páginas 27, 28, 29 e 30.

² De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

GÓMEZ-ADORNO, H.; ALEMÁN, Y.; AYALA, D. V.; SANCHEZ-PEREZ, M. A.; PINTO, D.; SIDOROV, G. Author clustering using hierarchical clustering analysis. In: *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, September 11-14, 2017*. Dublin, Ireland: [s.n.], 2017. Citado 2 vezes nas páginas 37 e 55.

GRANADOS, A.; CAMACHO, D.; RODRÍGUEZ, F. de B. Is the contextual information relevant in text clustering by compression? *Expert Syst. Appl.*, v. 39, n. 10, p. 8537–8546, 2012. Citado na página 47.

GRANADOS, A.; CEBRIÁN, M.; CAMACHO, D.; RODRÍGUEZ, F. de B. Reducing the loss of information through annealing text distortion. *IEEE Trans. Knowl. Data Eng.*, v. 23, n. 7, p. 1090–1102, 2011. Citado na página 47.

GRIEVE, J. Quantitative authorship attribution: An evaluation of techniques. *Digital Scholarship in the Humanities*, v. 22, n. 3, p. 251–270, 2007. Citado na página 20.

HINH, R.; SHIN, S.; TAYLOR, J. Using frame semantics in authorship attribution. In: *2016 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2016, October 9-12, 2016*. Budapest, Hungary: [s.n.], 2016. p. 4093–4098. Citado 3 vezes nas páginas 51, 52 e 55.

HOUVARDAS, J.; STAMATATOS, E. N-gram feature selection for authorship identification. In: *Artificial Intelligence: Methodology, Systems, and Applications, 12th International Conference, AIMSA 2006, September 12-15, 2006, Proceedings*. Varna, Bulgaria: [s.n.], 2006. p. 77–86. Citado na página 51.

HSIEH, F. C.; DIAS, R. F. S.; PARABONI, I. Author profiling from facebook corpora. In: *11th International Conference on Language Resources and Evaluation (LREC-2018)*. Miyazaki, Japan: ELRA, 2018. p. 2566–2570. Citado 5 vezes nas páginas 24, 56, 59, 61 e 68.

JOCKERS, M. L.; WITTEN, D. M. A comparative study of machine learning methods for authorship attribution. *Digital Scholarship in the Humanities*, v. 25, n. 2, p. 215–223, 2010. Citado na página 16.

JUOLA, P. Authorship attribution. *Foundations and Trends in Information Retrieval*, v. 1, n. 3, p. 233–334, 2008. Citado na página 19.

JUOLA, P. An overview of the traditional authorship attribution subtask. In: *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, September 17-20*. Rome, Italy: [s.n.], 2012. Citado na página 16.

JURAFSKY, D.; MARTIN, J. H. *Speech and language processing - an introduction to natural language processing, computational linguistics, and speech recognition*. [S.l.]: Prentice Hall, 2000. (Prentice Hall series in artificial intelligence). Citado na página 27.

KESELJ, V.; PENG, F.; CERCONI, N.; THOMAS, C. N-gram-based author profiles for authorship attribution. *Proceedings of the Conference Pacific Association for Computational Linguistics PACLING'03: 2003*, 09 2003. Citado 2 vezes nas páginas 33 e 34.

KESTEMONT, M.; TSCHUGGNALL, M.; STAMATATOS, E.; DAELEMANS, W.; SPECHT, G.; STEIN, B.; POTTHAST, M. Overview of the author identification task at PAN-2018: cross-domain authorship attribution and style change detection. In: *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, September 10-14*. Avignon, France: [s.n.], 2018. Citado 3 vezes nas páginas 16, 17 e 21.

KHONJI, M.; IRAQI, Y. A slightly-modified gi-based author-verifier with lots of features (ASGALF). In: *Working Notes for CLEF 2014 Conference, September 15-18*. Sheffield, UK: [s.n.], 2014. p. 977–983. Citado na página 40.

