Christopher Gabriel Fidelis de Oliveira Corrêa

MINERAÇÃO DE TEXTOS: ANÁLISE DE SENTIMENTOS DE TWEETS SOBRE O PROJETO DE LEI N° 2630/2020

Universidade Federal de Uberlândia
Faculdade de Gestão e Negócios

Uberlândia
2024

MINERAÇÃO DE TEXTOS: ANÁLISE DE SENTIMENTOS DE TWEETS SOBRE O PROJETO DE LEI N° 2630/2020

Trabalho de Conclusão de Curso apresentado à Faculdade de Gestão e Negócios da Universidade Federal de Uberlândia como parte dos requisitos para a obtenção do título de Bacharel em Gestão da Informação.

Área de concentração: Gestão da Informação

Orientador: Leandro Nogueira Couto

Uberlândia
2024

Agradecimentos

A minha mãe: Obrigado por sempre acreditar em mim, por me incentivar a estudar e dar o meu melhor em tudo o que faço, por todos os ensinamentos que me deu até hoje, por todos os esforços e sacrifícios que já fez e ainda faz para proporcionar uma vida digna para mim e meus irmãos. Agradeço por ser um exemplo de ser humano!

Aos meus irmãos: Obrigado por fazerem parte da minha vida e por me fazerem sorrir com assuntos bobos e aleatórios, que apenas vocês sabem trazer.

Aos meus familiares: Obrigado por todo o apoio e pelas palavras de incentivo.

Aos meus amigos: Obrigado por estarem ao meu lado nos momentos mais difíceis, por rirem e chorarem comigo. Agradeço por poder compartilhar tudo com vocês, pelos conselhos e pelos momentos que marcaram minha vida. Os desafios que enfrentei durante a graduação foram amenizados pela presença de vocês ao meu lado.

Ao meu orientador e aos professores da UFU: Obrigado por compartilharem seus conhecimentos comigo, pelas dicas que me deram durante toda a graduação. Obrigado pelo tempo e pela dedicação que foram essenciais para o meu desenvolvimento acadêmico e profissional.

A todos que estão presentes ou que passaram pela minha vida: Sempre que pensei em desistir ou estava exausto e estressado ao limite, sempre houve alguém para me oferecer palavras de encorajamento e dar forças para continuar. Muito obrigado!

"Os dados são o novo petróleo. É valioso, mas se não for refinado não pode realmente ser usado (...) então os dados devem ser decompostos e analisados para que tenham valor."

(Clive Humby)

Resumo

A internet está inserida em diversos aspectos da vida humana, facilitando a comunicação e a troca de informações. No âmbito político, as redes sociais são utilizadas para transmitir decisões e informações governamentais, aumentando o acesso dos indivíduos à política. Parlamentares e cidadãos recorrem a essas plataformas para expressar posições e influenciar questões sociais e políticas. No entanto, o uso inadequado das mídias sociais pode intensificar a polarização política e disseminar desinformação, ameaçando o regime democrático.

Este trabalho tem como finalidade realizar a análise dos sentimentos expressos pelos usuários da rede social Twitter/X em relação o Projeto de Lei nº 2630/2020, de autoria do Senado Federal brasileiro. Para isso, dados provenientes dessa rede foram coletados e, então, as etapas da Análise de Sentimentos foram realizadas. Diferentes algoritmos de classificação de textos foram aplicados à base de dados, antes e após a etapa de pré-processamento do texto. Após avaliar o desempenho dos classificadores, ficou evidente a melhora de performance quando aplicado o pré-processamento aos dados. Além disso, o SVM Linear foi o classificador que obteve a maior acurácia.

A partir dos resultados, notou-se que eventos do cotidiano que envolvem a disseminação de notícias falsas nas redes sociais colaboram para o aumento de tweets sobre o Projeto de Lei 2630/2020, sendo estes, predominantemente, negativos. Ademais, foi possível perceber que parlamentares que estão a direita do eixo político apresentam maior engajamento nas redes sociais.

Palavras-chave: Mineração de Dados. Classificação de Textos. Análise de Sentimentos. Redes Sociais.

Abstract

The internet is embedded in various aspects of human life, facilitating communication and information exchange. In the political realm, social media platforms are used to disseminate governmental decisions and information, enhancing individuals' access to politics. Both lawmakers and citizens turn to these platforms to express opinions and influence social and political issues. However, the misuse of social media can escalate political polarization and spread misinformation, posing a threat to democratic regime.

This work aims to analyze the sentiments expressed by users on the Twitter/X social network regarding Brazilian Senate Bill No. 2630/2020. To achieve this, data from this network were collected, and Sentiment Analysis steps were performed. Multiple text classification algorithms were applied to the dataset, both before and after text preprocessing. The performance evaluation of classifiers revealed a significant improvement when text preprocessing was applied. Additionally, Linear SVM was chosen as the classifier algorithm with the highest accuracy.

The results indicated that everyday events involving the dissemination of fake news on social media contribute to increased tweets about Bill 2630/2020, which were predominantly negative. Furthermore, it was observed that right-leaning legislators exhibit greater engagement on social media.

Keywords: Data Mining. Text Classification. Sentiment Analysis. Social Media.

Lista de ilustrações

Figura 1 – Fluxo do Processo do KDD. Tronchoni et al. (2010) 26
Figura 2 – Abordagens da Mineração na Web. Junior (2007) 28
Figura 3 – Etapas do processo de Descoberta de Conhecimento em Textos. Junior (2007) 29
Figura 4 – Processo de tokenização seguido por remoção de stopwords. Soares (2008) 31
Figura 5 – Classificador SVM. Adaptado de Tripathi (2021) 34
Figura 6 – Etapas de um analisador de sentimentos. Adaptado de Oliveira (2013) 39
Figura 7 – Consulta no Twitter/X realizada por meio do scraper Tweet Flash . . . 44
Figura 8 – Nuvem de palavras do mês de janeiro de 2024. Elaborado pelo autor . 49
Figura 9 – Quantidade de tweets publicados por mês. 51
Figura 10 – Quantidade de tweets por dia em Abril de 2023. 52
Figura 11 – Quantidade de tweets por dia em Maio de 2023. 52
Figura 12 – Quantidade de tweets por dia em Dezembro de 2023. 53
Figura 13 – Quantidade de tweets classificados em cada sentimento por semana nos meses de maior volume. 54
Figura 14 – Matriz de confusão Naive Bayes Bigrama. 56
Figura 15 – Matriz de confusão SVM RBF. 56
Figura 16 – Matriz de confusão SVM Linear. 57
Figura 17 – Nuvem de palavras contidas nos tweets coletados. 58
Figura 18 – Nuvem de palavras contidas nos tweets classificados como positivos. . . 59
Figura 19 – Nuvem de palavras contidas nos tweets classificados como negativos. . 60
Figura 20 – Nuvem de palavras contidas nos tweets classificados como neutros. . . . 60
Figura 21 – Mensagem veiculada na homepage do Google, em maio de 2023, que expõe a opinião da empresa sobre o PL 2630. 61
Figura 22 – Nota enviada pelo Telegram aos usuários sobre o Projeto de Lei 2630/2020. 61
Figura 23 – Conexão entre os usuários por meio de grafo. 62

Lista de tabelas

Tabela 1 – Matriz de confusão para três classes: positivo, negativo e neutro 36
Tabela 2 – Exemplos de tweets antes e após a etapa de pré-processamento 46
Tabela 3 – Trecho do data frame gerado pela função `get_nrc_sentiment` 47
Tabela 4 – Data Frame Tweets X Polaridade 47
Tabela 5 – Sumarização da base de tweets rotulados 48
Tabela 6 – Usuários com maior engajamento nos meses de abril, maio e dezembro de 2023. 54
Tabela 7 – Desempenho dos classificadores antes do pré-processamento do texto. . 55
Tabela 8 – Desempenho dos classificadores após o pré-processamento do texto. . . 55
Tabela 9 – Novos dados usados para testar o desempenho do classificador 57
Tabela 10 – Resultado da classificação dos novos dados 58

Lista de siglas

CART Árvore de Classificação e Regressão

DCBD Descoberta de Conhecimento em Bancos de Dados

DM Data Mining

KDD Knowledge Discovery in Databases

PLN Processamento de Linguagem Natural

STF Supremo Tribunal Federal

SVM Support Vector Machine

WWW World Wide Web

Sumário

1 INTRODUÇÃO 19
1.1 Justificativa 21
1.2 Objetivos 22
1.2.1 Objetivo Geral 22
1.2.2 Objetivos Específicos 22
1.3 Estrutura da Monografia 22

2 FUNDAMENTAÇÃO TEÓRICA 25
2.1 Mineração de Dados 25
2.2 Mineração na Web e Mineração de Textos 27
2.2.1 Mineração na Web 27
2.2.2 Mineração de Texto 27
2.3 Etapas do Processamento de Texto 28
2.3.1 Coleta 29
2.3.2 Pré-processamento 29
2.3.3 Indexação 32
2.3.4 Mineração 32
2.3.5 Análise Dos Resultados Obtidos 37
2.4 Análise de Sentimento / Mineração de Opinião 37
2.4.1 Etapas da Análise de Sentimentos 39
2.4.2 Desafios da Análise de Sentimentos em Redes Sociais 40

3 A SUA PROPOSTA 43
3.1 Coleta de Tweets 43
3.2 Pré-processamento 44
3.3 Criação da Base de Dados Rotulada 46
3.4 Classificação dos Tweets 48
3.5 Apresentação dos Resultados 48

3.5.1 Gráficos de Série Temporal 48
3.5.2 Nuvem de Palavras 48
3.5.3 Grafo 49

4 ANÁLISE DOS RESULTADOS 51
4.1 Quantidade de Tweets 51
4.2 Avaliação dos Classificadores 54
4.3 Classificação de Novos Dados 57
4.4 Nuvem de Palavras 58
4.5 Grafo 62

5 CONCLUSÃO 65
5.1 Principais Contribuições 66
5.2 Trabalhos Futuros 66

REFERÊNCIAS 69

Capítulo 1

Introdução

Na atualidade, a população manifesta seus sentimentos por meio de uma variedade de ferramentas online de mídia social, incluindo, blogs, redes sociais e plataformas de compartilhamento de vídeos e imagens, para narrar suas experiências e divulgar seus pontos de vista sobre diversos temas como, por exemplo, produtos, serviços, eventos e assuntos importantes (ROSA, 2015).

O crescimento de usuários de telefones celulares e tablets com acesso à internet possibilitou que indivíduos se mantivessem conectados ao longo do dia, aumentando assim, a quantidade de informações disponíveis na internet (FRANÇA et al., 2014). No Brasil, a população passa em média nove horas por dia conectada, de acordo com uma pesquisa realizada pelo site ElectronicsHub (FOGAÇA, 2023). Devido à predominância do uso da internet em praticamente todos os aspectos da vida humana contemporânea, Santos (2010), conclui que a internet representa o mais extenso depósito de informações no mundo, onde as pessoas interagem diariamente com uma vasta quantidade de dados.

Ao contrário das mídias de massa convencionais, como a televisão e o rádio, a internet possibilitou que cada usuário, além de receptor, fosse também produtor de conteúdo. Isso significa que, graças à capacidade de conexão em rede, as interações online têm impulsionado um aumento na produção e compartilhamento de uma variedade de informações, incluindo textos, imagens, vídeos e outros formatos (ALMEIDA, 2018). Sendo assim, Barbosa et al. (2012), pontuam que uma fonte promissora de informações reside em redes sociais, como o Twitter/X, onde inúmeras pessoas utilizam esse meio de comunicação para expor suas opiniões sobre experiências vividas no cotidiano (OLIVEIRA, 2013). Para se ter uma ideia da popularidade desta rede social, um levantamento feito pelo site The Social Shepherd, revela que pelo menos 500 milhões de tweets são enviados diariamente (SHEPHERD, 2024).

Lançado em 2006, o Twitter, ou X, como passou a ser chamado a partir de julho de 2023 (INFOMONEY, 2023), registrou, no início de 2022, cerca de 217 milhões de usuários ativos, sendo 19 milhões somente no Brasil (VALOR, 2022). Vale ressaltar que, devido à mudança do nome da rede social, ao longo do texto utilizaremos o termo Twitter/X sempre

que nos referirmos a ela. Barbosa et al. (2012) acrescentam que a dinâmica de interação no modelo do Twitter/X estimula os usuários a compartilharem e expressarem continuamente suas opiniões e sentimentos, os quais são disseminados entre seus seguidores. Além disso, os internautas têm a capacidade de publicar novas mensagens, denominadas tweets, ou compartilhar mensagens postadas por outros através de "retuítes".

Para Rosa (2015), nos últimos anos, o acompanhamento das redes sociais tem sido objeto de diversas pesquisas tanto no âmbito acadêmico quanto empresarial, cujo o objetivo é capturar comentários positivos e negativos relacionados a um determinado tema, e, através da análise dos textos, obter um conhecimento abrangente da opinião dos usuários sobre o assunto, evento ou fato em questão.

Para auxiliar nesse tipo de análise, a Mineração de Textos, também conhecida como Descoberta de Conhecimento em Textos, surge para oferecer um conjunto de técnicas capazes de automatizar o procedimento de coleta e organização dessas informações (FILHO, 2014). Além disso, uma área em ascensão na mineração de textos é a análise de sentimentos, que aborda as opiniões manifestadas por indivíduos em textos. Esse enfoque específico é denominado mineração de opinião ou análise de sentimento (SANTOS, 2010).

É possível observar o volume e variedade de trabalhos recentes de análise de sentimentos em postagens em redes sociais em revisões sistemáticas como Neri et al. (2012), que analisaram 1.000 postagens no Facebook sobre noticiários, comparando os sentimentos expressos ao serviço público de radiodifusão italiano em relação à empresa privada *La7* e destacou a importância do Facebook como plataforma de marketing online. Já Filho (2014), empregou uma versão do algoritmo de classificação Naive Bayes para analisar tweets sobre os jogos da seleção brasileira na Copa do Mundo de 2014 e conseguiu identificar os sentimentos dos usuários do Twitter/X em relação a esses jogos, apresentando os resultados por meio de mapas de calor e nuvens de palavras. Berto (2021), analisou os principais sentimentos expressos pelos brasileiros em tweets relacionados à Covid-19, abrangendo o espectro de oito emoções básicas humanas e concluiu que houve uma prevalência significativa de tweets com conteúdo negativo, predominantemente associados ao medo e à tristeza.

Dessa maneira, de acordo com Rosa (2015), as publicações feitas por usuários nas redes sociais desempenham um papel crucial ao viabilizar a troca de pensamentos e experiências, uma vez que, podem refletir sentimentos em relação a um determinado assunto, expressando opiniões favoráveis, desfavoráveis ou neutras. A mineração de texto associada à análise de sentimentos, possibilita que organizações e/ou entidades adquirem discernimento sobre os comentários dos usuários em seus perfis na web e utilizem os resultados para diversos propósitos, como por exemplo, estratégias de marketing, procedimentos de segurança, aprimoramento de serviços, entre outras finalidades (FILHO, 2014).

Por fim, vale ressaltar que, para Kaakinen et al. (2020), a seletividade social e os algoritmos das redes sociais encorajam os cidadãos a habitar "bolhas psicossociais", tam-

bém conhecidas como ”câmaras de eco”, onde a interação social dos usuários é limitada à comunicação com pessoas cujas ideias são semelhantes. Em outras palavras, os indivíduos buscam interações sociais com pessoas que compartilham e validam suas identidades. Segundo os autores, a identificação social e a tendência para interagir com pessoas que pensam da mesma forma (homofilia) fazem com que informações vindas de membros dessa bolha sejam vistas como mais confiáveis. No âmbito político, Eady et al. (2019) complementam dizendo que apesar da disponibilidade de informações ideologicamente diversas, os indivíduos tendem, no entanto, a consumir informações e interagir com indivíduos que se alinham ideologicamente com suas próprias crenças políticas.

1.1 Justificativa

Um estudo conduzido pelo Comitê Gestor da Internet no Brasil indicou que três em cada quatro brasileiros utilizam a internet, totalizando 134 milhões de pessoas. Este dado destaca a consolidação na sociedade de elementos como a praticidade e rapidez na troca de dados, tais como mensagens, fotos, vídeos e áudios, impulsionados pela revolução tecnológica promovida pela internet (QUAGLIO et al., 2021). Para Holanda e Teixeira (2023), as redes sociais representam uma das contribuições mais significativas da Internet desde sua origem, permeando diversas dimensões da vida do indivíduo, inclusive no âmbito político.

Segundo Teixeira, Reis e Fonseca (2023), é comum a transmissão de decisões e informações governamentais por meio das redes sociais. Tal prática, embora amplie o acesso e inclusão da população na vida política do país, apresenta aspectos negativos e prejudiciais. Notadamente, tornou-se comum a disseminação de notícias falsas, destinadas a atacar um grupo político ou ideologia específicos, especialmente durante períodos eleitorais. Além disso, em conformidade com Holanda e Teixeira (2023), parlamentares e candidatos podem disseminar amplamente suas posições para conquistar eleitores favoráveis, mas também os próprios cidadãos podem se manifestar politicamente e influenciar em determinadas questões na sociedade.

