UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Avaliação de Classificadores na Análise de Sentimentos em Redes Sociais Durante a Pandemia da COVID-19

Lucas Alexandre Malakin

Dissertação de Mestrado do Programa de Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria (MECAI)

ICMC USP
SÃO CARLOS

Lucas Alexandre Malakin

Avaliação de Classificadores na Análise de Sentimentos em Redes Sociais Durante a Pandemia da COVID-19

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria. *EXEMPLAR DE DEFESA*

Área de Concentração: Matemática, Estatística e Computação

Orientadora: Profa. Dra. Solange Oliveira Rezende

USP – São Carlos
Fevereiro de 2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

M236a

Malakin, Lucas Alexandre
Avaliação de Classificadores na Análise de Sentimentos em Redes Sociais Durante a Pandemia da COVID-19 / Lucas Alexandre Malakin; orientadora Solange Oliveira Rezende. -- São Carlos, 2024.
77 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. Mineração de Textos. 2. Aprendizado de Máquina. 3. Análise de Sentimento. 4. Redes Sociais. I. Oliveira Rezende, Solange, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Lucas Alexandre Malakin

Evaluation of Classifiers in Sentiment Analysis on Social Media During the COVID-19 Pandemic

Dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP – in accordance with the requirements of the Professional Master's Program in Mathematics Statistics and Computing Applied to Industry, for the degree of Master in Science. *EXAMINATION BOARD PRESENTATION COPY*

Concentration Area: Mathematics, Statistics and Computing

Advisor: Profa. Dra. Solange Oliveira Rezende

USP – São Carlos
February 2024

Este trabalho é dedicado para todos aqueles que sonharam e iluminaram durante a escuridão.

AGRADECIMENTOS

Queria agradecer muito à Dra Solange Oliveira Rezende não somente pelo conhecimento transmitido, mas também pelo apoio dado nos momentos difíceis. Ela esteve presente e me deu os conselhos sábios que eu precisava para superar os desafios e encontrar meu equilíbrio. Também, como ela entendia e acolhia minhas preocupações foram essenciais para atenuar os momentos mais complicados.

Para quem passou e resolveu ficar, expresso minha profunda gratidão. Sua presença não apenas continua iluminando meu caminho, mas também enriquece minha jornada. Obrigado por escolher estar ao meu lado, por continuar a compartilhar de momentos preciosos e por tornar cada dia mais especial. Sua permanência em minha vida é um presente que valorizo imensamente no meu cotidiano.

Aos que passaram e escolheram seguir seus próprios caminhos, desejo expressar minha profunda gratidão. Embora nossas jornadas tenham se separado, guardo com carinho os momentos compartilhados e as valiosas lições aprendidas juntos. Agradeço por fazerem parte da minha vida, deixando sua marca e contribuindo para o meu crescimento pessoal. Que o futuro lhes reserve sucesso e realizações em suas novas jornadas.

“Tudo é mais fácil na vida virtual, mas perdemos a arte das relações sociais e da amizade”
(Zygmunt Bauman)

RESUMO

MALAKIN, L. A. **Avaliação de Classificadores na Análise de Sentimentos em Redes Sociais Durante a Pandemia da COVID-19**. 2024. 77 p. Dissertação (Mestrado – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

O crescente uso das redes sociais ao longo de quase três décadas transformou de maneira significativa a interação humana e como compartilhamos informações. Com esse aumento na utilização das redes sociais, inevitavelmente, ocorre uma produção massiva de dados, predominantemente textuais, apresentando tanto desafios quanto oportunidades. A informação textual desempenha um papel central na comunicação nas mídias sociais, sendo crucial para plataformas como Twitter, Facebook e Instagram. Adicionalmente, esses dados textuais alimentam técnicas, como mineração textual e análise de sentimentos, permitindo uma compreensão mais profunda das preferências e tendências dos usuários. Com o surgimento da COVID-19, causada pelo novo coronavírus, desencadeou uma pandemia global que impactou milhões de pessoas. Diante desse cenário desafiador, muitos indivíduos recorreram às redes sociais para expressar suas opiniões, compartilhar ideias e obter informações sobre a doença. Nesse contexto, este estudo analisou o sentimento presente nas mensagens relacionadas à COVID-19 no Brasil, utilizando técnicas de mineração textual e análise de sentimento. A abordagem adotada envolveu a associação de técnicas léxicas com aprendizado de máquina para a classificação dos sentimentos expressos nas sentenças. Dos resultados obtidos se destacam a eficácia de algoritmos de aprendizado de máquina, notadamente BERT, SVM e LSTM, que demonstraram um desempenho superior na classificação de sentimentos em comparação com outros algoritmos. Além disso, as análises revelaram padrões temporais nos sentimentos relacionados à COVID-19 no Brasil, fornecendo uma visão aprofundada do impacto das crises de saúde pública nas dinâmicas das redes sociais.

Palavras-chave: Mineração de Textos; Aprendizado de Máquina; Análise de Sentimento; Redes Sociais.

ABSTRACT

MALAKIN, L. A. **Evaluation of Classifiers in Sentiment Analysis on Social Media During the COVID-19 Pandemic**. 2024. 77 p. Dissertação (Mestrado – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

The increasing use of social media over almost three decades has significantly transformed human interaction and the way we share information. With this rise in social media usage, inevitably, there is a massive production of data, predominantly textual, presenting both challenges and opportunities. Textual information plays a central role in communication on social media, being crucial for platforms like Twitter, Facebook, and Instagram. Additionally, this textual data fuels techniques such as text mining and sentiment analysis, allowing a deeper understanding of user preferences and trends. With the emergence of COVID-19, caused by the novel coronavirus, a global pandemic unfolded, impacting millions of people. Faced with this challenging scenario, many individuals turned to social media to express their opinions, share ideas, and obtain information about the disease. In this context, this study sought to analyze the sentiment present in messages related to COVID-19 in Brazil, employing text mining and sentiment analysis techniques. The adopted approach involved combining lexical techniques with machine learning for the classification of sentiments expressed in sentences. The results highlight the effectiveness of machine learning algorithms, notably BERT, SVM, and LSTM, which demonstrated superior performance in sentiment classification compared to other algorithms. Furthermore, the analyses revealed temporal patterns in sentiments related to COVID-19 in Brazil, providing an in-depth insight into the impact of public health crises on the dynamics of social media.

Keywords: Text Mining; Machine Learning; Sentiment Analysis; Social Media.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizado de Máquina
AS	Análise de Sentimentos
BERT	*Bidirectional Encoder Representations from Transformers*
BoW	*Bag-of-Words*
CBoW	*Continuous Bag-of-Words*
IDF	*Inverse Document Frequency*
LSTM	*Long Short-Term Memory*
MO	Mineração de Opinião
MT	Mineração Textual
OMS	Organização Mundial da Saúde
PLN	Processamento de Linguagem Natural
POS	*Part of Speech*
RS	Redes sociais
SG	*Skip-Gram*
SVM	*Support Vector Machine*
TF	*Term Frequency*
VADER	*Valence Aware Dictionary for sEntiment Reasoner*

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

Quase três décadas depois do seu surgimento, Redes sociais (RS) são usadas por mais de um terço da população mundial (ORTIZ-OSPINA; ROSER, 2023), representando também mais de dois terços dos usuários de internet. As mídias sociais digitais tornaram-se parte integrante da sociedade contemporânea, influenciando profundamente o comportamento das pessoas de diversas maneiras. Essas plataformas oferecem um espaço virtual de comunicação, permitindo que os indivíduos se expressem, compartilhem experiências e se conectem com outras pessoas em escala global (LEE, 2018). Esse ecossistema digital molda os padrões de comunicação, introduzindo novas formas de expressão. As redes sociais funcionam dinamicamente para a construção de identidade, em que os utilizadores fazem a curadoria dos seus perfis online, partilhando aspectos da sua vida pessoal e profissional.

O impacto vai além da auto apresentação, uma vez que as redes sociais servem como fonte primária de informação, influenciando opiniões. A busca pela validação social por meio de curtidas e comentários contribui para um ciclo de *feedback* que pode influenciar o conteúdo que as pessoas compartilham. No entanto, as preocupações com a privacidade, o potencial de assédio online e o impacto da conectividade constante no bem-estar mental sublinham a complexa interação entre as redes sociais e o comportamento humano (AGGARWAL, 2011).

Com o aumento da utilização e a presença dominante no cotidiano das pessoas, as redes geram uma enorme quantidade de dados, que na maioria são textuais e audiovisuais (VERMA *et al.*, 2016). A informação textual desempenha um papel central no domínio das mídias sociais, servindo como o principal meio pelo qual os usuários se comunicam, compartilham conteúdo e expressam opiniões. Plataformas de mídia social, como Twitter, Facebook, Instagram e LinkedIn, aproveitam informações textuais para diversos fins, desde atualizações de *status* e divulgação de notícias até marketing e narração de histórias (LEE, 2018).

Os dados textuais também servem como base para o Processamento de Linguagem Natural (PLN) e Análise de Sentimentos (AS), permitindo que as plataformas extraiam análises

sobre os sentimentos, preferências e tendências dos usuários. A natureza textual do conteúdo das RS apresenta oportunidades e desafios, desde permitir uma comunicação rica até apresentar complexidades na análise do contexto, tom e intenção (CHOWDHARY; CHOWDHARY, 2020). À medida que as RS continuam a evoluir, o tratamento e a compreensão eficazes da informação textual continuam a ser fundamentais para o envolvimento dos utilizadores, a moderação de conteúdos e a dinâmica geral das comunidades online.

Por meio da utilização de PLN em RS, se fazem possíveis o desenvolvimento de aplicações nos setores públicos e privados, como, por exemplo, na Índia, um grupo de pesquisadores, em colaboração com o governo, conduziu uma análise de sentimentos em tweets, publicações feitas na rede social Twitter; para compreender a opinião da população em relação às medidas adotadas para os projetos de digitalização do país. Essas informações contribuíram para a identificação das ações que os governantes deveriam priorizar, conforme a opinião da população local (MISHRA; RAJNISH; KUMAR, 2016).

Utilizando o renomado centro turístico de Las Vegas como ponto focal, pesquisadores aplicaram AS em mensagens de contas relacionadas aos *resorts* da rede hoteleira da cidade. Estas métricas de sentimento derivadas proveram de base para o *benchmarking* entre empresas, facilitando uma análise comparativa das suas tendências de desempenho. Nesta pesquisa, mostraram a aplicação prática da AS utilizando dados do Twitter para construir métricas de tempo real para avaliar as atitudes e percepções dos clientes do setor hoteleiro (PHILANDER; ZHONG, 2016).