- KLEIN, D.; MANNING, C. D. Accurate unlexicalized parsing. In: *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, 7-12 July 2003, Sapporo Convention Center*. Sapporo, Japan: [s.n.], 2003. p. 423–430. Citado na página 47.
- KOPPEL, M.; SCHLER, J.; ARGAMON, S. Authorship attribution in the wild. *Language Resources and Evaluation*, v. 45, n. 1, p. 83–94, 2011. Citado 2 vezes nas páginas 41 e 51.
- KOPPEL, M.; WINTER, Y. Determining if two documents are written by the same author. *JASIST*, v. 65, n. 1, p. 178–187, 2014. Citado na página 40.
- LAN, A.; PARABONI, I. Definite description lexical choice: taking speaker’s personality into account. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, May 7-12*. Miyazaki, Japan: [s.n.], 2018. Citado na página 25.
- MAIRESSE, F.; WALKER, M. A.; MEHL, M. R.; MOORE, R. K. Using linguistic cues for the automatic recognition of personality in conversation and text. *J. Artif. Intell. Res.*, v. 30, p. 457–500, 2007. Citado na página 24.
- MARKOV, I.; STAMATATOS, E.; SIDOROV, G. Improving cross-topic authorship attribution: The role of pre-processing. In: *Computational Linguistics and Intelligent Text Processing - 18th International Conference, CICLing 2017, April 17-23, 2017, Revised Selected Papers, Part II*. Budapest, Hungary: [s.n.], 2017. p. 289–302. Citado 2 vezes nas páginas 50 e 55.
- MARON, M. E. Automatic indexing: An experimental inquiry. *Journal of the ACM*, v. 8, n. 3, p. 404–417, 1961. Citado na página 31.
- MIKOLOV, T.; SUTSKEVER, I.; CHEN, K.; CORRADO, G. S.; DEAN, J. Distributed representations of words and phrases and their compositionality. In: *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013*. Lake Tahoe, Nevada, United States: [s.n.], 2013. p. 3111–3119. Citado 2 vezes nas páginas 24 e 29.
- MOSTELLER, F.; WALLACE, D. L. Inference in an authorship problem. *Journal of the American Statistical Association*, v. 58, n. 302, p. 275–309, 1963. Citado na página 20.
- NEAL, T. J.; SUNDARARAJAN, K.; FATIMA, A.; YAN, Y.; XIANG, Y.; WOODARD, D. L. Surveying stylometry techniques and applications. *ACM Comput. Surv.*, v. 50, n. 6, p. 86:1–86:36, 2018. Citado 2 vezes nas páginas 19 e 20.
- NGUYEN, D.; DOGRUÖZ, A. S.; ROSÉ, C. P.; JONG, F. de. Computational sociolinguistics: A survey. *Computational Linguistics*, v. 42, n. 3, p. 537–593, 2016. Citado 2 vezes nas páginas 17 e 22.
- NGUYEN, D.; TRIESCHNIGG, R.; DOGRUOZ, A.; GRAVEL, R.; THEUNE, M.; MEDER, T.; de Jong, F. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In: *Proceedings of the 25th International Conference on Computational Linguistics, COLING 2014*. Dublin, Ireland: Association for Computational Linguistics (ACL), 2014. p. 1950–1961. Citado 2 vezes nas páginas 17 e 23.
- PAHLAVAN, K.; LEVESQUE, A. H. *Wireless Information Networks*. [S.l.]: John Wiley Sons, Inc., 2005. (Wiley series in telecommunications and signal processing). Citado na página 41.
- PATCHALA, J.; BHATNAGAR, R. Authorship attribution by consensus among multiple features. In: *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, August 20-26, 2018*. Santa Fe, New Mexico, USA: [s.n.], 2018. p. 2766–2777. Citado 2 vezes nas páginas 51 e 55.

