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

GUILHERME RAMOS CASIMIRO

Atribuição de autoria em dados temporais utilizando a rede social Reddit

São Paulo
2022

GUILHERME RAMOS CASIMIRO

Atribuição de autoria em dados temporais utilizando a rede social Reddit

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 17 de outubro de 2022. 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. Luciano Antonio Digiampietri

São Paulo

2022

Ficha catalográfica elaborada pela Biblioteca da Escola de Artes, Ciências e Humanidades,
com os dados inseridos pelo(a) autor(a)
Brenda Fontes Malheiros de Castro CRB 8-7012; Sandra Tokarevicz CRB 8-4936

Ramos Casimiro, Guilherme
Atribuição de autoria em dados temporais utilizando a rede social Reddit / Guilherme Ramos Casimiro; orientador, Luciano Antonio Digiampietri. -- São Paulo, 2022.
77 p: il.

Dissertacao (Mestrado em Ciencias) - Programa de Pós-Graduação em Sistemas de Informação, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, 2022.
Versão corrigida

1. Redes sociais online. 2. Análise de autoria. 3. Mineração de texto. 4. Dados Temporais. I. Digiampietri, Luciano Antonio, orient. II. Título.

Dissertação de autoria de Guilherme Ramos Casimiro, sob o título **"Atribuição de autoria em dados temporais utilizando a rede social Reddit"**, 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 17 de outubro de 2022 pela comissão julgadora constituída pelos doutores:

Prof. Dr. Luciano Antonio Digiampietri
Universidade de São Paulo
Presidente

Prof. Dr. Flávio Eduardo Aoki Horita
Universidade Federal do ABC

Prof. Dr. Marcio Moretto Ribeiro
Universidade de São Paulo

Dedico este trabalho à minha filha Daniela.

Agradecimentos

Agradeço aos meus pais, pois sem todo o apoio, suporte, instrução e amor eu não teria conseguido concluir este trabalho.

Agradeço à minha namorada Marianna, por todo apoio, paciência e companheirismo durante todo o decorrer deste trabalho.

Agradeço aos meus familiares por todo o suporte e incentivo.

Agradeço ao professor e amigo Luciano Antonio Digiampietri pela paciência e pelos conselhos, os quais foram essenciais para que esse trabalho fosse concluído.

Agradeço também a todos amigos que de alguma forma colaboraram para a conclusão deste trabalho.

"Be the change you wish to see in the world."

(Mahatma Gandhi)

Resumo

Casimiro, Guilherme Ramos. **Atribuição de autoria em dados temporais utilizando a rede social Reddit**. 2022. 76 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2022.

A praticidade trazida pelo uso dos *smartphones* resultou, nos últimos anos, em uma maior interação através das redes sociais online. As redes sociais podem influenciar tanto positivamente quanto negativamente os usuários, sendo um dos impactos negativos a propagação de notícias falsas. Neste contexto, identificar a correta fonte de uma informação ou se a informação é verídica se tornam atividades extremamente relevantes. Desde 2009 o número de trabalhos envolvendo redes sociais online e análise de autoria tem aumentado. O presente projeto tem como objetivo utilizar os comentários da rede social Reddit, em conjunto com dados da data dos comentários, para propor uma abordagem de identificação do correto autor de um comentário ao se utilizar a rede neural LSTM para tratar a questão do aprendizado ao longo do tempo. Um estudo de caso foi realizado e publicado como artigo completo no SBSI 2020, contendo os resultados iniciais do projeto, os quais exploram diferentes técnicas de mineração de texto. Além disso, os resultados finais deste trabalho foram publicados como artigo completo no SBSI 2022, usando uma distribuição de dados próxima à realidade e obtendo, para 10 autores, uma acurácia na classificação entre 97% e 99,6% para todas as características e entre 100 autores todas as características atingiram mais de 70% de acurácia.

Palavras-chaves: Redes sociais online. Análise de autoria. Mineração de texto. Dados Temporais.

Abstract

Casimiro, Guilherme Ramos. **Authorship Attribution on temporal data using Reddit social media**. 2022. 76 p. Dissertation (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2022.

The practicality brought by the use of smartphones has resulted, in recent years, in greater interaction through online social networks. Social networks can influence users both positively and negatively, one of the negative impacts is the spread of fake news. In this context, identifying the correct source of information or whether the information is true becomes extremely relevant activities. Since 2009, the number of works involving online social networks and analysis of authorship has increased. This project aims to use the comments from the Reddit social network, together with date time data from the comments, to present a model that identifies the correct author of a comment using the neural network LSTM to address the learning over time. A case study was carried out and published as a full paper at SBSI 2020, containing the initial results of the project, which explored text mining techniques. Furthermore, the final results from this project was also published as a full paper at SBSI 2022, using a data distribution more close to reality, achieving among 10 authors more than 97% of accuracy with chars feature having more than 99.6% of accuracy, among 100 authors all features achieved more than 70% of accuracy.

Keywords: Online social media. Authorship analysis. Text mining. Temporal data.

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

AA	*Authorship Attribution*
AT	*Author Topic*
AT-FA	*Author Topic-Fictitious Author*
AT-FA-P1	Atribuição probabilística com AT-FA (classificação sem autores fictícios)
AT-FA-P2	Atribuição probabilística com AT-FA (classificação com autores fictícios)
AT-FA-SVM	SVM treinado em distribuições AT-FA (com e sem autores fictícios)
AT-P	Atribuição probabilística com *AT*
AT-SVM	SVM treinado com distribuição de AT
AUC	Área sobre a curva
AV	*Authorship Verification*
BIS-11	*The Barratt Impulsiveness Scale, version 11*
BIS/BAS	*The Behavioral Inhibition and the Behavioral Activation Scales*
BOW	*Bag-of-Words*
BZip2	Compressor de arquivos que utiliza o algoritmo *Burrows–Wheeler*
C4.5	Algoritmo utilizado para criar uma árvore de decisão
CBC	*Compression-based Cosine*
CFA	*Correspondence Factorial Analysis*
CLM	*Chen-Li metric*
CLUTO	*A Clustering Toolkit*
COPA	*CCSoft Okey Player Abuse*
ECSMiner	*Enhanced Classifier for Data Streams with novel class Miner*
DADT	*Disjoint Author-Document Topic*

DADT-P Atribuição probabilística com DADT

DADT-SVM SVM treinado em distribuições DADT

D-MRelCRP *Dynamic Multi-Relational Chinese Restaurant Process*

DXMiner *Dynamic feature based Enhanced Classifier for Data Streams with novel class Miner*

EM *Expectation-Maximization*

E-SVR *Epsilon Support Vector Regressor*

GLAD *Groningen Lightweight Authorship Detection*

GZip *GNU* Zip

HTML *HyperText Markup Language*

ICWSM2012 *International conference on weblogs and social media* de 2012

IDF *Inverse Document Frequency*

IMDb62 Conjunto de dados de revisões de filmes escritos por 62 revisores

IMDb1M Conjunto de dados de revisões de filmes escritos por 1 milhão de revisores

IRC *Internet Relay Chat*

KLD *Kullback-Leibler Divergence*

KNN *k-nearest neighbors algorithm*

LDA *Latent Dirichlet Allocation*

LDA-H LDA *Hellinger*

LDAH-M *Multi Document* LDA-H

LDAH-S *Single Document* LDA-H

LDA-SVM SVM treinado em distribuições LDA

LSTM *Long Short Term Memory*

LZW	Lempel-Ziv-Welch
MLP	*Multi Layer Perceptron*
NB	*Naive Bayes*
NCD	*Normalized Compression Distance*
NLP	*Natural Language Processing*
PAN	é uma série de eventos científicos e tarefas compartilhadas em textos digitais forenses e estilometria
PAN'11	Conjunto de dados disponibilizado pelo PAN 2011
PAN'13	Conjunto de dados disponibilizado pelo PAN 2013
PAN'14	Conjunto de dados disponibilizado pelo PAN 2013
PAN'15	Conjunto de dados disponibilizado pelo PAN 2015
PANAS	*The Positive Affect and Negative Affect Scales*
PPMd	*Prediction by Partial Matching*
POS	*Part-of-Speech*
ROC	*Receiver operating characteristic*
RLP	*Re-centered local profile*
RNN	Rede Neural Recorrente
RMS	*Root Mean Square*
SCAP	*Source Code Author Profile*
SGD	*Stochastic Gradient Descent*
SKM	*Simple k-means*
SPI	*Simplified Profile Intersection*
SMO SVM	*Sequential Minimal Optimization* SVM

SMS	*Short Message Service*
SVC	*Support Vector Classifier*
SVM	*Support Vector Machine*
SVM-RBF	SVM *Radial Basis Function kernel*
TF	*Term Frequency*
TF-IDF	*Term Frequency-Inverse Document Frequency*
TXM	Plataforma *open-source* destinada a analise de texto/corpus
URL	*Uniform Resource Locator*
Zip	Formato de arquivo que suporta compressão de dados sem perda

Sumário

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

1 Introdução

Nos últimos anos, a interação com as redes sociais online vem crescendo cada vez mais e a facilidade trazida pelo uso de *smartphones* fez com que mais pessoas consigam obter informações de maneira muito mais rápida e fácil. A utilização das diferentes mídias sociais tanto para entretenimento quanto para obtenção de informações jornalísticas sobre os mais variados assuntos faz com que seu público alvo, ao longo do tempo, cresça cada vez mais. A influência das redes sociais na vida das pessoas pode ser refletida de várias maneiras, o lado positivo possibilita uma maior interação entre usuários, acesso prático às informações desejadas e entretenimento; o lado negativo envolve a rápida propagação de notícias falsas, além de poder afetar a intercomunicação pessoal e isolamento do mundo real. Com isso, o impacto negativo pode gerar possíveis quadros de doenças psíquicas ou então a formação de bolhas ideológicas nas redes sociais fazendo com que o usuário receba ou veja somente informações que sejam do agrado dele, desestimulando o lado crítico das pessoas.

Uma das mídias sociais que tem se tornado bastante popular é o *Reddit*. O foco desta rede (ou fórum de discussão) é a formação de *subreddits*, sendo que uma *subreddit* é uma espécie de comunidade na qual os usuários interagem entre si no formato de fórum. Estas comunidades possuem como objetivo reunir usuários interessados em assuntos em comum, tanto em uma perspectiva geral, por exemplo a *subreddit* Brasil, quanto em uma mais específica, por exemplo a *subreddit* de alguma série de televisão. O *Reddit* também promove o anonimato dos usuários, isto é, o único campo que o usuário precisa disponibilizar é um *username*, não sendo necessário qualquer outro tipo de informação. Assim, a interação dos usuários nesta rede é, potencialmente, mais sincera. Dado que existem milhões de *subreddits*, analisar as postagens feitas por cada usuário, bem como cada comunidade como um todo, pode fornecer informações sobre a variação das interação do usuário com as *subreddits* em comparação com os anos analisados.

Com o aumento da utilização das redes sociais online, o fenômeno de divulgação de notícias falsas, conhecidas como *fake news*, tornou-se muito comum no dia-a-dia dos usuários. Tendo em vista esse fenômeno, a verificação da autoria das publicações nas redes sociais torna-se ainda mais necessária a fim de avaliar a autenticidade das notícias ou a correta atribuição de seu autor.

Além disso, saber lidar com o crescente volume de dados sobre a interação das pessoas com as mídias sociais tem se tornado importante. Para lidar com os textos dos usuários, diferentes abordagens de Mineração de Texto em conjunto com técnicas de Processamento de Linguagem Natural (NLP) vêm sendo aplicadas para detectar diferentes tipos de padrões em textos de usuários. Uma destas técnicas de NLP que vem sendo bastante empregada é a Análise de Sentimentos. Esta técnica é bastante utilizada em revisões de produtos ou serviços. O trabalho de O'Connor *et al.* (2010) mostrou que, a partir de 2009, com o maior uso do Twitter pelas pessoas, a utilização da análise de sentimentos em predição de séries temporais obteve melhores resultados para a predição do comportamento do ano seguinte em comparação com outras abordagens mais tradicionais. Isso corrobora com a hipótese de que um maior uso das redes sociais pelas pessoas, em consonância do seu uso sincero, ou seja, uma exposição mais verdadeira do usuário, a análise de sentimentos pode extrair diversos padrões das mensagens e, até mesmo, ser uma boa característica para a identificação de autoria ou para auxiliar a resolver outros problemas de classificação.

O conteúdo com o qual os usuários interagem, como interagem e com que frequência interagem podem fornecer indícios de comportamentos futuros do usuário, por exemplo, a próxima *subreddit* que o autor irá comentar ou as emoções contidas no próximo comentário. Com isso, analisar a exposição dos usuários no *Reddit* pode ajudar a construir uma boa forma de identificação de usuários e ajudar na predição de comportamento futuro.

Na área de atribuição de autoria foi identificado que muitas das pesquisas focam especificamente em microblogs como o Twitter. Apesar de ser uma rede social com muita informação, a limitação dos caracteres pode levar a uma perda de informações que outra rede social, sem limitação de caracteres, não possui. Foi identificado que, no estado da arte, poucos trabalhos envolvendo a rede social *Reddit* (bem como outras redes sem limites de caracteres nas postagens). Além disso, também foi identificado que poucos trabalhos lidam com a questão temporal, considerando que o modo como o usuário interage na rede social pode variar muito com o passar dos anos (CASIMIRO; DIGIAMPIETRI, 2020).

Este trabalho utilizou os dados dos comentários do *Reddit* dispostos pelo projeto *Open Source* Pushshift[1]. Além disso, o presente trabalho utilizou técnicas de mineração de texto, NLP e aprendizado de máquina para identificar automaticamente o autor de um comentário do *Reddit*, levando em conta aspectos temporais dos comentários. Além disso a

[1] https://pushshift.io/

questão temporal do trabalho foi tratada com algoritmos que consideram a informação temporal no treinamento.

1.1 Hipótese

A autoria do usuário está também ligada ao momento em que ele está (que é tanto o ponto de vista temporal da análise sendo feita em um contexto global, quanto pessoal e mais íntimo do usuário em si), com isso é importante detectar e diferenciar o perfil do usuário em datas diferentes. Por meio de um reconhecimento melhor do momento em que o autor escreveu uma mensagem, será possível identificar a autoria com maior eficácia, especialmente considerando grandes conjuntos de dados e sem sofrer grandes variações de desempenho nos anos analisados.

Esta hipótese foi testada utilizando as medidas precisão, revocação e medida f (*f-measure*) e comparando a solução proposta com variações do modelo segundo o estado da arte.

1.2 Objetivos

Com a finalidade de explorar a questão temporal na rede social, o presente projeto pretende aprimorar a atribuição de autoria utilizando informações do conteúdo das postagens dos autores, incluindo informações temporais (a data que a postagem foi realizada), sobre os comentários escritos por ele ao utilizar a RNN LSTM para tratar a questão de aprendizado ao longo do tempo.

Assim, este projeto teve como objetivo desenvolver e implementar um algoritmo para a atribuição de autoria que considere a informação temporal, explorando informações estatísticas do estilo de escrita oriundos do comentário do autor.

Para alcançar esse objetivo geral os seguintes objetivos específicos foram definidos:

- Organização e disponibilização de um conjunto de dados rotulado com informações detalhadas sobre as postagens [2];
- Avaliação do impacto na atribuição de autoria considerando diversas características;

[2] https://drive.google.com/drive/folders/1hCWyUrJUeCbg7yU3B7sCS22NmlkzOhbY?usp=sharing

- Desenvolvimento e validação do classificador em comparação com variações do modelo segundo o estado da arte.

1.3 Justificativa

Se for demonstrado que a autoria do usuário está intrinsecamente ligada ao momento em que ele está, então futuras abordagens de atribuição de autoria poderão tirar proveito dos dados temporais a fim de aumentar a acurácia na classificação dos autores.

1.4 Limitações

Uma das motivações do presente trabalho ao se utilizar o Reddit como base de dados, foi a questão da anonimidade dos usuários no Reddit. Por mais que isso tenha sido utilizado no presente trabalho como um dos motivos para uma exposição mais sincera do usuário, não foram avaliados os impactos que isso teria em uma investigação forense, a qual não será tratada no decorrer do trabalho.

1.5 Organização deste documento

O restante deste documento está organizado da seguinte forma. O capítulo 2 descreve os conceitos fundamentais para o entendimento do trabalho. O capítulo 3 apresenta a revisão de literatura realizada. O capítulo 4 descreve o conjunto de dados e as técnicas que foram utilizados. O capítulo 5 detalha e discute os resultados obtidos. Por fim, o capítulo 6 contém as considerações finais acerca do projeto.

2 Conceitos Básicos

Neste capítulo são apresentados alguns tópicos fundamentais para a compreensão do projeto. Todos os conceitos aqui abordados são pertinentes às áreas de Atribuição de autoria e Mineração de Texto. Com isso, primeiramente é realizada uma breve introdução ao tema de Atribuição de autoria em conjunto com a estilometria. Em seguida, é apresentado o tema de Mineração de Texto, apresentando os conceitos utilizados no trabalho, como: pré-processamento, técnicas de modelagem e classificação de texto.

2.1 Atribuição de autoria

A atribuição de autoria é a tarefa de identificar a autoria de um item. Este "item" pode ser tanto um texto em uma rede social, uma carta, capítulos de um livro, uma pintura, código fonte de algum *software* etc.

As motivações por trás da Atribuição de autoria podem ser diversas, porém elas geralmente englobam: identificação de plágio, garantindo o reconhecimento de *copyrights*; identificação de perfis previamente banidos em redes sociais que posteriormente criaram outros perfis; resolução de crimes através da linguística forense, identificando, por exemplo, o autor ou o conjunto de autores de mensagens de ameaças. No contexto do atual trabalho, a Atribuição de autoria será focada na análise textual de postagens feitas na rede social Reddit.

2.1.1 Estilometria

A estilometria é a análise estatística do estilo do autor. Assim como a Atribuição de autoria, essa análise pode ser feita em outros escopos que não sejam necessariamente postagens em redes sociais, como: estilo do autor em pinturas, *design* dos códigos fontes etc. Geralmente, a estilometria é uma das ferramentas empregadas na análise textual de textos anônimos ou textos cuja a autoria esteja sendo disputada.

Com o foco da estilometria em estilo de escrita do autor, (STAMATATOS, 2009) define as características da estilometria como: léxicas, caracteres, sintáticas, semânticas e de aplicação especifica.

As características léxicas é a forma de se ver um texto como uma sequência de *tokens*, sendo estes *tokens* palavras, números ou pontuações, com isso podemos extrair características como riqueza de vocabulário, tamanho das palavras, frequência do uso das palavras etc.

As características de caracteres é a forma de se ver um texto como uma sequência de caracteres, com isso características como contagem de dígitos, caracteres alfabéticos, letras maiúsculas e minúsculas podem ser extraídas.

As características sintáticas é baseada na estrutura das frases e funções das palavras, em que a ideia é que o autor inconscientemente tende a usar estruturas sintáticas similares ao longo do texto. Com isso, as características usadas nessa abordagem são: frequência do uso funções das palavras (adjetivo, verbo, nome etc), ordem das funções das palavras, erros sintáticos etc.

As características semânticas visa analisar o significado das palavras ou de trechos de um texto, com isso da pra se analisar o significado das palavras utilizadas e das sentenças, análise do contexto e/ou assunto que está sendo discutido.