Com o surgimento do novo coronavírus, causador da COVID-19, que resultou em um surto de pneumonia viral na China (BAI *et al.*, 2020), posteriormente sendo classificado como pandemia pela Organização Mundial da Saúde (OMS) (SOHRABI *et al.*, 2020), afetando milhões de pessoas ao redor do mundo. Alguns países, sobretudo o Brasil, adotaram a estratégia de confinamento social total ou parcial, com isso, muitas pessoas recorreram às RS para expressar suas opiniões, ideias e se informarem sobre a doença (DUBEY, 2020).

Dessa forma, justifica-se um estudo do sentimento presente nas mensagens de RS relacionadas à COVID-19, no Brasil, usando técnicas de Mineração Textual (MT) e AS com uso de processamento de linguagem natural. Também, a comparação de diferentes abordagens na tarefa de classificação do sentimento das sentenças.

Este trabalho visa investigar a composição do processamento de texto utilizando abordagens léxicas e classificadores de AM, com foco na análise de dados provenientes do Twitter durante a pandemia da COVID-19 no Brasil. A pesquisa tem por objetivo compreender como as técnicas de PLN e AS podem ser aplicadas eficazmente para extrair informações relevantes e representativas das mensagens em RS durante um evento de grande impacto como a pandemia. Além disso, faz parte dos objetivos comparar diferentes abordagens de classificação de sentimentos para avaliar sua eficiência e precisão na análise do conteúdo textual relacionado à COVID-19 nas RS brasileiras. Esta dissertação foi dividida em fundamentos e trabalhos relacionados, proposta de análise, avaliação experimental e conclusão.

CAPÍTULO

2

FUNDAMENTOS E TRABALHOS RELACIONADOS

Neste capítulo são abordadas as RS como plataformas de interações sociais, o processo de MT aplicados em problema de classificação e técnicas para representação e classificação de sentimento. Com isso, tem-se o rol de trabalhos práticos usando MT, Aprendizado de Máquina (AM) e AS em dados relacionados a RS. Dessa maneira, esse levantamento possibilita a pavimentação necessária para aplicação da metodologia, dos resultados e conclusões do trabalho desenvolvido nos próximos capítulos

2.1 Análise de sentimento

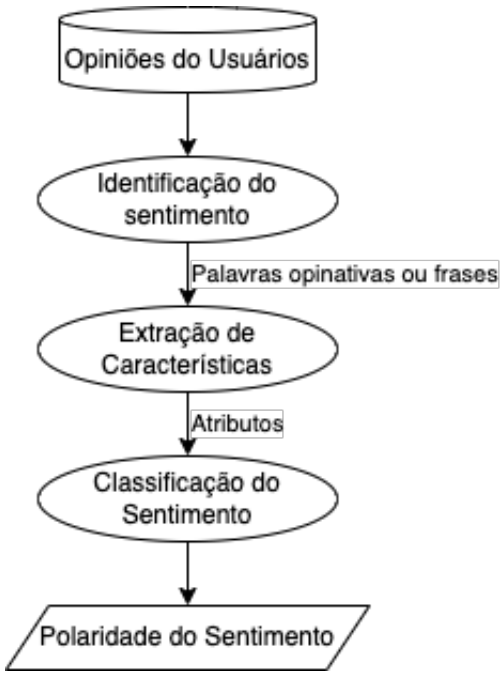
AS e Mineração de Opinião (MO) é o estudo computacional da opinião, atitudes e emoções numa entidade. A entidade pode ser representada por indivíduos, eventos ou tópicos. Embora ambas áreas expressam significados similares, a segunda extrai e analisa a opinião das pessoas, enquanto a primeira identifica o sentimento expresso em um texto, posteriormente analisando-o (LIU, 2022).

AS pode ser considerada um processo de classificação, existindo três principais níveis de granularidades: documento, observando o sentimento global expresso no texto; sentença, classificando a polaridade de cada sentença no texto; e característica, analisando a polaridade das opiniões sobre características ou atributos de objeto (WILSON; WIEBE; HOFFMANN, 2005).

O processo de AS pode ser dividido em quatro etapas, com pode ser visto na Figura 1. A primeira corresponde à identificação se a sentença é objetiva ou subjetiva. A segunda corresponde à extração de características, sendo responsável por extrair as características de um objeto, produto ou serviço em análise. A terceira etapa corresponde à classificação de sentimentos, determinando a polaridade do texto (positivo, neutro e negativo). A quarta e última etapa corresponde à visualização dos resultados, sendo responsável por demonstrar o resultado

da análise textual (MEDHAT; HASSAN; KORASHY, 2014; SILVA; LIMA; BARROS, 2012).

Figura 1 – Processo de AS



Fonte: Medhat, Hassan e Korashy (2014).

Outra grande área de importância está relacionada ao conjunto de dados para uso. A maioria dos conjuntos de dados estão relacionados às avaliações de produtos, sendo importantes para os negócios, ao apoiarem decisões estratégicas conforme as avaliações dos usuários. Embora os principais conjuntos de dados estejam associados as avaliações de produtos, AS também é empregada no mercado de ações (SOLOMON; TUTEN, 2017), notícias (XU; PENG; CHENG, 2012) e debates políticos (MAKS; VOSSEN, 2012). Em debates políticos, por exemplo, podemos mensurar a opinião das pessoas sobre candidatos, partidos ou ideais, dessa maneira, a polarização das eleições pode ser predita por publicações politizadas. As RS são consideradas uma boa fonte de informações, pois as pessoas compartilham e discutem suas opiniões sobre diversos assuntos (ORTIZ-OSPINA; ROSER, 2023).

2.1.1 Identificação do sentimento

A identificação do sentimento é dividida em duas classes, subjetiva e objetiva, em que a primeira expressa informações factuais, enquanto a segunda usualmente expressa opiniões ou visões. O sentimento pode expressar muitos tipos de informações, por exemplo opiniões, valores, emoções, especulações, dentre outros. Algumas indicam sentimento positivo ou negativo, porém outras são neutras. A pesquisa inicial costuma resolver de maneira autônoma esse problema de classificação (LIU, 2022).

A maioria das abordagens para solucionar este problema são baseadas em aprendizado supervisionado. Por exemplo, na utilização do classificador Naive Bayes associado com um

conjunto binarizado de atributos textuais, como a presença de pronomes, adjetivos, números ou advérbios (WIEBE; BRUCE; O'HARA, 1999). Também, há abordagens não supervisionada para classificação de sentimento, que usa da presença de expressões subjetivas em uma sentença para determinar o sentimento da sentença (WIEBE *et al.*, 2000).

2.1.2 Extração de características

A extração de características tem por função a identificação de atributos textuais, em que algumas das características são (LIU, 2022; AGGARWAL; AGGARWAL, 2015):

1. Extração baseada em substantivos frequentes: utiliza analisador gramatical para identificar os substantivos mais frequentes por meio de unigramas e n-gramas;

2. Extração através da relação entre as palavras, do inglês*Part of Speech* (POS): utiliza das relações e propriedades morfossintáticas das palavras;

3. Extração baseada em palavras sentimentais: algumas palavras são constantemente expressadas como sentimentos positivos ou negativos. Como, por exemplo, para sentimentos positivos bom, incrível e maravilhoso, já para sentimentos negativos ruim, péssimo e lixo;

4. Extração baseada em alternadores de sentimentos (negações): existem palavras que alteram a orientação do sentimento indo do positivo para o negativo e vice-versa. Como, por exemplo, nas frases "Eu gosto de chocolate"e "Eu não gosto de chocolate", em que o emprego do "não"ocasiona na alteração da polaridade;

2.1.3 Classificação do sentimento

Técnicas para classificação de sentimento são divididas em abordagens via AM, abordagens léxicas e híbridas. Abordagens de AM fazem o uso dos atributos e características linguísticas presentes em corpus textuais, que podem ser extraídos com algoritmos e técnicas de MT. Já abordagens léxicas se baseiam em coleções classificadas, também é subdividida em duas sub abordagens, em que a primeira se baseia no uso de dicionários e a segunda em corpus, que é uma coleção representativa de textos; das quais usam de estatística ou métodos semânticos para determinar a polaridade do sentimento. A abordagem híbrida combina as duas abordagens citadas anteriormente, entretanto, recorrendo à abordagem léxica majoritariamente (MEDHAT; HASSAN; KORASHY, 2014). Essas técnicas são exibidas na Figura 2.

Figura 2 – Abordagens para AS

Fonte: Medhat, Hassan e Korashy (2014).

2.2 Abordagens para classificação de sentimentos

Como visto anteriormente as técnicas para classificação de sentimentos são divididas em três grupos, neste trabalho serão exploradas as técnicas que fazem uso de abordagens com AM e léxicos.

2.2.1 Técnicas com aprendizado de máquina supervisionado

Os métodos baseados em algoritmos de AM supervisionado dependem da existência de um conjunto de dados rotulados (MICHALSKI; CARBONELL; MITCHELL, 2013). Existem muitas abordagens como pode ser visto na Figura 2, nessa subseção, é apresentada uma simples descrição dos classificadores mais comuns.

2.2.1.1 Classificadores baseados em árvores de decisão

São construídos a partir de dados de treinamento, consistindo entre nós de decisão e folhas. Cada nó de decisão na árvore representa uma condição ou critério aplicado aos dados de entrada. Com base nesse critério, os dados são divididos em diferentes caminhos para os nós filhos. Essa divisão é realizada recursivamente até que sejam alcançadas as folhas da árvore, que representam as classes ou categorias de saída (MAIMON; ROKACH, 2005; QUINLAN, 1986).

Durante a construção da árvore, o algoritmo visa otimizar a seleção das condições de divisão, procurando aquelas que melhor separam as classes e reduzem o erro de classificação.

Diferentes métricas, podem ser usadas para medir o erro dos dados em cada nó. Uma vez que a árvore de decisão é construída, ela pode ser usada para classificar novos exemplos de dados com base nas condições presentes nos nós, pelos quais os atributos ou características dos dados são avaliados seguindo o caminho na árvore até chegar a uma folha, em que a classe é atribuída (MAIMON; ROKACH, 2005; QUINLAN, 1986).