- PENG, F.; SCHUURMANS, D.; WANG, S.; KESELJ, V. Language independent authorship attribution using character level language models. In: *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics - Volume 1*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003. (EACL '03), p. 267–274. ISBN 1-333-56789-0. Disponível em: <https://doi.org/10.3115/1067807.1067843>. Citado na página 41.
- PENG, J.; CHOO, K. R.; ASHMAN, H. Astroturfing detection in social media: Using binary n-gram analysis for authorship attribution. In: *2016 IEEE Trustcom/BigDataSE/ISPA, August 23-26*. Tianjin, China: [s.n.], 2016. p. 121–128. Citado na página 16.
- PIERGALLINI, M.; DOGRUÖZ, A. S.; GADDE, P.; ADAMSON, D.; ROSÉ, C. P. Modeling the use of graffiti style features to signal social relations within a multi-domain learning paradigm. In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2014, April 26-30*. Gothenburg, Sweden: [s.n.], 2014. p. 107–115. Citado 2 vezes nas páginas 17 e 23.
- POTHA, N.; STAMATATOS, E. A profile-based method for authorship verification. In: *Artificial Intelligence: Methods and Applications - 8th Hellenic Conference on AI, SETN 2014, May 15-17, 2014. Proceedings*. Ioannina, Greece: [s.n.], 2014. p. 313–326. Citado 5 vezes nas páginas 33, 34, 43, 44 e 55.
- POTHA, N.; STAMATATOS, E. An improved impostors method for authorship verification. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 8th International Conference of the CLEF Association, CLEF 2017, September 11-14, 2017, Proceedings*. Dublin, Ireland: [s.n.], 2017. p. 138–144. Citado 2 vezes nas páginas 40 e 55.
- POTHA, N.; STAMATATOS, E. Intrinsic author verification using topic modeling. In: *Proceedings of the 10th Hellenic Conference on Artificial Intelligence, SETN 2018, July 09-12, 2018*. Patras, Greece: [s.n.], 2018. p. 20:1–20:7. Citado 3 vezes nas páginas 16, 45 e 55.
- POTTHAST, M.; PARDO, F. M. R.; TSCHUGGNALL, M.; STAMATATOS, E.; ROSSO, P.; STEIN, B. Overview of pan'17 - author identification, author profiling, and author obfuscation. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 8th International Conference of the CLEF Association, CLEF 2017, September 11-14, Proceedings*. Dublin, Ireland: [s.n.], 2017. p. 275–290. Citado na página 16.
- RAGHAVAN, S.; KOVASHKA, A.; MOONEY, R. J. Authorship attribution using probabilistic context-free grammars. In: *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Short Papers*. Uppsala, Sweden: [s.n.], 2010. p. 38–42. Citado na página 47.
- RAMOS, R.; NETO, G.; SILVA, B. B. C.; MONTEIRO, D. S.; PARABONI, I.; DIAS, R. Building a corpus for personality-dependent natural language understanding and generation. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, May 7-12, 2018*. Miyazaki, Japan: [s.n.], 2018. Citado 5 vezes nas páginas 18, 24, 25, 57 e 60.
- REDDY, P.; REDDY, T.; CHAND, M.; VENKANNABABU, A. A new approach for authorship attribution. *Advances in Intelligent Systems and Computing*, v. 701, p. 1–9, 2018. Citado 2 vezes nas páginas 36 e 55.
- RENNIE, J. D. M.; SHIH, L.; TEEVAN, J.; KARGER, D. R. Tackling the poor assumptions of naive bayes text classifiers. In: *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003*. Washington, DC, USA: [s.n.], 2003. p. 616–623. Citado na página 32.