Por fim, as características de cunho de aplicações especificas irá depender do problema em si que está sendo abordado. No caso de análise de *emails*, pode-se analisar estruturas das mensagens em si, como saudações, despedidas e assinaturas das mensagens. No caso de textos provenientes de HTMLs, pode-se também analisar as fontes e suas cores utilizadas, análise de repetições dos rótulos HTML etc.

2.2 Mineração de Texto

Segundo Silva, Peres e Boscarioli (2017) a Mineração de Dados é o processo de descoberta de conhecimento a partir dos dados. Com isso, a partir do conhecimento produzido, pode-se gerar melhorias em serviços no geral.

Os dados a serem trabalhados podem possuir naturezas distintas e, com isso, a maneira de manipulá-los pode ser diferente. Os dados podem ser divididos segundo suas estruturas em dois conjuntos: os dados estruturados, cujo armazenamento dá-se em estruturas tabulares, em que cada linha representa um evento específico a ser estudado e cada coluna as características de cada evento (SILVA; PERES; BOSCARIOLI, 2017); os dados não estruturados, ao contrário dos estruturados, não possuem uma estrutura

padrão, com isso a maneira de manipulá-los é diferente. Além dos dados textuais analisados neste trabalho, outros tipos de dados não estruturados são: imagens, sons e vídeo (SILVA; PERES; BOSCARIOLI, 2017).

Com isso, a Mineração de Textos pode ser tratada como uma subárea da Mineração de Dados, em que as abordagens de pré-processamento, modelagem e classificação irão variar de acordo com objetivo do estudo. As abordagens textuais utilizadas no presente trabalho estão relacionadas ao escopo da Atribuição de Autoria, logo, certos tipos de abordagem aqui empregadas podem não ser relevantes em outros escopos.

2.2.1 Pré-processamento de texto

A etapa de pré-processamento de dados textuais é um processo importante na transformação do formato não estruturado do texto para um formato estruturado (AGGARWAL, 2018). De modo geral, uma das primeiras técnicas a ser utilizada é a de *tokenização*, nela a extração do texto resulta em uma sequência de *tokens*, isto é, uma sequência de caracteres com algum significado semântico.

A depender do domínio da aplicação envolvida, antes da tokenização algumas abordagens devem ser utilizadas a fim de remover ou extrair informações dos textos brutos. Em dados textuais provenientes de fóruns, como no caso do Reddit ou blogs, é muito comum o texto bruto conter rótulos de HTML ou *markdowns*. Portanto a remoção ou contabilização desses rótulos, neste tipo de cenário, podem ser importantes para uma posterior etapa de tokenização e modelagem dos dados.

Algumas etapas comuns em pré-processamento de dados textuais, tais como remoção de *stopwords*, radicalização e lemantização não serão abordadas, já que tais etapas podem modificar ou excluir elementos do texto intrínsecos ao estilo de escrita de cada autor.

2.2.2 Modelagem dos dados textuais

Esta seção apresenta algumas das abordagens mais utilizadas para a modelagem (ou representação) de dados textuais utilizadas na mineração de textos.

BOW

Uma das representações mais comuns de um texto se dá utilizando BOW. O conjunto de palavras em um texto é convertido para uma representação multidimensional esparsa, em que o universo de palavras ou termos corresponde às dimensões dessa representação (AGGARWAL, 2018). Em um BOW binário, caso a palavra ou termo não esteja presente no texto é atribuído a este termo o valor 0, caso contrário é atribuído o valor 1.

N-grama

A representação de BOW apresenta uma solução simples e eficiente, porém esta abordagem descarta informações contextuais em uma frase (STAMATATOS, 2009). No caso das três frases seguintes: "*take care*", "*take a look*" e "*take a shower*", a palavra *take* seria contada três vezes, porém em contextos totalmente diferentes. Com isso, n-grama é uma forma de representar o texto como N ocorrências de palavras contínuas, mais conhecidas como colocações.

Além disso, o valor de N em si pode ser tanto um valor único quanto um intervalo de valores. Por exemplo, para a frase "Hoje está um lindo dia", a representação de um bigrama seria uma lista contendo: (Hoje, está), (está, um), (um, lindo) e (lindo, dia). No caso de N sendo um intervalo de valores, utilizando o mesmo exemplo anterior, porém para 1,3-grama a sua representação seria uma lista contendo: (Hoje), (está), (um), (lindo), (dia), (Hoje, está), (está, um), (um, lindo), (lindo, dia), (Hoje, está, um), (está, um, lindo) e (um, lindo, dia).

Por fim, o n-grama pode ser utilizado para para diferentes tipos de *tokens*, então ele pode ser utilizado para *tokens* que representem palavras, caracteres, caracteres apenas de palavras (excluindo pontuações, espaços etc), rótulos POS etc.

TF-IDF

Além de representar BOW ou n-grama de forma binária ou baseado em contagens, é possível utilizar a frequência relativa dos termos em suas representações. Uma das maneiras mais comuns de se utilizar é calculando o TF-IDF através da seguinte equação 1:

$$(TF * IDF)_t = f_t * log\left(\frac{D}{D_t}\right) \quad (1)$$

em que o valor $TF * IDF$ para o termo t é representado pela sua frequência f_t ponderada pelo logaritmo do total de documentos no corpus D dividido pelo total de documentos em que o termo t aparece D_t.

Uma das vantagens na utilização da normalização IDF é que ao invés de considerar apenas os termos mais frequentes no corpus, termos pouco utilizados e, com isso, bastante intrínsecos a cada autor e importantes na autoria, apresentarão valores próximos a 1, enquanto termos mais comuns e usados por muitos autores apresentarão valores próximos a 0.

POS *tagging*

POS *tagging* ou rótulos POS é uma técnica alternativa de representação de uma sentença. Assim como uma sentença pode ser vista como um conjunto de palavras (*tokens*) obedecendo uma certa estrutura, o POS *tagging* transforma cada *token* na sentença em sua respectiva classe gramatical.

Como o presente trabalho usará somente bases textuais na língua inglesa, um exemplo da representação POS *tagging* universal para a frase "*She was reading a book.*", seria: "NOUN VERB VERB DET NOUN .", sendo a frase representada, respectivamente, por nome (ou substantivo), verbo, verbo, artigo, nome (ou substantivo) e marca de pontuação.

Baseado em tópicos

Um tipo de modelagem bastante utilizado para conjunto de dados de tópico cruzado é a modelagem por tópicos. Um dos modelos mais utilizados é o LDA. O processo de geração do modelo simplificado LDA para o documento i D_i é apresentado a seguir:

1. Gerar o número n_i de *tokens*, incluindo repetições, do D_i a partir de uma distribuição de Poisson (AGGARWAL, 2018).
2. Gerar as frequências relativas de diferentes tópicos do D_i a partir de uma distribuição Dirichlet (AGGARWAL, 2018).

3. Para cada um dos n_i *tokens* do D_i, primeiro selecione o componente latente r com probabilidade $P(G_r|\overline{X}_i)$ e então gere o termo j com probabilidade $P(t_j|G_r)$ (AGGARWAL, 2018).

Os tópicos resultantes, formalmente chamados de fatores latentes, do LDA irão variar de acordo com os *tokens* que serviram de entrada. Caso os *tokens* utilizados como entrada não contenham *stopwords*, os fatores resultantes serão frequentemente relacionados a tópicos (SEROUSSI; ZUKERMAN; BOHNERT, 2011). Entretanto, se somente forem retidos *stopwords*, os fatores resultantes perdem sua interpretabilidade como tópicos e passam a ser vistos mais como marcas estilísticas (SEROUSSI; ZUKERMAN; BOHNERT, 2011).

2.2.3 Classificação textual

A etapa de classificação, no presente trabalho, visou a atribuir corretamente o autor de um dado comentário feito no *Reddit*, levando em conta os aspectos temporais refletidos nos estilos de escrita dos autores. Por isso, foi utilizado o classificador LSTM. Como o LSTM é um classificador de natureza mais complexa dentre outros mais comuns em classificação textual, como SVM, NB ou até mesmo a Rede Neural MLP, abaixo será dada uma explicação detalhada sobre os aspectos dessa rede e outros classificadores utilizados no decorrer do trabalho.

LSTM

O LSTM é um tipo de RNN fechada. As RNNs são baseadas na ideia de criação de caminhos através do tempo, cujas derivadas não explodem, com pesos de conexão que podem mudar a cada passo de tempo (GOODFELLOW; BENGIO; COURVILLE, 2016).

A estrutura de uma célula LSTM é mostrada na figura 1. As células são conectadas recorrentemente umas com as outras, substituindo as unidades escondidas de redes recorrentes comuns (GOODFELLOW; BENGIO; COURVILLE, 2016).

Uma célula de LSTM possui as unidades de *entrada*, *portão de entrada*, *estado*, *portão de esquecimento*, *saída* e *portão de saída*.

A unidade de *entrada* da célula é calculada normalmente como uma característica de uma rede neural artificial, recebendo como entrada valores provenientes da etapa de pré-processamento como o n-grama de palavras, caracteres etc, cujo valor pode ser acumulado para a unidade de estado caso o *portão de entrada* sigmoide permitir.

A unidade de *estado* possui um auto-laço (*self-loop*) cujos pesos são controlados pelo *portão de esquecimento*, sendo que cada quadrado preto da figura representa um atraso de um passo de tempo. O papel do *portão de esquecimento* com o atraso no tempo em conjunto com a unidade de *estado* é o que representa a "memória" da rede ao longo do tempo, sendo que o papel do *portão de esquecimento* é justamente controlar, ao longo do tempo, o que será esquecido ou não pela LSTM.

A unidade de *saída* da célula pode ser desligada através do *portão de saída* (GOODFELLOW; BENGIO; COURVILLE, 2016). Conforme observado na figura 1, a unidade de *estado* pode ser usada como uma entrada extra para as outras unidades.

Como a LSTM é baseada em modelos sequenciais, o seu uso no presente trabalho visa a aproveitar a característica da memória de longo prazo do modelo. Esta característica é alcançada pelas unidades de *estado* de células LSTM, cujas informações anteriormente passadas pelas unidades de *estado* podem ser retidas, utilizando combinações de operações parciais de esquecimento e incremento das unidades de *portão de esquecimento* e *portão de entrada*, respectivamente (AGGARWAL, 2018).

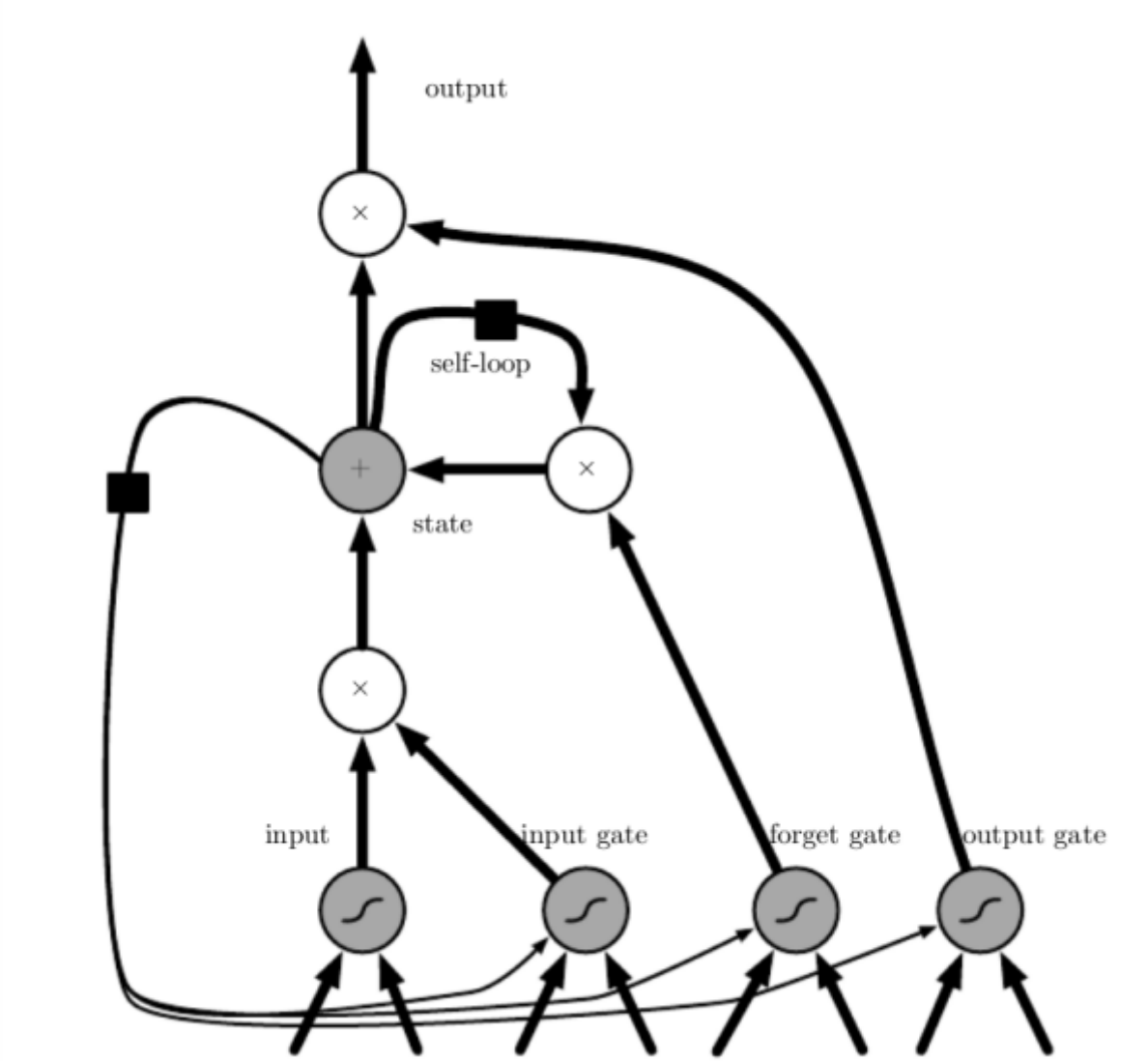
A unidade de *portão de esquecimento* $f_i^{(t)}$ (para cada passo de tempo t e célula i) controla os pesos do *self-loop* configurando-os para valores entre 0 e 1 através da unidade sigmoide (GOODFELLOW; BENGIO; COURVILLE, 2016):

$$f_i^{(t)} = \sigma\left(b_i^f + \sum_j U_{i,j}^f x_j^{(t)} + \sum_j W_{i,j}^f h_j^{(t-1)}\right) \tag{2}$$

em que $x^{(t)}$ é o vetor de entrada do passo atual e $h^{(t)}$ corresponde ao vetor da camada escondida do passo atual, contendo todas as saídas de todas as células LSTM. b^f, U^f e W^f são, respectivamente, os *bias*, pesos de entrada e pesos recorrentes para os portões de esquecimento. Os estados internos das células LSTM são atualizadas da seguinte maneira, porém com um peso *self-loop* condicional $f_i^{(t)}$ (GOODFELLOW; BENGIO; COURVILLE, 2016):

$$s_i^{(t)} = f_i^{(t)} s_i^{(t-1)} + g_i^{(t)} \sigma\left(b_i + \sum_j U_{i,j} x_j^{(t)} + \sum_j W_{i,j} h_j^{(t-1)}\right) \tag{3}$$

Figura 1 – Diagrama em bloco de uma célula da Rede Neural Recorrente LSTM



Fonte: Goodfellow, Bengio e Courville (2016)

em que b, U e W denotam, respectivamente, os *bias*, pesos de entrada e pesos recorrentes da célula LSTM. A unidade de portão de entrada externo $g_i^{(t)}$ é calculada de maneira similar ao portão de esquecimento (com a utilização da unidade sigmoide para obter um valor entre 0 e 1), porém com seus próprios parâmetros (GOODFELLOW; BENGIO; COURVILLE, 2016):

$$g_i^{(t)} = \sigma\left(b_i^g + \sum_j U_{i,j}^g x_j^{(t)} + \sum_j W_{i,j}^g h_j^{(t-1)}\right) \tag{4}$$

A saída $h_i^{(t)}$ da célula LSTM pode ser desligada através da unidade portão de saída $q_i^{(t)}$, que também utiliza unidade sigmoide da seguinte maneira (GOODFELLOW; BENGIO; COURVILLE, 2016):

$$h_i^{(t)} = tanh\left(s_i^{(t)}\right) q_i^{(t)} \tag{5}$$

$$q_i^{(t)} = \sigma\left(b_i^o + \sum_j U_{i,j}^g x_j^{(t)} + \sum_j W_{i,j}^o h_j^{(t-1)}\right) \tag{6}$$

em que os parâmetros b^o, U^o e W^o são os *bias*, pesos de entrada e pesos recorrentes, respectivamente (GOODFELLOW; BENGIO; COURVILLE, 2016).

SGD

O classificador SGD implementa a classificação de modelos lineares, como SVM ou regressão logística, mas utilizando o treinamento SGD. A implementação do SGD traz algumas vantagens no âmbito dos grandes conjuntos de dados, pois possibilita o uso de *code tuning*, isto é, configurações diferentes para poder ter um melhor resultado. Adicionalmente, em diferentes cenários esta técnica apresentou resultados bastante eficientes. Suas desvantagens são a quantidade de hiperparâmetros a serem configurados e sua sensibilidade à normalização das características.

Perceptron

A ideia do classificador Perceptron é muito semelhante ao classificar SGD, de fato, caso a função de perda seja configurada para ser *'perceptron'* os classificadores se tornam equivalentes. O que o torna diferente do classificador SGD é não necessitar de uma taxa de aprendizado e, por padrão, não haver uma função de penalização.

Passive-Aggressive

Os algoritmos *Passive-Aggressive*, assim como Perceptron e SGD são voltados para aprendizado de larga escala. Sua implementação é similar ao Perceptron, não requerendo taxa de aprendizado. A única diferença é a utilização de uma constante C para regularizar o tamanho do passo máximo que pode ser tomado, minimizando a influência de *outliers* no modelo.

3 Revisão de Literatura

Swain, Mishra e Sindhu (2017) apresentaram uma revisão de literatura abordando diferentes técnicas do campo de atribuição de autoria. Os autores pesquisaram trabalhos relativos às tarefas de: atribuição de autoria, caracterização de autoria, verificação de autoria, descriminação de autoria, detecção de plágio, indexação e segmentação de texto e desidentificação de autoria. Os trabalhos relatados pertencem a um domínio bem amplo, como mídias sociais, obras literárias, e-mails, códigos de *software*, etc. Com isso, esta revisão tem um objetivo bem similar a revisão abordada neste capítulo. Contudo, a Revisão de Literatura apresentada neste capítulo aborda métodos e técnicas de detecção de autoria em dados textuais provenientes de mídias sociais. Pelo fato do tema da revisão ser mais específico, ele é tratado de maneira mais profunda, explorando técnicas de classificação para a detecção de autoria, técnicas de clusterização para geração de características distintas entre usuários e a eficácia na utilização de novas características. Além disso, a Revisão de Literatura deste capítulo possui um escopo de tempo maior, analisando artigos publicados até Setembro de 2019 e que, portanto, não foram incluídos no trabalho de Swain, Mishra e Sindhu (2017).

3.1 Métodos de pesquisa

Com o objetivo de entender o cenário atual sobre métodos e técnicas de detecção de autoria em dados provenientes de mídias sociais, um protocolo de Revisão Sistemática foi criado a fim de explorar mais a fundo o tema. O protocolo foi elaborado seguindo as diretrizes propostas por Kitchenham e Charters (2007). Assim, a revisão foi organizada em três fases: planejamento, condução e extração.