Para Holanda e Teixeira (2023), a utilização inadequada das mídias sociais pode constituir uma ferramenta intensificadora da polarização política e da propagação de desinformação, representando assim dois impactos severamente prejudiciais ao regime democrático. Ainda segundo os autores, diversas medidas legais surgiram no mundo com o intuito de regulamentar as mídias sociais e enfrentar a polarização política e a desinformação como, por exemplo, o Projeto de Lei nº 2630/2020, de autoria do Senado Federal brasileiro, que visa dar origem à Lei Brasileira de Liberdade, Responsabilidade e Transparência.

Este estudo, portanto, visa aplicar técnicas da área de mineração de texto em uma base de dados de tweets referentes ao Projeto de Lei nº 2630/2020. Dessa forma, este trabalho

poderá servir como estudo de caso para análise de como ocorre o engajamento com temas controversos nas redes sociais. Após aplicar técnicas de Processamento de Textos e Análise de Sentimentos, será possível explorar como os usuários pensam a respeito da aprovação deste projeto.

1.2 Objetivos

1.2.1 Objetivo Geral

O objeto deste trabalho é identificar o sentimento de postagens no Twitter sobre o Projeto de Lei nº 2630/2020 por meio de técnicas de pré-processamento e Análise de Sentimentos, bem como, algoritmos de classificação de texto, em uma base de dados constituída por tweets, para categorizar os textos de acordo com sua polaridade (positivo, negativo ou neutro).

1.2.2 Objetivos Específicos

- Coletar dados textuais (tweets) do Twitter/X referentes ao Projeto de Lei nº 2630/2020;
- Construir uma base rotulada para ser utilizada em experimentos, classificando cada tweets de acordo com sua polaridade (positivo, negativo ou neutro);
- Realizar o pré-processamento dos dados coletados;
- Examinar os algoritmos de classificação na literatura que empregam variadas abordagens para categorizar os tweets, selecionando aquele que obtém melhor desempenho para o problema.
- Validar o modelo de acordo com o conjunto de teste;
- Representar por meio de nuvens de palavras os termos mais frequentes mencionados pelos internautas;
- Demonstrar a conexão entre os internautas por meio de um grafo;
- Investigar como eventos do cotidiano afetam os sentimentos dos internautas no que diz respeito ao projeto de lei 2036/2020.

1.3 Estrutura da Monografia

O restante deste trabalho está dividido da seguinte maneira: No capítulo 2 é apresentado o referencial teórico com toda a fundamentação necessária para o desenvolvimento e entendimento do estudo, incluindo uma visão geral sobre Mineração de Dados, Mineração

para Web e Mineração de Textos. Além da descrição das etapas de pré-processamento de textos e conceitos relacionados a análise de sentimentos. No capítulo 3 é apresentada a metodologia desenvolvida no trabalho, detalhando os passos necessários para alcançar os propósitos do projeto. No capítulo 4, são expostos os resultados obtidos a partir da classificação da base com o classificador escolhido e são observados e analisados dados como, volume de tweets, conexão entre os usuários e termos mais frequentes no texto. No capítulo 5 são apresentadas as conclusões e contribuições deste trabalho juntamente com recomendações para pesquisas futuras.

Capítulo 2

Fundamentação Teórica

2.1 Mineração de Dados

Ao longo do tempo, constatou-se que a velocidade de coleta de informações excedia a velocidade de processamento ou análise das mesmas (CARDOSO; MACHADO, 2008). De acordo com Witten et al. (2011), há uma dificuldade em encontrar padrões, correlações e tendências em grandes volumes de dados, uma vez que, embora os dados possam conter *insights* valiosos, a sua análise pode ser desafiadora devido à complexidade e a grande quantidade de informações disponíveis. Portanto, tornou-se vital o desenvolvimento e implementação de técnicas e ferramentas automatizadas para aperfeiçoar o processo de extração de informações relevantes de grandes volumes de dados. Um processo interativo, que visa solucionar este desafio, que ultrapassa as capacidades humanas, é a descoberta de conhecimento em bancos de dados (CARDOSO; MACHADO, 2008).

A Descoberta de Conhecimento em Bancos de Dados (DCBD), ou em inglês, Knowledge Discovery in Databases (KDD), também conhecida como mineração de dados (do inglês Data Mining (DM)), concentra-se na exploração computadorizada de extensos conjuntos de dados, visando identificar padrões interessantes dentro deles (FELDMAN et al., 1998). Segundo Galvão e Marin (2009), a expressão Mineração de Dados foi introduzida pela primeira vez como sinônimo de KDD. No entanto, ela constitui uma técnica inserida em uma das etapas no processo do KDD. Apesar disso, em algumas bibliografias o termo mineração de dados tornou-se mais popular do que KDD (TRONCHONI et al., 2010).

O processo de Descoberta de Conhecimento em Banco de Dados é mostrado na Figura 1 e consiste nos seguintes passos (TRONCHONI et al., 2010):

1. Limpeza dos dados: para remover ruídos, inconsistências e informações irrelevantes;

2. Integração dos dados: onde é feita a combinação de dados vindos de múltiplas fontes;

3. Seleção dos dados: é realizada a recuperação junto ao banco de dados dos dados pertinentes para a análise;

4. Transformação dos dados: preparação e formatação dos dados para que possam ser eficientemente utilizados na mineração;
5. Mineração de dados: aplicação de técnicas para extrair padrões de dados;
6. Avaliação e representação do conhecimento: técnicas de visualização são empregadas para apresentar as informações extraídas ao usuário.

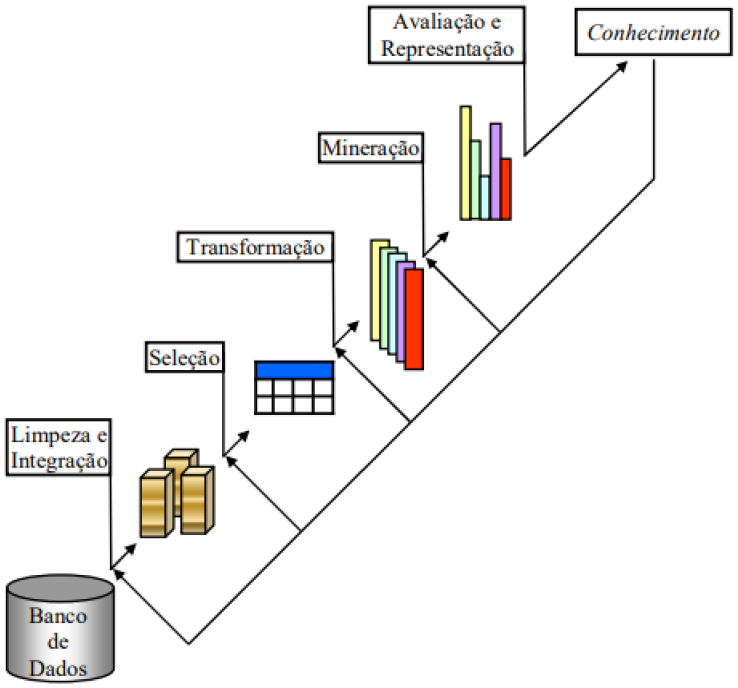

Figura 1 – Fluxo do Processo do KDD. Tronchoni et al. (2010)

Conforme Tronchoni et al. (2010), existem dois propósitos essenciais na mineração de dados: predição e descrição. A primeira busca prever valores futuros ou desconhecidos de variáveis específicas com base em outras variáveis presentes nos conjuntos de dados. Por outro lado, a descrição visa identificar padrões que descrevam os dados de modo que possam ser interpretáveis pelos usuários.

Assim, compreende-se que o objetivo principal da mineração de dados é descobrir relacionamentos entre dados, proporcionando suporte para realizar previsões de padrões futuros com base em eventos passados. E assim, facilitar a tomada de decisão (EVAGELISTA; POLETTO, 2018). Para Cardoso e Machado (2008), a mineração de dados tem como finalidade revelar, de maneira automática ou semi-automática, os padrões que se encontram "ocultos"na abundante quantidade de dados armazenados em bancos de dados.

Em suma, a mineração de dados é aplicada para apoiar a tomada de decisão a partir da descoberta de conhecimento por meio dos dados, e envolve inúmeras técnicas e aplicações

(GUPTA; CHANDRA, 2020). Por exemplo, conhecer os hábitos dos consumidores é crucial para profissionais de marketing, enquanto políticos podem se beneficiar de uma determinada compreensão dos padrões comportamentais da população para estabelecer estratégias eficazes (CORDEIRO, 2003).

2.2 Mineração na Web e Mineração de Textos

2.2.1 Mineração na Web

Atualmente, a relevância da World Wide Web (WWW) como fonte de informação é inquestionável, sendo considerada um imenso banco de dados onde a maior parte dessas informações é de natureza não estruturada, caracterizando-se pela ausência de uma estrutura ou organização explícita associada (CORDEIRO, 2003). Para Nandwani e Verma (2021), a grande quantidade de dados não estruturados na internet provém das plataformas de redes sociais. De tal forma, surge a necessidade de absorver essas informações, e reconhecendo as dificuldades e custos associados a essa tarefa por meios não computacionais surge o termo Mineração na Web, derivado da Mineração de Dados (SANTOS, 2010).

Assim a expressão Mineração na Web (do inglês, *Web Mining*) consiste na pesquisa de conhecimento no World Wide Web – é o Mineração de Dados orientado para a internet (CORDEIRO, 2003). Conforme a Figura 2, a Mineração na Web se divide em três abordagens distintas (JUNIOR, 2007):

1. Web Mining de conteúdo: preocupa-se com a estrutura da página em si, analisando os textos, imagens e outros componentes presentes nos documentos HTML;

2. Web Mining de estrutura: estuda o relacionamento entre páginas da web por meio de seus hiperlinks;

3. Web Mining de uso: concentra-se em descobrir os padrões de navegação dos usuários.

Este trabalho se situa na subárea da Mineração na Web denominada Web Mining de Conteúdo, uma vez que, serão analisados textos extraídos da rede social Twitter/X.

2.2.2 Mineração de Texto

A quantidade de dados cresce exponencialmente a cada dia devido ao aumento contínuo do volume de informações disponíveis online. Atualmente, a maior parte dos dados governamentais, empresariais e institucionais são armazenados eletronicamente, na forma de bancos de dados de texto. Esses dados geralmente são semi-estruturados, visto que, não são totalmente desestruturados nem completamente estruturados. Por exemplo, um

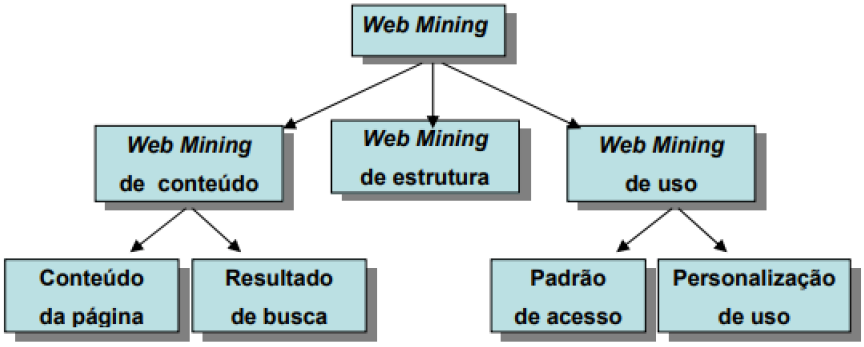

Figura 2 – Abordagens da Mineração na Web. Junior (2007)

documento pode incluir certos elementos estruturados como título, autores, data de publicação e categoria, entre outros, mas também pode conter partes significativas de texto não estruturado, como resumo e conteúdo (SAGAYAM; SRINIVASAN; ROSHNI, 2012).

Sendo assim, a mineração de texto, também conhecida como mineração de dados textuais ou descoberta de conhecimento em bancos de dados textuais, envolve o processo de análise de documentos de texto para extrair padrões e/ou informações significativas e não triviais (WITTEN et al., 2011). Logo, é um campo multidisciplinar baseado em recuperação de informações, mineração de dados, aprendizado de máquina, estatística e linguística computacional (FAN et al., 2006).

A mineração de texto pode ser aplicada em mecanismos de pesquisa, sistemas de gestão de relacionamento com o cliente, filtragem de e-mails, análise de sugestões de produtos, detecção de fraudes e análise de mídias sociais. Neste último caso, se utiliza a mineração de texto para extrair a opinião e/ou sentimento dos internautas e para realizar análise preditiva e de tendências (TALIB et al., 2016) (MÄNTYLÄ; GRAZIOTIN; KUUTILA, 2018). Dessa maneira, a principal área de estudo deste trabalho é a Mineração de Textos, que pode ser compreendida como uma extensão da Mineração de Dados (TAN et al., 1999).

2.3 Etapas do Processamento de Texto

Neste capítulo, serão analisadas e discutidas as etapas da metodologia para aquisição de conhecimento a partir de textos, proposta por Aranha (2007). De acordo com Junior (2007), os procedimentos e técnicas da metodologia analisada são consideradas como "estado da arte"e estão representadas na Figura 3.

De forma geral, Aranha (2007), define que o processo de mineração de texto pode ser realizado em cinco etapas: coleta, pré-processamento, indexação, mineração e análise dos resultados obtidos. A seguir, será detalhada cada uma dessas fases.

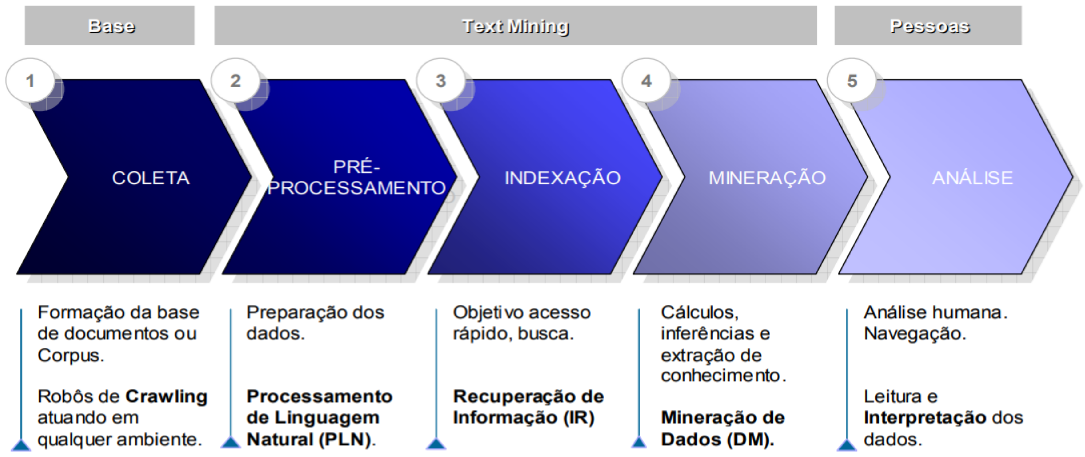

Figura 3 – Etapas do processo de Descoberta de Conhecimento em Textos. Junior (2007)

2.3.1 Coleta

Consiste na etapa inicial do processo onde serão coletadas as informações que farão parte da base textual de trabalho, conhecida como corpus. Um corpus é caracterizado como uma coleção de textos, que pode ser processada por computadores, representando uma ou várias linguagens naturais (GOMES, 2013). A coleta exige um considerável esforço para garantir a qualidade e adequação do material, essencial para a aquisição de conhecimento (JUNIOR, 2007).

O corpus pode ser estático, sendo extraído apenas uma vez, ou dinâmico. No último caso, os dados são atualizados a todo momento através de robôs autônomos que coletam novas informações, sendo que, neste caso, os dados antigos podem ser substituídos ou podem ser adicionados novos dados junto aos antigos (ARANHA, 2007).

Ainda de acordo com Aranha (2007), uma das principais dificuldades na coleta de dados na mineração de texto está em descobrir onde estes dados estão armazenados. Geralmente, os dados são extraídos de três cenários distintos: em diretórios do disco rígido, em tabelas de múltiplos bancos de dados e na Internet.

Para a coleta de dados na internet, como é o caso deste trabalho, é comum empregar ferramentas de suporte, como Motores de Busca Robóticos (Robotic Internet Search Engines) e Diretórios de Assunto (Subject Directories). Crawler ou Web Crawler é o nome dado a esses robôs que navegam na internet de forma autônoma coletando informações de documentos que contenham palavras-chave ou frases de seu interesse (PETERSON, 1997).

2.3.2 Pré-processamento

A etapa de pré-processamento, executada imediatamente após a coleta, visa conferir alguma estrutura à massa textual, ou seja, é nessa etapa que os textos são transformados

em dados estruturados. Além disso, essa fase é frequentemente a mais desafiadora em Mineração de Texto, pois não há uma técnica única que sirva para obter uma representação satisfatória em todos os domínios, sendo que, cada técnica se adequa melhor em determinado contexto (JUNIOR, 2007).