- ROCHA, A.; SCHEIRER, W. J.; FORSTALL, C. W.; CAVALCANTE, T.; THEOPHILO, A.; SHEN, B.; CARVALHO, A.; STAMATATOS, E. Authorship attribution for social media forensics. *IEEE Trans. Information Forensics and Security*, v. 12, n. 1, p. 5–33, 2017. Citado 4 vezes nas páginas 19, 31, 38 e 55.
- ROOKER, T. Review of neurocomputing: Foundations of research. *AI Magazine*, v. 10, n. 4, p. 64–66, 1989. Citado na página 29.
- ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, v. 20, p. 53 – 65, 1987. Citado na página 38.
- SALTON, G.; WONG, A.; YANG, C. A vector space model for automatic indexing. *Commun. ACM*, v. 18, n. 11, p. 613–620, 1975. Citado na página 28.
- SANTOS, V. G. dos; PARABONI, I.; SILVA, B. B. C. Big five personality recognition from multiple text genres. In: *Text, Speech, and Dialogue - 20th International Conference, TSD 2017, August 27-31, 2017, Proceedings*. Prague, Czech Republic: [s.n.], 2017. p. 29–37. Citado na página 25.
- SANTOS, W. R. dos; PARABONI, I. Moral Stance Recognition and Polarity Classification from Twitter and Elicited Text. In: *Recent Advances in Natural Language Processing (RANLP-2019)*. Varna, Bulgaria: [s.n.], 2019. p. 1069–1075. Citado 5 vezes nas páginas 17, 18, 26, 60 e 61.
- SAPKOTA, U.; BETHARD, S.; MONTES-Y-GÓMEZ, M.; SOLORIO, T. Not all character n-grams are created equal: A study in authorship attribution. In: *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, May 31 - June 5, 2015*. Denver, Colorado, USA: [s.n.], 2015. p. 93–102. Citado 3 vezes nas páginas 38, 41 e 50.
- SARI, Y.; STEVENSON, M. Exploring word embeddings and character n-grams for author clustering. In: *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, 5-8 September, 2016*. Évora, Portugal: [s.n.], 2016. p. 984–991. Citado 4 vezes nas páginas 20, 38, 53 e 55.
- SARI, Y.; STEVENSON, M.; VLACHOS, A. Topic or style? exploring the most useful features for authorship attribution. In: *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, August 20-26, 2018*. Santa Fe, New Mexico, USA: [s.n.], 2018. p. 343–353. Citado 3 vezes nas páginas 16, 41 e 55.
- SAVOY, J. Authorship attribution based on specific vocabulary. *ACM Trans. Inf. Syst.*, v. 30, n. 2, p. 12:1–12:30, 2012. Citado na página 19.
- SCHLER, J.; KOPPEL, M.; ARGAMON, S.; PENNEBAKER, J. W. Effects of age and gender on blogging. In: *Computational Approaches to Analyzing Weblogs, Papers from the 2006 AAAI Spring Symposium, Technical Report SS-06-03, Stanford, March 27-29, 2006*. California, USA: [s.n.], 2006. p. 199–205. Citado 4 vezes nas páginas 18, 26, 51 e 60.
- SEIDMAN, S. Authorship verification using the impostors method notebook for PAN at CLEF 2013. In: *Working Notes for CLEF 2013 Conference, September 23-26*. Valencia, Spain: [s.n.], 2013. Citado na página 40.
- SEROUSSI, Y.; SMYTH, R.; ZUKERMAN, I. Ghosts from the high court’s past: Evidence from computational linguistics for dixon ghosting for mctiernan and rich. *University of New South Wales Law Journal*, 2011. Citado 2 vezes nas páginas 41 e 47.