3.1.1 Planejamento

A fase de planejamento é responsável por planejar a forma como será implementado o protocolo da revisão sistemática. Nesta fase, são levantados o objetivo da revisão, as fontes a serem pesquisadas, os critérios de inclusão e exclusão e os dados a serem extraídos dos trabalhos.

Objetivo da Revisão

Esta revisão visa a entender o estado da arte em métodos e técnicas utilizadas na detecção de autoria em dados textuais provenientes de mídias sociais, a fim de avaliar as lacunas existentes no estado da arte e se justifica por não ter sido encontrada na literatura uma revisão tão atualizada e específica sobre este assunto.

Questões de pesquisa

Esta revisão visa responder às seguintes perguntas:

1. Quais são os métodos e as técnicas existentes para o pré-processamento de dados textuais de mídias/redes sociais para a detecção de autoria?
2. Quais são os métodos e as técnicas para a classificação/identificação/detecção da autoria propriamente dita?
3. Quais são as técnicas e métricas utilizadas para avaliar esses métodos?

Critérios de Inclusão e Exclusão

A fim de avaliar se o trabalho é relevante para a revisão elaborada, os seguintes critérios de inclusão e exclusão de trabalhos foram elaborados:

Critérios de inclusão:

1. Serão incluídos trabalhos publicados e disponíveis integralmente em base de dados científicas.
2. Serão incluídos trabalhos que já possuam aprovação pela comunidade científica.
3. Serão incluídos trabalhos que abordam métodos e técnicas para detecção de autoria a partir de bases textuais de mídias/redes sociais.

Critérios de exclusão:

1. Serão excluídos trabalhos que não se referem à tarefa de identificação de autoria.
2. Serão excluídos trabalhos que não utilizem informações textuais para a detecção de autoria.
3. Serão excluídos trabalhos que não utilizem bases textuais de mídias/redes sociais.

4. Serão excluídos trabalhos de revisão (estudos secundários).
5. Serão excluídos trabalhos que não estejam escritos na língua inglesa.

Para ser aceito para a revisão, um artigo deve atender a todos os critérios de inclusão e a nenhum dos de exclusão.

Protocolos de Busca

As bases de dados utilizadas para a obtenção dos artigos foram ACM DL[1] e IEEE Xplore[2]. A string de busca utilizada em ambas as bases é apresentada a seguir:

Palavras-chaves: "*Authorship*" relacionada com os termos "*detection*", "*classification*", "*identification*", "*analysis*" ou "*attribution*" , e "*text mining*" ou "*data mining*" ou "*temporal series*".

3.1.2 Condução

Nesta etapa, todos os títulos, resumos e palavras-chaves dos 371 artigos foram analisados, àqueles que atendiam a todos os critérios de inclusão, comentados anteriormente, foram selecionados para uma leitura integral. Com isso, foram 40 artigos inicialmente selecionados. Todavia, conforme apresentado na figura 2, após uma leitura integral, 9 artigos foram excluídos por não atenderem plenamente aos critérios.

3.1.3 Extração

Os 31 artigos restantes foram lidos na íntegra e seus dados foram extraídos. As Tabelas 1, 2 e 3 sumarizam as características de cada um dos 31 artigos.

3.2 Resultados

Em uma análise geral dos artigos extraídos, a figura 3 mostra que o número de trabalhos relacionados à atribuição de autoria em mídias sociais cresceu e ficou mais

[1] ACM Digital Library: ⟨https://dl.acm.org/⟩, acessado em 24/01/2020.
[2] IEEE Xplore Digital Library: ⟨https://ieeexplore.ieee.org/Xplore/home.jsp⟩, acessado em 24/01/2020.

Tabela 1 – Resumo dos trabalhos analisados

Referência	Ano	Conjunto de dados	Características	Técnicas	Foco
Abbasi e Chen (2005)	2005	'*White Knights*' (Fórum) '*Pales-tinian Al-Aqsa Martyrs*' (Fórum)	Léxico, Caracteres e Aplicação específica	C4.5, SVM e agrupamento (ROECK; AL-FARES, 2000)	AA
Koppel *et al.* (2006)	2006	Blog: 18,000 conjunto de textos de usuários	Léxico e Caracteres	Linear SVM e Similaridade Cosseno	AA
Tan e Tsai (2010)	2010	Projeto Gutenberg (E-books), BizBlogs(Blog) e Fórum não especificado	Léxico	NB	AA
Layton, Watters e Dazeley (2010)	2010	Twitter: 200 *tweets* de 14,000 usuários	Caracteres	SCAP e *K-means*	AA
Pillay e Solorio (2010)	2010	Fórum do portal '*Chronicle of Higher Education*'	Léxico, Caracteres e Aplicação específica	C4.5, NB, Redes Bayesianas e CLUTO (agrupamento)	AA
Seroussi, Zukerman e Bohnert (2011)	2011	Judgment, IMDb62 e Blog	Léxico e Tópicos	SVM e LDA	AA
Lakkaraju, Bhattacharya e Bhattacharyya (2012)	2012	Twitter e Facebook	Tópicos, Baseado em tempo Customizada	KNN e D-MRelCRP	Novo modelo
Cristani *et al.* (2012)	2012	Skype: conversas de chat diárias	Léxico, Caracteres e Aplicação específica	*Cumulative Match Characteristic*	AA
Seroussi, Bohnert e Zukerman (2012)	2012	PAN'11 e Blog	Léxico	Linear SVM, LDA+Hellinger, AT, AT-FA e DADT	AA
Donais *et al.* (2013)	2013	SMS	Customizada	ChatSafe, Discrete NB e Classificador Gaussiano	AA
Roffo *et al.* (2013)	2013	Skype	Léxico, Caracteres e Aplicação específica	*Cumulative Match Characteristic*	AA
Perez *et al.* (2013)	2013	Twitter	Aplicação específica	SPOT(SVM)	AA
Bouanani e Kassou (2013)	2013	Fórum	Customizada	*Correspondence Factorial Analysis* (CFA) e TXM	Nova característica

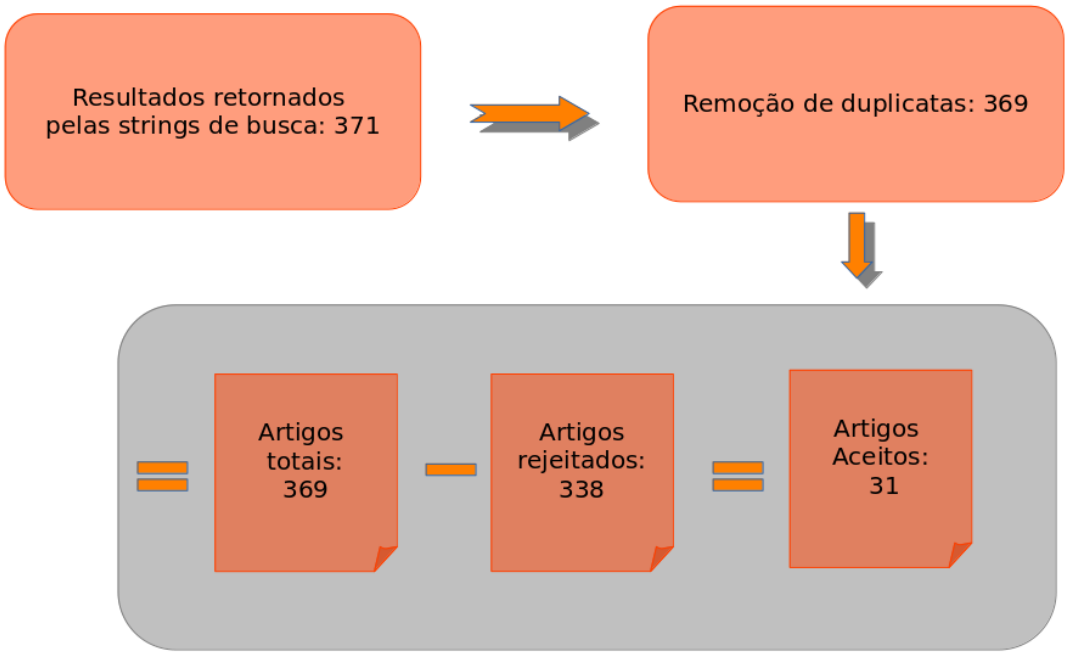
Tabela 2 – Resumo dos trabalhos analisados

Referência	Ano	Conjunto de dados	Características	Técnicas	Foco
Seker, Al-Naami e Khan (2013)	2013	IMDB62	Léxico e Sintático	SVM, (SAHA *et al.*, 2018), ECSMiner e DXMiner	AA
Inches, Harvey e Crestani (2013)	2013	IRC logs ("krjin" e "irclogs")	Léxico	Distância Qui-Quadrado e *Kullback-Leibler Divergence* (KLD)	AA
Roffo *et al.* (2014)	2014	*Kimble*	Léxico, Caracteres, Aplicação Específica	*Cumulative Match Characterisct* e *Epsilon Support Vector Regressor* (E-SVR)	AA e Traços
Seroussi, Zukerman e Bohnert (2014)	2014	Judgment, PAN'11, IMDb62, IMDb1M e Blog	Tópicos	AT-P, AT-FA-P1, AT-FA-P2 DADT-P, Token SVM, LDA-SVM, AT-SVM, AT-FA-SVM, DADT-SVM	AA
Albadarneh *et al.* (2015)	2015	Twitter	Léxico	NB	AA
Azarbonyad *et al.* (2015)	2015	Twitter e Enron Email	Caracteres	SCAP, Regressor Linear (decaimento)	AA
Igawa *et al.* (2015)	2015	Twitter	Léxico	*Simplified Profile Intersection* (SPI)	AA
Spitters *et al.* (2015)	2015	Fórum *Black Market Reloaded*	Léxico, Caracteres e Baseado em tempo	SVM (LIBLINEAR) e Similaridade Cosseno	AA AV
Kim, Noh e Park (2015)	2015	*Bulletin Board 'a park for everyone'*	Caracteres, Aplicação específica e Baseado em tempo	SVM (*kernel* linear)	*Multi-id users*
Kuzu, Balci e Salah (2016)	2016	*CCSoft Okey Player Abuse* (COPA)	Léxico, Caracteres e Aplicação específica	Similaridade Cosseno e *re-centered local profile* (RLP)	AA
Yan e Matthews (2016)	2016	Twitter: 100 *tweets* chineses de dez usuários	Léxico e Caracteres	SKM e EM	AA
Halvani, Winter e Graner (2017)	2017	PAN'13, PAN'14, PAN'15, Koppel Blogs, Amazon Product Data e Reddit Cross-Topic	Caracteres	*Normalized Compression Distance* (NCD), *Compression-based Cosine* (CBC) e *Chen-Li metric* (CLM)	AV
Sultana, Polash e Gavrilova (2017)	2017	Twitter: 200 *tweets* de 70 usuários	Léxico, Caracteres e Aplicação específica	*Cumulative Match Characteristic*	AA

Tabela 3 – Resumo dos trabalhos analisados

Referência	Ano	Conjunto de dados	Características	Técnicas	Foco
Banga e Mehndiratta (2017)	2017	Twitter: 500 *tweets* de 10 usuários	Léxico, Caracteres e Sintático	SVM, MLP, Regressão Logística, SVC Linear e NB	AA
Petrasova, Khairova e Lewoniewski (2018)	2018	Blog *Authorship Corpus* (SCHLER *et al.*, 2006)	Customizada	Precisão	Nova característica
Le e Safavi-Naini (2018)	2018	Twitter	Léxico, Caracteres e Customizada	SMO SVM	AA
Litvinova, Litvinova e Panicheva (2019)	2019	KavkazChat	Léxico, Caracteres e Sintático	SVC Linear	AA
Ding *et al.* (2019)	2019	PAN' 14 e ICWSM2012	Customizada	Regressão logística	Nova característica

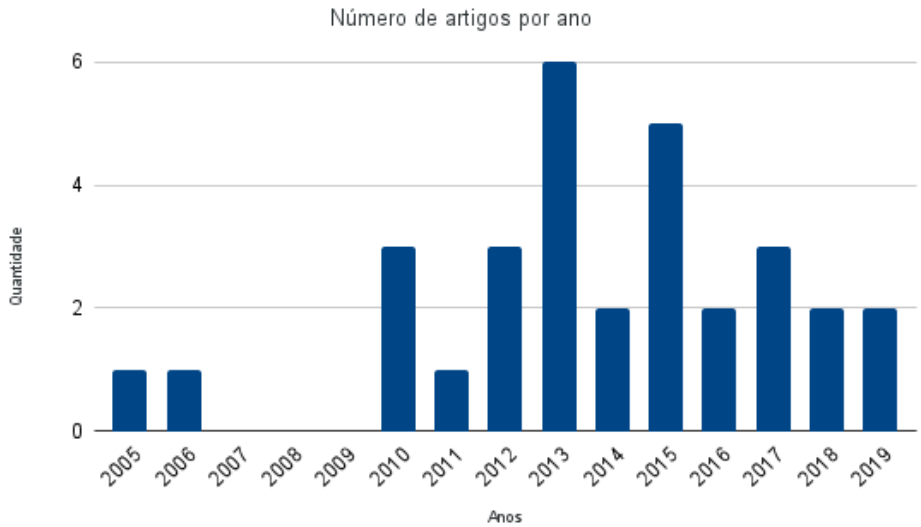
Figura 2 – Diagrama da seleção dos trabalhos



Fonte: Guilherme Casimiro, 2022

frequente a partir de 2010, provavelmente impactada pelo crescimento de redes sociais como *Twitter*, *Facebook* e *Reddit*.

Figura 3 – Distribuição da quantidade de publicações ao longo dos anos

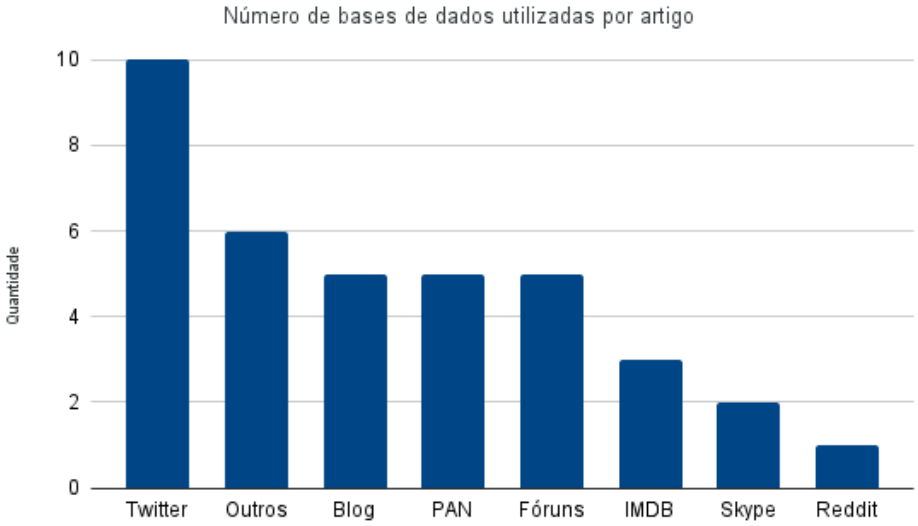


Fonte: Guilherme Casimiro, 2022

Dentre os conjuntos de dados dos trabalhos analisados, a figura 4 evidencia um grande uso do Twitter, PAN, Blog e fóruns no geral, sendo responsáveis por mais de 67%

das bases utilizadas nos trabalhos. Vale ressaltar apenas uma utilização do Reddit como base de dados em um trabalho que comparou bases de dados muito variadas.

Figura 4 – Distribuição das bases de dados utilizadas

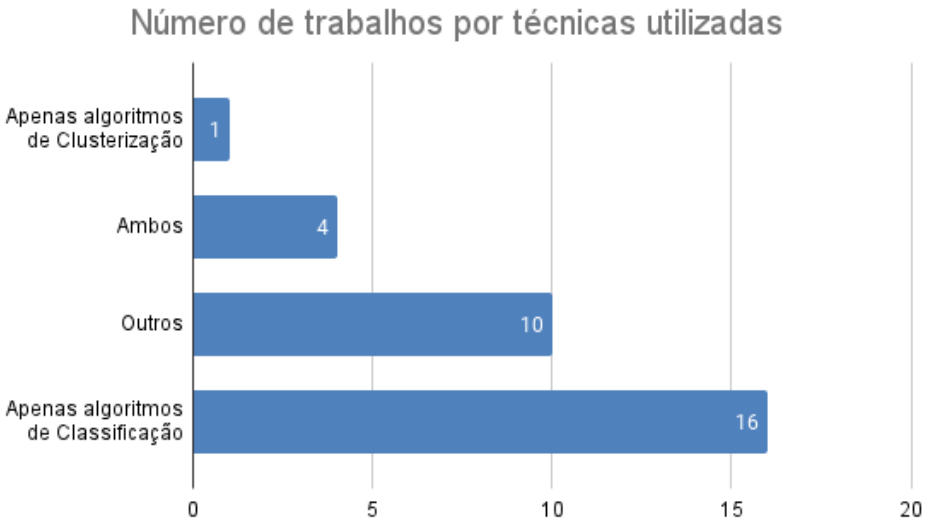


Fonte: Guilherme Casimiro, 2022

Houve uma maior utilização de algoritmos de classificação como pode ser visto na figura 5, visto que a maioria dos trabalhos focaram na identificação dos autores de um dado texto, porém observa-se a utilização de algoritmos de clusterização tanto para auxiliar na classificação quanto na análise de características. Alguns trabalhos não utilizaram algoritmos de classificação ou clusterização, estes trabalhos focaram em propor diferentes formas de utilização de características em cada cenário, com isso houve um grande uso de *Cumulative Match Characteristic*, Similaridade Cosseno e LDA.

Para melhor compreensão dos detalhes de cada artigo, estes foram agrupados em três tópicos: Atribuição de autoria - modelagem geral, que consiste na utilização de métodos e técnicas consolidados em redes sociais e desenvolvimento de novas abordagens em Atribuição de autoria; Atribuição de autoria em *chats*, técnicas e métodos exclusivos de artigos que se basearam em *chats* ou mensagens instantâneas; Atribuição de autoria baseado em tópicos, artigos que utilizaram ou propuseram um abordagem contendo a utilização de tópicos como característica discriminante do autor.

Figura 5 – Distribuição dos algoritmos utilizados



Fonte: Guilherme Casimiro, 2022

3.2.1 Atribuição de autoria - modelagem geral

Abbasi e Chen (2005) descrevem um dos primeiros modelos de Atribuição de Autoria para mídias sociais, em específico, voltado a fóruns de grupos extremistas. Os autores modelaram o problema tanto para fóruns na língua árabe quanto na língua inglesa. As características envolvendo o problema foram: de natureza léxicas envolvendo contagem simples de tamanho de palavras e sentenças; riqueza de vocabulário; caracteres, utilizando medidas simples de contagem de caracteres e frequência de caracteres; aplicação específica, envolvendo tamanho de fonte, *hyperlinks*, inserção de imagens e informações de contato. Além dessas características, os autores tiveram que adaptar os tipos de características utilizadas do inglês para o árabe, preservando o problema de alongamento de palavras (a mesma palavra só que com tamanho maior) e utilizando palavas raízes através de um algoritmo de clusterização. Para resolver o problema de Autoria, o algoritmo SVM superou a Árvore de decisão C4.5 em todos os cenários.