De maneira geral, o pré-processamento tem como intuito aprimorar a qualidade dos dados existentes e organizá-los. As atividades executadas nessa fase têm como objetivo preparar os dados para serem submetidos a algoritmos de indexação ou mineração de dados (ARANHA, 2007).

Nesse processo, palavras são extraídas do documento, e geralmente, um subconjunto delas não é considerado (stopwords), visto que seu papel está relacionado à organização estrutural das sentenças e não tem poder discriminatório sobre diferentes classes. Ademais, para reduzir termos semanticamente relacionados à mesma raiz, um lematizador é aplicado (GONÇALVES et al., 2006).

Nas próximas subseções, será apresentada uma breve explicação de algumas técnicas utilizadas para realizar o pré-processamento.

2.3.2.1 Processamento de Linguagem Natural

Processamento de Linguagem Natural (PLN) refere-se a um conjunto de técnicas teórico-computacionais, projetadas para analisar e representar textos que naturalmente ocorrem em diferentes níveis de análise linguística. O objetivo principal do PLN é alcançar um processamento de linguagem semelhante ao humano, para diversas tarefas (LIDDY, 2001).

Técnicas PLN podem ser utilizadas para auxiliar no pré-processamento dos dados textuais como, por exemplo, a eliminação de palavras irrelevantes (stopwords), a segmentação de palavras, a lematização, entre outras (FILHO, 2014).

2.3.2.2 Tokenização

Nessa fase, ocorre a divisão de um texto em suas unidades elementares, ou seja, é necessário separar cada palavra presente no texto, bem como a pontuação, entre outros elementos. Essas partes elementares são chamadas de tokens (SANTOS, 2010). Como exemplo, a frase "O Projeto de Lei 2630/20 institui a Lei Brasileira de Liberdade, Responsabilidade e Transparência na Internet." possui dezoito tokens, conforme mostra o exemplo abaixo:

[O] [Projeto] [de] [Lei] [2630/20] [institui] [a] [Lei] [Brasileira] [de] [Liberdade] [,] [Responsabilidade] [e] [Transparência] [na] [Internet] [.]

A tokenização é viabilizada pela presença de caracteres de controle de arquivo ou formatação, como espaços ou sinais de pontuação, que, em determinadas situações, podem

ser reconhecidos como tokens delimitadores (SOARES, 2008).

Contudo, a atividade de reconhecimento de tokens, que é relativamente simples para os seres humanos, pode ser complexa ao ser realizada por um computador. Essa dificuldade decorre do amplo leque de funções que os delimitadores podem desempenhar. Por exemplo, o "ponto"pode indicar o final de uma sentença, mas também é empregado em abreviações e números (JUNIOR, 2007).

2.3.2.3 Remoção de Stopwords

Um dos grandes problemas relacionados ao processamento de linguagem natural está no elevado número de dimensões de análise. Dessa forma, para reduzir a dimensionalidade e consequentemente diminuir o tempo de processamento, podemos remover as palavras não discriminantes. Normalmente, as palavras que aparecem com maior frequência nos textos, adicionam pouco valor à análise. Estas são frequentemente chamadas de palavras não discriminantes ou, em inglês, stopwords (GOMES, 2013) e quando agrupadas formam uma lista de palavras sem relevância para a análise, esse grupo recebe o nome de *stoplist* (SANTOS, 2010).

Geralmente, as listas de palavras não discriminantes incluem preposições, substantivos e determinantes. Ademais, o processo de obtenção das stopwords pode ser realizada de maneira manual ou automática (ARANHA, 2007). A Figura 4 ilustra um exemplo de um processo de tokenização seguido da remoção de stopwords.

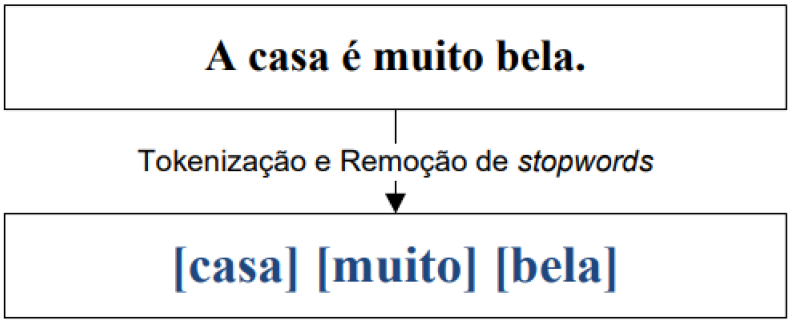

Figura 4 – Processo de tokenização seguido por remoção de stopwords. Soares (2008)

2.3.2.4 Normalização

Refere-se ao processo de Redução de Léxico que se baseia no agrupamento de tokens com o intuito de torná-los mais consistentes e comparáveis. Isso envolve a aplicação de técnicas para lidar com variações morfológicas e problemas de sinonímia. A aplicação dessa técnica emprega uma melhora significativa nos algoritmos de mineração de texto, principalmente para sistemas de classificação (JUNIOR, 2007).

Segundo Santos (2010), dentre as técnicas disponíveis para normalizar as palavras no texto, merecem destaque os métodos de radicalização (stemming) e lematização (lemmatization). O processo de radicalização se dedica a reduzir cada palavra do vocabulário até alcançar sua raiz. Isso proporciona principalmente a remoção de sufixos que indicam variações na forma da palavra, como plural e tempos verbais (JUNIOR, 2007). Na lematização ocorre a substituição das diversas formas de representação da palavra pela forma primitiva. Por exemplo, as palavras “livro”, “livros” e “livraria” compartilham a mesma palavra de origem, “livro” (ARANHA, 2007).

Além das etapas descritas acima, Gomes (2013), acrescenta os processos de eliminação de acentuação e a conversão de letras maiúsculas. Em alguns idiomas, como o inglês, a remoção de acentos não possui grande impacto, uma vez que, a compreensão da palavra não se altera, porém, o mesmo não ocorre com o espanhol. Já a conversão de letras maiúsculas é uma prática comum e possibilita que palavras iniciadas com letra maiúscula não sejam interpretadas pelo software como termos distintos quando estão escritas integralmente em minúsculas.

Ademais, considerando o contexto deste trabalho, Filho (2014) acrescenta que no processo de pré-processamento também podemos descartar conteúdos irrelevantes dos tweets como, por exemplo, links, nomes de usuário (marcados pelo caractere "@') e caracteres não alfabéticos, com exceção do caractere "#", que é utilizado para marcar uma palavra como hashtag.

2.3.3 Indexação

Após a etapa de Pré-Processamento, começa-se a fase de Indexação. Esse processo envolve a criação de estruturas complementares chamadas índices, que asseguram uma recuperação rápida e eficiente de documentos e seus termos (SOARES, 2008). As técnicas de indexação de documentos tornaram-se amplamente difundidas devido à crescente demanda e expansão da área de Recuperação de Informações (ARANHA, 2007).

Esta fase não foi considerada crucial para os objetivos deste trabalho, já que não foi necessário efetuar pesquisas nos textos que formam a base de dados. Portanto, esta etapa não será abordada com a mesma profundidade das demais.

2.3.4 Mineração

Durante a fase de mineração, algoritmos são aplicados sobre os dados para abstrair o conhecimento implícito (SOARES, 2008) e identificar padrões interessantes (FILHO, 2014). De acordo com Aranha (2007), nesse estágio é possível empregar algoritmos provenientes de diversas áreas do conhecimento, como Aprendizado de Máquina, Estatística, Redes Neurais e Banco de Dados. Porém, algoritmos que apresentam comportamento do tipo “caixa preta” podem dificultar a interpretação dos resultados.

Algoritmos de classificação e clusterização, além de processos de sumarização e extração de características podem ser empregados neste processo. Os algoritmos de classificação servem para classificar cada documento, como pertencente ou não a uma classe específica. Já a clusterização envolve o agrupamento de textos em conjuntos de textos semelhantes, conhecidos como clusters. No caso deste último, a dificuldade se encontra no fato de que não se tem conhecimento prévio da quantidade de clusters que uma coleção de textos contém. A sumarização tem como objetivo resumir automaticamente o conteúdo do texto de forma a obter um texto mais conciso que original, mas, mantendo seu significado principal. Por fim, a extração de informação visa obter dados estruturados a partir de dados não estruturados (formato em que os textos, geralmente, estão representados). Os principais dados textuais são convertidos em formato tabular, permitindo a aplicação de métodos de mineração de dados conhecidos (SANTOS, 2010).

Em suma, se o usuário busca identificar relações entre documentos, avaliando a semelhança e a formação natural de grupos, deve-se optar pela clusterização. Por outro lado, se esses grupos de documentos já estão definidos, seja por algoritmos ou conhecimento prévio, a classificação será indicada quando se quer determinar em qual grupo um novo documento se encaixa. Embora a clusterização e a classificação sejam tarefas compartilhadas entre Mineração de Textos e Mineração de Dados, a sumarização e extração de características são específicas da primeira (JUNIOR, 2007).

2.3.4.1 Naive Bayes

O algoritmo de classificação Naive Bayes usa métodos probabilísticos e estatísticos baseados no Teorema de Bayes, representado pela Equação 1, considerando fortes suposições de independência entre as variáveis (DEY et al., 2016). O Naïve Bayes é um dos algoritmos mais populares em mineração de dados devido ao seu rápido processamento, fácil implementação e alto nível de eficácia (NOVENDRI et al., 2020). Este algoritmo calcula a probabilidade de uma classe com base em seus atributos e determina a classe que tem a maior probabilidade, assumindo que cada atributo nos dados é mutuamente exclusivo. Especificamente, em mineração de texto, o Naive Bayes irá calcular a probabilidade de um documento pertencer a uma determinada classe (categoria) com base nas frequências das palavras no documento (DEY et al., 2016).

$$P(C|X) = \frac{P(X|C) * P(C)}{P(x)} \tag{1}$$

Além disso, de acordo Dey et al. (2016), podemos assumir que, para um conjunto de dados x = x1,x2,...xj, onde vários atributos serão utilizados para determinar o resultado de uma classe, podemos reescrever a equação acima como:

$$P(C|\mathbf{X}) \propto P(C) \prod_{i=1}^{n} P(X_i|C) \tag{2}$$

2.3.4.2 Support Vector Machine

Em tarefas de classificação, o Support Vector Machine (SVM) é uma técnica de aprendizado de máquina discriminante que visa encontrar, com base em um conjunto de dados de treinamento independente e identicamente distribuído (iid), uma função discriminante que possa prever corretamente rótulos para novas instâncias adquiridas (AWAD; KHANNA, 2015). Em outras palavras, o algoritmo produz um ótimo hiperplano que categoriza novas instâncias com base em dados de treinamento rotulados (aprendizado supervisionado) (TRIPATHI, 2021).

Ao contrário das abordagens de aprendizado de máquina generativas, que requerem cálculos de distribuições de probabilidade condicional, uma função discriminante de classificação recebe um ponto de dados x e o atribui a uma das diferentes classes que fazem parte da tarefa de classificação. De uma perspectiva geométrica, aprender um classificador é equivalente a encontrar a equação para uma superfície multidimensional que melhor separa as diferentes classes no espaço de características (AWAD; KHANNA, 2015).

Os principais componentes do SVM podem ser vistos na Figura 5. Nela estão a equação dos vetores de suporte, separando o hiperplano, e também a equação que representa a distância entre vetores de suporte e a distância entre um plano de suporte e um vetor de suporte.

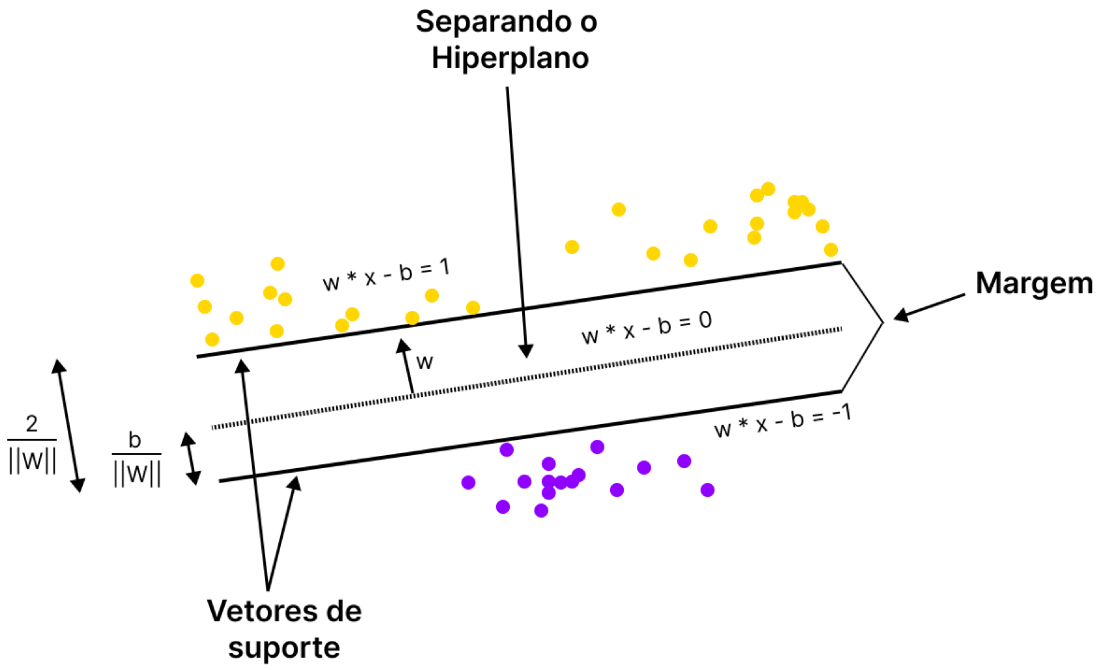

Figura 5 – Classificador SVM. Adaptado de Tripathi (2021)

2.3.4.3 Árvore de Decisão

A Árvore de Decisão (do inglês, Decision Tree) é um algoritmo que pode ser usado tanto para regressão quanto para classificação, podendo lidar com dados categóricos e também numéricos (AHUJA et al., 2019). Dessa maneira, uma árvore de decisão é uma estrutura de dados hierárquica em que cada nó interno representa uma condição em um atributo, cada ramo representa o resultado dessa condição, e cada folha representa a classe de saída ou valor de regressão (SURESH; BHARATHI, 2016).

O algoritmo básico para construir uma árvore de decisão é um algoritmo ganancioso que constrói a árvore de forma recursiva, começando do topo e dividindo os dados de treinamento de acordo com os valores dos testes selecionados. O processo começa com o conjunto de dados de treinamento, onde o algoritmo seleciona a melhor característica e a usa para dividir os dados, maximizando a informação para classificação. Um nó de teste é criado para representar essa divisão, e então o processo de construção da árvore continua, recursivamente, dividindo as tuplas de dados restantes com base nos valores dos testes selecionados (SURESH; BHARATHI, 2016). Pode-se utilizar o índice de Gini e também o parâmetro de ganho de informação para decidir qual atributo será usado para divisão adicional do conjunto de dados. Se usado o índice de Gini, a árvore de decisão é chamada Árvore de Classificação e Regressão (CART), e, se usarmos ganho de informação então ele é chamado de ID3 (AHUJA et al., 2019).

2.3.4.4 Avaliação do Modelo de Predição

Após o modelo de predição ser desenvolvido com o intuito de predizer dados diferentes daqueles utilizados em sua construção, existe a necessidade de avaliar sua capacidade de realizar predições precisas. Normalmente, essa avaliação ocorre dividindo-se os dados em dois conjuntos distintos: treinamento e teste. O modelo é então construído usando o conjunto de treinamento e, posteriormente, aplicado aos dados do conjunto de teste para realizar previsões. Essa abordagem permite avaliar o desempenho do modelo e estimar a incerteza de suas previsões (FONTOURA, 2011).

Levando em consideração o contexto de que os dados serão classificados em três classes "positivo", "negativo"ou "neutro", e, sabendo que o modelo foi construído utilizando parte da base de dados como conjunto de treinamento, é possível obter as seguintes medidas (CORRÊA et al., 2017):

- VP (verdadeiros positivos): número de instâncias positivas classificadas corretamente;
- VN (verdadeiros negativos): número de instâncias negativas classificadas corretamente;
- VE (verdadeiros neutros): número de instâncias neutras classificadas corretamente;

- FP (falsos positivos): número de instâncias negativas ou neutras classificadas como positivas;
- FN (falsos negativos): número de instâncias positivas ou neutras classificadas como negativas;
- FE (falsos neutros): número de instâncias positivas ou negativas classificadas como neutras.

A partir dessas medidas cria-se uma matriz de confusão que é uma forma de visualização dos resultados utilizada para avaliar modelos de predição. que utilizam algoritmos de classificação. Cada linha da matriz representa as instâncias preditas de uma classe, enquanto cada coluna da matriz representa as instâncias reais de uma classe (FONTOURA, 2011), como mostrado na Tabela 1.