- SEROUSSI, Y.; ZUKERMAN, I.; BOHNERT, F. Collaborative inference of sentiments from texts. In: *User Modeling, Adaptation, and Personalization, 18th International Conference, UMAP 2010, June 20-24, 2010. Proceedings*. Big Island, HI, USA: [s.n.], 2010. p. 195–206. Citado na página 41.
- SOLORIO, T.; ROSSO, P.; MONTES-Y-GÓMEZ, M.; SHRESTHA, P.; SIERRA, S.; GONZÁLEZ, F. A. 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, EACL 2017, April 3-7, Volume 2: Short Papers*. Valencia, Spain: [s.n.], 2017. p. 669–674. Citado 2 vezes nas páginas 16 e 41.
- STAMATATOS, E. Author identification using imbalanced and limited training texts. In: *18th International Workshop on Database and Expert Systems Applications (DEXA 2007), 3-7 September 2007*. [S.l.: s.n.]. Citado na página 51.
- STAMATATOS, E. Author identification: Using text sampling to handle the class imbalance problem. *Inf. Process. Manage.*, v. 44, n. 2, p. 790–799, 2008. Citado na página 41.
- STAMATATOS, E. A survey of modern authorship attribution methods. *JASIST*, v. 60, n. 3, p. 538–556, 2009. Citado 4 vezes nas páginas 16, 19, 20 e 21.
- STAMATATOS, E. On the robustness of authorship attribution based on character n-gram features. *Journal of Law and Policy*, 2013. Citado na página 47.
- STAMATATOS, E. Authorship attribution using text distortion. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, April 3-7, 2017, Volume 1: Long Papers*. Valencia, Spain: [s.n.], 2017. p. 1138–1149. Citado 4 vezes nas páginas 47, 48, 54 e 55.
- STEVENSON, M.; VLACHOS, A.; SARI, Y. Continuous n-gram representations for authorship attribution. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, April 3-7, 2017, Volume 2: Short Papers*. Valencia, Spain: [s.n.], 2017. p. 267–273. Citado na página 41.
- SUNDARARAJAN, K.; WOODARD, D. L. What constitutes “style” in authorship attribution? *Proceedings of the 27th International Conference on Computational Linguistics*, p. 2814—2822, 2018. Citado 2 vezes nas páginas 46 e 55.
- TEAHAN, W. J.; HARPER, D. J. Using compression-based language models for text categorization. In: *Language Modeling for Information Retrieval*. [S.l.]: Springer Netherlands, 2003. Citado 2 vezes nas páginas 39 e 47.
- TSCHUGGNALL, M.; STAMATATOS, E.; VERHOEVEN, B.; DAELEMANS, W.; SPECHT, G.; STEIN, B.; POTTHAST, M. Overview of the 5th author profiling task at PAN 2017: Style breach detection and author clustering. In: *In Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin*. [S.l.: s.n.], 2017. Citado na página 24.
- TURNEY, P. D.; PANTEL, P. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, v. 37, n. 48, p. 141–188, 2010. Citado na página 28.
- UCELAY, M. J. G.; VILLEGAS, M. P.; FUNEZ, D. G.; CAGNINA, L. C.; ERRECALDE, M. L.; RAMÍREZ-DE-LA-ROSA, G.; VILLATORO-TELLO, E. Profile-based approach for age and gender identification. In: *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, 5-8 September, 2016*. Évora, Portugal: [s.n.], 2016. p. 864–873. Citado 2 vezes nas páginas 42 e 44.

VARELA, P.; JUSTINO, E. J. R.; BRITTO, A.; BORTOLOZZI, F. A computational approach for authorship attribution of literary texts using sintatic features. In: *2016 International Joint Conference on Neural Networks, IJCNN 2016, July 24-29, 2016*. Vancouver, BC, Canada: [s.n.], 2016. p. 4835–4842. Citado na página 20.

VEL, O. Y. de; ANDERSON, A.; CORNEY, M.; MOHAY, G. M. Mining email content for author identification forensics. *SIGMOD Record*, v. 30, n. 4, p. 55–64, 2001. Citado 2 vezes nas páginas 20 e 21.

VERHOEVEN, B.; DAELEMANS, W.; PLANK, B. TwiSty: a multilingual Twitter Stylometry corpus for gender and personality profiling. In: *10th International Conference on Language Resources and Evaluation (LREC-2016)*. Portoroz, Slovenia: ELRA, 2016. p. 1632–1637. Citado 3 vezes nas páginas 18, 27 e 60.

WANG, P.; XU, J.; XU, B.; LIU, C.; ZHANG, H.; WANG, F.; HAO, H. Semantic clustering and convolutional neural network for short text categorization. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Volume 2: Short Papers*. Beijing, China: [s.n.], 2015. p. 352–357. Citado na página 38.

YAGER, R. R.; LIU, L. (Ed.). *Classic Works of the Dempster-Shafer Theory of Belief Functions*. [S.l.]: Springer, 2008. v. 219. (Studies in Fuzziness and Soft Computing, v. 219). Citado na página 51.