Alguns dos algoritmos de árvores de decisão são (GUPTA *et al.*, 2017):

- CART: é um algoritmo de árvore de decisão amplamente utilizado que pode ser usado tanto para tarefas de classificação quanto de regressão. Funciona dividindo recursivamente os dados em subconjuntos com base no atributo mais significativo.
- ID3: é um dos primeiros algoritmos de árvore de decisão. Funciona selecionando o melhor atributo para dividir os dados em cada etapa com base no ganho de informação.
- C4.5: é uma melhoria sobre o ID3 e pode lidar com dados discretos e contínuos. Utiliza uma técnica chamada poda para evitar o *overfitting*.
- C5.0: é um aprimoramento do C4.5, oferecendo melhor desempenho e eficiência.
- CHAID: é um algoritmo de árvore de decisão projetado especificamente para dados categóricos. Utiliza o teste qui-quadrado para determinar o atributo mais significativo em cada etapa.
- MARS: é um tipo de algoritmo de árvore de decisão usado para tarefas de regressão. Funciona ajustando regressões lineares por partes aos dados.

As árvores aleatórias são um tipo de algoritmo de AM derivado das árvores de decisão que se baseia na técnica de aprendizado por agrupamento. Esse método combina vários modelos de predição ou classificação mais simples para uma tarefa específica, originando modelos agregados mais complexos. Ao contrário de depender de um único modelo de árvore de decisão, as florestas aleatórias podem construir múltiplas árvores de decisão durante o treinamento, combinando suas previsões para obter uma saída final mais precisa e estável (CUTLER; CUTLER; STEVENS, 2012). Em resumo, as árvores de decisão são os blocos de construção das florestas aleatórias, e as florestas aleatórias utilizam múltiplas árvores de decisão para fazer previsões mais precisas e robustas, especialmente em conjuntos de dados complexos.

2.2.1.2 Classificadores lineares

Os classificadores lineares são algoritmos de aprendizado de máquina essenciais para tarefas de classificação, amplamente reconhecidos por sua simplicidade, interpretabilidade e eficiência. Esses algoritmos operam sob a premissa de que os dados podem ser divididos por

fronteiras de decisão lineares, como uma linha em um espaço bidimensional ou um hiperplano em espaços de dimensões superiores (MICHALSKI; CARBONELL; MITCHELL, 2013).

Um exemplo popular de classificador linear é a Máquinas de Vetores de Suporte, do inglês *Support Vector Machine* (SVM); que tem em vista encontrar um hiperplano no espaço de características que melhor separe os dados em diferentes classes. No caso da classificação, as SVM buscam encontrar o hiperplano de separação que maximiza a margem, ou seja, a distância entre os pontos de dados mais próximos de diferentes classes, conhecidos como vetores de suporte (PISNER; SCHNYER, 2020).

A ideia por trás das SVM é que, ao maximizar a margem, o classificador terá melhor generalização para novos dados (MAMMONE; TURCHI; CRISTIANINI, 2009; HEARST *et al.*, 1998). As SVM são muito utilizadas em dados textuais devido à natureza esparsa do texto, considerando também que alguns atributos são irrelevantes, mas estão correlacionados com outros de maior relevância (JOACHIMS *et al.*, 1997).

Redes neurais são modelos computacionais inspirados pelo funcionamento do cérebro humano. Elas são compostas por unidades de processamento chamadas de neurônios artificiais ou perceptrons, interconectados por conexões ponderadas. Cada neurônio recebe entradas, realiza um cálculo com base nessas entradas e em um valor de peso associado a cada conexão, e produz uma saída. Essa saída pode ser enviada para outros neurônios como entrada, formando uma rede interconectada de neurônios (ZURADA, 1992).

Uma rede neural é geralmente organizada em camadas, com uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. As camadas ocultas realizam o processamento interno dos dados, enquanto a camada de saída produz o resultado final. Durante o treinamento de uma rede neural, os pesos das conexões são ajustados com base em um algoritmo de aprendizado, como o *backpropagation*, por exemplo. O objetivo é encontrar os pesos ótimos que permitam à rede neural aprender padrões nos dados de treinamento e fazer previsões precisas em novos dados (NIELSEN, 2015).

Em PLN existem diversas aplicações de redes neurais (GOLDBERG, 2022), como:

1. Classificação de texto: as redes neurais podem classificar textos em categorias pré-definidas, como detecção de spam, AS e classificação de notícias.

2. Reconhecimento de entidades nomeadas: Redes neurais podem ser usadas para identificar e extrair informações específicas em textos, como nomes de pessoas, organizações, datas e locais.

3. AS: as redes neurais podem analisar a polaridade do sentimento em textos, determinando se uma declaração é positiva, negativa ou neutra.

4. Tradução automática: redes neurais são aplicadas no desenvolvimento de sistemas de tradução automática, permitindo a tradução entre diferentes idiomas.

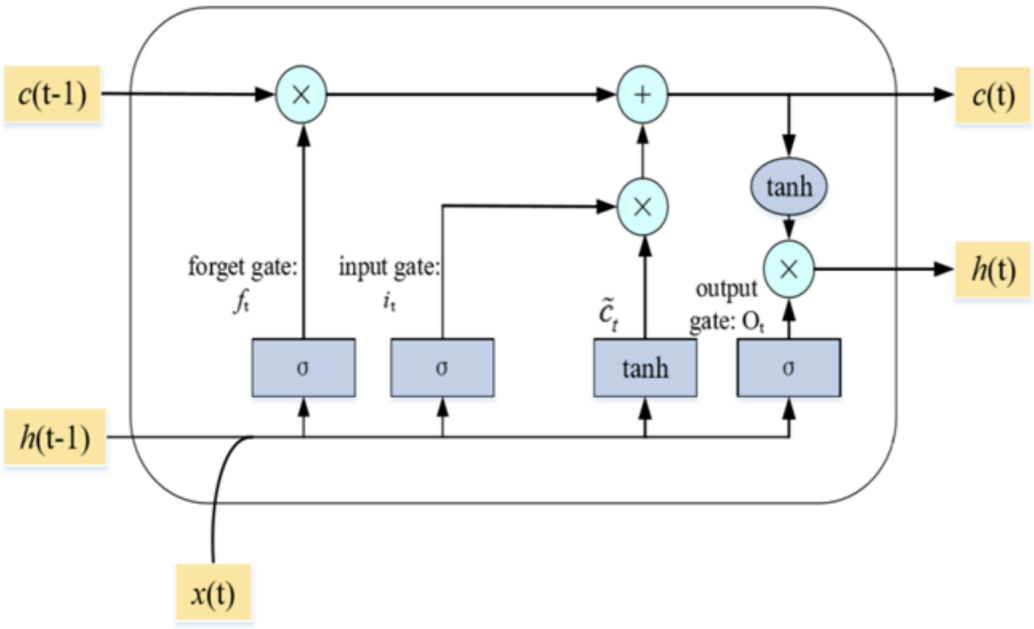
5. Geração de texto: as redes neurais podem ser utilizadas para gerar texto coerente e relevante, como em chatbots, resumos automáticos e criação de histórias.

6. Análise de tópicos: redes neurais podem ser aplicadas na identificação e classificação de tópicos em grandes volumes de textos, auxiliando na organização e categorização de informações.

As *Long Short-Term Memory* (LSTM) correspondem a uma arquitetura de rede neural recorrente (RNN) especializada em lidar com problemas de sequência, tais como previsão de séries temporais, tradução de texto, reconhecimento de fala, dentre outros. As LSTM podem aprender dependências de longo prazo em dados sequenciais, o que as torna especialmente eficazes em tarefas na qual a compreensão do contexto é crucial (YUAN; LI; WANG, 2019).

Os pontos de memória de uma rede LSTM são denominados células, que conseguem identificar os dados a serem armazenados ou descartados pela rede. É dividida em três grandes componentes, o portão de esquecimento, responsável por quais partes do estado de longo prazo devem ser apagadas; o portão de entrada, controla quais partes devem ser adicionadas ao estado de longo prazo; e o portão de saída, que controla as partes que devem ser lidas e exibidas no estado de tempo atual(YU *et al.*, 2019). A Figura 3 apresenta a estrutura básica da célula LSTM, no qual o vetor $x(t)$, entrada atual; e o estado $h(t-1)$, de curto prazo anterior; acionam quatro camadas diferentes (SHERSTINSKY, 2020).

Figura 3 – Célula LSTM



Fonte: Yu *et al.* (2019).

O funcionamento da célula LSTM é descrito a seguir:

- O portão de esquecimento (*forget gate*) controla, a partir de sua função logística σ, quais partes do estado de longo prazo devem ser apagadas, dessa maneira, retornando valores entre 0 e 1.
- O portão de entrada (*input gate*) controla quais partes dever ser adicionadas ao estado de longo prazo.
- O portão de saída (*output gate*) controla as partes que devem ser lidas e exibidas no estado de tempo atual.

2.2.1.3 Classificadores por regras de associação

Regras de associação são um tipo de técnica de mineração de dados utilizada para descobrir relações existentes entre variáveis em grandes conjuntos de dados. Essas regras identificam associações frequentes entre itens em conjuntos de dados. A ideia de minerar essas regras surgiu da análise em cestas de compras dos clientes, nos quais os clientes que compram determinados produtos são mais propensos a adquirirem um produto específico, por exemplo, os clientes que compram cerveja, tendem frequentemente a comprar amendoim (CARVALHO, 2007).

As regras de associação são representadas como uma implicação na forma $LHS \Rightarrow RHS$, em que *LHS* e *RHS* são respectivamente, o lado esquerdo (*Left Hand Side*), antecedente; e o lado direito (*Right Hand Side*), consequente. Por exemplo, em uma regra de associação $cerve\,ja \Rightarrow amendoim$, $cerve\,ja$ seria o antecedente e $amendoim$ seria o consequente. As medidas mais empregadas para regras de associação são o suporte, que representa a frequência dos padrões associativos; e a confiança, a proporção de vezes que o consequente ocorre junto ao antecedente (ZHANG; ZHANG, 2002).

Regras de obtenção de regras de associação são obtidas a partir de dois passos (AGRAWAL; IMIELIŃSKI; SWAMI, 1993):

1. Encontrar todos os conjuntos de itens que possuam suporte maior ou igual ao suporte mínimo especificado pelo usuário.
2. Utilizar todos os conjuntos de itens frequentes para gerar as regras de associação com confiança maior ou igual à confiança mínima estabelecida pelo usuário

Com base no conjunto de itens obtidos no passo 1, o objetivo do passo 2 é gerar todas as regras de associação (ZHANG; ZHANG, 2002).