Koppel *et al.* (2006) propuseram uma das primeiras abordagens para lidar com Atribuição de Autoria em larga escala, para isso foi utilizada uma base de 10,000 blogueiros. O trabalho apresentou dois experimentos: o primeiro utiliza informações léxicas como TF-IDF de palavras de conteúdo, funções de palavras e *strings* de caracteres não alfanuméricos e IDF de palavras de conteúdo; o segundo experimento corresponde à utilização de meta-aprendizado, como: similaridade absoluta do *snippet* para o autor mais bem ranqueado,

Gap em graus de similaridade absoluta entre o autor mais bem ranqueado e o k-autor ranqueado, o *rank* do autor mais bem ranqueado usando as outras duas representações entre outras em conjunto com as representações anteriores. O segundo experimento visa a responder se é possível dar a resposta de um autor, através de um par de sucesso, em que todas as três representações identifiquem o mesmo autor mais bem ranqueado. Para o primeiro experimento, a chance do autor mais bem ranqueado ser de fato o autor do *snippet* é de menos 40%, enquanto que para a segunda tentativa a chance de dar uma resposta conclusiva se é ou não o autor é de 31,3%, sendo que 88,2% o autor é corretamente classificado.

Tan e Tsai (2010) apresentaram uma comparação de atribuição de autoria envolvendo textos online de blogs e *e-books*. Para isso, foram utilizadas algumas características léxicas de contagem simples de palavras, sentenças, funções de palavras e algumas características de caracteres para contagem de algumas marcas de pontuações. Em ambos os conjuntos de dados a etapa de autoria foi realizada com apenas dois usuários. O classificador NB se mostrou adequado para textos com grandes conjuntos de treinamento e cujos textos não sejam muito longos. A utilização de funções de palavras atingiu ótimos resultados para textos com mais de 5.000 palavras, enquanto que características léxicas e de caracteres de contagem simples são boas para textos mais curtos, com menos de 5.000 palavras.

Layton, Watters e Dazeley (2010) abordaram o problema da limitação de caracteres do Twitter, sabendo que os modelos tradicionais de Atribuição de Autoria lidam com um número mínimo de palavras suficientemente grandes, o limite máximo de 140 caracteres imposto pelo Twitter, até então, era uma grande barreira. Para a realização dos experimentos foi utilizado o SCAP e n-grama de caracteres, como características, para $1 < N < 8$. Além dos valores de N diferentes, os experimentos foram realizados em quatro cenários de pré-processamento diferentes: texto bruto, substituindo todas as citações pelo caractere @, substituindo todas as *hashtags* pelo caractere # e a combinação de ambas as substituições anteriores. Os resultados mostraram que as *hashtags* compartilhadas pelo usuário não são efetivas para identificação da autoria, porém ao retirar as citações a acurácia caiu 26%, indicando que identificar as pessoas que o usuário responde nas redes é efetivo para identificar a autoria. Além disso, houve uma verificação do número de *tweets* para cada usuário para se obter um bom resultado, em que até 120 *tweets* a acurácia foi aumentando e a partir de 140 *tweets* estabilizou.

Pillay e Solorio (2010) lidaram com identificação de autorias em um fórum para um número máximo de 100 usuários. Neste trabalho, além de lidar com um número grande de usuários e características léxicas, caracteres e sintáticas comuns de atribuição de autoria, este trabalho propõe o uso de agrupamento de forma que cada usuário pertença somente a um grupo. A inserção do agrupamento melhorou a acurácia em todos os casos.

Perez *et al.* (2013) desenvolveram um *framework* que combina análise de conteúdo e comportamental de perfis do Twitter, para detecção de redes suspeitas. O objetivo deste trabalho focou em primeiro detectar perfis suspeitos, após a detecção é identificado o relacionamento entre esses perfis. Por último, um algoritmo de agrupamento é utilizado para evidenciar essas campanhas suspeitas. Este trabalho identificou que perfis suspeitos utilizam muitas *hashtags* para ganhar visibilidade e tendem a utilizar muito mais URLs.

Diante do problema de *Big Data* e dados *streaming*, abordagens tradicionais de treinamento de classificadores podem ser um problema. Com isso, Seker, Al-Naami e Khan (2013) abordaram o problema de Atribuição de Autoria em dados *streming*, os autores mostraram que a taxa de erro tende a diminuir com o aumento do tamanho dos *chunks* apresentado ao classificador.

Ainda na questão de *Big Data*, Albadarneh *et al.* (2015) trazem esse tema para o problema de autenticação de autoria em *tweets* árabes, cuja língua possui uma série de dialetos e recursos de NLP escassos. Foi utilizada a representação de BOW em conjunto com TF-IDF. Os perfis utilizados para extração de autoria foram perfis de celebridades árabes, com isso os resultados atingidos não foram altos, a hipótese dos autores é que os *tweets* postados estavam relacionados a temas de outros *tweets* de celebridades (dado que a obtenção dos *tweets* foram no mesmo período), com um enfoque muito grande no público alvo/seguidores.

Bouanani e Kassou (2013) utilizam técnicas de construção de vocabulários para identificar assinaturas de perfis de autores na internet. Para isso, é utilizado o método de CFA, este método permite uma visualização gráfica em um espaço de baixa dimensão de dados categóricos.

Com o crescimento das redes sociais e sua interação ao longo do tempo, algumas questões referentes à escrita dos usuário ao longo do tempo são levantadas. O trabalho Azarbonyad *et al.* (2015) levantou três questões em conjuntos de dados do Twitter e de E-mails: o estilo de escrita do autor muda com o passar do tempo? Caso mude, a mudança acontece na mesma frequência?; Como a mudança temporal do estilo de escrita afeta na

Atribuição de Autoria?; Como essas mudanças podem ser capturadas e incluídas?. Os resultados mostraram que os usuários mudam significativamente seus estilos de escrita ao longo do tempo, porém em taxas e tempos diferentes. O modelo de amostragem de característica baseado em tempo superou os modelos de amostragem de característica não baseado em tempo e SCAP baseado e não baseado em tempo.

Igawa *et al.* (2015) abordaram a questão de contas que foram comprometidas no Twitter, ou seja, contas legítimas que foram tomadas e usadas para publicar conteúdo falso ou ameaçador. Para simular que a conta foi comprometida, os autores colocaram no conjunto de dados de um usuário, dados de teste aleatórios de outros usuários. Aplicar remoção de *hashtag* e citação pioraram os resultados, como visto em outros trabalhos. Um número suficientemente grande de palavras em um corpus também foi avaliado, com 50 palavras num corpus obtendo resultados ruins por não conseguir extrair muitos n-gramas e com 100 palavras obtendo o melhor resultado. Além disso, a abordagem utilizada tem resultados muito bons em precisão e revocação, possuindo uma taxa de falsos negativos baixa.

Em um cenário de fórum de mercado na *Dark Web*, Spitters *et al.* (2015) levantam duas temáticas de autoria: identificação de *alias*, similar ao problema de múltiplos ids de usuário, e atribuição de autoria. Os autores criaram os *alias* a partir de um subconjunto de postagens de um usuário. Assim como em outros trabalhos, foi utilizado um vetor de tempo para capturar a hora do dia em que foi postado, em conjunto com características léxico e caracteres. Para classificação de *alias*, ao reduzir o número de *alias* o desempenho aumenta. O resultado no caso da atribuição de autoria, a chance do correto autor estar no top cinco usuários prováveis é de 97%.

Yan e Matthews (2016) utilizou os algoritmos de agrupamento, SKM e EM, para determinar a autoria de *tweets* chineses. Foram utilizados 10 usuários no conjunto de dados e variando $k = 3$, $k = 5$ e $k = 10$. Os resultados mostraram que SKM foi melhor nos cenários em que todas as características foram utilizadas, e utilizando somente função de palavras o algoritmo EM foi o que obteve o melhor desempenho.

Um outro tipo de características que dá pra ser extraídos baseado em caracteres são os métodos de compressão, estes métodos são geralmente utilizados por programas de compressão de arquivos. Halvani, Winter e Graner (2017) utilizaram quatro tipos de compressores como característica para verificação de autoria: PPMd, GZip, BZip2, Zip e LZW. Foram utilizadas três medidas de dissimilaridade: NCD, CBC e CLM. Os autores

testaram diferentes tipos de características em variados conjunto de dados, sendo um deles do Reddit. Em um primeiro experimento, os autores testaram os cinco tipos de características, em conjunto com as três diferentes medidas de dissimilaridade em quatro corpus do PAN a fim de achar a melhor combinação para o experimento de reconhecimento de autoria. No segundo experimento, a combinação de PPMd e CBC foi utilizada a fim de determinar o *threshold* do método proposto, para comparação com baselines. Os baselines utilizados foram GLAD, *Profile Based Method* e *Impostors Method*. Em dois cenários, o método proposto superou os três *baselines* propostos e atingiu resultados similares ao GLAD nos conjuntos PAN'13 e PAN'14.

Banga e Mehndiratta (2017) fizeram um estudo exploratório de diversos cenários da Atribuição de Autoria no Twitter, envolvendo: diferentes tipos de classificação, características e tamanho do corpus. Os classificadores utilizados no trabalho foram: SVM-RBF, SVM-Sigmóide, NB, Regressão Logística, SVC e MLP. Os tipos de características utilizadas foram: n-grama de palavras, caracteres e rótulos POS, e frequência de distribuição de palavras. Regressão logística e SVC obtiveram os melhores resultados gerais, a MLP se mostrou eficiente na utilização de n-grama de caracteres.

Sultana, Polash e Gavrilova (2017) propõem um método de construção de perfil sociocomportamental, baseado em vetores de resposta de amigos, amigos *retweetados* ou mencionados, *hashtags* compartilhadas e domínios ou *links* compartilhados. Todos os vetores possuem um elemento com uma tupla contendo um termo referente ao amigo, *hashtags* ou *link* e o peso desse termo utilizando TF para os amigos e TF-IDF para *hashtags* e *links*. O método proposto mostrou-se estável e com taxa de reconhecimento acima de 90%.

Petrasova, Khairova e Lewoniewski (2018) apresentou um modelo de construção de colocação de similaridade semântica, sendo uma colocação uma combinação de duas ou mais palavras frequentemente usadas juntas e sendo sintaticamente e semanticamente integradas. Esse modelo combina informações estatísticas, sintáticas e semânticas das palavras. O modelo resultante propõe três colocações: Adjetivo-nome, Nome-nome e Verbo-nome.

Le e Safavi-Naini (2018) propõe um algoritmo de informação mútua para refino de características discriminantes de cada autor. O algoritmo de *k-signature*, um caso específico de informação mútua, foi utilizado como comparação. O algoritmo proposto seleciona as top-l características mais discriminantes de cada autor baseado em uma

série de probabilidades. Os resultados mostraram que o algoritmo é estável, superando o *k-signature* em todos os casos.

Litvinova, Litvinova e Panicheva (2019) analisam diferentes tipos de características para Atribuição de Autoria em um fórum de textos russos, a fim de avaliar a eficácia de n-grama de caracteres para o dialeto russo; validar se tipos diferentes de n-gramas de caracteres codificam informações em termos de estilo e conteúdo; verificar se n-gramas de caracteres podem ser supridos com características genéricas representando padrões de discurso. Foram utilizados um corpus de tópico simples e outro de múltiplos tópicos. As características utilizadas no trabalho seguiram trabalhos anteriores, utilizando *affix n-grams* (prefixo, sufixo, prefixo espaçado e sufixo espaçado), *word n-grams* (palavra toda, meio da palavra e multi palavras) e *punctuation n-grams* (*beg-punct*, *mid-punct*, *end-punct*). Além destas características, outras quatro características foram propostas: *Discourse marker n-grams* (DMarker) n-grama de palavras com pelo menos uma palavra como marcador de discurso e sem rótulo, as outras palavras são representadas pelo seus rótulos POS; *Combined n-grams which capture syntactic information on POS and punctuation levels* (PunctPOS) n-grama de *tokens* (palavras e pontuações) com pelo menos um dos *tokens* representados por uma pontuação, todas os outros *tokens* são representados pelos rótulos POS; *Combined n-grams with words replaced with ** (StarMark), igual ao PunctPos, porém ao invés dos rótulos POS, as palavras são marcadas por *; *Combined n-grams with punctuation marks replaced by PNCT and words replaced with ** (StarPunct), igual ao StarMark, porém as marcas de pontuação são trocadas por 'PUNCT'.

Ding *et al.* (2019) apresenta um modelo de aprendizado utilizando redes neurais para propor três tipos de características: Tópico e Léxico, Caracteres e Sintático. O Tópico e Léxico é baseado na escolha dos *tokens* léxicos, pelo autor, utilizando uma sequência do seu vocabulário para construir uma sentença e expressar seus interesses. Para Caracteres a fim de capturar diferentes escolhas morfológicas na escolha e escrita dos *tokens* léxicos. O Sintático se baseia nos vizinhos de *token* léxico como bigramas de rótulos POS.

3.2.2 Atribuição de Autoria em chats

Apesar de poderem ser inclusas no âmbito das mídias sociais, as mensagens provenientes de *chats* ou mensagens instantâneas possuem uma natureza diferente de mensagens

postadas em fóruns ou Twitter, por exemplo. Como é o caso das características baseadas em turnos, em que cada turno é medido baseado na última mensagem do interlocutor com o usuário. Cristani *et al.* (2012) apresentaram características estilométricas voltadas à natureza das conversas online. Foram abordadas tanto características do tipo caracteres já bem estabelecidas em Atribuição de Autoria, quanto durações de turno, velocidade de escrita, graus de mimetismo e número de caractere de retorno. Ao utilizar características desta natureza houve aumento significativo no desempenho da Atribuição de Autoria.

Donais *et al.* (2013) fizeram um estudo comparativo da ferramenta *ChatSafe* e outros classificadores, esta ferramenta recebe os textos como entrada e usa o conceito de evidências na construção de perfis de diferentes usuários.

Inches, Harvey e Crestani (2013) apresentaram uma nova abordagem para o problema de atribuição de autoria em *logs* de *chat* IRC. Os autores propuseram um método que seleciona as palavras mais discriminantes para cada autor, baseado em suas conversações. O método proposto se baseia na proposição que uma mensagem de um usuário pode impactar todas as futuras mensagens naquela conversa. Com isso, o método proposto identifica o vocabulário específico de um usuário dentre todos os usuários com quem ele conversou.

Roffo *et al.* (2013) propuseram uma nova modelagem para conversas online. Ao invés de considerar características baseadas na conversa como um todo, a modelagem proposta baseia-se nos turnos. Além de características léxicas, sintáticas e características baseadas em turno já abordadas, foram utilizados também o tempo de resposta do turno anterior e categoria dos *emoticons* (positivo, negativo, outro). Foi verificado que com o aumento do número de turnos, o reconhecimento do autor aumentou.

Roffo *et al.* (2014) abordaram a relação de personalidade e identidade usando o chat da rede social *Kimble*. A participação dos experimentos foi feita sempre com um interlocutor no chat e ao final da conversa os participantes preencheram três questionários para avaliar fatores psicológicos: BIS-11, BIS/BAS e PANAS. Ao ligar os resultados de correlação entre características estatísticas e características de personalidade descobriu-se que afetividade positiva é fortemente relacionado ao estilo da pessoa, pois é correlacionada com quatro características estatísticas. Utilizando-as, a detecção de autoria atingiu 72% de AUC ROC; impulsividade planejada se correlaciona com cinco características que sozinhas respondem por 69,4% de AUC, assim como as quatro características correlacionadas à

evasão de punição; afetividade negativa possuí três características correlacionadas, obtendo 65,9% de AUC.

Kuzu, Balci e Salah (2016) abordaram o reconhecimento de autoria em um cenário de múltiplas participações em *chats* de jogo turco. As características utilizadas neste trabalho são muito parecidas com os outros trabalhos envolvendo *chats*, com a adição de características não textuais específicas da aplicação: número de sessões de *chat* e pontuações por usuário. A utilização da normalização dos textos piorou os resultados, dado que o erro de escrita pode ser um traço estilístico.

3.2.3 Atribuição de Autoria baseada em tópicos

A utilização de tópicos é tratada a parte, devido à proposição de variados modelos empregando este conceito, podendo ser utilizados em diversas bases, sejam elas tópico cruzado ou não. Seroussi, Zukerman e Bohnert (2011) utilizaram a abordagem de LDA para atribuição de autoria, com o intuito de construir modelos de tópicos a partir dos textos dos usuários. As distribuições de tópicos foram utilizadas como entrada de um algoritmo SVM e em conjunto com a distância LDA-H para achar o mais provável autor. Foram utilizadas duas variações das distribuições de tópicos: LDAH-M, construído a partir de todos os documentos de treinamento; LDAH-S, todos os documentos de um autor são concatenados em um documento simples e o modelo LDA aprende dos documentos de perfis. Em um cenário balanceado e com bastante autores, o modelo LDAH-S conseguiu superar modelos *baselines*.

Lakkaraju, Bhattacharya e Bhattacharyya (2012) implementaram o D-MRelCRP, um modelo baseado em influência de redes sociais, levando em conta que os usuários são influenciados por: fatores globais, como tópicos muito discutidos na internet; fatores geográficos, preferências pessoais do usuário; e pela rede de amigos. Além de assumir essas características, o modelo implementado é dinâmico, isto é, estas características não são estáticas, e sim evoluem com o tempo, tanto moldando como o usuário é influenciado nas redes sociais, tanto no fator que certos tópicos têm uma influência muito temporária.

Seroussi, Bohnert e Zukerman (2012) avançaram na proposição de um novo modelo de tópicos, que considera tanto os tópicos de cada documento quanto os tópicos dos autores. O modelo DADT parece similar ao modelo AT-FA, porém no modelo DADT os tópicos

de documentos e de autores são separados, há diferentes prioridades na distribuição das palavras, há aprendizado da razão entre palavras do documento e do autor, e, por último, DADT define o processo que irá gerar os autores. O modelo foi comparado com SVM e modelos de distribuição de tópicos, como: LDA-H, AT e AT-FA. DADT superou todos os outros modelos e no conjunto de dados PAN'11 obteve o terceiro melhor resultado. Como o modelo é totalmente supervisionado, em comparação com os modelos totalmente supervisionados do PAN'11, este obteve o melhor resultado.

Seroussi, Zukerman e Bohnert (2014) apresentam uma extensão de seus trabalhos anteriores (SEROUSSI; ZUKERMAN; BOHNERT, 2011; SEROUSSI; ZUKERMAN; BOHNERT, 2014), focando em mais experimentos em conjunto de dados de tamanhos variados. O trabalho além de lidar com bases de dados variadas, utiliza variações dos modelos empregados em Seroussi, Bohnert e Zukerman (2012), para cada modelo AT, AT-FA e DADT, é empregado a sua versão de atribuição probabilística e a versão de modelo de entrada para uma SVM. Além disso, o artigo reporta aplicações baseado em tópicos. O modelo DADT-P obteve os melhores resultados no geral, superando o *baseline* Token-SVM. Ao utilizar um conjunto de dados desbalanceado com milhares de autores, os modelos baseados em tópicos sofreram uma queda maior que o Token-SVM, justificado pelo SVM ajustar sempre uma instância por classe, enquanto os modelos baseados em tópicos utilizam a mesma instância para todas as classes.