Dados classificados como				
Positivo	Negativo	Neutro		
VP	FN	FE	Positivo	
FP	VN	FE	Negativo	Valor real
FP	FN	VE	Neutro	

Tabela 1 – Matriz de confusão para três classes: positivo, negativo e neutro

De acordo com Schmitt (2013), a partir da matriz de confusão é possível avaliar o desempenho do modelo a partir de métricas de qualidade como: Medida de Precisão (Precision), Revocação (Recall), Medida-F (F-measure) e Acurácia (Acurracy). Abaixo, estão as definições e as fórmulas de cada medida:

Acurácia: Calcula a proporção entre verdadeiros positivos e negativos e sobre falsos positivos e negativos, ou seja, calcula quanto percentual de casos está classificado corretamente. Sua fórmula é:

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

Precisão: A proporção entre observações positivas previstas e o número total de observações positivas é conhecida como precisão. Isso é calculado da seguinte forma:

$$Precição = \frac{VP}{VP + FP}$$

Recall: Descreve a proporção de observações positivas previstas corretamente.

$$Revocação = \frac{VP}{VP + FN}$$

F-score: É a média harmônica entre os valores de Recall e Precision. Pode ser obtida pela seguinte formula:

$$F\text{-}Score = 2 \times \frac{Precição \times Revocação}{Precição + Revocação}$$

2.3.5 Análise Dos Resultados Obtidos

A etapa de Análise, ou Pós-Processamento, como também é chamada, refere-se à avaliação da eficiência da aplicação dos algoritmos na fase anterior (JUNIOR, 2007). Para avaliar os resultados do processo de Mineração de Textos, são empregadas métricas de desempenho (SOARES, 2008). Estas podem ser quantitativas como a taxa de erro, o tempo de CPU e a complexidade do modelo ou qualitativa, uma vez que, um especialista também pode verificar a compatibilidade dos resultados através do conhecimento prévio do domínio (JUNIOR, 2007). Além disso, cabe ao usuário fazer uma avaliação final sobre a aplicabilidade dos resultados (ARANHA, 2007).

Essa etapa merece uma atenção especial uma vez que as informações extraídas podem não estar claras para o usuário. Dessa maneira, para facilitar a interpretação da saída do processo, gráficos, tabelas e outros elementos gráficos podem ser explorados através de ferramentas de visualização. Neste caso, a utilização de cores e distâncias em gráficos, auxilia na compreensão do sentido de grandes e complexos conjuntos de dados, tornando-os mais acessíveis e compreensíveis (JUNIOR, 2007). Por fim, é importante ressaltar que ao finalizar esta etapa, pode-se optar por retroceder para qualquer fase anterior do processo com o intuito de corrigir qualquer problema encontrado (SANTOS, 2010).

2.4 Análise de Sentimento / Mineração de Opinião

As opiniões desempenham um papel crucial em diversas atividades humanas, influenciando significativamente nossos comportamentos, uma vez que, estamos sempre buscando a opinião dos outros quanto precisamos tomar alguma decisão. Empresas, estão constantemente atrás da opinião dos consumidores sobre os seus produtos e serviços. Os consumidores também querem saber a opinião de outras pessoas antes de comprar determinado produto. Ademais, opiniões de outros sobre candidatos políticos podem influenciar na decisão de voto em uma eleição política. Com o avanço das mídias sociais na internet, o conteúdo gerado, como análises, discussões em fóruns e postagens em redes sociais, tornou-se uma fonte essencial para a tomada de decisões, refletindo a crescente importância das opiniões na sociedade atual (LIU, 2022).

A busca e análise podem ser conduzidas manualmente quando há poucas opiniões. Contudo, à medida que a quantidade de informações aumenta, essa abordagem torna-se trabalhosa. Para facilitar a tarefa, surgiu uma nova área de pesquisa conhecida como Análise de Sentimentos e/ou Opiniões (OLIVEIRA, 2013). Liu (2022), define a Análise de Sentimentos (AS), também conhecida como Mineração de Opinião, como um campo de

estudo que examina as opiniões, sentimentos, avaliações, atitudes e emoções das pessoas em relação a entidades como produtos, serviços, organizações, indivíduos, eventos, tópicos e seus atributos. Em suma, a análise de sentimento se preocupa em determinar a polaridade das expressões em textos, geralmente classificando-as como positivas ou negativas em relação a uma entidade (OLIVEIRA, 2013).

Matos, Magalhães e Souza (2020), definem Análise de Sentimentos como um campo que investiga a extração de opiniões utilizando Recuperação de Informação (RI), Inteligência Artificial (IA) e técnicas de Processamento de Linguagem Natural (PLN). Para França e Oliveira (2014), o propósito da análise de sentimentos é determinar, por meio da avaliação dos termos presentes no texto, se o documento em análise expressa uma opinião positiva, negativa ou neutra.

Para Silva, Lima e Barros (2012), a Análise de Sentimentos, entre outras funções, dedica-se a categorizar as opiniões manifestadas em textos sobre um objeto específico (produto, serviço, instituição ou pessoa) como positivas (por exemplo, o Twitter/X é ótimo! (+)) ou negativas (por exemplo, o Twitter/X é a pior rede social que existe (-)). Em geral, a polaridade de um texto é revelada por meio de palavras opinativas, como adjetivos (bom, ruim), advérbios (pouco, muito) e alguns substantivos (amigo, inimigo).

As palavras opinativas, do inglês opinion words, são palavras que estão associadas a um sentimento como, por exemplo, “bom”, “maravilhoso"e “incrível” são palavras positivas, enquanto, “ruim”, “medíocre” e “terrível” são palavras negativas. Uma lista de palavras opinativas é chamado de opinion lexicon ou dicionário de opiniões, Contudo, um dicionário pode ser ineficaz para realizar a análise de opinião uma vez que as palavras podem ter sentimentos diferentes quando considerados domínios diferentes. Algumas frases não contêm avaliações (por exemplo, perguntas, condicionais, comparativas, etc.), enquanto sentenças irônicas e/ou sarcásticas podem introduzir interferências nos dados (KAUER, 2016).

A Análise de Sentimento pode ser conduzida em três níveis de granularidade: (1) no nível do documento, avaliando o sentimento global expresso no documento; (2) no nível da sentença, classificando a polaridade de cada frase no texto; e (3) no nível de entidades e aspectos em que se considera que cada opinião inclui um sentimento (positivo ou negativo) e um alvo (da opinião). Um documento pode abordar opiniões sobre vários aspectos de várias entidades, conferindo assim, o máximo de detalhes sobre as preferências e desagrados do usuário (SILVA; LIMA; BARROS, 2012; KAUER, 2016).

Em conformidade com Liu (2022) e Kauer (2016), para a determinar a polaridade de cada opinião, existem duas formas principais:

1. Atribuição baseada em aprendizado supervisionado: Esta técnica utiliza um método de aprendizado supervisionado em nível de sentença para definir a polaridade de uma opinião. Nesse caso, a sentença pode representar o escopo (ou seja, as palavras relacionadas ao aspecto de interesse) da expressão de sentimento associada à opinião.

Essa abordagem torna o modelo dependente dos dados de treinamento, resultando em possíveis desempenhos inferiores ao ser aplicado em domínios distintos.

2. Atribuição baseada em informações léxicas: Esses métodos, geralmente supervisionados, utilizam dicionários de opiniões e recursos de processamento, como analisadores gramaticais, árvores de dependências, para determinar a polaridade de uma opinião.

A análise de sentimento é comumente abordada como um problema de classificação, dessa forma ela pode ser utilizada para classificar dados textuais (FILHO, 2014). Assim, dois subtemas têm sido extensivamente pesquisados: (i) categorizar as opiniões expressas em um documento como negativas ou positivas e, (ii) classificar uma sentença como objetiva ou subjetiva, e, para as sentenças subjetivas, categorizar as opiniões como negativas, positivas ou neutras (GOMES, 2013). Vale ressaltar que textos que expressam opiniões são considerados subjetivos, ao contrário dos textos objetivos, que consistem apenas em fatos (SILVA; LIMA; BARROS, 2012).

2.4.1 Etapas da Análise de Sentimentos

De acordo com Oliveira (2013), a análise de sentimento geralmente possui as seguintes fases: seleção da fonte de dados e a busca relacionada à entidade de interesse, processamento dos dados, análise e apresentação dos resultados, conforme mostra a Figura 6.

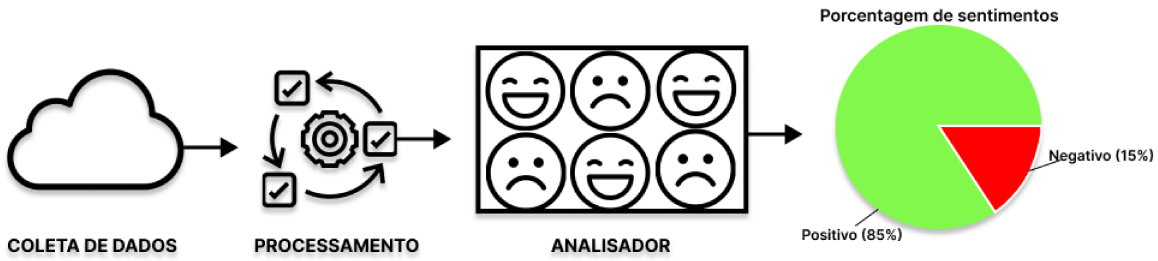

Figura 6 – Etapas de um analisador de sentimentos. Adaptado de Oliveira (2013)

As etapas são definidas por:

1. Coleta: A escolha da fonte de dados para a análise de sentimentos varia de acordo com o tipo de aplicação desejada. Se o objetivo é capturar opiniões do público em geral de forma online, redes sociais como Facebook, Instagram e Twitter/X são fontes ricas em informações. Geralmente, essas plataformas oferecem APIs para facilitar o acesso aos dados, que geralmente são comentários gerados pelos usuários;

2. Processamento: Também chamada de pré-processamento, é nesse estágio que os dados são tratados para corrigir problemas de escrita e redução de dimensionalidade como, por exemplo, remoção de stopwords, normalização, tratamento de gírias e abreviaturas, entre outros;

3. Análise: É a principal fase do processo, onde ocorre a classificação dos textos, em positivo, negativo ou neutro, através de algoritmos de aprendizado de máquina ou técnicas baseadas em recursos léxicos;

4. Apresentação dos resultados: É o último estágio da análise de sentimentos, onde os resultados são apresentados de forma clara e objetiva ao usuário através de textos ou recursos visuais, como gráficos.

2.4.2 Desafios da Análise de Sentimentos em Redes Sociais

A aplicação da análise de sentimentos em redes sociais é útil para compreender o senso comum sobre determinado acontecimento como, eventos, economia e política (KANSAON; BRANDÃO; PINTO, 2018). Além disso, a AS em redes sociais geralmente realiza análise em nível de documento, dada a complexidade associada à análise em nível de aspectos e ao considerável volume de dados (LIU, 2022; KAUER, 2016). A categorização de opiniões em textos provenientes das redes sociais apresenta desafios devido ao elevado grau de informalidade, utilização de gírias e à expressão de diversas emoções por meio de desenhos e símbolos (ROSA, 2015).

Além das dificuldades inerentes aos sistemas de análise de sentimentos ao lidar com textos convencionais, a análise de sentimentos por meio do Twitter/X, como abordado neste estudo, enfrenta desafios adicionais, conforme destacado por Silva (2016) e Oliveira (2013). Eles são:

1. O uso de dialetos regionais: Como a análise é realizada em textos livres escritos por diversos usuários, não há um padrão linguístico uniforme, o que gera dificuldades na extração dos termos a serem analisados. Isso ocorre frequentemente devido ao uso de dialetos que variam conforme a cultura e a localização dos usuários;

2. Ambiguidade dos comentários ou termos: Esse é um problema muito comum, uma vez que, termos podem ter diferentes significadas de acordo com o contexto, causando erro no resultado apresentado ao usuário;

3. Detecção de ironia e sarcasmo nos comentários: O uso de ironia e sarcasmo dificulta a análise dos textos uma vez que inverte o sentido dos termos usados (uma mensagem positiva torna-se negativa ou vice-versa).

4. O tamanho do texto a ser analisado: Quando há limite de caracteres do texto, como no caso do Twitter/X, os usuários empregam muitas abreviaturas e gírias para

expressar suas opiniões. Tais palavras podem não estar no dicionário de palavras ou podem não ter ocorrido nos dados de treinamento.

5. Variação na ortografia: Devido à natureza espontânea, o contexto informal e as limitações de comprimento das mensagens, a ortografia nos tweets apresenta maior variabilidade em comparação com outros gêneros textuais, como páginas da web, blogs e jornais. Diversos fenômenos decorrem dessa variação ortográfica, incluindo erros ortográficos, abreviações, utilização de maiúsculas e repetições de letras e sílabas para atribuir ênfase.

6. Esparsidade dos dados: Tweets apresentam considerável interferência devido à utilização frequente de erros de ortografia. Esse fenômeno resulta em uma escassez de dados e afeta o desempenho global da análise de sentimentos. A principal causa dessa escassez é a constatação de que uma parcela significativa dos termos presentes nos tweets ocorre em menos de 10 vezes em todo o conjunto de dados.

7. Stop Words: Geralmente as listas de stopwords não são adequadas para o Twitter/X. Por exemplo, a palavra "like"em inglês, embora frequentemente incluída nas listas de palavras irrelevantes, possui uma significativa capacidade de discriminação de sentimentos.

8. Símbolos especiais (Special tokens): *Emoticons* e URLs podem dificultar o uso de ferramentas de processamento de linguagem natural.

9. Quantidade de Dados: Apesar do tamanho dos textos serem limitados, milhares de tweets são publicados por minuto. Em um ambiente de grande escala como esse, uma opinião específica sobre um tópico fica totalmente oculta, tornando praticamente inviável para um usuário comum extrair conteúdo útil através de diversas fontes.

10. Contexto Multilingual: Usuários do Twitter/X podem se expressar em vários idiomas, até mesmo na mesma publicação.

Capítulo 3

Metodologia

3.1 Coleta de Tweets

A primeira etapa deste trabalho consiste na coleta de dados. Inicialmente, a coleta seria realizada utilizando a API oficial do Twitter/X. Contudo, após analisar a documentação, constatou-se que essa ferramenta não seria adequada para este estudo, uma vez que, ela apenas recupera tweets publicados nos sete dias anteriores à data de consulta e retorna um número limitado de instâncias. Como para este estudo, o período de interesse compreende o intervalo entre o fim da eleição presidencial em 2022 até o mês de janeiro de 2024, foi necessário buscar uma ferramenta que permitisse recuperar tweets mais antigos. Portanto, para este trabalho, optou-se por utilizar um scraper da plataforma Apify, uma solução de automação e web scraping, chamado Tweet Flash, que permite recuperar tweets publicados em qualquer data na rede social. Por meio de sua interface, é possível fazer consultas ao Twitter/X e recuperar tweets conforme os argumentos especificados, como ilustrado na Figura 7.

Os argumentos utilizados nas consultas são descritas a seguir:

- Search Terms: são as palavras-chaves que serão utilizadas nas buscas dos tweets. As palavras-chaves usadas como parâmetro foram “PL2036”, “PL2036SIM” e “PL2036NÃO”

- Max Tweets Per Query/Profile/URL: o número máximo de tweets a serem raspados por palavra-chave. Foi definido o número de 20 mil tweets por palavra-chave.

- Tweet Language: definição do idioma dos tweets, neste contexto, português.

- Start: indica a data mais antiga (inclusive) a partir da qual será realizada a consulta. A data deve estar no formato ’AAAA-MM-DD’. A data “2022-10-30” foi utilizada como parâmetro.

- End: Indica a data mais recente (exclusiva) a ser considerada na consulta. A data deve estar no formato 'AAAA-MM-DD'. A data "2024-01-31" foi utilizada como parâmetro.