2.2.1.4 Classificadores probabilísticos

O classificador Naive Bayes é o mais simples e de uso comum (MEDHAT; HASSAN; KORASHY, 2014). Ele recorre ao cálculo da probabilidade condicional de uma classe com

base na distribuição das palavras em um documento, atribuindo para a entrada a classe de maior probabilidade condicional (RISH *et al.*, 2001). Faz o uso do teorema de Bayes para esse cálculo, podendo ser expressa na equação 2.1:

$$P(C|X) = \frac{P(C)P(X|C)}{P(X)} \tag{2.1}$$

em que:

- $P(C|X)$ é a probabilidade condicional da classe C dado um conjunto de atributos X. Isso representa a probabilidade de um exemplo pertencer à classe C dado que possui os atributos X.
- $P(X|C)$ é a probabilidade de observar o conjunto de atributos X dado que a classe é C. Isso representa a probabilidade de ocorrerem os atributos X, assumindo que a classe é C.
- $P(C)$ é a probabilidade a priori da classe C. Isso representa a probabilidade de um exemplo aleatório pertencer à classe C, independentemente dos atributos.
- $P(X)$ é a probabilidade de observar o conjunto de atributos X. Isso representa a probabilidade de ocorrerem os atributos X, independentemente da classe.

Redes Bayesianas, também conhecidas como redes probabilísticas ou redes de crença, são modelos gráficos probabilísticos que representam e inferem relações de dependência probabilística entre variáveis. Elas combinam conceitos da teoria de grafos e da teoria das probabilidades para modelar relacionamentos complexos entre variáveis e realizar inferências sobre elas.

Uma rede Bayesiana é composta por nós que representam as variáveis e por arestas direcionadas que indicam as dependências probabilísticas entre elas. Cada nó representa uma variável aleatória e possui uma distribuição de probabilidade condicional que descreve a relação entre essa variável e suas variáveis pai (variáveis das quais ela depende diretamente), em que as probabilidades condicionais são estimadas a partir de dados ou conhecimento especializado.

São construídas com base no princípio de que uma variável é condicionalmente independente de todas as outras variáveis, dado o conhecimento de suas variáveis pai. Essa propriedade permite representar complexas relações de dependência de forma modular e eficiente, também, permitindo interpretabilidade, possibilitando entender o raciocínio probabilístico (MEDHAT; HASSAN; KORASHY, 2014; FRIEDMAN; GEIGER; GOLDSZMIDT, 1997).

O classificador de máxima entropia, que também é um classificador probabilístico, se baseia no princípio da entropia máxima, que encontrará o modelo mais equilibrado e imparcial dado um conjunto de restrições. No contexto da classificação, o objetivo do classificador é atribuir uma etiqueta de classe a uma instância de entrada. Durante o treinamento, o classificador busca encontrar os parâmetros que maximizam a entropia do modelo, ou seja, que geram uma

distribuição de probabilidade mais uniforme e imparcial. Isso é feito por meio da aplicação de restrições baseadas nas características do conjunto de dados. Essas restrições são utilizadas para calcular as probabilidades condicionais das classes dado um conjunto de características (NIGAM; LAFFERTY; MCCALLUM, 1999).

2.2.2 Técnicas com aprendizado de máquina não supervisionado

Técnicas de AM não supervisionado são métodos no qual o algoritmo tenta encontrar padrões nos dados sem a necessidade de rótulos ou supervisão externa. Essas técnicas são úteis quando não se tem acesso a dados rotulados ou quando se deseja explorar a estrutura subjacente dos dados, entretanto, podem ser difíceis de serem interpretados devido à descoberta de padrões abstratos dos dados, apresentam elevada complexidade computacional e necessitam de validação da rotulação para avaliação da qualidade dos resultados (MICHALSKI; CARBONELL; MITCHELL, 2013). Técnicas de AM não supervisionadas não serão abordadas neste trabalho, porém, podem ser aplicadas a trabalhos futuros relacionados ao conjunto de dados obtido por este trabalho.

2.2.3 Abordagens léxicas com uso de dicionário

A abordagem baseada em dicionário é um método utilizado no PLN e análise de texto que depende do uso de um dicionário ou léxico pré-definido. Nessa abordagem, um dicionário contém uma lista de palavras ou frases juntamente com suas etiquetas correspondentes, como etiquetas de sentimento (positivo, negativo, neutro) ou etiquetas de tópicos (por exemplo, esportes, política, entretenimento) (LIU, 2022).

Durante o processo de análise, o texto é comparado com as entradas do dicionário. Se uma palavra ou frase do texto corresponder a uma entrada, ela é atribuída à etiqueta correspondente. Essa abordagem pode ser aplicada a várias tarefas, tais como AS, classificação de tópicos e MO (SEBASTIANI; ESULI, 2006).

As abordagens baseadas em dicionário podem ser simples de implementar, oferecendo vantagens como interpretabilidade, pois os resultados são baseados em mapeamentos explícitos de palavras etiquetadas. Também, podem ser computacionalmente eficientes, especialmente ao lidar com grandes volumes de dados textuais (LIU, 2022; CAMBRIA *et al.*, 2013).

No entanto, as abordagens baseadas em dicionário possuem limitações. Elas podem não capturar nuances e significados dependentes do contexto das palavras. Também, podem ser sensíveis a variações no uso da linguagem, incluindo gírias, abreviações ou erros ortográficos que podem não estar incluídos no dicionário (SANTOS; LADEIRA, 2014). Portanto, a seleção e curadoria cuidadosa do dicionário são essenciais para obter resultados precisos. Alguns dos conjuntos de dados léxicos rotulados são WorldNet, OpLexicon e SentiLex.

A WorldNet é uma base de dados lexical e um recurso linguístico utilizado no PLN. É um

lexicon (dicionário) semântico que agrupa palavras em conjuntos de sinônimos chamados synsets (conjuntos sinônimos). Cada synset em WordNet representa um conceito semântico e contém um conjunto de palavras relacionadas que são consideradas sinônimos dentro desse contexto específico. Por exemplo, o synset "carro"pode conter palavras como "automóvel", "veículo"e "automotor", todas relacionadas ao mesmo conceito. Além das relações de sinônimos, WordNet também captura outras relações semânticas entre palavras, como hiperônimos (palavras mais gerais que englobam um conceito) e hipônimos (palavras mais específicas incluídas em um conceito) (MILLER, 1995).

O OpLexicon e o Sentilex são bases de dados específicas para a língua portuguesa, utilizadas em tarefas de AS e MO. Trata-se de dicionários de polaridade de palavras, em que cada palavra é atribuída a uma polaridade ou emoção específica. É construído com base em anotações manuais realizadas por especialistas, que atribuem as polaridades às palavras de acordo com seu contexto e uso comuns. Os dicionários contém uma extensa lista de palavras, incluindo adjetivos, advérbios, verbos e substantivos, sendo frequentemente associados as expressões de opinião e sentimentos, possuindo mais de 15.000 palavras classificadas para o OpLexicon e 82.000 para o Sentilex (MACHADO; PARDO; RUIZ, 2018; FILHO; PARDO; ALUÍSIO, 2013).

Hutto e Gilbert (2014) propuseram o *Valence Aware Dictionary for sEntiment Reasoner* (VADER) que é um recurso léxico e uma ferramenta de AS em língua inglesa. Ele é projetado para identificar e medir o sentimento expresso em um texto, atribuindo uma pontuação de polaridade para cada palavra ou expressão. O VADER foi desenvolvido considerando as especificidades da linguagem e o contexto das RS, em que as emoções e sentimentos são frequentemente expressos de maneiras sutis e com uso de gírias e emojis. Ele consegue capturar nuances e ambiguidades de sentimentos, além de lidar com a negação, intensidade e sarcasmo presentes em muitas mensagens online. O dicionário subjacente ao VADER é composto por um conjunto de palavras pré-classificadas com pontuações de polaridade. Essas pontuações variam de -1 a +1, indicando a intensidade do sentimento negativo ou positivo associado a cada palavra com base nas pontuações das palavras, é possível calcular uma pontuação geral de sentimento para todo o texto analisado, em que a pontuação individual de cada componente é normalizada pela equação 2.2.

$$\frac{x}{\sqrt{x^2 - \alpha}} \tag{2.2}$$

em que:

- x é a soma das pontuações sentimentais individuais de cada componente.
- α é o parâmetro de normalização.

2.2.4 Abordagens léxicas com uso de corpus

A abordagem baseada em corpus é um método utilizado em PLN e linguística baseada na análise e exploração de grandes conjuntos de dados textuais. Nessa abordagem, o corpus é utilizado como fonte de informações para diferentes tarefas linguísticas, como AS, extração de informações, tradução automática, dentre outras. Essa abordagem se baseia na coleta e preparação do conjunto de dados, que pode incluir textos de diferentes domínios (LIU, 2022). Em seguida, técnicas de PLN podem ser aplicadas para extrair informações relevantes, identificar padrões, construir modelos estatísticos e realizar análises linguísticas. Essa abordagem é mais empírica e orientada ao domínio para o PLN, pois considera a variação linguística, contextos de uso e padrões observáveis nos textos do domínio. Dessa maneira, ajuda a solucionar problemas relacionados a opinião com contexto em domínio específico [40].

2.2.4.1 Abordagens lexicais estatísticas com uso de corpus

As abordagens lexicais baseadas em estatística são métodos utilizados no PLN que se baseia em estatísticas e frequências de palavras em um corpus para analisar e extrair informações relevantes. Nessa abordagem, as estatísticas são usadas para atribuir pesos ou pontuações às palavras com base em sua frequência de ocorrência no corpus. Isso pode incluir a contagem de palavras, a frequência relativa de palavras em relação ao tamanho do corpus ou outros cálculos estatísticos, um exemplo é o *Bag-of-Words* (BoW) que será abordado em outra seção (ZHANG; JIN; ZHOU, 2010).

A abordagem baseada em estatísticas pode envolver o uso de técnicas como a frequência de termos, em inglês *Term Frequency* (TF); a frequência inversa de documento, em inglês *Inverse Document Frequency* (IDF); a medida TF-IDF (que combina as duas), bem como outras métricas estatísticas. Essa abordagem é vantajosa porque considera a importância relativa das palavras com base em sua frequência e distribuição nos dados, também permite que as palavras mais frequentes e informativas tenham um papel maior na análise, enquanto palavras menos relevantes são consideradas com menos peso (RAMOS *et al.*, 2003). O coeficiente da TF-IDF pode ser visto na equação 2.3.

$$TF - IDF(x,y) = f_{x,y} \log \frac{N}{Df_x} \tag{2.3}$$

em que:

- $f_{x,y}$ é frequência do termo x no documento y.
- N número de documentos que contém o termo x.
- Df_x é o número total de documentos.