No trabalho de Kim, Noh e Park (2015) foi abordado o problema de múltiplos ids de usuário, isto é, a posse de vários ids pertencerem a um único usuário. Neste trabalho foram utilizadas características léxicas, caracteres e tópicos da linguagem coreana. Além disso, foi utilizado um vetor de característica baseado em tempo, a fim de identificar a hora do dia em que o usuário posta. O modelo de tópicos utilizado foi o LDA, com 100 tópicos utilizados. Os resultados indicaram que o espaçamento de palavras (característica discriminante do coreano) foi a que obteve o melhor desempenho e a característica de tempo obteve o segundo melhor resultado.

3.3 Discussão

Como visto nos resultados da revisão sistemática, muitos trabalhos focaram no Twitter. Sabendo da limitação de caracteres imposta por esta rede social, não é possível

generalizar completamente as mesmas configurações em um cenário no qual não haja tal restrição, como é o caso do Reddit. Porém, alguns trabalhos serviram para ilustrar técnicas e métodos que não se prendem à natureza da limitação de caracteres. O trabalho de Pillay e Solorio (2010) foi o único que utilizou a ideia de algoritmos de agrupamento, como meta-característica, para gerar somente um grupo (*cluster*) por usuário. Outros trabalhos lidaram com a problemática de dados *streaming* (SEKER; AL-NAAMI; KHAN, 2013), podendo ser facilmente utilizados para o caso de *Big Data* em que o conjunto de dados é muito grande para ser carregado de uma só vez, ou então em casos de modelos baseado no tempo (AZARBONYAD *et al.*, 2015), em que o estilo de escrita do autor tende a mudar significativamente.

Em problemáticas cujo conteúdo do texto é sensível, modelos que substituem palavras por um rótulo POS ou *, e pontuações por "PUNCT" podem ser bons discriminantes de autores (LITVINOVA; LITVINOVA; PANICHEVA, 2019). Da mesma forma, a utilização de modelos de compressão não trata o conteúdo do texto em si e pode ser utilizada facilmente para diversos idiomas (HALVANI; WINTER; GRANER, 2017).

A área de Atribuição de Autoria se consolidou fortemente em torno de textos curtos, como visto em importante trabalhos utilizando o Twitter. Porém, muitos tipos de modelos se mostraram bem consolidados em casos genéricos. A utilização de tópicos como características se mostrou bem robusta, em especial os modelos baseado em DADT (SEROUSSI; BOHNERT; ZUKERMAN, 2012; SEROUSSI; ZUKERMAN; BOHNERT, 2014). Os trabalhos que envolveram diretamente a classificação de um autor ou seu ranqueamento probabilístico, fizeram o uso de técnicas muito similares como é o caso das variações do classificador SVM e Similaridade Cosseno ou CBC, respectivamente.

Mesmo com a diversidade dos trabalhos analisados, nenhum trabalho em si propôs uma modelagem de *Big Data* levando em conta aspectos temporais na escrita do autor, o trabalho mais próximo utilizou apenas 4-grama de caracteres em um conjunto de dados não tão grande, cujos dados provenientes do Twitter possuem limitação de caracteres (AZARBONYAD *et al.*, 2015). Além disso, nenhum trabalho abordou o fator emocional ou afetivo na escrita do autor. Roffo *et al.* (2014) abordam o tema de traços de personalidade e correlação com as características, porém os resultados foram obtidos após a aplicação de um questionário, o que no caso de autores anônimos de rede social é inviável.

Contudo, uma grande parte dos trabalhos analisados mostraram técnicas já consolidadas que serão utilizados no trabalho como é o caso dos n-gramas balanceados com

o TF-IDF, a utilização dos n-gramas de caracteres, palavras e de classes gramaticais e, por fim, o fator temporal em que (SEKER; AL-NAAMI; KHAN, 2013) mostraram a importância do tamanho do *batch* utilizado no treinamento sendo inversamente proporcional à taxa de erro do classificador e que o estilo de escrita do autor tende a mudar drasticamente ao comparar os anos (AZARBONYAD *et al.*, 2015).

Por fim, nenhum dos trabalhos analisados utilizou a rede social Reddit diretamente, que vem crescendo com os anos e tem o fator discriminante da anonimidade dos usuários, podendo ser este um fator em que os usuários se sintam muito mais confortáveis em se expressar e possa facilitar na detecção de estilo de escrita de cada autor.

4 Materiais e Métodos

O conjunto de dados utilizado no trabalho foi extraído do projeto *Open Source* Pushshift. Este projeto tem como intenção disponibilizar dados do *Reddit* para pesquisadores e instituições acadêmicas, seguindo as diretrizes de privacidade[1] do Reddit, em que dados pessoais não serão compartilhados, porém as informações referentes à postagem do usuário e tempo serão compartilhadas.

Este capítulo está organizando em duas seções. Na primeira (seção 4.1) o conjunto de dados utilizado é apresentado, incluindo algumas estatísticas descritivas. Já a seção 4.2 detalha as etapas de pré-processamento, modelagem dos dados, técnica de treinamento e pós processamento dos dados.

4.1 Conjunto de dados

O conjunto de dados utilizado neste trabalho é o mesmo utilizado por Casimiro e Digiampietri (2020), porém compreende os comentários feitos nos anos de 2009 a 2015 no *Reddit*. A extração dos dados no presente trabalho difere do trabalho anterior, sendo que neste foi utilizada a interface do Pushshift (BAUMGARTNER, 2019). Após a extração, os dados passaram por filtros a fim de remover instâncias com campos de comentário ou usuário marcados como removidos. Outros campos irrelevantes à análise do trabalho foram descartados, por exemplo: *edited*, *author_flair_text* e *author_flair_css_class*. Os únicos campos dos comentários relevantes para o trabalho são: *author*, *body*, *subreddit* e *created_utc*, em que o campo *author* contém o nome de usuário do autor do comentário, *body* o texto da postagem em formato *Markdown*, *subreddit* o nome da *subreddit* em que a postagem foi feita e *created_utc* o *timestamp* do momento em que a postagem foi feita. Por fim, as postagens que foram feitas em outra língua que não fosse a inglesa, foram também descartados.

Após a filtragem básica das instâncias, foi realizada uma nova filtragem na base dados a fim de selecionar somente os comentários dos N autores que mais comentaram, sendo $N <= 100$. Com isso, a base final consiste dos comentários feitos pelos N autores que mais comentaram no Reddit entre os anos de 2009 e 2015. Os conjuntos de dados

[1] https://www.redditinc.com/policies/privacy-policy-may-25-2018

resultantes consistem em 619.0 GB e 850.4 GB de dados para as bases de 10 e 100 autores que mais comentaram no Reddit, respectivamente.

4.2 Técnicas utilizadas

Após a etapa de extração dos dados, foi realizada a etapa de pré-processamento de dados. Como explicado na seção anterior, a primeira etapa do pré-processamento dos dados consiste em remover instâncias de comentários cujos campos *body* ou *author* estão marcados como removidos, através da *string* "*removed*". Após este filtro, um segundo filtro foi aplicado de forma a permitir apenas comentários na língua inglesa. Como o comentário escrito no Reddit é registrado como um texto *markdown*, então a próxima etapa do pré-processamento foi a conversão do texto de *markdown* para HTML, em seguida a remoção dos rótulos HTML foi realizada.

Com a filtragem dos dados de acordo com o escopo do trabalho, os comentários passaram pelas modelagens descritas no capítulo 2, a fim de serem apresentados de maneira correta ao classificador. Conforme os resultados preliminares, vistos em Casimiro e Digiampietri (2020), o n-grama de caracteres dentro de palavras se mostrou menos eficiente do que n-grama de caracteres. Com isso, neste trabalho foram utilizados n-gramas, normalizados pelo TF-IDF, de palavras, caracteres e *POS tagging*. Para palavras foi considerado $1 \leq n \leq 3$; foi considerado $3 \leq n \leq 4$, para caracteres; por fim, para *POS tagging*, foi considerado $3 \leq n \leq 4$ e $n = 3\text{-}5$.

O classificador utilizado foi o LSTM, porque ele aborda tanto a questão do grande volume de dados (*Big Data*) quanto a questão temporal do trabalho, em que os dados entre os anos de 2009 e 2015 são inseridos de maneira cronológica, como entrada em lotes de tamanho fixado, para o treinamento, apresentando sempre dados do passado do ponto de vista temporal e testando com dados do futuro, e no decorrer do treinamento essa rede vai esquecendo os dados mais antigos através da unidade do portão de esquecimento das células da LSTM. Além disso, durante a etapa de treinamento, foi utilizada como validação a técnica de validação cruzada, para $K = 10$.

Durante o pós-processamento dos dados, os resultados de cada cenário foram avaliados de acordo com a acurácia, precisão, revocação e medida F1.

A construção do modelo LSTM foi feita utilizando a API do Keras (CHOLLET *et al.*, 2015) em conjunto com o Tensorflow (ABADI *et al.*, 2015). Além disso, utilizou-se o *framework* do *scikit-learn* (BUITINCK *et al.*, 2013) para codificação dos autores utilizando-se o método de *one hot encoder*, divisão dos conjuntos de dados de testes e treinamento, vetorização em TF-IDF para extração de características textuais como n-gramas de palavras e caracteres. Para a manipulação das características textuais, utilizou-se a API do NLTK (BIRD; KLEIN; LOPER, 2009) para tokenização de palavras, remoção de *stopwords* e conversão para rótulos POS. Para manipulações tabulares envolvendo CSV com os dados dos comentários em si, utilizou-se a biblioteca do Pandas (MCKINNEY, 2010) em conjunto com o Numpy (HARRIS *et al.*, 2020). Por fim, para a construção dos gráficos de resultados foi utilizada a biblioteca Matplotlib (HUNTER, 2007).

5 Solução Proposta

O presente capítulo tem por objetivo descrever a solução proposta, apresentando e discutindo os resultados alcançados. Com este objetivo, o capítulo foi dividido em duas seções: 5.1 e 5.2, seções em que foram tratados, respectivamente, uma estudo de caso, explorando diferentes abordagens de características, métricas e classificadores, e os resultados finais com a solução proposta por este trabalho.

5.1 Estudo de Caso - Análise Preliminar

Como forma de analisar o contexto da rede social Reddit, foi realizado um estudo de uma *subreddit Science* e os comentários dos dez autores que mais comentaram nesta *subreddit* entre os anos de 2007 e 2009 (CASIMIRO; DIGIAMPIETRI, 2020). Os autores que mais comentaram nesta subreddit, respectivamente, desde o primeiro até o décimo: *IConrad*, *redditcensoredme*, *MarshallBanana*, *matts2*, *NoMoreNicksLeft*, *otakucode*, *mutatron*, *ferdinand*, *mexicodoug* e *judgej2*. Neste estudo foram avaliados diversos cenários para três contextos diferentes: classificação binária entre os dois autores que mais comentaram, classificação binária entre o autor que mais comentou contra todos e classificação multiclasse entre os dez autores. Os cenários de classificação binária foram avaliados utilizando curva ROC e para classificação multiclasse foram utilizas precisão, revocação e medida F1.

As representações utilizadas variaram os valores de n-gramas, para $N = 1, 2, 3, 4, 5$ e 1-5, com diferentes tipos de extração de característica: caracteres, caracteres de palavras, *POS tagging* e palavras. Como textos em rede sociais são mais informais que textos jornalísticos, por exemplo, é muito comum conter erros gramaticais, gírias, abreviações, etc. Com isso, utilizar extração de palavras pode não ser tão eficiente quanto extrações de caracteres ou caracteres de palavras. Porém, tratando-se de uma *subreddit* de Ciências, a extração de caracteres pode representar muito bem o estilo de escrita do autor.

Para explorar melhor cada cenário resultante, foram utilizados três classificadores: SGD, *Perceptron*, *Passive-Aggressive*, todas estas implementações não focaram em nenhum tipo de *code tuning* e utilizaram parametrizações padrões.

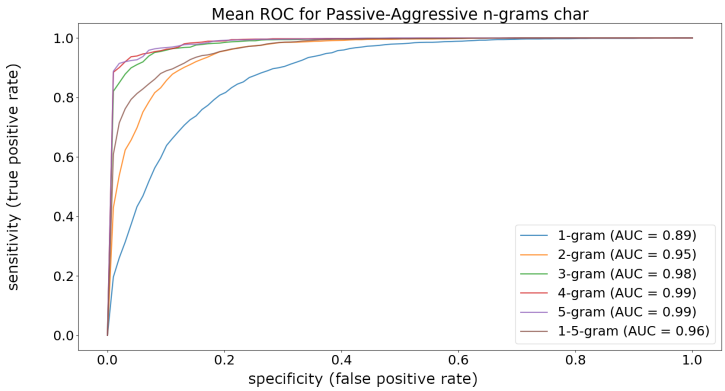
5.1.1 Análise Classificação Binária

A medida ROC tem como objetivo avaliar o balanceamento entre os falsos e verdadeiros positivos, com isso ao analisar as figuras da curva ROC pode-se avaliar se os classificadores sofreram qualquer tipo de viés, isto é, um número desproporcional entre falsos positivos e verdadeiros positivos. Para a classificação binária, são analisados com mais detalhe a classificação entre os dois autores e a classificação entre o autor que mais comentou contra todos, destacando o desempenho de cada uma das quatro características mencionadas anteriormente.

Classificação entre os dois autores que mais comentaram

Utilizando a representação de caracteres, as figuras 6, 7 e 8 mostraram que para todos os classificadores o comportamento foi padrão em que o aumento no valor de N dos n-gramas aumentou a acurácia dos classificadores. Este comportamento pode ser explicado pelo fato que com um valor maior de N há mais detalhes da escrita do autor. A figura 8 mostrou que houve um desbalanceamento maior no resultado de unigrama para o classificador Perceptron, em que foi constatado um resultado de 89% de verdadeiros negativos e apenas 62% de verdadeiros positivos.

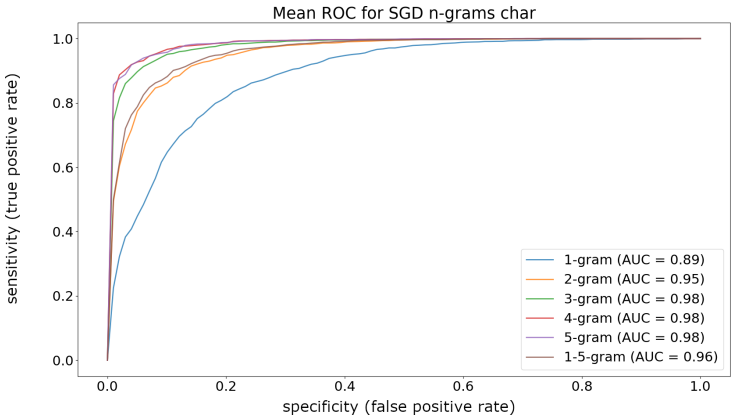
Figura 6 – Curva ROC do classificador *Passive-Aggressive* para dois autores utilizando caracteres



Fonte: Casimiro e Digiampietri (2020)

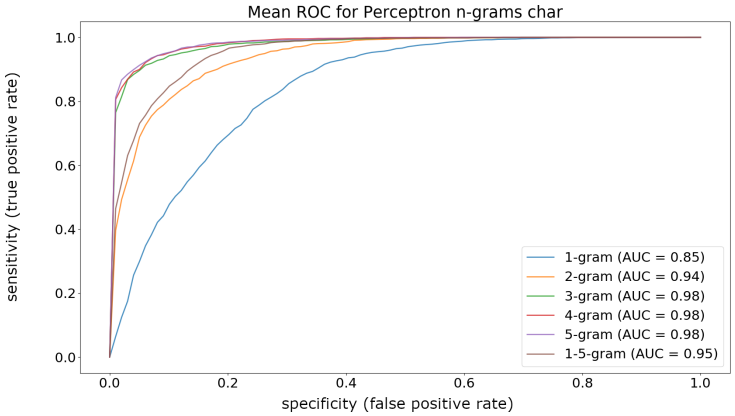
No caso de caracteres apenas de palavras, os resultados foram similares aos de caracteres no geral, porém um pouco inferiores, por exemplo, para N =1-5, o classificador *Passive-Aggressive* teve uma perda de 5% de acurácia. A perda de efetividade pode ser explicada porque os comentários do Reddit podem conter tabulações, pontuações, adição

Figura 7 – Curva ROC do classificador *SGD* para dois autores utilizando caracteres



Fonte: Casimiro e Digiampietri (2020)

Figura 8 – Curva ROC do classificador *Perceptron* para dois autores utilizando caracteres



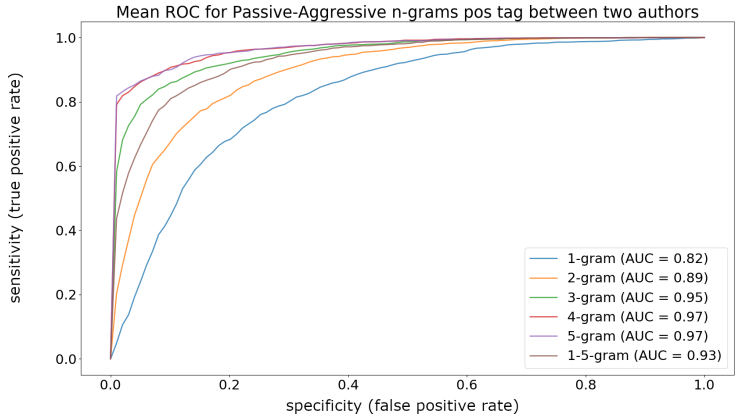
Fonte: Casimiro e Digiampietri (2020)

de negrito e itálico às palavras, listas desordenadas e ordenadas, etc. Essa característica peculiar do Reddit permite com que haja muito mais extração de características em caracteres gerais do que somente caracteres de palavras.

Utilizando a representação das postagens de acordo com *POS Tagging*, o classificador *Passive-Aggressive* mostrou bons resultados para a curva ROC, especialmente para $n = 4$ e $n = 5$, conforme pode ser observado na figura 9.

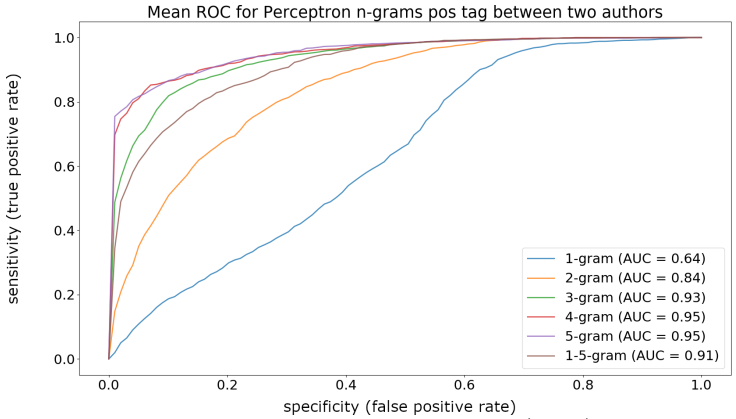
O classificador *Perceptron* mostrou limitações na distinção entre verdadeiros positivos e negativos com unigramas, a curva da figura 10 mostra esse desbalanceamento. Esta diferença é resultado da classificação utilizando 1-grama, que atingiu 49% de verdadeiros positivos e 68% de verdadeiros negativos. Além dos unigramas, para n = 1-5 o classificador também apresentou uma grande diferença de classificação entre verdadeiros positivos e negativos. A classificação resultou em, respectivamente, 74% e 92% de verdadeiros positivos e negativos, fazendo com que a curva 1-5-grama não ficasse balanceada.