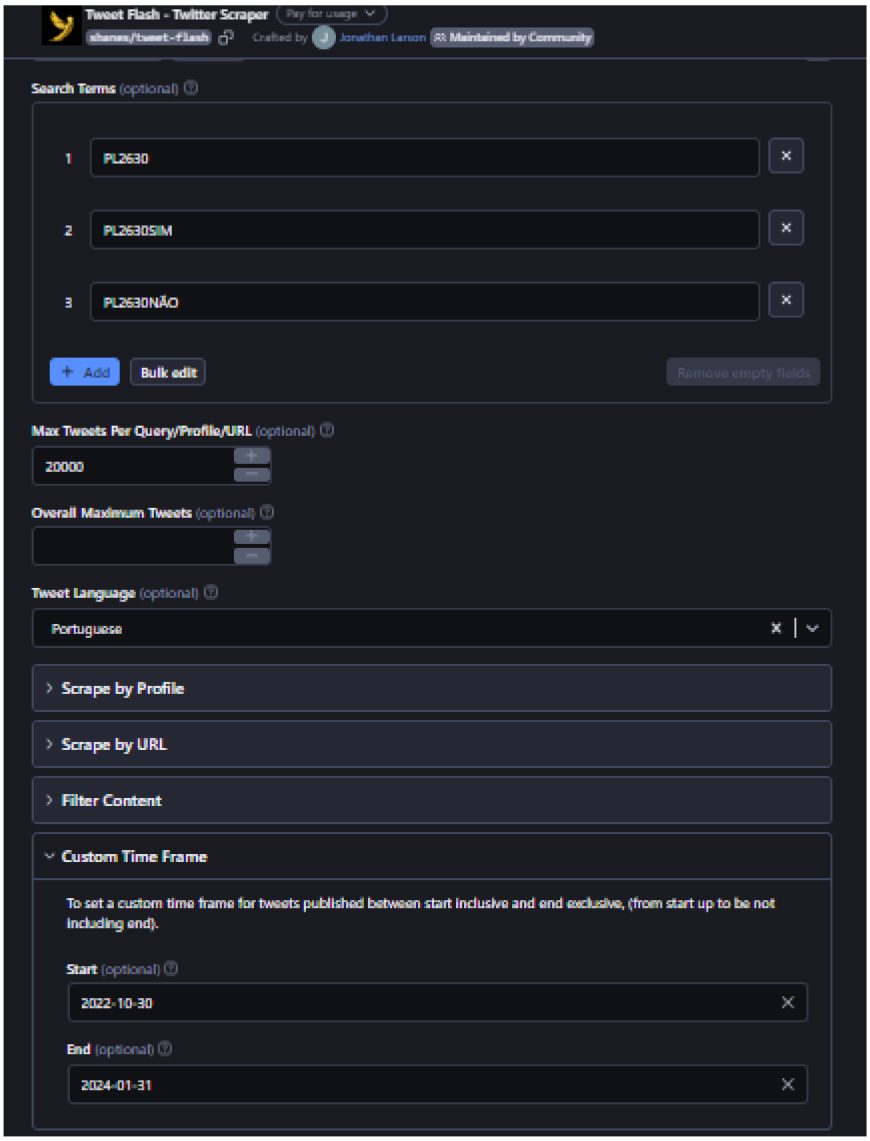

Figura 7 – Consulta no Twitter/X realizada por meio do scraper Tweet Flash

Ao final da consulta, um arquivo CSV *(Character-separated values)* foi gerado contendo os tweets e seus metadados como, por exemplo, número de curtidas, retweets, respostas, nome do usuário, data da publicação, hashtags utilizadas, usuários mencionados e o link do tweet.

3.2 Pré-processamento

Após a coleta dos tweets é necessário aplicar técnicas de pré-processamento que envolvem a transformação, limpeza, seleção e redução de volume dos dados, antes da etapa

de mineração (MORAIS; AMBRÓSIO, 2007). As etapas utilizadas neste trabalho foram implementadas por meio da linguagem de programação Python e estão listadas abaixo:

- Transformação de letras maiúsculas em minúsculas com o intuito de padronizar o texto e evitar que palavras sejam interpretadas de forma diferente;

- Remoção de caracteres não alfabéticos e pontuação, uma vez que, não contribuem para a classificação.

- Remoção de stopwords, isto é, eliminação de palavras irrelevantes que não alteram o sentido do texto como, pronomes, artigos, preposições e outras palavras auxiliares. Para isso foi utilizada a biblioteca NLTK *(Natural Language Toolkit)* que contém listas de stopwords em diversos idiomas;

- Aplicação de lematização por meio da biblioteca NLP (Processamento de Linguagem Natural) para agrupar as formas flexionadas de uma palavra para que possam ser analisadas como um único item, identificado pelo "lema"da palavra.

- Tratamento de gírias, abreviações, correção ortográfica e remoção de emojis através da biblioteca *Enelvo*.

- Conversão dos textos em vetores numéricos. Para isso, foi empregada a técnica de TF-IDF cujo objetivo é identificar quais palavras têm mais importância em um documento (SHOUZHONG; MINLIE, 2016). Portanto, o TF-IDF é influenciado pela quantidade de vezes que uma palavra aparece no texto, ou seja, sua frequência. No entanto, para balancear esse número, o algoritmo considera também a frequência do termo em todo o corpus. Essa técnica é dividida em duas etapas: o cálculo da frequência do termo *(TF ou Term Frequency)* e o cálculo da frequência inversa *(IDF ou Inverse Document Frequency)* (CAPOBIANCO, 2016);

- Remoção de links.

A Tabela 2 faz uma comparação de como ficou uma amostra de tweets após a aplicação do pré-processamento.

Tabela 2 – Exemplos de tweets antes e após a etapa de pré-processamento

Tweets antes do pré-processamento	**Tweets após o pré-processamento**
Processo vem contra quem difama e publica mentiras sobre, irmão. Acredito que vc não seja burro (espero) Leia sobre a PL2630 e verás que o governo atual que quer censura. Ou viva na sua ignorância né... Meu limite para te responder já acabou. Foi bom o papo	processo difama publicar mentira irmao acreditar nao ser burro esperar ler vera governo atual censura vivo ignoranciar limete responder acabar papo
O dia de ontem serviu para dar munição para o congresso dar andamento a pl2630, o evento de ontem não foi aleatório, tem método, essa galera não estão nem aí para o povo e as coisas importantes para o País, é um projeto de poder.	dia servir municao congresso andamentar evento nao aleatorio metodo galerar nao estao coisa importante pai projeto
Vcs também deveriam pedir urgência na votação do PL das FAKE News, ahh lembrei SEM MENTIRA A EXTREMA DIREITA SOME NÉ?	voces tambem dever pedir urgência votacao Fake news lembrar mentira extremo some

3.3 Criação da Base de Dados Rotulada

Algoritmos de classificação como o SVM *(Support Vector Machine)*, *Naive Bayes* e *Decision Tree* são modelos supervisionados. Logo, se faz necessário ter uma base rotulada para treinamento. Em outras palavras, para que esses algoritmos aprendam a classificar novos dados, eles precisam ser treinados em um conjunto de dados em que as entradas estejam associadas a rótulos conhecidos.

Para criar esta base utilizou-se o método de *NRC Emotion Lexicon* (MOHAMMAD; TURNEY, 2013) que consiste em uma lista de palavras e suas associações com oito emoções básicas (raiva, medo, antecipação, confiança, surpresa, tristeza, alegria e nojo) e dois sentimentos (negativo e positivo). Este dicionário possui cerca de 14.000 palavras em 105 idiomas, incluindo o português. Quando uma palavra é encontrada no dicionário, sua associação com uma emoção é verificada, permitindo a identificação da relação entre a palavra e uma emoção específica (MARTINS; ARAÚJO, 2021). No. caso deste trabalho, somente utilizaremos os sentimentos (negativo e positivo).

Para o processo de identificação dos sentimentos utilizou-se a função `get_nrc_sentiment` presente no software R por meio do pacote *syuzhet*. Esta função lê o documento, com-

parando cada palavra com as presentes no dicionário NRC e atribui pesos às emoções correspondentes com base nessa associação. Esses pesos podem ser usados para determinar a intensidade das emoções presentes no texto. Ao final, a função gera um arquivo no formato de um data frame, com os sentimentos organizados em colunas, uma para cada sentimento. Cada linha desse documento representa um tweet. Para facilitar a interpretação dos dados, as colunas referentes aos sentimentos foram traduzidas para o português do Brasil. A Tabela 3 ilustra a saída da função.

Tabela 3 – Trecho do data frame gerado pela função `get_nrc_sentiment`

Tweet	**Negativo**	**Positivo**
engracar mundo brincar zoar programa rir edicao engracar claaaro vcs precisar arranjar minima pra manchar imagem idol admirar educacao humildade industria urgente Fake news	1	2
arthur lira responsavel aberracao passado camar politico tratorar democracia interesse de ele ganhar comporta tambem aprovar custo Fake news	0	6
deitar abestar falar grave cometer inocente desconexo defender paria inutel igual cadeia verme urgente votacao Fake news	6	3
Internet nao terra ninguem continuar cometer ato ilicito falso bandeira liberdade expressao regular proteger usar rede mundial computador pra big techs nao lucr dor video slpnggiantspt	8	6
olhar ninguem gostar acusar falsamente nao olhar legal nao precisar justica arcar responsabilidade ato resolver	3	3

Por fim, foi adicionada a coluna “Polaridade”, que indica se o comentário tem um sentimento predominantemente positivo, neutro ou negativo. Uma amostra da base de dados é exemplificada na Tabela 4.

Tabela 4 – Data Frame Tweets X Polaridade

Tweet	**Polaridade**
engracar mundo brincar zoar programa rir edicao engracar claaaro vcs precisar arranjar minima pra manchar imagem idol admirar educacao humildade industria urgente Fake news	Positivo
arthur lira responsavel aberracao passado camar politico tratorar democracia interesse de ele ganhar comporta tambem aprovar custo Fake news	Positivo
deitar abestar falar grave cometer inocente desconexo defender paria inutel igual cadeia verme urgente votacao Fake news	Negativo
Internet nao terra ninguem continuar cometer ato ilicito falso bandeira liberdade expressao regular proteger usar rede mundial computador pra big techs nao lucr dor video slpnggiantspt	Negativo
olhar ninguem gostar acusar falsamente nao olhar legal nao precisar justica arcar responsabilidade ato resolver	Neutro

A Tabela 5 sumariza os dados contidos na base de tweets rotulados

Tabela 5 – Sumarização da base de tweets rotulados

Polaridade	**Quantidade de Tweets**
Positivo	17.918
Negativo	18.688
Neutro	18.164

3.4 Classificação dos Tweets

Foram escolhidos três algoritmos de classificação, Naive Bayes multinomial, SVM e Decision Tree, para classificar os tweets como "positivo", "negativo" ou "neutro". Em suma, o funcionamento dos três classificadores ocorre da seguinte maneira: o corpus é dividido em dois conjuntos, denominados treino e teste. A partir dos dados de treinamento, é criado o modelo de classificação dos tweets. O modelo, então, associa as palavras às categorias em que ocorrem nos dados de treinamento, permitindo, dessa forma, que o algoritmo aprenda a rastrear a possibilidade de associar um tweet a um determinado sentimento a partir da presença das palavras associadas a esse tweet. Posteriormente, o modelo de classificação é avaliado utilizando o conjunto de testes para validar sua acurácia. Ademais, para mensurar o impacto da etapa de pré-processamento na classificação dos tweets, optou-se por aplicar os algoritmos antes e após esta etapa. Métricas como acurácia, tempo de execução e matriz de confusão foram utilizadas para mensurar o desempenho dos classificadores.

3.5 Apresentação dos Resultados

3.5.1 Gráficos de Série Temporal

Após a classificação dos tweets, foram desenvolvidos gráficos de linha para avaliar o comportamento dos sentimentos dos internautas, em relação ao Projeto de Lei 2630/2020, ao longo do período de coleta dos dados. Além disso, foi realizada a associação entre os picos na quantidade de tweets em determinado período com fatos do cotidiano, bem como, a avaliação de como esses acontecimentos impactaram no sentimento dos usuários.

3.5.2 Nuvem de Palavras

Este trabalho apresenta nuvens de palavras, uma representação visual das palavras mais frequentes em um texto, sendo que, quanto maior for o número de ocorrências de uma palavra, maior a mesma será na nuvem de palavras. Este estudo apresenta quatro nuvens de palavras, sendo uma gerada usando toda a base de dados e outras três geradas somente usando tweets de cada um dos sentimentos (positivo, negativo e neutro). Um exemplo de nuvem de palavras é apresentado abaixo, na Figura 8. Na ocasião, a nuvem foi gerada a partir dos tweets coletados no mês de janeiro de 2024.

Figura 8 – Nuvem de palavras do mês de janeiro de 2024. Elaborado pelo autor

3.5.3 Grafo

Por fim, gerou-se um grafo para entender como os usuários estão conectados com o objetivo de reconhecer padrões de comunicação e/ou grupos de internautas que estão interagindo mais frequentemente. Para isso, o grafo foi criado utilizando a NetworkX, uma biblioteca python comumente utilizada para manipulação de grafos e redes complexas. Após gerado, o grafo foi armazenado em um arquivo no formato GraphML. Por sua vez, o software Gephi foi empregado para melhorar a visualização do grafo e aplicar algoritmos de clusterização para formação dos grupos de usuários.

Capítulo 4

Análise dos Resultados

4.1 Quantidade de Tweets

A Figura 9 mostra a quantidade de tweets recuperada a partir da coleta de dados de acordo com o mês. Pode-se perceber que a maior ocorrência de tweets ocorreu nos meses de abril, maio e dezembro de 2023.

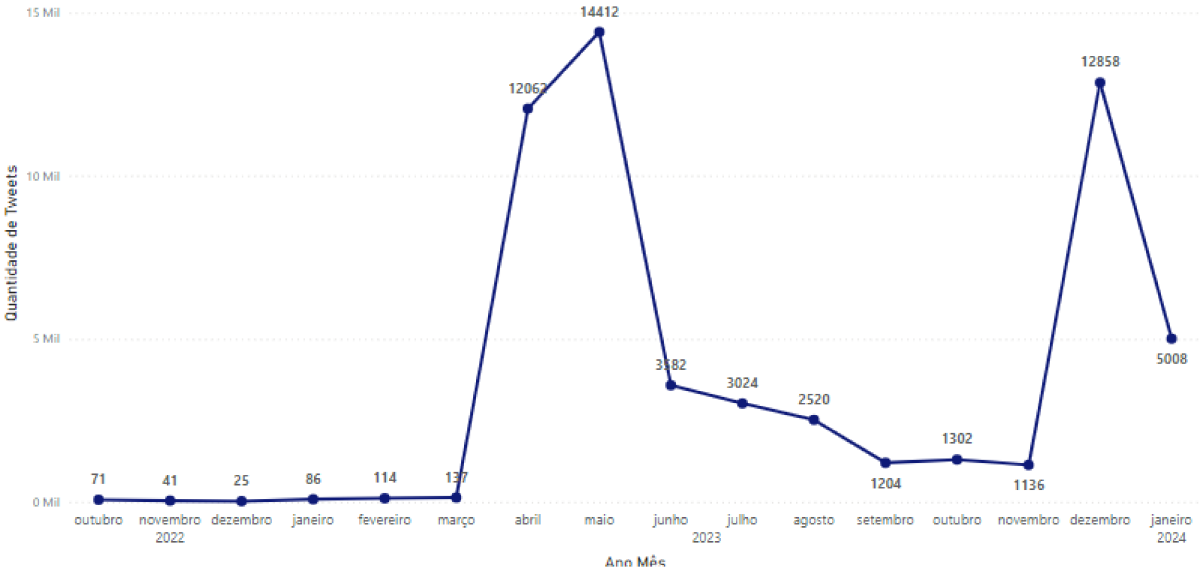

Figura 9 – Quantidade de tweets publicados por mês.

Conforme exposto na subseção 3.5.1, buscou-se associar os picos de tweets a acontecimentos do cotidiano. Dessa forma, diminui-se a granularidade dos dados nos meses que apresentaram maior número de tweets, para analisar como o volume de publicações se comportou nos dias dos meses em questão , bem como, qual fato proporcionou o aumento da discussão sobre a PL 2630. Assim, as Figuras 10, 11 e 12, trazem o número de tweets por dia, para os meses de abril, maio e dezembro de 2023, respectivamente.

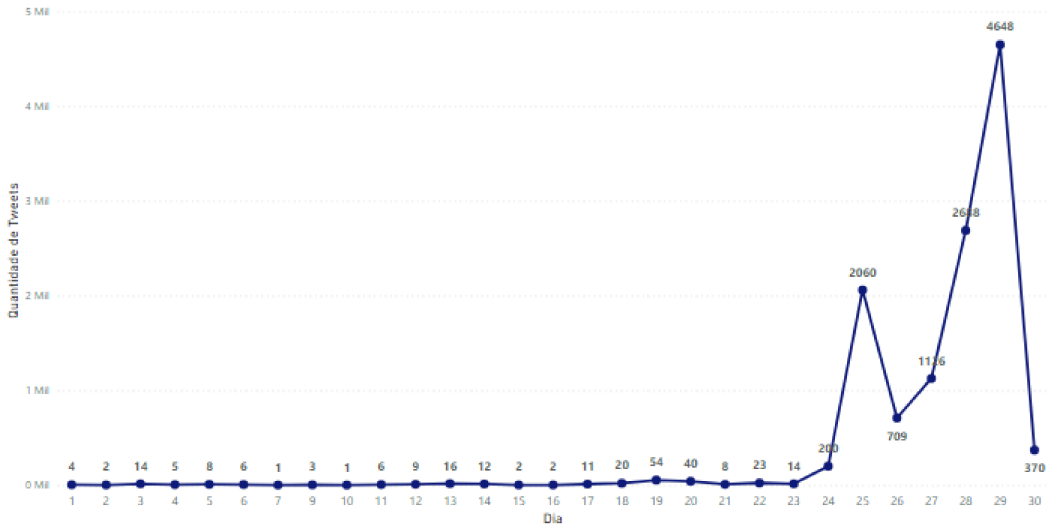

Figura 10 – Quantidade de tweets por dia em Abril de 2023.