2.2.4.2 Abordagens lexicais semânticas com uso de corpus

As abordagens lexicais baseadas em semântica são métodos utilizados no PLN que se concentra na análise e exploração do significado das palavras em um contexto linguístico. Nessa abordagem, as informações semânticas das palavras são utilizadas para entender e extrair o sentido dos textos (SUN; LUO; CHEN, 2017). Essa abordagem envolve o uso de recursos lexicais, como dicionários semânticos, WorldNet, OpLexicon, SentiLex, dentre outros; que fornecem informações sobre os significados e as relações entre as palavras.

Ao utilizar a abordagem lexical baseada em semântica, as palavras em um texto são mapeadas para seus respectivos significados e relações semânticas. Isso permite uma compreensão mais profunda do conteúdo textual, incluindo a detecção de ambiguidades, a identificação de termos-chave e a inferência de informações implícitas. Além dos recursos lexicais, a abordagem também pode se beneficiar de técnicas de PLN, como análise de dependência, análise semântica e modelagem de tópicos, para capturar e representar o significado das palavras em um contexto mais amplo (LIU, 2022; SUN; LUO; CHEN, 2017).

2.3 As redes sociais e o Twitter

Ao longo das últimas duas décadas, as RS se tornaram uma parte integral da vida de mais de um terço da população mundial. Plataformas como Facebook, Twitter, Instagram, dentre outras, transformaram a maneira pela qual pessoas expressam suas opiniões, emoções e compartilham os seus cotidianos (GALLAGHER, 2017).

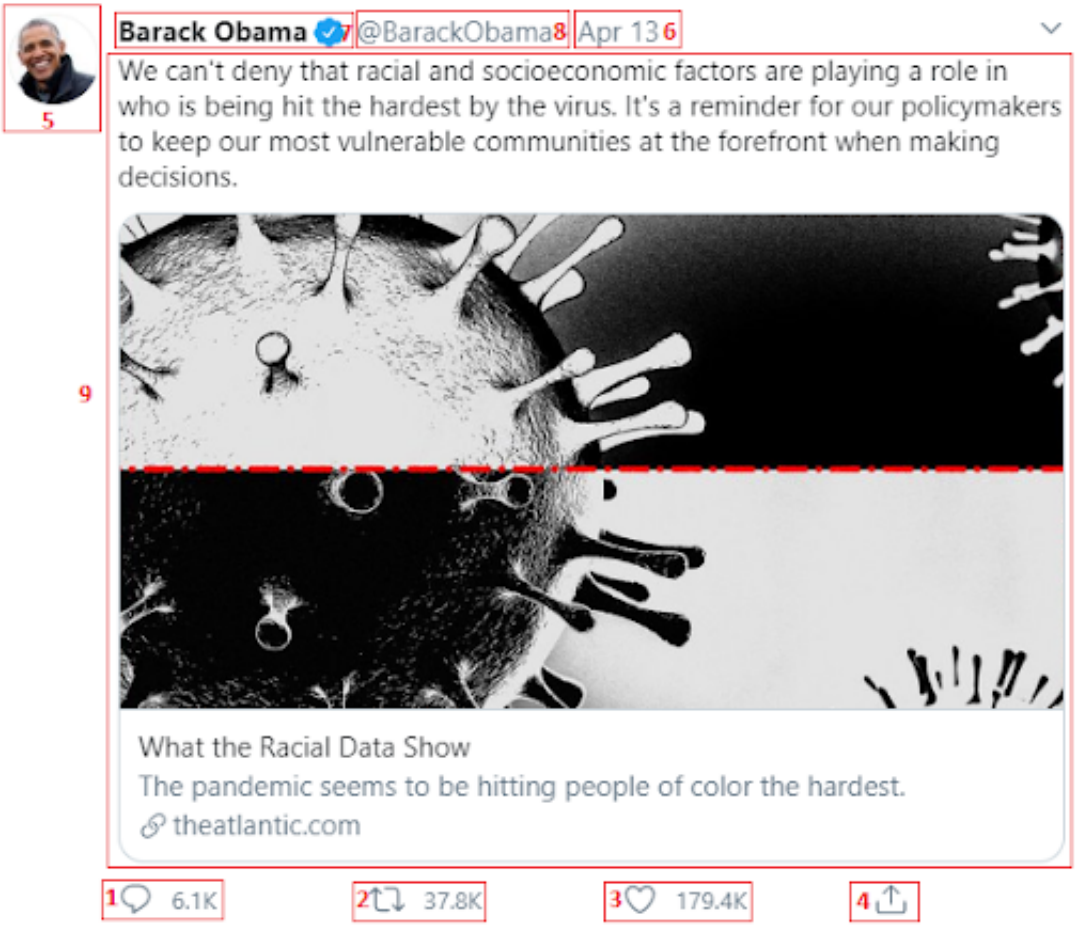
Essas redes têm sido amplamente adotadas devido à sua capacidade de conectar pessoas, mesmo que estejam distantes geograficamente; no compartilhamento de informações e interesses, ao permitirem que usuários partilhem de fotos, vídeos, links e *hobbies* com seus contatos; na comunicação, como mensagens privadas ou de bate-papo; na construção de comunidades, ao permitirem que pessoas que têm interesses específicos debatam e compartilhem de ideias; como ferramentas de promoção e marketing, pois empresas fazem o uso para promoção da venda de produtos e serviços; e como fonte de informação, pois agências de notícias fazem o uso para compartilhamento de eventos e atualizações em tempo real, permitindo que usuários sejam informados de maneira rápida sobre os acontecimentos (LEE, 2018; SOLOMON; TUTEN, 2017).

O Twitter é uma plataforma de RS em formato de *microblogging* que contém notícias, informações e opiniões dos usuários (MISHRA; RAJNISH; KUMAR, 2016). Lançado em 2006, o Twitter se tornou uma das RS mais populares e influentes do mundo. Ele permite que os usuários publiquem mensagens, sigam outros usuários e sejam seguidos por eles. Dessa maneira, criando uma rede de interações, permitindo que as pessoas compartilhem pensamentos, notícias, opiniões e atualizações pessoais em tempo real. O serviço alcançou a marca de 330 milhões de usuários mensais em 2019, com uma média diária de 152 milhões de usuários ativos e 500

milhões de tweets (ASLAM, 2023).

Um tweet é uma mensagem curta publicada no Twitter, sendo a unidade básica de conteúdo no Twitter, limitada a 280 caracteres de texto (até 2017 foi limitada a 140 caracteres). Pode conter diferentes tipos de conteúdo, incluindo texto, links, imagens e vídeos. A página do perfil de cada usuário, contém informações como o nome, menções, respostas, *retweet*, sendo a replicação da mensagem por outros usuários; e *like*, o qual é um tipo de aprovação dos usuários da rede (TWITTER, 2020). A estrutura de um tweet é apresentada na Figura 4.

Figura 4 – Corpo de um *tweet* - Exemplo Barack Obama



Fonte: Elaborada pelo autor.

Na Figura 4, tem-se circulado e numerado em vermelho:

1. Quantidade de respostas
2. Quantidade de *retweets*
3. Quantidade de *likes*
4. Opções extra do *tweet*, tais como: link, adicionar a favoritos, dentre outros
5. Foto do perfil

6. Data de publicação
7. Nome do perfil com verificação, denotada pela *check mark*
8. Nome do usuário
9. Corpo da mensagem, limitada em 280 caracteres

Nota-se que para figuras públicas existe uma verificação que prova a autenticidade do perfil. Também, toda resposta tem a mesma estrutura de um *tweet*, assim então, cada resposta pode ser considerada um *tweet*. No perfil do usuário, existem outras informações, como pode ser visto na Figura 5.

Figura 5 – Perfil de um usuário do Twitter - Exemplo Barack Obama

Fonte: Elaborada pelo autor.

Na Figura 5, tem-se circulado e numerado em vermelho:

1. Nome do perfil com a quantidade de *tweets*
2. Uma pequena biografia do usuário, localização, data que entrou na rede, dentre outras

3. Dados gerais, como quantidade de seguidores e de pessoas que seguem
4. Nome do perfil com a quantidade de *tweets*
5. Opções gerais relacionadas ao usuário
6. Mural onde são salvas as postagens, respostas, *tweets* que gostou, dentre outros

Ao longo das últimas duas décadas, as RS, como o Facebook, Twitter e Instagram, revolucionaram como as pessoas expressam suas opiniões, compartilham suas vidas e se conectam entre si. O Twitter, em particular, se destacou como uma plataforma de microblogging que se tornou uma das RS mais populares e influentes do mundo desde seu lançamento. Além de permitir que os usuários compartilhem notícias, informações e opiniões em tempo real, o Twitter também oferece ferramentas poderosas para pesquisa e desenvolvimento, tornando-se uma fonte valiosa de dados para análises em diversos campos. O estudo proposto neste trabalho visa investigar como técnicas de PLN e AS podem ser aplicadas para extrair informações relevantes das mensagens no Twitter durante eventos de grande impacto, como a pandemia da COVID-19. Essa pesquisa é essencial não apenas para entender o comportamento e a comunicação das pessoas durante crises, mas também para desenvolver estratégias eficazes para lidar com essas situações no futuro.

2.4 Mineração de textos

A MT pode ser definida como um processo de extração de informações úteis a partir de documentos textuais escritos em linguagem natural (REZENDE, 2003), a qual é realizada por uma infinidade de algoritmos de AM, que podem resultar em aplicações a partir da informação gerada (AGGARWAL; AGGARWAL, 2018). O *framework* proposto por Rezende (2003) contém cinco etapas genéricas que podem ser aplicadas a distintos problemas, as quais são a Identificação do Problema, o Pré-Processamento, a Extração de Padrões, o Pós-Processamento e a Utilização do Conhecimento, que pode ser visto na Figura 6.

Figura 6 – Processo de mineração de textos

Fonte: Rezende (2003).

1. Identificação do problema: corresponde ao estudo do domínio da aplicação tal qual a definição de objetivos a serem alcançados no processo de MT (AGGARWAL; AGGARWAL, 2018). Parte do sucesso da extração de conhecimento depende da aptidão e contexto dos analistas para com o domínio explorado na tarefa de localizar padrões. Dessa maneira, é de grande importância um estudo prévio a fim de se adquirir destreza com relação ao domínio (RUEDEN *et al.*, 2021).

2. Pré-processamento: é comum que dados disponibilizados para análise não estejam em um formato adequado para a extração de conhecimento, com isso, fazem-se necessários os usos de métodos de tratamento, limpeza e redução do volume de dados antes de iniciar a etapa do enriquecimento de informação. A partir da mesma fonte de informação são providos três conjuntos de dados, dessa maneira, é importante uma etapa de fusão de informação, ou integração, na qual unificamos todos os conjuntos em uma única fonte de dados para ser utilizada na Extração de Padrões (AGGARWAL; AGGARWAL, 2018).