Figura 9 – Curva ROC do classificador *Passive-Aggressive* para dois autores utilizando classes gramaticais



Fonte: Casimiro e Digiampietri (2020)

Figura 10 – Curva ROC do classificador *Perceptron* para dois autores utilizando classes gramaticais

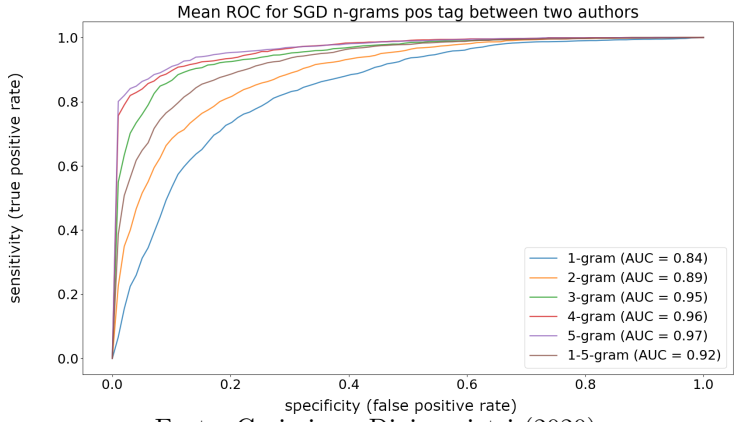


Fonte: Casimiro e Digiampietri (2020)

Com o classificador *SGD*, houve apenas uma desproporção maior entre resultados positivos para $n = 1$ como mostra a figura 11. Essa desproporção é evidenciada nas taxas de 67% de verdadeiros negativos e 81% de verdadeiros positivos.

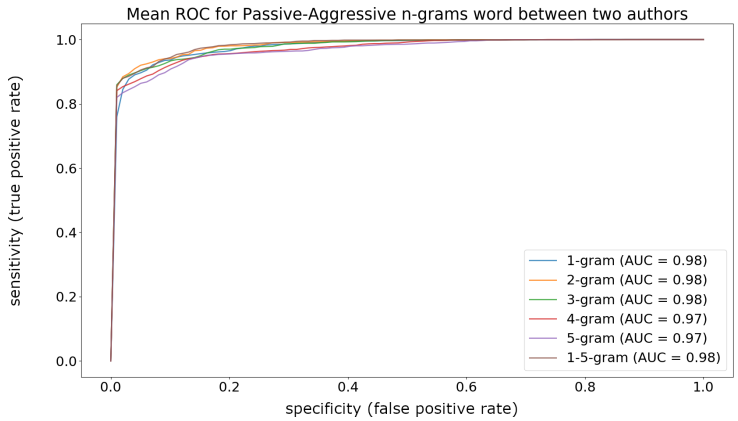
Por fim, a representação textual por meio de n-gramas de palavras obteve excelentes resultados. As figuras 12, 13 e 14 mostraram resultados de área sob a curva ROC maiores que 94% em todos os cenários. Além disso, observa-se nos três gráficos que os melhores resultados vão de unigrama até trigrama e 1-5-grama, tendo uma pequena piora ao se utilizar tetragrama e pentagrama de palavras. Este comportamento vai ao encontro aos resultados de outros estudos da área de Atribuição de Autoria, em que os melhores resultados de autoria se concentram na utilização de bigramas e trigramas, mostrando que valores maiores de N trazem excesso de especificidade dos n-gramas.

Figura 11 – Curva ROC do classificador *SGD* para dois autores utilizando classes gramaticais



Fonte: Casimiro e Digiampietri (2020)

Figura 12 – Curva ROC do classificador *Passive-Aggressive* para dois autores utilizando palavras



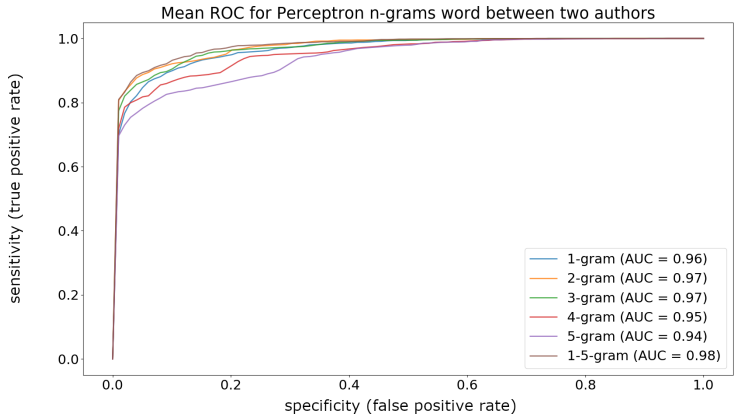
Fonte: Casimiro e Digiampietri (2020)

Classificação entre o autor que mais comentou contra todos

O segundo contexto será a classificação binária entre *IConrad*, autor que mais comentou, e todos os outros nove autores que mais comentaram. Esse contexto é mais desafiador, já que há mais chances de um dos nove autores terem certos estilos de escrita parecidos com o autor que mais comentou.

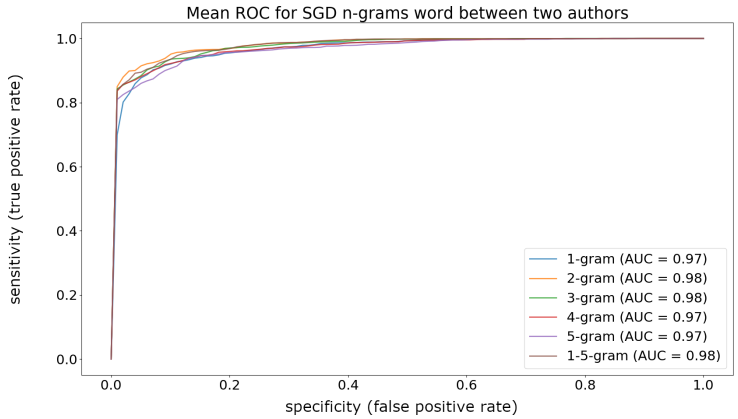
O classificador *Passive-Aggressive*, no cenário de extração de caracteres, para $n = 1$ obteve resultados bem inferiores em comparação com os outros valores de n, como mostra a figura 15. Isto deve-se ao fato que utilizando apenas unigrama de caracteres e, em um contexto de um contra todos, os traços estilísticos gerados são muito gerais, com isso há maior probabilidade de confusão entre o autor em questão e qualquer outro autor da *subreddit*. Para $n > 1$ os resultados foram satisfatórios, obtendo área sob a curva de mais

Figura 13 – Curva ROC do classificador *Perceptron* para dois autores utilizando palavras



Fonte: Casimiro e Digiampietri (2020)

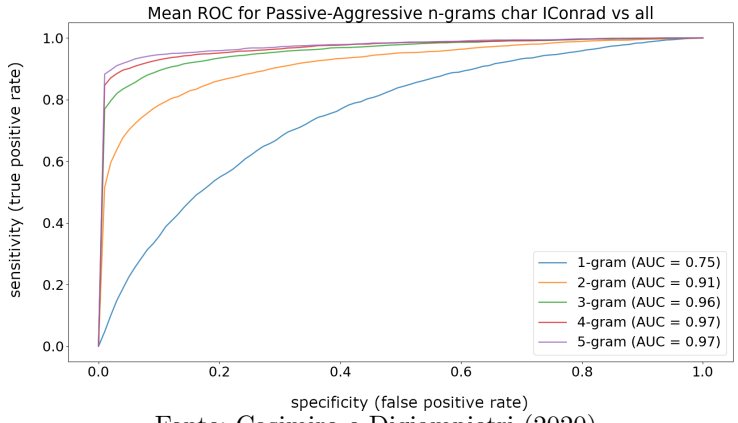
Figura 14 – Curva ROC do classificador *SGD* para dois autores utilizando palavras



Fonte: Casimiro e Digiampietri (2020)

de 90% em todos os casos. Ao se considerar n-gramas de caracteres de palavras, assim como no primeiro contexto, os resultados foram semelhantes, porém um pouco inferiores.

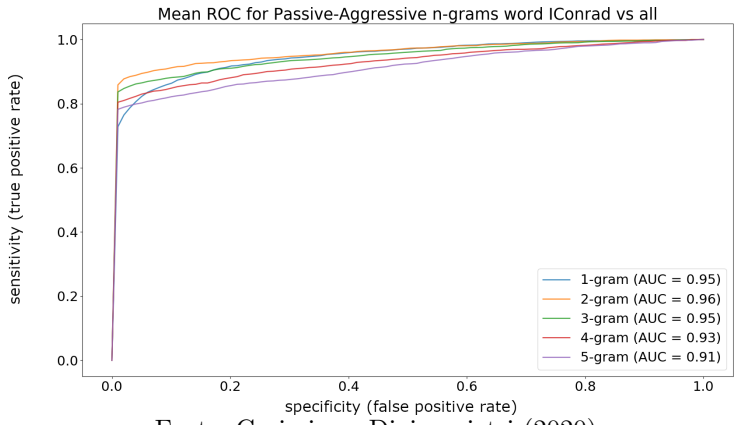
Figura 15 – Curva ROC do classificador *Passive-Aggressive* para o autor *IConrad* contra todos utilizando caracteres



Fonte: Casimiro e Digiampietri (2020)

No cenário de n-gramas de palavras, a figura 16 mostra que este tipo de característica é eficiente na identificação do autor no cenário analisado, com resultado de área sob a curva de no mínimo 90% em todos os casos. Além disso, observa-se uma perda de desempenho para $n > 2$.

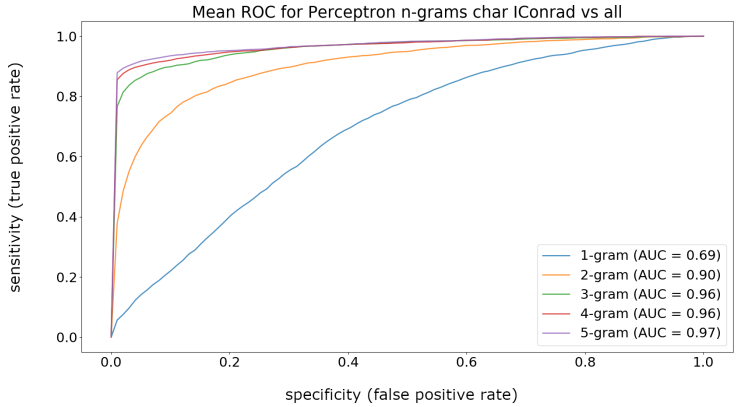
Figura 16 – Curva ROC do classificador *Passive-Aggressive* para o autor *IConrad* contra todos utilizando palavras



Fonte: Casimiro e Digiampietri (2020)

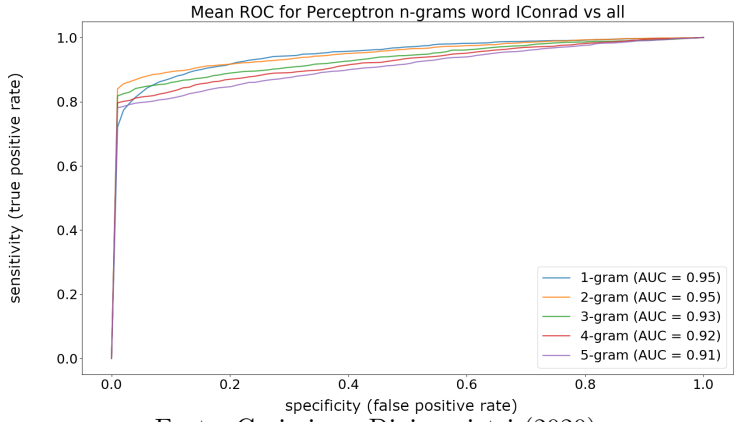
O classificador *Perceptron* também obteve resultados satisfatórios nos cenários de caracteres e palavras, como observa-se nas figuras 17 e 18, respectivamente. Os resultados obtidos por este classificador se assemelham aos obtidos pelo *Passive-Aggressive*, exceto nos cenários de extração de caracteres e caracteres de palavras utilizando unigrama em que obteve 69% e 66% de área sob a curva, mostrando uma maior dificuldade de distinguir entre o autor e o resto da *subreddit*.

Figura 17 – Curva ROC do classificador *Perceptron* para o autor *IConrad* contra todos utilizando caracteres



Fonte: Casimiro e Digiampietri (2020)

Figura 18 – Curva ROC do classificador *Perceptron* para o autor *IConrad* contra todos utilizando palavras



Fonte: Casimiro e Digiampietri (2020)

5.1.2 Análise Classificação Multiclasse

Ao analisar os resultados das acurácias dos cenários da classificação multiclasse, observou-se que os resultados obtidos da extração de caracteres, caracteres de palavras e *POS Tagging* são muito semelhantes. Com isso, o cenário de n-gramas de palavras será visto mais detalhadamente, enquanto que n-gramas de caracteres será discutido mais brevemente.

No cenário de n-gramas de caracteres, os classificadores *Passive Aggressive*, *Perceptron* e *SGD* corroboraram a tendência já observada na medida de acurácia: aumentando-se o valor de n, o modelo se torna mais eficiente. Além disso, outro resultado obtido foi que com o aumento do valor de n, mais equilibrados são os resultados para cada autor, Por exemplo, para $n = 1$ o melhor modelo foi do autor *redditcensoredme* com valor da medida F1 de $0,38$, o pior resultado foi do autor *mexicodoug* com valor da medida F1 de $0,14$, enquanto que para $n = 5$ os melhores resultados atingiram $0,91$ de medida F1, o pior resultado foi para o autor *mexicodoug* que obteve $0,86$ de medida F1.

A diferença entre o melhor e o pior resultado para cada n-grama reforça a ideia de que com o aumento do n-grama o classificador consegue captar mais marcas de estilo de cada autor, conseguindo assim criar vários perfis únicos, com isso não só há o aumento da medida F1 e a diminuição da diferença entre o melhor e pior resultado, mas como, também, os valores da precisão e revocação de cada modelo ficam mais próximos com o aumento do valor de n, evidenciando a consistência de cada modelo construído.

Os classificadores apresentaram resultados muito próximos para precisão e revocação, porém para a medida F1, que avalia a relação entre a precisão e a revocação, o classificador *SGD* mostrou valores inferiores que os outros dois classificadores, mesmo com o aumento do valor de n, apesar dos resultados serem bons. Isso indica que o classificador *SGD*, em função da taxa de aprendizado e função de penalização, é mais sensível que os outros dois classificadores.

Como mostrado nas tabelas 4, 5 e 6, os classificadores desempenharam resultados satisfatórios para todos n-gramas.

No cenário de n-gramas de caracteres, ao aumentar o valor de n, mais eficiente e mais consistente o modelo do autor se torna, porém no caso de n-gramas de palavras, valores intermediários de n, como $n = 2$ e $n = 3$, atingem os melhores resultados. Neste caso os valores de $n = 1$ e $n = 5$ apresentam os piores resultados do cenário. Além disso, a combinação dos n-gramas (no caso, n =1-5) produziu uma pequena melhora nos resultados, evidenciando a ideia de que para n-gramas de palavras colocar valores variáveis não acrescenta muitos ruídos, como no caso de n-gramas de caracteres.

Tabela 4 – Precisão, revocação e medida F1 para o classificador *Passive-Aggressive* para os n-gramas de palavras

	1			2			3			4			5			1-5		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
IConrad	71	84	77	80	91	85	67	91	77	63	88	74	57	88	69	82	92	87
redditcensoredme	71	90	79	74	96	84	66	96	78	61	96	75	60	95	73	79	96	87
ferdinand	83	81	82	91	88	90	93	84	88	88	83	85	88	81	84	92	89	91
NoMoreNicksLeft	81	75	78	92	87	90	90	83	86	88	81	84	89	79	84	93	88	90
judgej2	80	69	74	93	83	88	97	79	87	95	78	86	97	76	85	95	83	89
otakucode	81	81	81	94	87	90	89	85	87	90	80	85	92	79	85	94	89	91
mutatron	81	68	74	92	82	87	96	79	87	98	76	86	98	75	85	92	84	88
MarshallBanana	72	76	74	85	88	87	85	86	85	81	85	83	80	83	81	85	90	87
mexicodoug	80	75	77	89	83	86	96	79	87	98	76	85	98	74	85	91	85	88
matts2	83	80	81	91	89	90	93	86	89	95	81	88	93	79	86	90	91	91

Fonte: Casimiro e Digiampietri (2020)

Como mostrado no cenário anterior dos caracteres, a tabela 4 reforça a ideia de que o classificador *Passive Aggressive* obtém os melhores resultados. A tabela 5 apresenta resultados semelhantes, porém um pouco inferiores, do classificador *Perceptron*. Por fim, a tabela 6 contém os resultados do classificador *SGD*, que neste caso são inferiores.

Estas comparações permitem elucidar a ideia que o classificador *SGD*, apesar de sua robustez, com inúmeros hiperparâmetros, dentre eles a taxa de aprendizado e função de penalização, acarreta em um classificador mais sensível, facilitando a geração de *overfit*.

Tabela 5 – Precisão, revocação e medida F1 para o classificador *Perceptron* para os n-gramas de palavras

	1			2			3			4			5			1-5		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
IConrad	68	82	75	81	89	85	77	87	82	67	85	75	63	85	72	78	90	84
redditcensoredme	73	85	79	85	91	88	81	88	84	77	86	81	72	84	77	83	92	87
ferdinand	79	75	77	88	86	87	89	83	86	86	80	83	87	79	83	89	86	88
NoMoreNicksLeft	81	70	75	89	84	87	88	83	86	87	79	83	84	78	81	89	85	87
judgej2	75	71	73	88	82	85	87	81	84	83	79	81	86	77	81	92	82	86
otakucode	81	73	77	90	85	87	85	83	84	89	80	84	89	78	83	91	85	88
mutatron	69	69	69	86	83	84	79	81	80	87	77	82	83	76	80	83	84	84
MarshallBanana	67	75	71	80	87	83	79	85	82	77	84	80	76	82	79	83	86	84
mexicodoug	79	74	76	81	84	82	82	80	81	86	77	81	85	75	80	86	83	85
matts2	81	75	78	90	87	89	92	85	88	81	82	81	81	80	81	91	88	90

Fonte: Casimiro e Digiampietri (2020)

Tabela 6 – Precisão, revocação e medida F1 para o classificador *SGD* para os n-gramas de palavras

	1			2			3			4			5			1-5		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
IConrad	57	78	66	73	89	80	75	89	81	70	87	78	66	85	74	73	90	81
redditcensoredme	64	87	74	77	94	84	75	93	83	72	91	80	69	87	77	77	95	85
ferdinand	74	74	74	86	86	86	91	85	88	88	82	85	84	81	83	88	87	88
NoMoreNicksLeft	75	62	68	90	83	86	88	84	86	86	82	84	88	78	83	89	83	86
judgej2	74	59	66	91	79	85	90	81	85	89	79	84	88	77	82	93	78	85
otakucode	74	72	73	90	84	87	91	84	87	87	82	84	85	79	82	92	85	88
mutatron	72	54	62	89	79	84	87	80	83	85	80	82	90	76	83	90	79	84
MarshallBanana	63	63	63	82	84	83	83	86	84	82	83	83	76	82	79	82	85	84
mexicodoug	70	64	67	89	80	84	88	81	85	90	77	83	88	76	81	89	82	85
matts2	72	72	72	88	87	87	88	88	88	86	83	85	82	81	82	88	88	88

Fonte: Casimiro e Digiampietri (2020)

Por outro lado, como os outros dois classificadores, *Passive-Aggressive* e *Perceptron*, não necessitam de uma taxa de aprendizado e função de penalização, assim, problemas de *overfit* são mais difíceis de acontecerem (no sentido que não é necessário uma configuração específica de parâmetros para se evitar problemas de *overfit*), com isso estes classificadores apresentaram os melhores resultados, sendo o *Passive-Aggressive* o melhor classificador no geral. Apesar dos dois apresentarem resultados muito parecidos, é provável que a diferença consiste no tratamento de *outliers* do classificador *Passive Aggressive*, configurável por meio da constante C de regularização do tamanho máximo, com isso este classificador limita a influência dos *outliers*.