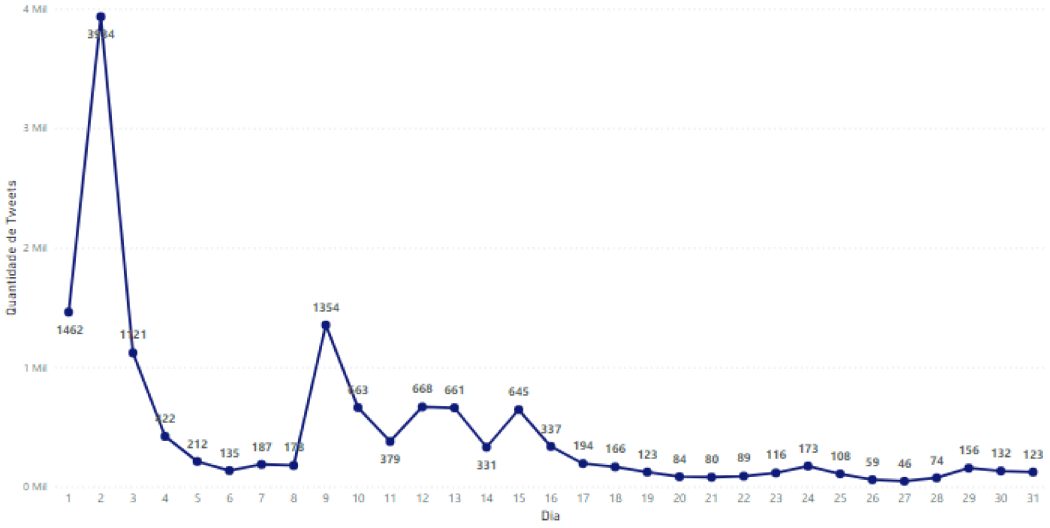

Figura 11 – Quantidade de tweets por dia em Maio de 2023.

Na Figura 10, é possível ver que a quantidade de tweets sobre a PL 2630 aumenta de forma significa a partir do dia 27 de abril, data esta, em que o então deputado Orlando Silva, relator do projeto, protocolou a versão final do PL das fake news (2.620 de 2020). Após isso, a Figura 11 traz novamente um aumento no número de tweets a partir do dia 1 de maio, às vésperas da votação do projeto no plenário da Câmara dos Deputados, que ocorreu em 2 de maio.

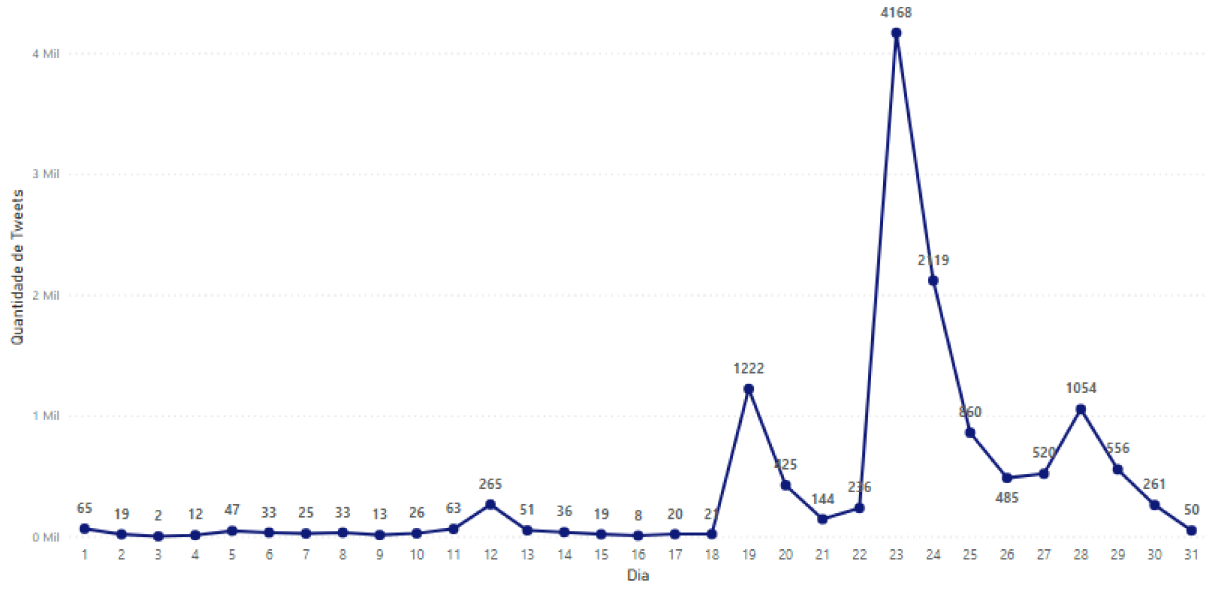

Figura 12 – Quantidade de tweets por dia em Dezembro de 2023.

Outro aumento significativo na quantidade de tweets publicados ocorreu em 23 de dezembro, conforme a Figura 12. O aumento ocorrido foi motivado pelo falecimento de uma jovem após a divulgação de notícias falsas sobre ela nas plataformas online. Além disso, a morte de um influenciador digital, no dia 28 de dezembro de 2023, trouxe à tona a discussão sobre a importância de estabelecer limites para discursos de ódio e responsabilizar as plataformas digitais pela disseminação de desinformação. Ambos os fatos levaram a uma reação por parte dos membros do governo que incentivaram o apoio à aprovação do Projeto de Lei 2630/2020.

Exemplificado alguns eventos que impactaram no número de tweets, buscou-se explorar a quantidade de tweets classificados para cada sentimento nos meses de abril, maio e dezembro de 2023, sendo os dados organizados por semana. A Figura 13, revela que, até a protocolização do projeto, em 27 de abril de 2023, os tweets positivos e neutros superavam os negativos. Contudo, na semana de votação, houve uma queda no número de tweets e a polaridade deles se manteve balanceada. O grande volume nos tweets na semana em que o projeto foi protocolado mostra a pressão exercida pelos usuários sobre os parlamentares. Ademais, o caso da morte da jovem e do influenciador digital em dezembro de 2023, fez crescer o número de tweets negativos, que superaram em aproximadamente 1,5 vezes os positivos.

Por fim, investigou-se quais são os usuários que obtiveram maior engajamento, considerando o número de curtidas, retweets e respostas, nos meses de abril, maio e dezembro de 2023. A Tabela 6, mostra que a organização de ativistas digitais que combate discursos de ódio e desinformação na internet, *Sleeping Giants*, ocupou o primeiro lugar, seguida da página de fofocas “Choquei”. Da terceira até a sexta posição, ocuparam parlamentares

que se encontram à direita do eixo político, sendo eles, Mário Frias, Carla Zambelli, Sérgio Moro e Eduardo Bolsonaro. Ademais, o relator da PL 2630, Orlando Silva, ocupou a oitava posição dentre os usuários com maior engajamento.

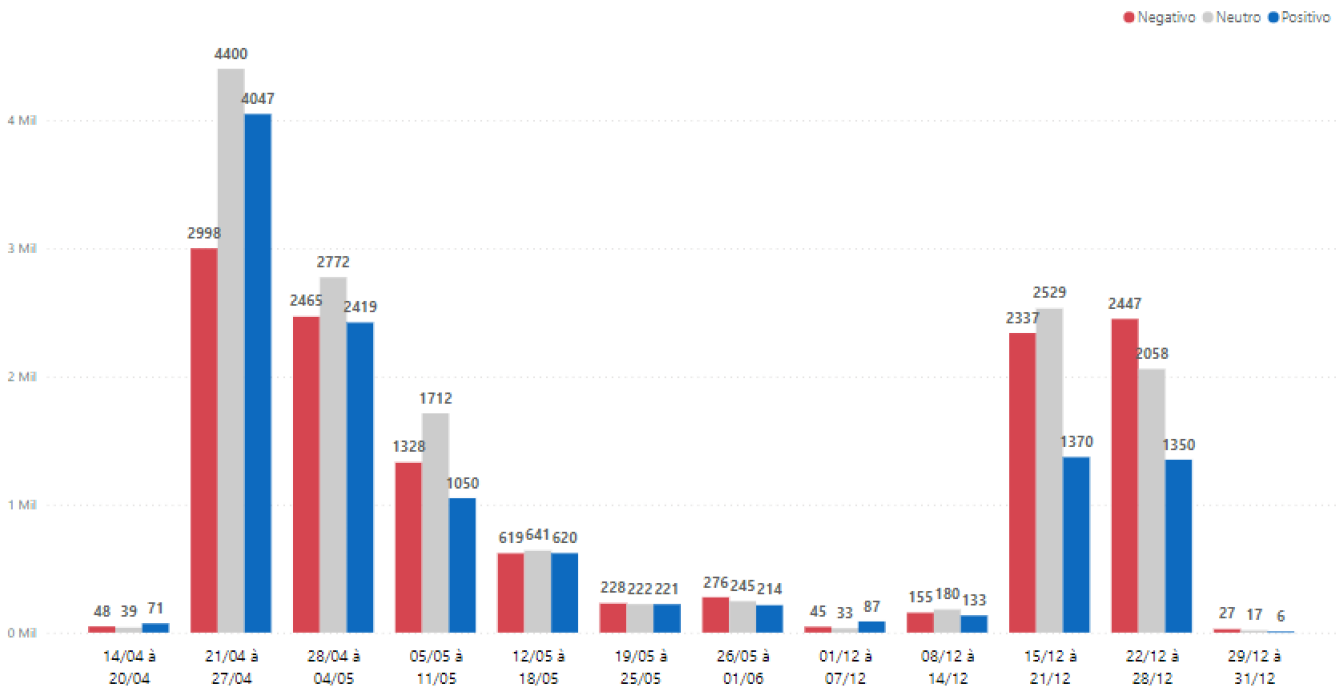

Figura 13 – Quantidade de tweets classificados em cada sentimento por semana nos meses de maior volume.

Tabela 6 – Usuários com maior engajamento nos meses de abril, maio e dezembro de 2023.

Usuário	**Curtidas**	**Retweets**	**Respostas**
Sleeping Giants Brasil	813.105	182.173	17.017
CHOQUEI	343.943	34.124	7.523
MarioFrias	312.759	68.339	3.731
Carla Zambelli	266.856	64.636	3.005
Sergio Moro	248.064	39.183	2.674
Eduardo Bolsonaro	202.197	54.087	2.411
Peter Jordan	155.708	15.774	1.167
Orlando Silva	154.640	36.197	4.606
Bia Kicis	111.694	31.226	1.766
Randolfe Rodrigues	104.827	23.202	1.951

4.2 Avaliação dos Classificadores

Conforme abordado na subseção 3.4, os classificadores foram aplicados antes e após a etapa de processamento para avaliar o impacto dessa fase na classificação dos dados,

avaliando métricas como tempo de execução e acurácia do modelo. As Tabelas 7 e 8, mostram o desempenho de cada classificador antes e após o pré-processamento do texto.

Tabela 7 – Desempenho dos classificadores antes do pré-processamento do texto.

Classificadores	**Tempo de execução (em segundos)**	**Acurácia (%)**
Naive Bayes - Multinomial	3,40	71
Naive Bayes (alpha=0)	2,95	73
Naive Bayes - Bigrama	6,68	75
SVM - Linear	679,43	81
SVM - RBF	1140,00	85
Decision Three	23,00	55

Tabela 8 – Desempenho dos classificadores após o pré-processamento do texto.

Classificadores	**Tempo de execução (em segundos)**	**Acurácia (%)**
Naive Bayes - Multinomial	2,60	74
Naive Bayes (alpha=0)	2,98	72
Naive Bayes - Bigrama	6,00	77
SVM - Linear	179,00	86
SVM - RBF	360,00	81
Decision Three	22,00	55

De acordo com as figuras acima, nota-se que, com exceção do algoritmo Naive Bayes multinomial com parâmetro alpha igual a zero, todos os classificadores apresentaram melhora no seu tempo de execução após a etapa de pré-processamento. Como a base de dados deste trabalho contém aproximadamente 58 mil tweets, podemos dizer que com uma quantidade maior de dados o impacto desta fase seria ainda mais evidente. Levando em consideração a acurácia do modelo, os classificadores Decision Three, Naive Bayes multinomial com parâmetro alpha igual a zero e o SVM com kernel RBF não obtiveram melhora após o pré-processamento.

Evidente a importância do pré-processamento, optou-se por treinar os modelos com a base tratada. Assim, as Figuras 14, 15 e 16, mostram a matriz de confusão dos três algoritmos de classificação que obtiveram maior percentual de acurácia, sendo eles, Naive Bayes Bigrama, SVM Linear e SVM RBF, após o pré-processamento. Como o algoritmo SVM linear foi o que apresentou melhor desempenho, este foi escolhido para este trabalho.

A Figura 14, obtida após a aplicação do algoritmo Naive Bayes Bigrama, em que no lugar de vetorizar o texto "por palavra", a vetorização ocorreu por cada duas palavras, mostra a matriz de confusão para tal algoritmo.

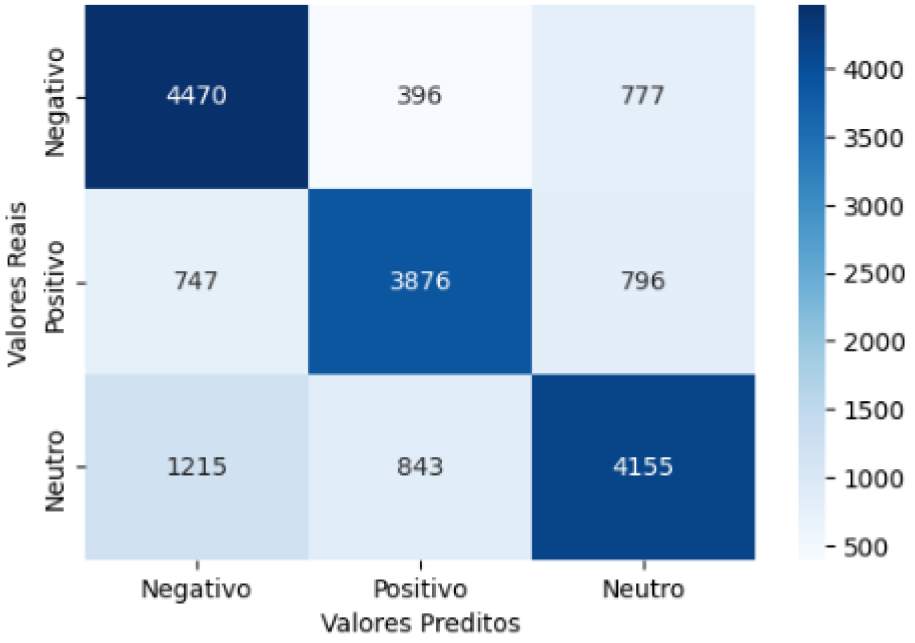

Figura 14 – Matriz de confusão Naive Bayes Bigrama.

A Figura 15, obtida após a aplicação do algoritmo SVM linear, mostra a matriz de confusão para tal algoritmo.

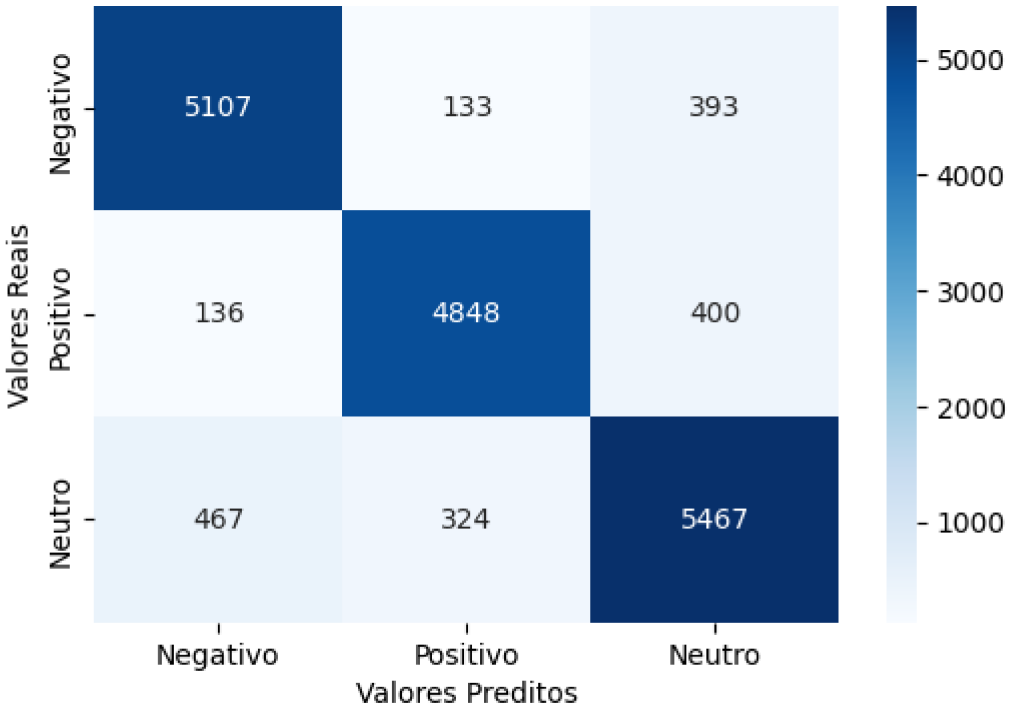

Figura 15 – Matriz de confusão SVM RBF.