3. Extração de padrões: corresponde a etapa destinada a direcionar ao cumprimento dos objetivos definidos na Identificação do Problema, em que serão realizadas escolhas relacionadas às configurações de execuções ou aplicações de algoritmos com a finalidade da

extração de conhecimento (AGGARWAL; AGGARWAL, 2018).

4. Pós-processamento: os resultados da extração de padrões são avaliados utilizando métricas específicas definidas durante a etapa de identificação do problema. A escolha da métrica depende do tipo de aplicação utilizada, como agrupamento, associação, regressão ou classificação.

5. Utilização do conhecimento: a partir das etapas anteriores e da validação das técnicas aplicadas, o conhecimento extraído é fundamental para os processos de classificação e comparação dos resultados entre os modelos. A etapa de utilização do conhecimento deve ser constantemente monitorada para garantir a boa qualidade das métricas de validação na etapa de pós-processamento, caso haja degradação das métricas, o processo pode retroceder para a etapa responsável pela extração de padrões ou pós-processamento. Também, se houver alteração de escopo dos objetivos, o processo deve retornar para a etapa de identificação do problema.

2.5 Modelos de representação

Modelos de representação são técnicas ou estruturas que permitem representar informações de maneira organizada e adequada para a realização de tarefas específicas. Em AM e PLN, os modelos de representação são usados para converter dados não-estruturados, como texto, em uma forma que possa ser processada por algoritmos de aprendizado. A escolha do modelo de representação é um passo crucial no processo de MT, pois uma representação adequada pode reduzir o consumo de recursos computacionais, diminuir o tempo de processamento e melhorar a capacidade de aprendizado dos algoritmos (SINOARA *et al.*, 2019; ROSSI, 2016).

2.5.1 Bag-of-Words

O BoW é um modelo de representação textual no âmbito do espaço vetorial. Ele trata um documento de texto como uma sacola de palavras, sem considerar a ordem ou a estrutura gramatical. Nesse modelo, cada documento é representado por um vetor numérico, em que cada elemento corresponde a uma palavra única encontrada no corpus. A contagem ou frequência das palavras é usada para determinar os valores dos elementos do vetor (AGGARWAL; AGGARWAL, 2018).

Para construir o modelo, é necessário criar um vocabulário único a partir do conjunto de documentos de treinamento. Cada palavra única é atribuída a uma posição específica no vetor de representação. Em seguida, cada documento é analisado e a frequência de ocorrência de cada palavra no vocabulário é contada. Essas contagens são usadas para preencher os elementos do vetor de representações correspondentes (AGGARWAL; AGGARWAL, 2018).

Textos podem apresentar distribuição assimétrica da frequência das palavras, as quais são consideradas *outliers* para as de alta frequência, que podem ocasionar perda da qualidade de informação. Com isso, técnicas para normalização dos valores são empregadas para equalizar os termos na representação (AGGARWAL; AGGARWAL, 2018; SINOARA *et al.*, 2019; ROSSI, 2016). A TF-IDF, já citada anteriormente, é uma ferramenta léxica que suporta essa equalização para a representação do BoW.

2.5.2 Modelos de word embeddings

O BoW supre muitos dos problemas em PLN, no entanto, para aplicações nas quais a representação sequencial do texto se faz necessária, exigindo a compreensão semântica, ele não é aplicável. Com isso, modelos de word embeddings são técnicas utilizadas para representar palavras em um espaço vetorial contínuo, em que esses modelos capturam relações semânticas e sintáticas entre as palavras com base em sua distribuição em um corpus de texto (AGGARWAL; AGGARWAL, 2018).

Existem diferentes tipos de modelos de word embeddings, sendo os mais comuns o Word2Vec, o GloVe (Global Vectors for Word Representation em inglês) e o FastText. Esses modelos são treinados em grandes quantidades de texto não rotulado, com o objetivo de aprender representações vetoriais densas para as palavras (DHARMA *et al.*, 2022). O Word2Vec utiliza uma abordagem baseada em duas arquiteturas principais: *Continuous Bag-of-Words* (CBoW) e *Skip-Gram* (SG). Ele prevê uma palavra com base em seu contexto, enquanto o SG prevê o contexto dado uma palavra. Esses modelos aprendem a representação vetorial de cada palavra de forma que palavras semelhantes tenham vetores próximos no espaço (JANG; KIM; KIM, 2019).

O GloVe, por sua vez, utiliza uma abordagem que combina informações de co-ocorrência global e local das palavras em um corpus. Ele constrói uma matriz de co-ocorrência que registra a frequência de palavras ocorrendo próximas umas das outras. Em seguida, utiliza técnicas de decomposição de matriz para obter as representações vetoriais das palavras (PENNINGTON; SOCHER; MANNING, 2014).

O FastText é uma extensão do Word2Vec que considera a estrutura de subpalavras das palavras. Em vez de tratar cada palavra como uma unidade indivisível, o FastText divide as palavras em n-gramas menores, aprendendo as representações para essas subpalavras, permitindo que o modelo lide melhor com palavras raras ou desconhecidas (BOJANOWSKI *et al.*, 2017).

2.5.3 Modelos de transformers

Os modelos baseados em *transformers* são diferentes dos modelos de linguagem mais simples, nos quais as palavras são apresentadas sequencialmente. Nos *transformers*, todas as palavras de uma sequência podem ser processadas simultaneamente, permitindo a construção de *embeddings* no domínio da sentença. Uma das vantagens da sua utilização se dá pela paralelização

de treinamento, otimizando a capacidade de processamento (VASWANI *et al.*, 2017). Outra vantagem dos *transformers* é a implementação do mecanismo de atenção no decodificador. Dessa maneira, cada palavra possui um vetor de valores que indica quais outras palavras da sentença são mais relevantes para o processamento da palavra atual. Isso permite capturar relações semânticas mais complexas e melhora a qualidade das representações. Dessa maneira, uma mesma palavra em diferentes posições da sentença pode possuir *embeddings* diferentes, refletindo sua contextualização, contribuindo para uma melhor representação.

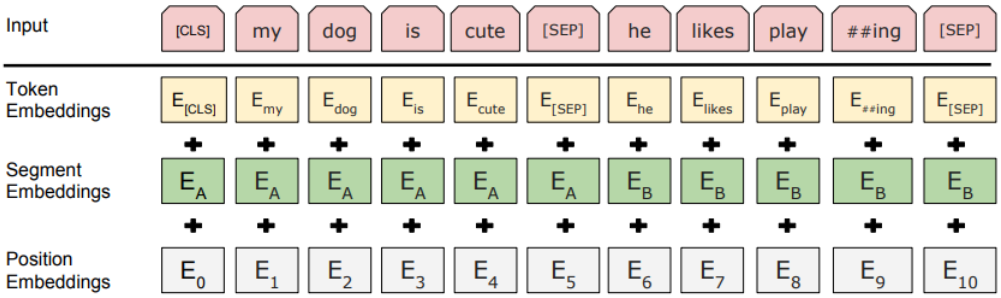
Bidirectional Encoder Representations from Transformers (BERT) é um modelo de linguagem pré-treinado desenvolvido pela Google. Ele se baseia na arquitetura *transformers*. A principal inovação do BERT é a sua capacidade de aprender representações de palavras contextualizadas, ou seja, as palavras são codificadas considerando o contexto em que estão inseridas (DEVLIN *et al.*, 2018).

Também, emprega o princípio de transferência de aprendizado, que se refere ao processo em que um modelo treinado para resolver um problema é aplicado de alguma maneira a um segundo problema relacionado (MARCACINI *et al.*, 2018). Essa abordagem é particularmente valiosa em aplicações que envolvem imagens, áudio e grandes conjuntos de dados textuais, uma vez que os modelos são treinados extensivamente em vastas bases de dados ao longo de dias ou semanas, utilizando recursos de processamento exclusivos de instituições especializadas. Posteriormente, esses modelos podem ser disponibilizados para uso geral. A partir desse treinamento inicial, uma ou mais camadas do modelo são ajustadas para se adequarem ao problema de interesse, otimizando assim o desempenho.

O BERT é estruturado em várias camadas de codificadores *transformers*, no qual duas arquiteturas distintas do modelo diferem na quantidade de *transformers*, núcleos de atenção e tamanho da camada oculta. O $\text{BERT}_{\text{BASE}}$ é constituído por 12 camadas, 12 núcleos de atenção e uma camada oculta com tamanho de 768, ao passo que o $\text{BERT}_{\text{LARGE}}$ apresenta 24 camadas, 16 núcleos de atenção e uma camada oculta com tamanho de 1024.ance e a capacidade de generalização, totalizando respectivamente 110 e 340 milhões de parâmetros (ACHEAMPONG; NUNOO-MENSAH; CHEN, 2021).

A representação de entrada do BERT para a frase *"my dog is cute. he likes playing"* é ilustrada na Figura 7. O primeiro *token* de cada sequência é sempre um *token* de classificação especial. Se um *token* não estiver presente no vocabulário do WordPieces, que é um mecanismo que faz a separação dos *tokens* que podem ser palavras, parte de palavras ou pontuação; o modelo procura o *token* mais próximo para o prefixo e cria um novo *token*. Além dos vetores de *embeddings*, são adicionados um vetor que indica a qual sentença o *token* pertence e outro vetor que indica a posição do *token* na sentença. A representação de entrada final do BERT é obtida pela soma dos vetores de *word embeddings*, de posição e de segmento.

Figura 7 – Representação de entrada do BERT.



Fonte: Devlin *et al.* (2018).

Ao contrário de outros modelos que treinam palavras de forma unidirecional (da esquerda para a direita ou da direita para a esquerda), o BERT utiliza um treinamento bidirecional, analisando o contexto completo da sentença, considerando as palavras anteriores e posteriores a cada palavra, permitindo uma melhor compreensão das nuances e relações de sentido (DEVLIN *et al.*, 2018; HABIMANA *et al.*, 2020).

O treinamento do BERT envolve duas etapas principais: pré-treinamento e ajuste fino. Na etapa de pré-treinamento, o modelo é treinado em abundância de texto não rotulado, aprendendo a prever palavras mascaradas e a compreender a relação entre pares de sentenças. Na etapa de ajuste fino, o modelo é refinado em tarefas específicas usando dados rotulados (DEVLIN *et al.*, 2018).

2.6 Métricas para validação de resultados

Nesta seção são apresentadas às métricas utilizadas para avaliação das diferentes abordagens de AM.