5.1.3 Análise Geral

O estudo de caso analisado focou em investigar e tentar responder às três seguintes questões: a atribuição de autoria consegue ser validada no contexto da rede social Reddit?; há uma influência significativa nos resultados ao se utilizar diferentes formas de extração de n-gramas, como caracteres, caracteres de palavras, palavras ou classes gramaticais?; utilizar três classificadores semelhantes, diferindo apenas nas questões dos parâmetros a serem configurados (que podem ser ajustados para tratar questões específicas da classificação, como evitar *overfit* e *outliers*) influenciará de maneira significativa nos resultados?

Para essas perguntas o Reddit possui um estilo de rede social muito diferente e pouco explorado, no âmbito da Atribuição de Autoria, ao se comparar com o Twitter. Duas das maiores diferenças entre essas redes sociais são o estilo de edição de comentário, podendo utilizar as peculiaridades do texto *markdown* com negrito, itálico, cabeçalhos, listas, etc, e sem limitação explicita da quantidade de caracteres.

Este estudo permitiu que para certos tipos de características no contexto do Reddit, algumas parametrizações do classificador utilizado impactarão de maneira positiva ou negativa na classificação. Além disso, o cenário de n-gramas de palavras mostrou-se bastante eficiente ao utilizar bigramas, trigramas e até 1-5-gramas, enquanto que unigrama, tetragrama e pentagrama mostraram os piores resultados. Por fim, n-grama de caracteres alcançou resultados satisfatórios, permitindo aferir que o valor de N ideal em que se alcança o melhor resultado possível pode ser um pouco maior que 5, já que o aumento do valor de N melhorou a eficiência dos classificadores.

5.2 Resultados Finais

Os resultados foram organizados em doze cenários, combinando cada alvo, os 10 e 100 autores que mais comentaram, com cada característica: caractere, rótulos POS e palavras. Os resultados das classificações mostram o aumento da acurácia por época para cada cenário, alguns resultados estabilizaram mais cedo do que outros, pois uma condição de parada antecipada foi alcançada, usando três épocas de paciência. Além da acurácia para todas as épocas, para cada cenário foram extraídas as métricas de precisão, revocação e medida F1 utilizando a média ponderada. Dado que os dados para 10 e 100 autores são

desbalanceados, para manter a proporção dos dados mais realística, foi utilizada a média ponderada ao invés da média macro.

Tabela 7 – Medidas de precisão, revocação e medida F1 para todos os cenários utilizando a média ponderada

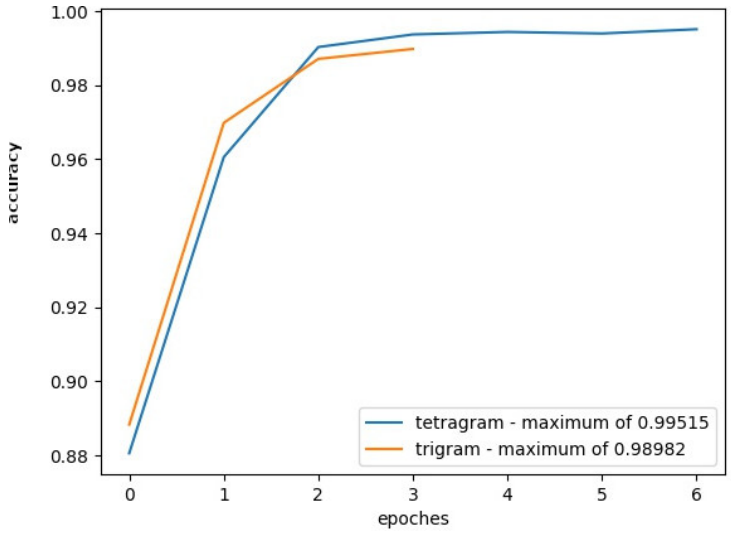
Alvos	Características	n-grama	Precisão	Revocação	Medida F1
10 autores	palavras	1-grama	98,51%	98,52%	98,51%
10 autores	palavras	3-gramas	98,67%	98,67%	98,67%
10 autores	palavras	2-gramas	98,93%	98,94%	98,93%
10 autores	caracteres	4-gramas	99,6%	99,6%	99,6%
10 autores	caracteres	3-gramas	99,15%	99,16%	99,16%
10 autores	rótulos POS	3-gramas	97,78%	97,8%	97,76%
10 autores	rótulos POS	4-gramas	97,39%	97,43%	97,37%
10 autores	rótulos POS	{3,5}-gramas	98,5%	98,51%	98,48%
100 autores	palavras	2-gramas	68,77%	72,8%	69,99%
100 autores	palavras	3-gramas	68,99%	72,59%	69,94%
100 autores	palavras	1-grama	69,14%	72,88%	70,21%
100 autores	rótulos POS	3-gramas	65,97%	71,11%	67,57%

Fonte: Casimiro e Digiampietri (2022)

Utilizando caracteres como características, apenas dois cenários foram produzidos para 10 autores. Isso foi planejado, pois para alcançar bons resultados de classificações a revisão de literatura e estudos prévios indicaram que o valor de N precisa ser maior ou igual a 3, com tetragrama atingindo resultados melhores em comparação aos produzidos com trigrama e pentagrama alcançando resultados melhores do que com tetragrama, por exemplo. Porém, para extrair valores maiores de N, em um grande conjunto de dados, mais poder computacional é necessário, e isso foi além do escopo do trabalho atual. O mesmo se aplica a não utilização dos 100 autores como alvos da classificação para caracteres como característica.

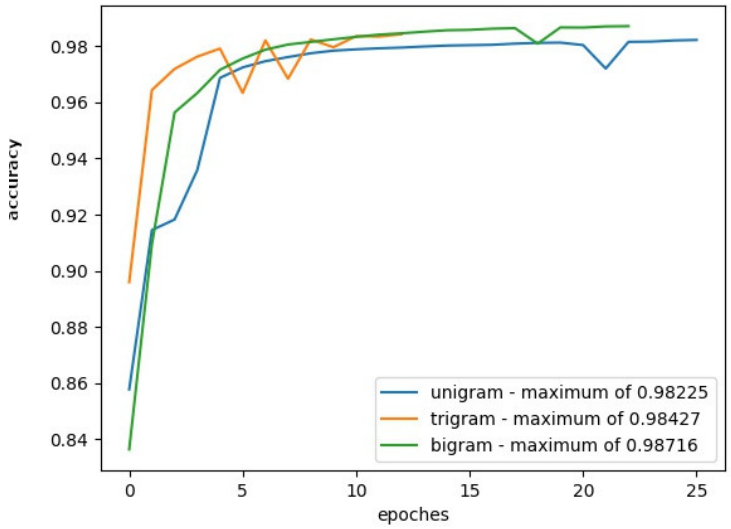
Os resultados para caracteres na figura 19 mostram um bom desempenho para 10 autores. Para trigrama, a acurácia alcançou quase 99% e para a precisão o resultado é um pouco maior que 99,5%. Estes resultados mostraram que caracteres como características não são apenas boas características em redes sociais com limitação de caracteres como Twitter, mas também são uma ótima estratégia para redes sociais sem limitação de caracteres, como o Reddit. Além de atingir ótimos resultados de acurácia, as medidas de precisão, revocação e medida F1 também mostraram ótimos resultados acima de 99%, como mostra a tabela 7.

Figura 19 – Acurácia média entre 10 autores utilizando caracteres como característica



Fonte: Casimiro e Digiampietri (2022)

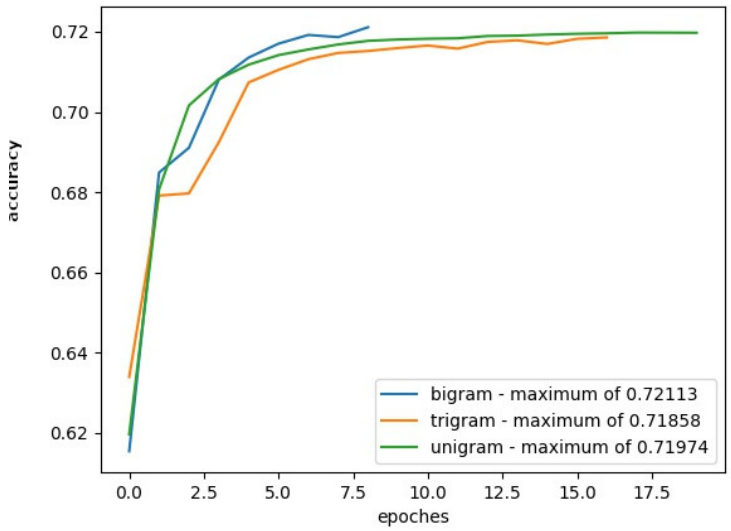
Figura 20 – Acurácia média entre 10 autores utilizando palavras como características



Fonte: Casimiro e Digiampietri (2022)

Na revisão da literatura, foi indicado que a característica mais estável e que atinge os melhores resultados são os n-gramas de caracteres. De fato, para redes sociais com limitação de caracteres, o uso de características baseadas em palavras dos comentários pode não alcançar bons resultados, pois os autores têm menos caracteres para expor o que eles desejam compartilhar e isso pode levar a um encurtamento das palavras. Dado isto, os autores podem não usar as palavras (ou na combinação delas) que eles originalmente pensaram, isto pode afetar o desempenho da atribuição de autorias com base nas palavras (ou a combinação delas).

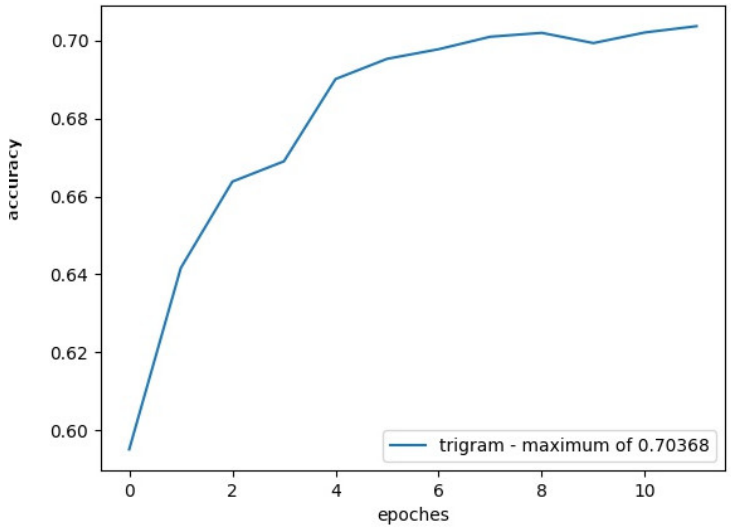
Figura 21 – Acurácia média entre 100 autores utilizando palavras como características



Fonte: Casimiro e Digiampietri (2022)

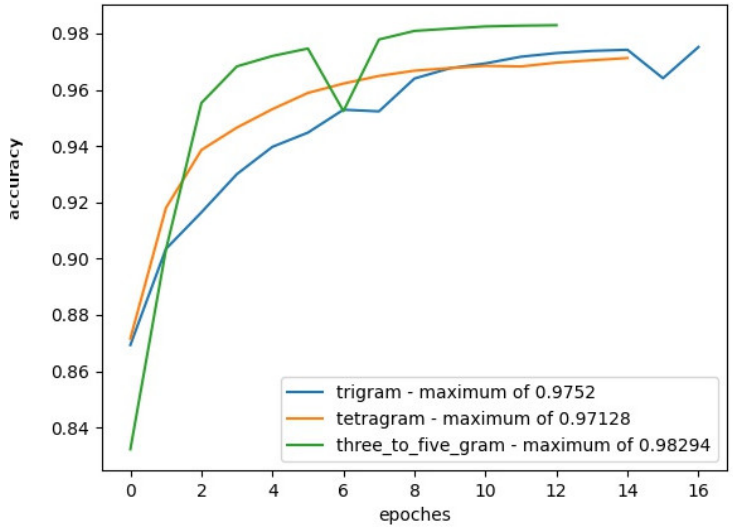
Em uma rede social sem a limitação dos caracteres, os resultados indicaram que as palavras como característica atingem resultados muito bons, com uma acurácia geral de mais de 98% para 10 autores e mais de 71% para 100 autores como mostrado, respectivamente, nas figuras 20 e 21. A tabela 7 mostra que todos os n-gramas de palavras apresentaram bons resultados para ambos os alvos, 10 e 100 autores. Para os resultados do cenário de 10 autores, o bigrama de palavras teve os melhores resultados de precisão, revocação e medida F1, porém para 100 autores, o unigrama de palavras apresentou os melhores resultados para precisão, revocação e medida F1.

Figura 22 – Acurácia media entre 100 autores utilizando rótulos POS como característica



Fonte: Casimiro e Digiampietri (2022)

Figura 23 – Acurácia média entre 10 autores utilizando rótulos POS como característica



Fonte: Casimiro e Digiampietri (2022)

Na revisão da literatura, nenhum dos trabalhos usou rótulos POS como característica, isso pode estar relacionado ao fato que para redes sociais com limitação de caracteres, o uso correto de palavras sem o seu encurtamento ou erro de digitação e pontuação é mais difícil, e essa característica pode se tornar pouco útil. Por outro lado, essa característica em redes sociais sem limitação de caracteres pode se tornar útil, porque com TF-IDF de n-grama de rótulos POS é possível identificar algumas construções gramaticais intrínsecas ao autor. Os resultados da acurácia para 10 autores (figura 23) mostrou que bons resultados foram atingidos utilizando trigrama, tetragrama e uma combinação de trigrama, tetragrama e pentagrama, com uma acurácia acima dos 97%. Com a combinação de n-gramas a acurácia aumentou em mais de 0,5% quando comparada ao trigrama e mais de 1% quando comparada ao tetragrama. Devido ao custo computacional, para 100 autores, apenas trigrama foi utilizado como mostra a figura 22, com uma acurácia de mais de 70%.

Como visto na tabela 7, os resultados de precisão, revocação e medida F1 para rótulos POS são um pouco inferiores do que os resultados de palavras como característica, porém foram obtido bons resultados para precisão (quase 66%), revocação (71,11%) e medida F1 (67,57%). Para 10 autores a combinação de n-gramas (com n variando de 3 a 5) apresentou os melhores resultados com uma acurácia geral maior que 98%. Estes resultados são muito próximos dos resultados atingidos para unigrama de palavras.

6 Considerações finais

O presente trabalho abordou a atribuição automática de autoria considerando dados de redes sociais.

O escopo deste projeto se limitou apenas ao contexto temporal de mensagens extraídas do *Reddit* (fórum de discussão online sem limitação de caracteres), pertencentes à língua inglesa, utilizando técnicas para representação do texto baseadas em n-gramas de caracteres, palavras ou presentações de rótulos *POS*, explorando duas situações diferentes mas relacionadas em redes sociais: grande volume de dados e dados temporais para lidar com a atribuição de autoria. Quando um problema envolve análise temporal em um contexto de rede social obedecendo a distribuição dos dados no mundo real, geralmente envolve uma enorme quantidade de dados. Um dos principais desafios que apareceram no decorrer do desenvolvimento do modelo de classificação para esse problema foi o custo computacional para o desenvolvimento. É comum durante o desenvolvimento analisar e testar diferentes valores para os hiper-parâmetros e achar o valor mais apropriado, ou até mesmo para o pré-processamento dos dados, porém no contexto de grandes conjuntos de dados, estas tarefas acabam por consumir um grande tempo da pesquisa.

Dados do Reddit foram selecionados para este projeto, pois neste fórum existe certa anonimidade do usuário e não há limitação de caracteres. Estas características, presentes apenas em algumas redes sociais, podem permitir um espaço mais adequado para que os usuários possam usar o seu estilo próprio de escrita, sem forçar o encurtamento de palavras ou uso de gírias para caber na limitação de caracteres, e a exposição potencialmente sincera na rede social sem receios que usuários próximos (ou até mesmo amigos) possam seguir e ver as interações do usuário em si.

O uso das três principais características a partir das quais os textos foram representados (palavras, caracteres e rótulos POS) validaram que elas podem ser usadas para classificar corretamente o autor de um comentário anônimo dentro de um conjunto de autores. O n-grama de palavras mostrou ser uma boa característica porque pode capturar a construção do vocabulário do usuário dentro de uma sentença, com o TF-IDF valorizando os n-gramas que são mais intrínsecos ao usuário recebendo um valor mais alto, e n-gramas mais comuns recebendo um valor mais baixo. Além disso, no contexto de uma rede social sem limitação de caracteres, os resultados de unigrama, bigrama e trigrama de palavras

estarem próximos aos resultados do n-grama de caracteres é justamente pelo fato do usuário poder utilizar as palavras que ele normalmente utilizaria, evitando gírias e encurtamentos desnecessários.

Para n-grama de caracteres, uma situação interessante é que os erros gramaticais não são "penalizados", o que poderia ser usado como uma característica discriminante para um autor que possui esse padrão de erro, além disso, alguns *emoticons* ou caracteres especiais podem ser usados como características discriminantes. Além disso, o bom resultado dessa característica já é previsto tendo em vista que ela consegue capturar, com o aumento do valor de N, mais detalhes do que as outras características em si que analisam a palavra como um todo, que é o caso dos n-gramas de palavras e rótulos POS, já que o n-grama de caracteres pode analisar tanto uma ou mais partes de um palavra (caso o valor de N seja menor que o número de caracteres da palavra em si) ou o conjunto de palavras (caso o valor de N seja maior do que o número de caracteres de mais de uma palavra).

No caso de n-grama de rótulos POS, é esperado um resultado inferior aos outros dois n-gramas tendo em vista que a expressividade das classes gramaticais são inferiores ao de palavras e caracteres, já que uma combinação de diferentes palavras pode levar a uma mesma combinação de classes gramaticais.

Todos os cenários avaliados atingiram resultados considerados bastante satisfatórios. Tendo 10 autores como alvos, os resultados foram acima de 97% de acurácia, precisão, revocação e medida F1. Para 100 autores como alvos, o pior resultado foi alcançado pelo trigrama de rótulos POS como característica, com 70% de acurácia, quase 66% de precisão, 71% de revocação e 67% de medida F1, mas outros cenários atingiram um intervalo de valores similares, com acurácia ficando entre 71% e 72%, precisão entre 68% e 69%, revocação com 72% e medida F1 entre 69% e 70%. Esses resultados mostraram que apesar do cenário de 100 autores, os resultados são estáveis.

Para esses resultados, três situações interessantes foram identificadas. Inicialmente, destacamos que o único cenário que n-grama de caracteres foi utilizado foi para 10 autores como alvos, porém mostrou os melhores resultados dentre outras características, como o uso de tetragrama superou de trigrama em acurácia, precisão, revocação e medida F1 em 0,5% nessas quatro métricas. Talvez um n-grama com $n > 4$ pudesse atingir resultados melhores em comparação ao tetragrama. Em relação ao uso de rótulos POS, para a classificação de 10 autores, o melhor resultado dentre os rótulos POS como característica foi uma combinação de trigrama, tetragrama e pentagrama como característica, porém se

combinássemos n-grama de palavras ou caracteres o mesmo padrão de melhor resultado será atingido? Isso ainda está aberto à investigação.