A Figura 16, obtida após a aplicação do algoritmo SVM linear, mostra a matriz de confusão para tal algoritmo.

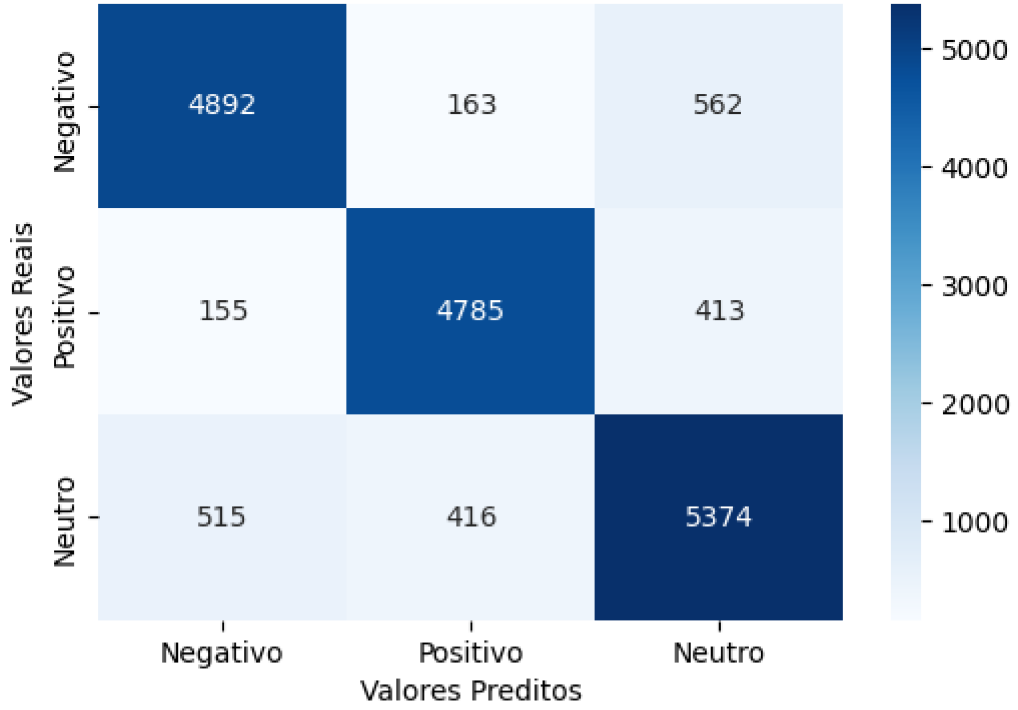

Figura 16 – Matriz de confusão SVM Linear.

4.3 Classificação de Novos Dados

Para testar o desempenho do SVM linear, foram criadas seis sentenças, sendo uma delas irônica, já que, conforme exposto na subseção 2.4.2, um dos desafios em realizar a análise de sentimentos por meio do Twitter/X inclui a detecção de ironias. Na Tabela 9, estão as novas frases e como cada uma deveria ser categorizada pelo classificador.

Novos Tweets	**Polaridade**
A PL 2630 é fundamental para manter a internet segura	Positivo
Um absurdo essa PL, os deputados devem votar contra	Negativo
As big techs estão censurando quem é a favor da PL 2630	Neutro
Com a aprovação da PL estarem cada vez mais próximos de uma ditadura	Negativo
Como eu estou feliz em ser censurado por essa PL	Negativo
Votação da PL 2630 deve acontecer essa semana em Brasília	Neutro

Tabela 9 – Novos dados usados para testar o desempenho do classificador

Na Tabela 10 está o resultado da classificação, onde, temos as sentenças após o processo de pré-processamento, a probabilidade que o SVM atribuiu para cada classe e o resultado da classificação.

De acordo com a tabela 10, podemos ver que o algoritmo obteve dificuldade em classificar a frase neutra “Votação da PL 2630 deve acontecer essa semana em brasília”, atribuindo com 84% a probabilidade do sentimento expresso na frase ser negativo. Além disso, o algoritmo também não conseguiu classificar a frase irônica “Como eu estou feliz

Texto Pré-processado	Negativo (%)	Neutro (%)	Positivo (%)	Classificação
fundamental manter Internet seguro	0	0	100	Positivo
absurdo deputado votar	100	0	0	Negativo
big techs estão censurar	13	82	5	Neutro
aprovação estar próximo ditadura	96	4	0	Negativo
feliz censurar	3	7	90	Positivo
votação acontecer semana Brasília	84	15	1	Negativo

Tabela 10 – Resultado da classificação dos novos dados

em ser censurado por essa PL", atribuindo com 90% a probabilidade do sentimento expresso na frase ser positivo. Isso ocorre porque o modelo se baseia em palavras isoladas e em seu uso frequente em contextos positivos ou negativos. Assim, a frase foi classificada como positiva devido a presença do termo "feliz", mas, se analisarmos o contexto, observamos que a frase expressa o oposto do que as palavras significam literalmente e que o usuário está descontente com à possibilidade de ser censurado pelo Projeto de Lei.

4.4 Nuvem de Palavras

Conforme expresso na subseção 3.5.2, foram produzidas quatro nuvens de palavras, sendo uma delas, utilizando toda a base de dados e outras três considerando apenas os tweets classificados em cada sentimento.

Figura 17 – Nuvem de palavras contidas nos tweets coletados.

A Figura 17, mostra os termos mais frequentes nos tweets coletados, sendo possível ver a presença das palavras que circundam a discussão em torno deste projeto como, por

exemplo, “PL”, “Fake News”, “PL 2630", "censura ","liberdade", "contra", "redes sociais"e "internet".

Na Figura 18, que contém a nuvem de palavras referente aos tweets classificados como positivos, pode-se perceber a presença de termos de natureza positiva como “defender”, “salva”, “verdade”, “direito” e “liberdade”. Além disso, analisando o contexto das discussões em torno deste projeto, é possível ver claramente a presença dos termos “criança” e “salva vidas”, palavras frequentemente utilizadas por usuário que saem em defesa da aprovação deste projeto, com a argumentação de que a regulamentação das redes sociais acarretaria em um ambiente virtual mais seguro para as crianças. Essa argumentação ganhou ainda mais força após a divulgação de uma reportagem no programa "Fantástico", da Rede Globo, em junho de 2023, em que mostrou como o aplicativo Discord havia se transformado em uma ferramenta para praticar crimes contra menores de idade.

Figura 18 – Nuvem de palavras contidas nos tweets classificados como positivos.

Na Figura 19, está a nuvem de palavras formada pelos termos mais frequentes em tweets classificados como negativos. O termo “censura” ganha grande destaque nessa nuvem de palavras, seguido de outros termos que expressam sentimentos negativos como, por exemplo “ataque”, “contra”, “crime”, “ódio” e “punir”. Dessa maneira, é possível resumir que os tweets com sentimentos negativos seguem a ideia de serem contrários ao PL 2630, uma vez que, ele fere a liberdade de expressão e a regulamentação das redes sociais seria uma forma utilizada pelo governo para censurar opiniões de opositores.

Por fim, a Figura 20, traz a nuvem de palavras gerada a partir de tweets classificados como neutros. As palavras mais frequentes são aquelas utilizadas por veículos de comunicação para informar a população sobre os fatos em torno desse projeto. Dessa maneira, os termos “PL”, “PL 2630”, “censura”, “redes sociais”, “regulamentação”, “deputado”,

"internet", "contra" e "aprovação", são as palavras mais presentes em tweets com esse sentimento. Ademais, vale ressaltar a presença do nome do relator do projeto, Orlando Silva.

Figura 19 – Nuvem de palavras contidas nos tweets classificados como negativos.

Figura 20 – Nuvem de palavras contidas nos tweets classificados como neutros.

De modo geral, os termos centrais que rodeiam a discussão em torno do projeto estão presentes em todas as nuvens de palavras. Além deles, podemos ver também o nome de algumas plataformas digitais como Google, Telegram, Twitter/X e Youtube. A justificativa da presença desses nomes está no fato delas serem diretamente impactadas por esse

projeto, mas, também porque frequentemente utilizaram seus meios de comunicação para se expressarem sobre o tema, conforme mostram as figuras 21 e 22, e assim alimentaram a discussão no Twitter/X.

Figura 21 – Mensagem veiculada na homepage do Google, em maio de 2023, que expõe a opinião da empresa sobre o PL 2630.

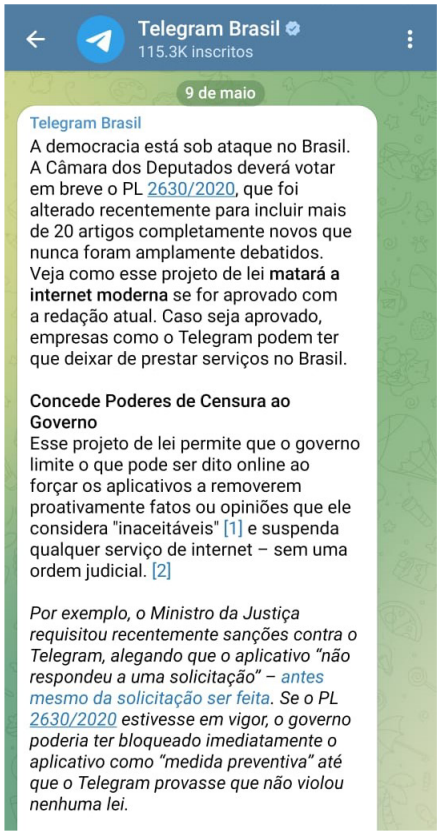

Figura 22 – Nota enviada pelo Telegram aos usuários sobre o Projeto de Lei 2630/2020.

4.5 Grafo

A Figura 23, mostra como os usuários estão conectados entre si. Após a clusterização, é possível perceber a existência de seis grupos, ou seja, grupos em que os nós estão mais conectados entre si, e menos conectados com os nós dos demais grupos. Analisando os grupos, temos que o grupo laranja tem como usuário central o deputado Nikolas Ferreira (`@nikolas_dm`), sendo também, parte desse grupo composto de parlamentares que se situam à direita do eixo político, como Carla Zambelli, Eduardo Bolsonaro, Flavio Bolsonaro, Carlos Jordy, dentre outros.

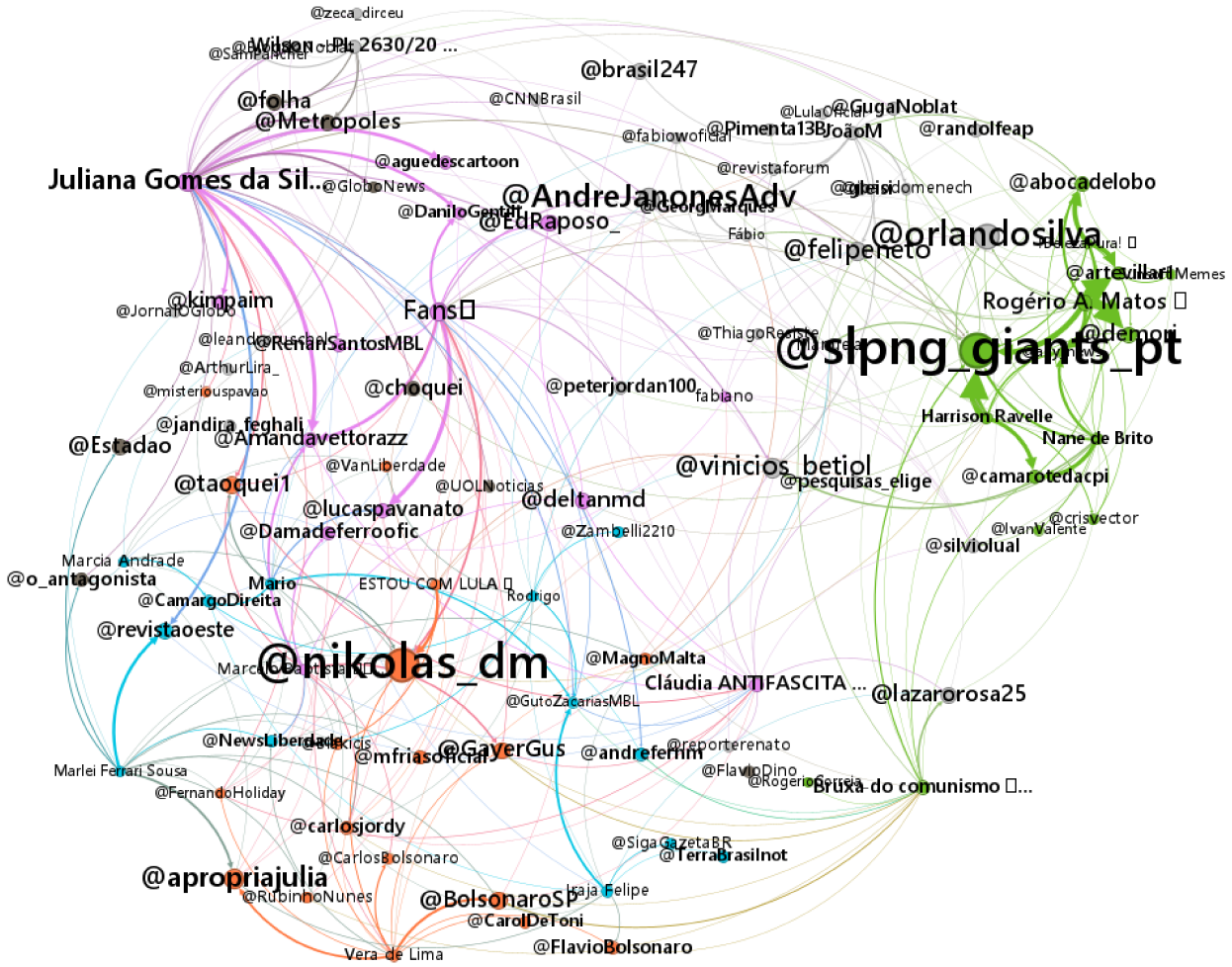

Figura 23 – Conexão entre os usuários por meio de grafo.

O grupo cinza escuro é composto principalmente por veículos de comunicação, como os jornais Folha de São Paulo, Metrópoles, Estadão e Antagonista. O grupo em cinza mais claro é composto por parlamentares de esquerda como André Janones (@AndreJanonesAdv), Gleise Hoffman (@gleise) e Randolfe Rodrigues (@randolfeap), além do atual

presidente do Brasil Luiz Inácio Lula da Silva (@LulaOficial) e do influenciador digital Felipe Neto (@FelipeNeto).

No grupo verde, está o usuário com maior número de conexão com outros internautas, a página Sleeping Giants (`@slpng_giants_pt`). Além disso, fazem parte desse grupo, comunidades de jornalismo independentes como os usuários @abocadelobo e @camarote-dacpi.

Capítulo 5

Conclusão

Neste trabalho, foi apresentada uma breve contextualização sobre as redes sociais e seu papel na viabilização da troca de ideias e experiências entre os internautas. O modelo adotado pelo Twitter/X incentiva os usuários a compartilharem e expressarem continuamente seus pensamentos e emoções, os quais são disseminados entre os seus seguidores. Sendo assim, uma rede social interessante de ser explorada nas áreas de Mineração de Texto e Análise de Sentimentos.

Considerando isso, o objetivo deste estudo foi realizar a análise dos sentimentos expressos nos tweets associados ao Projeto de Lei 2630/2020. Para alcançar esse objetivo, foram seguidas as fases da análise de sentimentos: coleta dos tweets, pré-processamento dos textos, construção de uma base de dados rotulada, classificação dos textos, análise e apresentação dos resultados. Três algoritmos de classificação foram estudados e aplicados na base de dados: Naive Bayes, SVM e Decision Tree. Sendo que para os dois primeiros, trabalhou-se com variações nos parâmetros do modelo. Além disso, buscando compreender o impacto da etapa de pré-processamento no desempenho dos algoritmos, estes foram aplicados antes e após esta etapa. Após avaliar o desempenho de cada algoritmo em relação à classificação da base de dados rotulada, observou-se que a maioria deles apresentou tempo de execução e acurácia melhores quando aplicados após o pré-processamento dos dados. Assim, o SVM com kernel Linear foi eleito como o melhor classificador neste trabalho, uma vez que obteve um nível de acurácia de aproximadamente 86%.