A matriz de confusão, exibida na Tabela 1, é uma ferramenta fundamental na avaliação de algoritmos de classificação. A matriz de confusão mostra a performance de um modelo de classificação, permitindo a visualização dos acertos e erros em diferentes classes. Com base nos valores da matriz, é possível calcular diversas métricas de avaliação, como precisão (2.5), acurácia (2.4), sensibilidade (2.6) e F1-score (2.7).

Tabela 1 – Matriz de confusão

		Valor Predito	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Fonte: Chowdhary e Chowdhary (2020).

- Verdadeiro positivo (VP): segmentos que pertencem à classe positiva sendo classificados corretamente como positivos.
- Falso negativo (FN): segmentos que pertencem à classe negativa sendo classificados como positivos.
- Falso positivo (FP): segmentos que pertencem à classe positiva sendo classificados como negativos.
- Verdadeiro negativo (VN): segmentos que pertencem à classe negativa sendo classificados corretamente como negativos.

A acurácia representa a proporção de predições corretas (verdadeiros positivos e verdadeiros negativos) em relação ao número total de predições.

$$Acuracia = \frac{VP+VN}{VP+VN+FP+FN} \tag{2.4}$$

A precisão representa a proporção de predições corretas feitas pelo modelo em relação ao total de predições positivas.

$$Precisao = \frac{VP}{VP+FP} \tag{2.5}$$

A sensibilidade mede a proporção de predições positivas que foram corretamente classificadas pelo modelo.

$$Sensibilidade = \frac{VP}{VP+FN} \tag{2.6}$$

A métrica F1-Score combina precisão e sensibilidade para gerar uma métrica de qualidade global da avaliação do modelo.

$$F1 = 2 \times \frac{Precisao \times Sensibilidade}{Precisao + Sensibilidade} \tag{2.7}$$

2.7 Trabalhos relacionados

Nesta seção é feito um resumo sobre trabalhos relacionados a AS em RS com as técnicas de AM citadas anteriormente.

Giachanou e Crestani (2016) fizeram uma visão geral de trabalhos relacionados ao PLN em RS anteriores ao ano de 2016. Os estudos foram realizados no Twitter e abordaram assuntos como AS, MO, detecção de ironia, dentre outros.

Rahman, AlOtaibi e AlShehri (2019) propuseram a combinação de algoritmos de AM supervisionado e não supervisionado para AS no Twitter. Por meio de *tweets* extraídos por uma API da plataforma com assuntos relacionados às marcas McDonalds e KFC para encontrar o mais popular, puderam aplicar diferentes algoritmos e técnicas estatísticas. Inicialmente os dados não rotulados foram classificados com o uso de abordagens lexicais, sendo posteriormente explorados por técnicas supervisionadas, em que a abordagem com o uso de entropia máxima apresentou melhor desempenho. O trabalho também concluiu o McDonalds como mais popular com relação a avaliações positivas e negativas.

Zimbra *et al.* (2018) avaliaram o estado-da-arte para AS em RS, avaliando 28 abordagens em cinco conjuntos de dados distintos relacionados ao Twitter. Os resultados revelaram que, no geral, as técnicas apresentaram desempenho insatisfatório, com uma média na precisão de classificação de 61%. Também, houve uma grande variação nas precisões das abordagens, chegando a 31%, no qual as técnicas específicas de domínio obtiveram melhores resultados, superando as abordagens mais generalistas em 11%.

Wagh e Punde (2018) fizeram uma revisão das técnicas utilizadas na extração de sentimentos de *tweets*. O estudo mostrou que em uma associação da WordNet com abordagens de AM como SVM, Naive Bayes e Máxima Entropia, podem obter uma melhoria no resultado. No trabalho essa abordagem híbrida teve um aumento na precisão de até 5%.

Villavicencio *et al.* (2021) fizeram um estudo com dados de RS para entender o sentimento presente na população das Filipinas um mês após o início das vacinações contra a COVID-19. Esse estudo auxiliou o governo a analisar sua resposta perante a crise sanitária. Os sentimentos foram classificados com a abordagem de Naive Bayes, obtendo uma precisão superior a 81%.

Murthy *et al.* (2020) fizeram a análise de sentimentos com o uso de LSTM para a classificação binária entre positivo e negativo em opiniões textuais dos usuários em RS. Obtiveram resultado extremamente satisfatório, atingindo acurácia superior a 85% para a tarefa de classificação.

Carvalho e Plastino (2021) realizaram uma revisão das técnicas supervisionadas consideradas estado-da-arte em 22 conjuntos de dados do Twitter, utilizando diferentes abordagens e combinações na classificação da polaridade do sentimento. Concluíram que o conjunto de atributos e ferramentas léxicas foram os que apresentaram maior impacto para os modelos.

Endsuy (2021) utilizaram de uma análise exploratória e do VADER para analisar dados das eleições presidenciais dos Estados Unidos da América no ano de 2020 em RS, com aprofundamento quanto a localização e a variação do sentimento. Os autores obtiveram boa precisão e foi possível identificar as nuances relacionadas aos partidos e suas regionalizações.

Jianqiang e Xiaolin (2017) compararam os efeitos de diferentes métodos de pré-processamento de texto no desempenho da classificação de sentimentos em cinco conjuntos de dados do Twitter.

Foram realizadas combinações e associações entre técnicas de pre processamento, no qual os modelos mais sensíveis quanto a escolha do pré-processamento foram os baseados em árvores de decisão.

2.8 Considerações finais

O surgimento e a proliferação das RS nas últimas décadas transformaram profundamente como as pessoas se comunicam e interagem globalmente. Com mais de um terço da população mundial utilizando essas plataformas e representando mais de dois terços dos usuários da internet, as RS se tornaram uma parte essencial da sociedade contemporânea, oferecendo um espaço virtual para expressão, compartilhamento e conexão. Nesse contexto, o Twitter, uma das principais redes sociais, tem sido especialmente relevante durante eventos de grande impacto, como a pandemia da COVID-19.

A pandemia, causada pelo novo coronavírus, resultou em um aumento significativo no uso das redes sociais, à medida que as pessoas buscavam informações, expressavam suas opiniões e buscavam conexão em meio ao distanciamento social. Com a ascensão das redes sociais como uma fonte primária de informação e interação durante a pandemia, surge a necessidade de compreender não apenas o conteúdo compartilhado, mas também os sentimentos e opiniões expressos pelos usuários.

Nesse contexto, a mineração de textos no Twitter torna-se uma ferramenta crucial para entender os padrões de comportamento e sentimentos dos usuários durante a pandemia. A MT, que envolve a extração de informações úteis de documentos escritos em linguagem natural, é fundamental para analisar a enorme quantidade de dados gerados pelas interações dos usuários nas RS. AS tem em vista extrair a subjetividade e a polarização presentes nos textos, possibilitando uma compreensão mais profunda das opiniões e emoções dos usuários.

Para realizar a AS em *tweets* relacionados à COVID-19 no Brasil, é necessário um processo cuidadoso de pré-processamento dos dados coletados do Twitter. Isso inclui a limpeza, normalização e transformação dos dados em um formato adequado para a análise. Além disso, a extração de padrões a partir desses dados requer aplicar algoritmos de aprendizado de máquina, como árvores de decisão, árvores aleatórias, LSTM, BERT, Naive Bayes, regressão logística e SVM, para classificar os sentimentos expressos nos tweets como negativos, neutros ou positivos.

Com a aplicação dessas técnicas de MT e AS, é possível não apenas entender melhor o impacto e evolução da pandemia da COVID-19 em interações das pessoas nas RS, mas também gerar análises comparativas entre os classificadores de AM a partir de dados rotulados por algoritmos léxicos nesse conjunto de dados proveniente do Twitter durante a pandemia da COVID-19 no Brasil.

CAPÍTULO

3

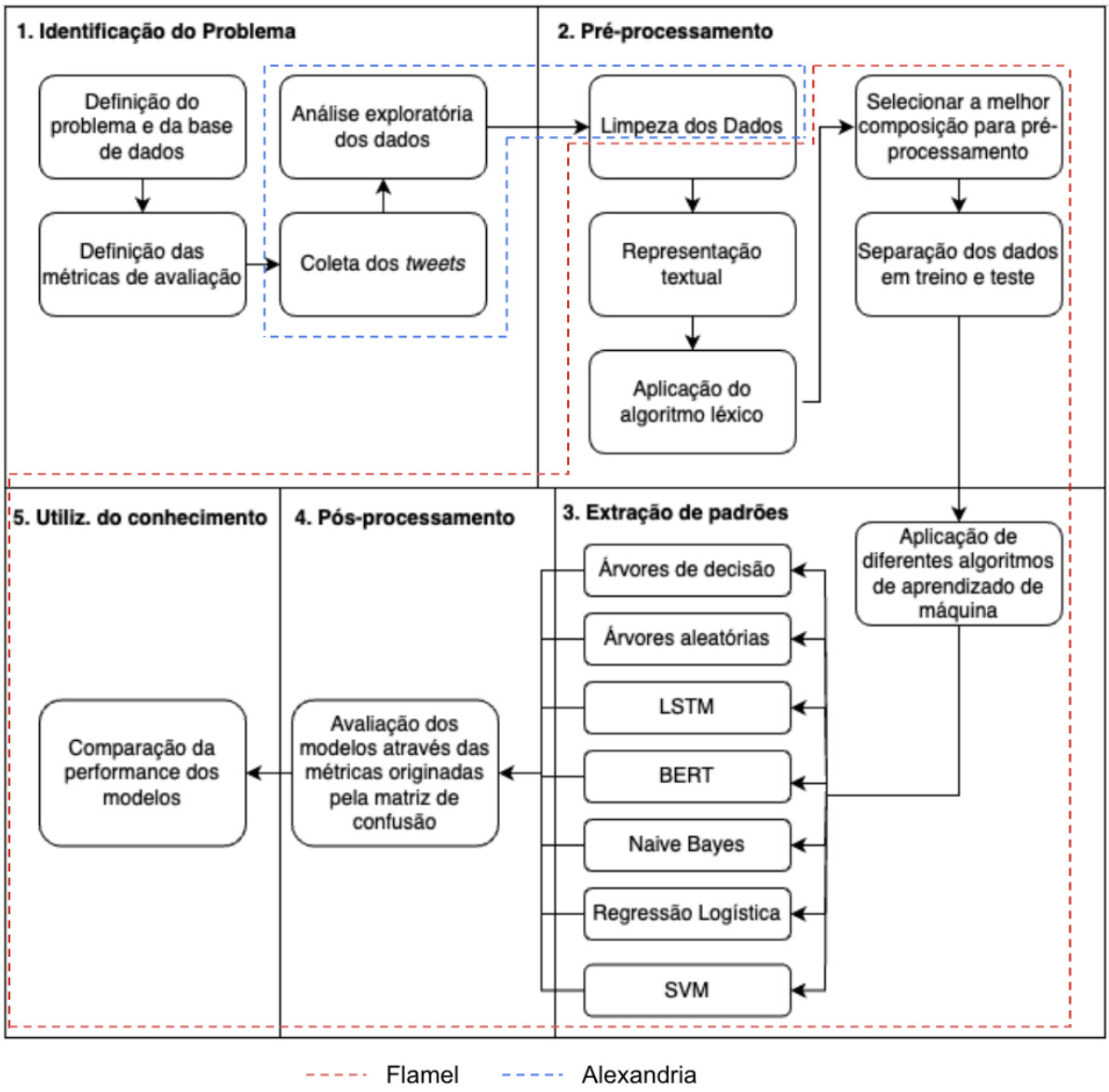
ANÁLISE DE SUBJETIVIDADE E COMPARAÇÃO DE CLASSIFICADORES COM DADOS DO TWITTER

Este capítulo visa realizar uma análise comparativa entre algoritmos de AM na tarefa de classificação do sentimento em tweets. Para alcançar esse objetivo, propomos uma abordagem que combina métodos léxicos para rotular o sentimento dos tweets como negativo, neutro ou positivo, com a utilização de algoritmos de AM. Esta análise se fundamenta nos tópicos apresentados no capítulo 2, cujos resultados são demonstrados no capítulo 4.