Terceiro, considerando palavras como características, para 10 autores como alvo, o bigrama atingiu os melhores resultados em todas as métricas, porém para a classificação de 100 autores, o unigrama atingiu os melhores resultados, o que gerou um resultado interessante e inesperado, dado que os bigramas podem extrair mais do estilo de escrita do usuário já que os unigramas não podem extrair padrão de colocação. Esta situação pode ter ocorrido por causa dos seguintes motivos: repetição dos bigramas entre os autores e esparsividade dos dados.

Destaca-se ainda que uma rede neural LSTM de camada simples foi implementada com otimizador RMS para validar a atribuição de autoria no contexto de grandes conjuntos de dados e análise temporal. Nenhuma otimização de hiper-parâmetros foi implementada por causa do custo computacional, porém uma otimização poderia, com uma implementação mais robusta e complexa da LSTM, conduzir a melhores resultados para 100 ou mais autores como alvos.

Contudo, pelo fato deste trabalho abordar questões que outros trabalhos, revisados na seção de revisão de literatura, não abordaram, não houve comparações dos resultados obtidos com um *baseline*, já que qualquer comparação entre resultados obtidos entre trabalhos cujas questões abordadas são diferentes não seria uma boa comparação. Além disso, não foi possível criar um *baseline* (rodando os mesmos cenários, só que em uma rede que não tratasse da questão temporal como a LSTM trata), pelo falta de tempo que outras questões do presente trabalho exigiram, como a busca de hiper-parâmetros adequados e o treinamento da rede com grande volume de dados.

Como resultados deste trabalho, dois artigos científicos foram apresentados e publicados em anais de eventos científicos: no SBSI 2020 (Simpósio Brasileiro de Sistemas de Informação) (CASIMIRO; DIGIAMPIETRI, 2020) e no SBSI 2022 (Simpósio Brasileiro de Sistemas de Informação) (CASIMIRO; DIGIAMPIETRI, 2022).

Para trabalhos futuros, pretende-se: observar o impacto de informações de sentimentos e tópicos como características; utilizar combinações de n-gramas de caracteres e palavras como características, observando se o mesmo comportamento visto nos rótulos POS se repetirá; utilizar combinações de pelo menos dois tipos diferentes de n-gramas (caracteres e rótulos POS, palavras e rótulos POS ou caracteres e palavras) juntos como características e avaliar os resultados em comparação com outras características; por fim,

desenvolver um *ensemble* de classificadores para analisar as melhorias nos resultados de classificação.

Referências[1]

ABADI, M.; AGARWAL, A.; BARHAM, P.; BREVDO, E.; CHEN, Z.; CITRO, C.; CORRADO, G. S.; DAVIS, A.; DEAN, J.; DEVIN, M.; GHEMAWAT, S.; GOODFELLOW, I.; HARP, A.; IRVING, G.; ISARD, M.; JIA, Y.; JOZEFOWICZ, R.; KAISER, L.; KUDLUR, M.; LEVENBERG, J.; MANé, D.; MONGA, R.; MOORE, S.; MURRAY, D.; OLAH, C.; SCHUSTER, M.; SHLENS, J.; STEINER, B.; SUTSKEVER, I.; TALWAR, K.; TUCKER, P.; VANHOUCKE, V.; VASUDEVAN, V.; VIéGAS, F.; VINYALS, O.; WARDEN, P.; WATTENBERG, M.; WICKE, M.; YU, Y.; ZHENG, X. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. 2015. Software available from tensorflow.org. Disponível em: ⟨https://www.tensorflow.org/⟩. Citado na página 52.

ABBASI, A.; CHEN, H. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, v. 20, n. 5, p. 67–75, Sep. 2005. Citado 2 vezes nas páginas 34 e 39.

AGGARWAL, C. C. *Machine learning for text*. [S.l.]: Springer, 2018. Citado 5 vezes nas páginas 24, 25, 26, 27 e 28.

ALBADARNEH, J.; TALAFHA, B.; AL-AYYOUB, M.; ZAQAIBEH, B.; AL-SMADI, M.; JARARWEH, Y.; BENKHELIFA, E. Using big data analytics for authorship authentication of arabic tweets. In: *Proceedings of the 8th International Conference on Utility and Cloud Computing*. Piscataway, NJ, USA: IEEE Press, 2015. (UCC '15), p. 448–452. ISBN 978-0-7695-5697-0. Disponível em: ⟨http://dl.acm.org/citation.cfm?id=3233397.3233483⟩. Citado 2 vezes nas páginas 35 e 41.

AZARBONYAD, H.; DEHGHANI, M.; MARX, M.; KAMPS, J. Time-aware authorship attribution for short text streams. In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 2015. (SIGIR '15), p. 727–730. ISBN 978-1-4503-3621-5. Disponível em: ⟨http://doi.acm.org/10.1145/2766462.2767799⟩. Citado 4 vezes nas páginas 35, 41, 48 e 49.

BANGA, R.; MEHNDIRATTA, P. Authorship attribution for textual data on online social networks. In: *2017 Tenth International Conference on Contemporary Computing (IC3)*. [S.l.: s.n.], 2017. p. 1–7. Citado 2 vezes nas páginas 36 e 43.

BAUMGARTNER, J. *Pushshift Reddit comments*. 2019. Disponível em: ⟨https://files.pushshift.io/reddit/comments/⟩. Citado na página 50.

BIRD, S.; KLEIN, E.; LOPER, E. *Natural language processing with Python: analyzing text with the natural language toolkit*. [S.l.]: "O'Reilly Media, Inc.", 2009. Citado na página 52.

BOUANANI, S. E. M. E.; KASSOU, I. Using lexicometry and vocabulary analysis techniques to detect a signature for web profile. In: *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*. [S.l.: s.n.], 2013. p. 1494–1498. Citado 2 vezes nas páginas 34 e 41.

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

BUITINCK, L.; LOUPPE, G.; BLONDEL, M.; PEDREGOSA, F.; MUELLER, A.; GRISEL, O.; NICULAE, V.; PRETTENHOFER, P.; GRAMFORT, A.; GROBLER, J.; LAYTON, R.; VANDERPLAS, J.; JOLY, A.; HOLT, B.; VAROQUAUX, G. API design for machine learning software: experiences from the scikit-learn project. In: *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. [S.l.: s.n.], 2013. p. 108–122. Citado na página 52.

CASIMIRO, G. R.; DIGIAMPIETRI, L. Authorship attribution using data from reddit forum. In: *SBSI 2020*. [S.l.: s.n.], 2020. (aceito par publicação). Citado 14 vezes nas páginas 19, 50, 51, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62 e 70.

CASIMIRO, G. R.; DIGIAMPIETRI, L. A. Authorship attribution with temporal data in reddit. In: *XVIII Brazilian Symposium on Information Systems*. New York, NY, USA: Association for Computing Machinery, 2022. (SBSI). ISBN 9781450396981. Disponível em: ⟨https://doi.org/10.1145/3535511.3535515⟩. Citado 5 vezes nas páginas 64, 65, 66, 67 e 70.

CHOLLET, F. *et al. Keras*. 2015. ⟨https://keras.io⟩. Citado na página 52.

CRISTANI, M.; ROFFO, G.; SEGALIN, C.; BAZZANI, L.; VINCIARELLI, A.; MURINO, V. Conversationally-inspired stylometric features for authorship attribution in instant messaging. In: *Proceedings of the 20th ACM International Conference on Multimedia*. New York, NY, USA: ACM, 2012. (MM '12), p. 1121–1124. ISBN 978-1-4503-1089-5. Disponível em: ⟨http://doi.acm.org/10.1145/2393347.2396398⟩. Citado 2 vezes nas páginas 34 e 45.

DING, S. H. H.; FUNG, B. C. M.; IQBAL, F.; CHEUNG, W. K. Learning stylometric representations for authorship analysis. *IEEE Transactions on Cybernetics*, v. 49, n. 1, p. 107–121, Jan 2019. Citado 2 vezes nas páginas 36 e 44.

DONAIS, J. A.; FROST, R. A.; PEELAR, S. M.; RODDY, R. A. Summary: A system for the automated author attribution of text and instant messages. In: *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*. [S.l.: s.n.], 2013. p. 1484–1485. Citado 2 vezes nas páginas 34 e 45.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep learning*. [S.l.]: MIT press, 2016. Citado 4 vezes nas páginas 27, 28, 29 e 30.

HALVANI, O.; WINTER, C.; GRANER, L. On the usefulness of compression models for authorship verification. In: *Proceedings of the 12th International Conference on Availability, Reliability and Security*. New York, NY, USA: ACM, 2017. (ARES '17), p. 54:1–54:10. ISBN 978-1-4503-5257-4. Disponível em: ⟨http://doi.acm.org/10.1145/3098954.3104050⟩. Citado 3 vezes nas páginas 35, 42 e 48.

HARRIS, C. R.; MILLMAN, K. J.; WALT, S. J. van der; GOMMERS, R.; VIRTANEN, P.; COURNAPEAU, D.; WIESER, E.; TAYLOR, J.; BERG, S.; SMITH, N. J.; KERN, R.; PICUS, M.; HOYER, S.; KERKWIJK, M. H. van; BRETT, M.; HALDANE, A.; RÍO, J. F. del; WIEBE, M.; PETERSON, P.; GÉRARD-MARCHANT, P.; SHEPPARD, K.; REDDY, T.; WECKESSER, W.; ABBASI, H.; GOHLKE, C.; OLIPHANT, T. E. Array programming with NumPy. *Nature*, Springer Science and Business Media LLC, v. 585, n. 7825, p. 357–362, set. 2020. Disponível em: ⟨https://doi.org/10.1038/s41586-020-2649-2⟩. Citado na página 52.

HUNTER, J. D. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, IEEE COMPUTER SOC, v. 9, n. 3, p. 90–95, 2007. Citado na página 52.

IGAWA, R. A.; ALMEIDA, A. M. G. de; ZARPELAO, B. B.; BARBON JR, S. Recognition of compromised accounts on twitter. In: *Proceedings of the Annual Conference on Brazilian Symposium on Information Systems: Information Systems: A Computer Socio-Technical Perspective - Volume 1*. Porto Alegre, Brazil, Brazil: Brazilian Computer Society, 2015. (SBSI 2015), p. 2:9–2:14. Disponível em: ⟨http://dl.acm.org/citation.cfm?id=2814058.2814061⟩. Citado 2 vezes nas páginas 35 e 42.

INCHES, G.; HARVEY, M.; CRESTANI, F. Finding participants in a chat: Authorship attribution for conversational documents. In: *2013 International Conference on Social Computing*. [S.l.: s.n.], 2013. p. 272–279. Citado 2 vezes nas páginas 35 e 45.

KIM, K.; NOH, Y.; Park, S. Detecting multiple userids on korean social media for mining tv audience response. In: *TENCON 2015 - 2015 IEEE Region 10 Conference*. [S.l.: s.n.], 2015. p. 1–4. Citado 2 vezes nas páginas 35 e 47.

KITCHENHAM, B.; CHARTERS, S. Guidelines for performing systematic literature reviews in software engineering. Citeseer, 2007. Citado na página 31.

KOPPEL, M.; SCHLER, J.; ARGAMON, S.; MESSERI, E. Authorship attribution with thousands of candidate authors. In: *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 2006. (SIGIR '06), p. 659–660. ISBN 1-59593-369-7. Disponível em: ⟨http://doi.acm.org/10.1145/1148170.1148304⟩. Citado 2 vezes nas páginas 34 e 39.

KUZU, R. S.; BALCI, K.; SALAH, A. A. Authorship recognition in a multiparty chat scenario. In: *2016 4th International Conference on Biometrics and Forensics (IWBF)*. [S.l.: s.n.], 2016. p. 1–6. Citado 2 vezes nas páginas 35 e 46.

LAKKARAJU, H.; BHATTACHARYA, I.; BHATTACHARYYA, C. Dynamic multi-relational chinese restaurant process for analyzing influences on users in social media. In: *2012 IEEE 12th International Conference on Data Mining*. [S.l.: s.n.], 2012. p. 389–398. Citado 2 vezes nas páginas 34 e 46.

LAYTON, R.; WATTERS, P.; DAZELEY, R. Authorship attribution for twitter in 140 characters or less. In: *2010 Second Cybercrime and Trustworthy Computing Workshop*. [S.l.: s.n.], 2010. p. 1–8. Citado 2 vezes nas páginas 34 e 40.

LE, H.; SAFAVI-NAINI, R. On de-anonymization of single tweet messages. In: *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics*. New York, NY, USA: ACM, 2018. (IWSPA '18), p. 8–14. ISBN 978-1-4503-5634-3. Disponível em: ⟨http://doi.acm.org/10.1145/3180445.3180451⟩. Citado 2 vezes nas páginas 36 e 43.

LITVINOVA, T.; LITVINOVA, O.; PANICHEVA, P. Authorship attribution of russian forum posts with different types of n-gram features. In: *Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval*. New York, NY, USA: ACM, 2019. (NLPIR 2019), p. 9–14. ISBN 978-1-4503-6279-5. Disponível em: ⟨http://doi.acm.org/10.1145/3342827.3342834⟩. Citado 3 vezes nas páginas 36, 44 e 48.

MCKINNEY Wes. Data Structures for Statistical Computing in Python. In: WALT Stéfan van der; MILLMAN Jarrod (Ed.). *Proceedings of the 9th Python in Science Conference*. [S.l.: s.n.], 2010. p. 56 – 61. Citado na página 52.

O'CONNOR, B.; BALASUBRAMANYAN, R.; ROUTLEDGE, B. R.; SMITH, N. A. From tweets to polls: Linking text sentiment to public opinion time series. In: *Fourth International AAAI Conference on Weblogs and Social Media*. [S.l.: s.n.], 2010. Citado na página 19.

PEREZ, C.; BIRREGAH, B.; LAYTON, R.; LEMERCIER, M.; WATTERS, P. Replot: Retrieving profile links on twitter for suspicious networks detection. In: *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*. [S.l.: s.n.], 2013. p. 1307–1314. Citado 2 vezes nas páginas 34 e 41.

PETRASOVA, S.; KHAIROVA, N.; LEWONIEWSKI, W. Building the semantic similarity model for social network data streams. In: *2018 IEEE Second International Conference on Data Stream Mining Processing (DSMP)*. [S.l.: s.n.], 2018. p. 21–24. Citado 2 vezes nas páginas 36 e 43.

PILLAY, S. R.; SOLORIO, T. Authorship attribution of web forum posts. In: *2010 eCrime Researchers Summit*. [S.l.: s.n.], 2010. p. 1–7. Citado 3 vezes nas páginas 34, 41 e 48.

ROECK, A. N. D.; AL-FARES, W. A morphologically sensitive clustering algorithm for identifying arabic roots. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*. [S.l.], 2000. p. 199–206. Citado na página 34.

ROFFO, G.; GIORGETTA, C.; FERRARIO, R.; RIVIERA, W.; CRISTANI, M. Statistical analysis of personality and identity in chats using a keylogging platform. In: *Proceedings of the 16th International Conference on Multimodal Interaction*. New York, NY, USA: ACM, 2014. (ICMI '14), p. 224–231. ISBN 978-1-4503-2885-2. Disponível em: ⟨http://doi.acm.org/10.1145/2663204.2663272⟩. Citado 3 vezes nas páginas 35, 45 e 48.

ROFFO, G.; SEGALIN, C.; Vinciarelli, A.; Murino, V.; Cristani, M. Reading between the turns: Statistical modeling for identity recognition and verification in chats. In: *2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance*. [S.l.: s.n.], 2013. p. 99–104. Citado 2 vezes nas páginas 34 e 45.

SAHA, D.; HAQUE, M. M.; SARKAR, A.; ALAM, F. *Novel class detection in concept drifting data streams using decision tree leaves*. Tese (Doutorado), 2018. Citado na página 35.

SCHLER, J.; KOPPEL, M.; ARGAMON, S.; PENNEBAKER, J. *Effects of age and gender on blogging. AAAI Spring Symposium on computational approaches for analyzing weblogs*. [S.l.]: Stanford, CA, 2006. Citado na página 36.

SEKER, S. E.; AL-NAAMI, K.; KHAN, L. Author attribution on streaming data. In: *2013 IEEE 14th International Conference on Information Reuse Integration (IRI)*. [S.l.: s.n.], 2013. p. 497–503. Citado 4 vezes nas páginas 35, 41, 48 e 49.

SEROUSSI, Y.; BOHNERT, F.; ZUKERMAN, I. Authorship attribution with author-aware topic models. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012. (ACL '12), p. 264–269. Disponível em: ⟨http://dl.acm.org/citation.cfm?id=2390665.2390728⟩. Citado 4 vezes nas páginas 34, 46, 47 e 48.

SEROUSSI, Y.; ZUKERMAN, I.; BOHNERT, F. Authorship attribution with latent dirichlet allocation. In: *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011. (CoNLL '11), p. 181–189. ISBN 978-1-932432-92-3. Disponível em: ⟨http://dl.acm.org/citation.cfm?id=2018936.2018957⟩. Citado 4 vezes nas páginas 27, 34, 46 e 47.

SEROUSSI, Y.; ZUKERMAN, I.; BOHNERT, F. Authorship attribution with topic models. *Comput. Linguist.*, MIT Press, Cambridge, MA, USA, v. 40, n. 2, p. 269–310, jun. 2014. ISSN 0891-2017. Disponível em: ⟨http://dx.doi.org/10.1162/COLI_a_00173⟩. Citado 3 vezes nas páginas 35, 47 e 48.

SILVA, L. A. da; PERES, S. M.; BOSCARIOLI, C. *Introdução à mineração de dados: com aplicações em R*. [S.l.]: Elsevier Brasil, 2017. Citado 2 vezes nas páginas 23 e 24.

SPITTERS, M.; KLAVER, F.; KOOT, G.; STAALDUINEN, M. v. Authorship analysis on dark marketplace forums. In: *2015 European Intelligence and Security Informatics Conference*. [S.l.: s.n.], 2015. p. 1–8. Citado 2 vezes nas páginas 35 e 42.

STAMATATOS, E. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, Wiley Online Library, v. 60, n. 3, p. 538–556, 2009. Citado 2 vezes nas páginas 22 e 25.

SULTANA, M.; POLASH, P.; GAVRILOVA, M. Authorship recognition of tweets: A comparison between social behavior and linguistic profiles. In: *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. [S.l.: s.n.], 2017. p. 471–476. Citado 2 vezes nas páginas 35 e 43.

SWAIN, S.; MISHRA, G.; SINDHU, C. Recent approaches on authorship attribution techniques — an overview. In: *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*. [S.l.: s.n.], 2017. v. 1, p. 557–566. Citado na página 31.

TAN, R. H. R.; TSAI, F. S. Authorship identification for online text. In: *2010 International Conference on Cyberworlds*. [S.l.: s.n.], 2010. p. 155–162. Citado 2 vezes nas páginas 34 e 40.

YAN, J.; MATTHEWS, S. J. Applying clustering algorithms to determine authorship of chinese twitter messages. In: *2016 IEEE MIT Undergraduate Research Technology Conference (URTC)*. [S.l.: s.n.], 2016. p. 1–4. Citado 2 vezes nas páginas 35 e 42.