Também foi possível concluir que fatos do cotidiano ampliaram o número de publicações sobre a PL 2630/2020, que atingiu o maior número de tweets nos dias entre a protocolização e votação do projeto nas câmara dos deputados. Nos meses que se seguiram, a discussão diminuiu, voltando a ter um novo pico em dezembro de 2023 com a morte de uma jovem e de um influenciador digital vítimas de notícias falsas.

Através das nuvens de palavras foi possível perceber que os tweets classificados como positivos tendem a ser favoráveis ou projeto, enquanto os negativos, são contrários a PL. Ademais, com o grafo foi possível ver que os usuários que opiniões políticas de esquerda ou de direita tentem a formar grupos, uma vez que, estão mais conectados. Além disso,

percebe-se uma alta influência de perfis de ativistas digitais no debate, como o perfil *Sleeping Giants*.

5.1 Principais Contribuições

As principais contribuições deste trabalho foram:

- Construção de uma base de dados de tweets em português sobre o Projeto de Lei 2630/2020, composta por 57.582 tweets, sendo essa base, possível de ser usada como conjunto de treinamento para algoritmos de aprendizagem supervisionada em outros trabalhos. A base se encontra disponível para download[1];
- Criação de um script em linguagem Python capaz de realizar diversas etapas de pré-processamento, bem como, a construção de nuvens de palavras e grafos;
- Comparação entre classificadores para esta tarefa de mineração de textos, elegendo o classificador SVM com kernel linear como sendo propício para a classificação do sentimento através de tweets;
- Comprovação de que a maioria dos classificadores obtém desempenho melhor ao trabalhar com dados pré-processados.

5.2 Trabalhos Futuros

Uma das maiores dificuldades encontradas neste trabalho foi a coleta dos dados. Apesar da base de dados conter aproximadamente 58 mil tweets, é interessante que em trabalhos futuros a base de dados seja incrementada com mais tweets sobre o tema, para que, com uma quantidade significativa de dados de treinamento distribuídos por um período mais longo de postagens, o classificador seja menos propenso ao sobreajuste, e assim, possa obter melhor desempenho e acurácia. Além disso, outra dificuldade encontrada foi a classificação manual dos dados de treinamento. Para se ter um conjunto de treino significativo, optou-se por trabalhar com 10% da base, o que resultaria na classificação manual de 5.800 tweets, demandando assim, uma quantidade significativa de tempo e recursos humanos. Por isso, a classificação foi feita utilizando a função `get_nrc_sentiment` presente no software R por meio do pacote “syuzhet".

Uma base de dados como essa também pode ser complementada com informações de geolocalização do usuário que publicou o tweet. Assim, é possível analisar em que regiões do Brasil a PL 2630/2020 está sendo mais discutida, bem como, quais regiões apresentam mais sentimentos positivos, negativos ou neutros sobre o assunto. Além do

[1] https://www.kaggle.com/christopherfidelis

que foi mencionado, outra abordagem seria realizar a análise de sentimento considerando apenas as hashtags contidas em cada tweet e não o tweet como um todo.

Além disso, o perfil dos usuários do Twitter pode não ser o mesmo dos usuários de outras plataformas como, Facebook, Instagram e Youtube. Dessa maneira, uma base mais rica pode ser construída levando em consideração as publicações nestas redes sociais e, posteriormente, pode-se analisar se há uma diferença entre os sentimentos expressos sobre a PL 2630/2020 em cada rede social.

Outros algoritmos de classificação podem ser utilizados em trabalhos futuros, como, por exemplo, o BERTimbau, que é uma adaptação do BERT (Bidirectional Encoder Representations from Transformers), desenvolvido pelo Google para aprimorar o mecanismo de busca da plataforma. O BERTimbau foi especialmente treinado para a língua portuguesa, o que o torna uma ferramenta promissora para tarefas de processamento de linguagem natural em português. Dessa maneira, pode-se comparar o desempenho deste classificador com os outros utilizados neste trabalho.

Todas as técnicas usadas neste trabalho podem ser também aplicadas à análise de sentimentos em outros temas modernos polarizantes. A mineração de dados aplicadas no domínio das redes sociais pode oferecer compreensão e visão ampla do panorama da opinião pública acerca de questões complexas.

Até a conclusão deste trabalho a PL 2630/2020 ainda não foi aprovada, é interessante completar a base utilizada neste estudo com tweets publicados até a sua aprovação/reprovação no congresso. E assim, analisar o sentimento dos usuários da rede social até o desfecho deste assunto. Como exemplo de tweets que podem ser incorporados a base de dados, estão aqueles publicados após o desentendimento entre o ministro do Supremo Tribunal Federal (STF), Alexandre de Moraes, e o dono do Twitter/X, Elon Musk, no mês de abril de 2024. Nesta polêmica, Elon Musk ameaçou descumprir decisões judiciais brasileiras e fez ataques ao ministro do supremo, acusando-o de censura. Esse acontecimento trouxe de volta o debate sobre a PL 2630 para o centro das discussões, visto que, os deputados da base do atual governo começaram a defender sua votação, para regular as plataformas de redes sociais.

Referências

AHUJA, R. et al. The impact of features extraction on the sentiment analysis. **Procedia Computer Science**, Elsevier, v. 152, p. 341–348, 2019.

ALMEIDA, S. M. Uso de big data em mídias sociais: panorama atual na ciência. PUC-Campinas, 2018.

ARANHA, C. N. Processamento automático para mineração de textos em português: sob o enfoque da inteligência computacional. **Rio de Janeiro: PUC-Rio**, 2007.

AWAD, M.; KHANNA, R. **Efficient learning machines: theories, concepts, and applications for engineers and system designers**. [S.l.]: Springer nature, 2015.

BARBOSA, R. et al. Characterizing the effectiveness of twitter hashtags to detect and track online population sentiment. In: **CHI'12 Extended Abstracts on Human Factors in Computing Systems**. [S.l.: s.n.], 2012. p. 2621–2626.

BERTO, B. B. Análise de sentimentos de tweets sobre a pandemia de covid-19. Universidade Presbiteriana Mackenzie, 2021.

CAPOBIANCO, K. Avaliação da etapa de pré-processamento na mineração de texto em redes sociais digitais. **Universidade Estadual de Londrina, Londrina–PR**, 2016.

CARDOSO, O. N. P.; MACHADO, R. T. M. Gestão do conhecimento usando data mining: estudo de caso na universidade federal de lavras. **Revista de administração pública**, SciELO Brasil, v. 42, p. 495–528, 2008.

CORDEIRO, J. P. da C. Extracção de elementos relevantes em texto/páginas da world wide web. **Master's thesis, Departamento de Ciência de Computadores Faculdade de Ciências da Universidade do Porto**, 2003.

CORRÊA, I. T. et al. Análise dos sentimentos expressos na rede social twitter em relação aos filmes indicados ao oscar 2017. Universidade Federal de Uberlândia, 2017.

DEY, L. et al. Sentiment analysis of review datasets using naive bayes and k-nn classifier. **arXiv preprint arXiv:1610.09982**, 2016.

EADY, G. et al. How many people live in political bubbles on social media? evidence from linked survey and twitter data. **Sage Open**, SAGE Publications Sage CA: Los Angeles, CA, v. 9, n. 1, p. 2158244019832705, 2019.

EVAGELISTA, T.; POLETTO, A. S. R. d. S. Algoritmos e tècnicas para a mineração de dados. In: . [S.l.: s.n.], 2018.

FAN, W. et al. Tapping the power of text mining. **Communications of the ACM**, ACM New York, NY, USA, v. 49, n. 9, p. 76–82, 2006.

FELDMAN, R. et al. Knowledge management: A text mining approach. In: **Proc. the 2nd Int. Conf. on Practical Aspects of Knowledge Management (PAKM98)**. [S.l.: s.n.], 1998. p. 9–1.

FILHO, J. A. C. Mineração de textos: análise de sentimentos utilizando tweets referentes à copa do mundo 2014. 2014.

FOGAÇA, A. **Brasileiros ficam mais de 9 horas olhando o celular, perdemos só para um país**. 2023. <https://olhardigital.com.br/2023/06/29/reviews/brasileiros-ficam-mais-de-9-horas-olhando-o-celular-perdemos-so-para-um-pais/>. Acesso em: 18 abr. de 2024.

FONTOURA, V. D. **Predição de falhas em projetos de software livre baseadas em métricas de redes sociais**. Dissertação (B.S. thesis) — Universidade Tecnológica Federal do Paraná, 2011.

FRANÇA, T. C. et al. Big social data: princípios sobre coleta, tratamento e análise de dados sociais. **XXIX Simpósio Brasileiro de Banco de Dados–SBBD**, v. 14, 2014.

FRANÇA, T. C. de; OLIVEIRA, J. Análise de sentimento de tweets relacionados aos protestos que ocorreram no brasil entre junho e agosto de 2013. In: SBC. **Anais do III Brazilian workshop on social network analysis and mining**. [S.l.], 2014. p. 128–139.

GALVÃO, N. D.; MARIN, H. d. F. Técnica de mineração de dados: uma revisão da literatura. **Acta Paulista de Enfermagem**, SciELO Brasil, v. 22, p. 686–690, 2009.

GOMES, H. J. C. **Text mining: Análise de sentimentos na classificação de notícias**. Dissertação (Mestrado) — Universidade NOVA de Lisboa (Portugal), 2013.

GONÇALVES, T. et al. Analysing part-of-speech for portuguese text classification. In: SPRINGER. **Computational Linguistics and Intelligent Text Processing: 7th International Conference, CICLing 2006, Mexico City, Mexico, February 19-25, 2006. Proceedings 7**. [S.l.], 2006. p. 551–562.

GUPTA, M. K.; CHANDRA, P. A comprehensive survey of data mining. **International Journal of Information Technology**, Springer, v. 12, n. 4, p. 1243–1257, 2020.

HOLANDA, L. N.; TEIXEIRA, F. C. A importância da regulamentação de mídias sociais em estados democráticos: Uma análise de direito comparado entre o projeto de lei nº 2630/2020 e a legislação portuguesa. **REVISTA FOCO**, v. 16, n. 02, p. e1021–e1021, 2023.

INFOMONEY, E. **Elon Musk muda o nome do Twitter para X; mudança vem como resposta ao Threads, da Meta**. 2023. <https://www.infomoney.com.br/negocios/elon-musk-muda-o-nome-do-twitter-para-x-mudanca-vem-como-resposta-ao-threads-da-meta/>. Acesso em: 18 mar. de 2024.

JUNIOR, J. R. C. Desenvolvimento de uma metodologia para mineração de textos. **Pontificia Universidad Catolica de Rio de Janeiro: Rio de janeiro, Brasil**, 2007.

KAAKINEN, M. et al. Shared identity and shared information in social media: Development and validation of the identity bubble reinforcement scale. **Media Psychology**, Taylor & Francis, v. 23, n. 1, p. 25–51, 2020.

KANSAON, D. P.; BRANDÃO, M. A.; PINTO, S. A. de P. Análise de sentimentos em tweets em português brasileiro. In: SBC. **Anais do VII Brazilian Workshop on Social Network Analysis and Mining**. [S.l.], 2018.

KAUER, A. U. Análise de sentimentos baseada em aspectos e atribuições de polaridade. 2016.

LIDDY, E. D. Natural language processing. 2001.

LIU, B. **Sentiment analysis and opinion mining**. [S.l.]: Springer Nature, 2022.

MÄNTYLÄ, M. V.; GRAZIOTIN, D.; KUUTILA, M. The evolution of sentiment analysis—a review of research topics, venues, and top cited papers. **Computer Science Review**, Elsevier, v. 27, p. 16–32, 2018.

MARTINS, L. D.; ARAÚJO, F. P. O. Mineração de texto para a análise do perfil emocional de usuários de jogo empático. **Anais do Computer on the Beach**, v. 12, p. 370–377, 2021.

MATOS, F. F.; MAGALHÃES, L. H. d.; SOUZA, R. R. Recuperação e classificação de sentimentos de usuários do twitter em período eleitoral. 2020.

MOHAMMAD, S. M.; TURNEY, P. D. Nrc emotion lexicon. **National Research Council, Canada**, v. 2, p. 234, 2013.

MORAIS, E. A. M.; AMBRÓSIO, A. P. L. Mineração de textos. **Relatório Técnico–Instituto de Informática (UFG)**, 2007.

NANDWANI, P.; VERMA, R. A review on sentiment analysis and emotion detection from text. **Social network analysis and mining**, Springer, v. 11, n. 1, p. 81, 2021.

NERI, F. et al. Sentiment analysis on social media. In: IEEE. **2012 IEEE/ACM international conference on advances in social networks analysis and mining**. [S.l.], 2012. p. 919–926.

NOVENDRI, R. et al. Sentiment analysis of youtube movie trailer comments using naïve bayes. **Bulletin of Computer Science and Electrical Engineering**, v. 1, n. 1, p. 26–32, 2020.

OLIVEIRA, C. Análise de sentimentos de comentários em português utilizando sentiwordnet. **Monografia de Especialização em Desenvolvimento de Sistemas para Web, Departamento de Informática, Universidade Estadual de Maringá**, 2013.

PETERSON, R. E. Eight internet search engines compared. Munksgaard International Publishers Ltd., Copenhagen, 1997.

QUAGLIO, L. O. et al. Jurisdição internacional e as fake news na era da pós-verdade: uma análise das leis no âmbito do direito digital vigentes no brasil e o pl nº 2630/2020. Universidade Federal de Uberlândia, 2021.

ROSA, R. L. **Análise de sentimentos e afetividade de textos extraídos das redes sociais.** Tese (Doutorado) — Universidade de São Paulo, 2015.

SAGAYAM, R.; SRINIVASAN, S.; ROSHNI, S. A survey of text mining: Retrieval, extraction and indexing techniques. **International Journal of Computational Engineering Research**, v. 2, n. 5, p. 1443–1446, 2012.

SANTOS, M. Protótipo para mineração de opinião em redes sociais: estudo de casos selecionados usando o twitter. **Monografia. Departamento de Ciência da Computação, Universidade Federal de Lavras**, 2010.

SCHMITT, V. F. Uma análise comparativa de técnicas de aprendizagem de máquina para prever a popularidade de postagens no facebook. 2013.

SHEPHERD, J. **23 Essential Twitter (X) Statistics You Need to Know in 2024**. 2024. <https://thesocialshepherd.com/blog/twitter-statistics#:~:text=At%20Least%20500%20Million%20Tweets>. Acesso em: 18 abr. de 2024.

SHOUZHONG, T.; MINLIE, H. Mining microblog user interests based on textrank with tf-idf factor. **The Journal of China Universities of Posts and Telecommunications**, Elsevier, v. 23, n. 5, p. 40–46, 2016.

SILVA, N. F. F. d. **Análise de sentimentos em textos curtos provenientes de redes sociais**. Tese (Doutorado) — Universidade de São Paulo, 2016.

SILVA, N. R.; LIMA, D.; BARROS, F. Sapair: Um processo de análise de sentimento no nível de característica. In: **4nd International Workshop on Web and Text Intelligence (WTI'12), Curitiba**. [S.l.: s.n.], 2012. p. 2.

SOARES, F. d. A. **Mineração de Textos na Coleta Inteligente de Dados na Web**. Tese (Doutorado) — PUC-Rio, 2008.

SURESH, A.; BHARATHI, C. Sentiment classification using decision tree based feature selection. **Ijcta**, v. 9, n. 36, p. 419–425, 2016.

TALIB, R. et al. Text mining: techniques, applications and issues. **International journal of advanced computer science and applications**, Science and Information (SAI) Organization Limited, v. 7, n. 11, p. 414–418, 2016.

TAN, A.-H. et al. Text mining: The state of the art and the challenges. In: **Proceedings of the pakdd 1999 workshop on knowledge disocovery from advanced databases**. [S.l.: s.n.], 1999. v. 8, p. 65–70.

TEIXEIRA, N.; REIS, J. R.; FONSECA, K. M. Os limites entre a liberdade e a regulação das redes sociais. In: **Anais do Congresso Nacional Universidade, EAD e Software Livre**. [S.l.: s.n.], 2023. v. 1, n. 15.

TRIPATHI, M. Sentiment analysis of nepali covid19 tweets using nb svm and lstm. **Journal of Artificial Intelligence**, v. 3, n. 03, p. 151–168, 2021.

TRONCHONI, A. B. et al. Descoberta de conhecimento em base de dados de eventos de desligamentos de empresas de distribuição. **Sba: Controle & Automação Sociedade Brasileira de Automatica**, SciELO Brasil, v. 21, p. 185–200, 2010.

VALOR, G. **Brasil tem a quarta maior base de usuários do Twitter no mundo**. 2022. <https://valorinveste.globo.com/internacional-e-commodities/noticia/2024/01/10/brasil-tem-a-quarta-maior-base-de-usuarios-do-twitter-no-mundo.ghtml>. Acesso em: 10 jan. 2024.

WITTEN, I. H. et al. Practical machine learning tools and techniques. In: ELSEVIER AMSTERDAM, THE NETHERLANDS. **Data mining**. [S.l.], 2011. v. 3.