3.1 Visão geral da proposta de mineração de textos com dados do Twitter

A MT refere-se à extração de informações e conhecimento úteis de documentos escritos em linguagem natural (REZENDE, 2003). Esse processo envolve uma variedade de algoritmos estatísticos e de AM, os quais podem resultar em aplicações práticas baseadas no conhecimento obtido (AGGARWAL; AGGARWAL, 2018). O *framework* proposto por Rezende (2003) ilustrado na Figura 6, compreende cinco etapas genéricas que podem ser aplicadas a diferentes problemas: identificação do problema, pré-processamento, extração de padrões, pós-processamento e utilização do conhecimento. A Figura 8 apresenta os processos para o desenvolvimento dessas etapas, baseando-se no *framework* adaptado e detalhado por Peixoto (2021) para a MT. Cada processo será descrito nas subseções seguintes.

Figura 8 – *Framework* adaptado para mineração de textos do Twitter



Fonte: Elaborada pelo autor.

Na Figura 8, observam-se os processos de identificação do problema, em que são realizados estudos do domínio da aplicação e definição dos objetivos; o pré-processamento, sendo responsável pela adequação do conjunto de dados; a extração de padrões, a qual é responsável pela extração de conhecimento; o pós-processamento, o qual avalia a extração de padrões por meio das métricas escolhidas; e a etapa de utilização do conhecimento, que permite análises adicionais e aplicações práticas com o conhecimento adquirido.

Em comparação com o *framework* adaptado e detalhado por Peixoto (2021) há pequenas alterações referentes ao domínio para a identificação do problema e etapas adicionais para pré-processamento, em função da necessidade da aplicação de um algoritmo léxico para a rotulação dos dados, assim como a seleção da melhor composição de técnicas para pré-processamento, porque os algoritmos escolhidos para a etapa de extração de padrões possuem esse requisito.

Também, há diferenças quanto ao conhecimento utilizado, em que os modelos são comparados em termos de desempenho.

3.2 Identificação do problema

Com a ascensão das redes sociais, houve proliferação de opiniões, como consequência, também a subjetividade presente nessas mídias. Eventos globais ou de grandes impactos regionais são mais perceptíveis dentro desses ecossistemas digitais, ao impactarem maior número de pessoas e usuários (PETRESCU; TRUICĂ; APOSTOL, 2019). Como domínio para ser explorado, optou-se pelo Twitter, ao ser a rede social que melhor disponibiliza o ferramental necessário para a extração de informações, podendo selecionar o conteúdo de estudo pela data, assunto, idioma, nacionalidade, dentre outros.

Para a tarefa de extração dos dados a partir da API disponibilizada pelo Twitter, fez-se necessário a obtenção de uma conta de desenvolvedor. Para esse processo foi necessário responder diversas perguntas e aceitar termos, como, por exemplo, da finalidade, interesses econômicos e sociais, da anonimização de usuários não verificados, dentre outros. Dessa maneira, foi desenvolvida a aplicação ***Alexandria***.

Alexandria é o primeiro produto de dados dentro desse projeto e tem a finalidade de fazer o serviço de carga da informação do Twitter para o banco de dados. Os dados são obtidos por meio de requisições a API disponibilizada pela RS, em que são especificados os parâmetros de busca, como palavras-chave, idioma, localização geográfica, filtro por usuário, dentre outros. Dessa maneira, pode-se refinar os resultados da busca conforme as necessidades específicas. Para as construções dos conjuntos de dados as palavras-chave utilizadas foram: covid, covid-19, corona, corona vírus, coronga e vírus. Após a coleta, os dados são divididos entre mensagens, perfil dos usuários e localidade.

Dessa maneira, para o período em estudo, optou-se por trabalhar com as mensagens compartilhadas pelos usuários na rede social com relação à pandemia da COVID-19 no período de 1 de outubro de 2019 até 31 de março de 2022, criando dois domínios: *i*) o primeiro contém as mensagens compartilhadas em território brasileiro (por volta de 800.000 mensagens), das quais se destacam a cidade ou unidade federativa; *ii*) o segundo contém as mensagens compartilhadas em língua portuguesa (cerca de 9.000.000 de mensagens).

A importância da distinção para o primeiro domínio se dá pelo combate à pandemia ter sido de maneira heterogênea no Brasil. Dessa maneira, faz-se possível uma visibilidade de eventos locais, mas de alto impacto regional, como visto no colapso sanitário em Manaus. Já para o segundo domínio, é importante entender o sentimento presente de maneira global em língua portuguesa. Nota-se que existe uma intersecção entre os domínios, da maneira que ambos podem ser utilizados em união.

Os dados do perfil dos usuários podem conter informações como a descrição do usuário, datas de nascimento e entrada na plataforma, verificação, dentre outros. Dados das mensagens contém a quantidade de respostas e compartilhamentos, o conteúdo, que pode ser textual, vídeo ou foto; data de publicação, dentre outros. Já os dados de localização contém a cidade, estado ou região ao qual o usuário pertence.

3.3 Pré-processamento dos *tweets*

É comum que dados disponibilizados para análise não estejam em um formato adequado para a extração de conhecimento, com isso, fazem-se necessários os usos de métodos de tratamento, limpeza e redução do volume de dados antes de iniciar a etapa do enriquecimento de informação (REZENDE, 2003). A partir da mesma fonte de informação são providos três conjuntos de dados: *i*) o primeiro contém as mensagens compartilhadas; *ii*) o segundo inclui as informações dos perfis dos usuários; *iii*) o terceiro possui os dados geográficos dos usuários. Dessa maneira, é importante uma etapa de fusão de informação, ou integração, em que unificamos todos os conjuntos em uma única fonte de dados para ser utilizada na Extração de Padrões (REZENDE, 2003).

Com a junção das informações são necessárias transformações para adequar os conjuntos, tais como as padronizações dos formatos de datas, identificadores das mensagens e dos perfis, também, limpezas de dados para quando foram retornadas informações vazias durante as requisições ou de atributos com valores inválidos. Para isso, foi criado um segundo produto de dados nomeado ***Flamel***, responsável pela limpeza, enriquecimento, agregação e análise dos resultados.

Por se tratar de informação textual proveniente de uma RS, pode haver a necessidade da normalização dos dados para uma representação textual adequada, tais como a remoção do excesso de espaçamento entre palavras, palavras em caixa baixa e adequações de pontuação e símbolos, entretanto, como serão aplicadas diferentes técnicas de AM, esse processo varia conforme a abordagem aplicada, tendo como base o trabalho desenvolvido por Jianqiang e Xiaolin (2017). Dessa maneira, também podem ser aplicados algoritmos de tokenização, lematização, *stemming* e remoção de *stop words*. Com isso, se fazem possíveis representações textuais, em que serão aplicados algoritmos de AM que buscam a extração de padrões.

A partir da elaboração da representação textual pode-se aplicar algoritmos léxicos para a AS, que está relacionada à extração da subjetividade e da polarização em um texto (LIU, 2022) e pode ser realizada a partir de algoritmos de AM supervisionados e não supervisionados. Inicialmente, a abordagem não supervisionada será utilizada, devido ao grande número de documentos para serem rotulados, portanto, abordagens léxicas com uso de dicionário, citadas na seção 2.2.3; serão utilizadas para a rotulação do conjunto de dados.

Com isso, a partir das palavras rotuladas com sua polarização positiva, negativa ou neutra,

pode-se compor modelos para a tarefa de classificação. Serão utilizados o OpLexicon, que é um léxico de sentimento para português que utiliza múltiplas fontes totalizando mais de 15.000 palavras classificadas (MACHADO; PARDO; RUIZ, 2018); o SentiLex, que também é um léxico sentimental para português que possui mais de 82.000 palavras classificadas (FILHO; PARDO; ALUÍSIO, 2013); e o LeIA, que é uma adaptação do VADER para português (ALMEIDA, 2018).

Devido ao grande volume de dados e à variedade de técnicas de AM a serem avaliadas, optou-se por utilizar o método de validação cruzada *holdout*, que consiste em dividir o conjunto total de dados em dois subconjuntos mutualmente exclusivos, um destinado para treino e outro para teste. Esses subconjuntos possuem as frações de 80% dos dados para treino e 20% para teste, porém, quando necessário são divididos em 65% dos dados para treino, 15% para validação e 20% para teste.

3.4 Extração de padrões

Corresponde a etapa destinada a direcionar ao cumprimento dos objetivos definidos na Identificação do Problema, em que serão realizadas escolhas relacionadas às configurações de execuções ou aplicações de algoritmos com a finalidade da extração de conhecimento (REZENDE, 2003). Os algoritmos de AM selecionados para a tarefa de classificação multivariada que terão seu desempenho comparado ao final desse trabalho são árvores de decisão, árvores aleatórias, LSTM, BERT, Naive Bayes, regressão logística e SVM. Esses algoritmos juntamente com os pré-processamentos requiridos integram a aplicação Flamel, já citado anteriormente. O Flamel tem por objetivo fazer as etapas finais dos pré-processamentos requeridos pelas técnicas de AM, a extração de padrões com a aplicação dos algoritmos e a avaliação das abordagens.

3.5 Pós-processamento

A partir da Extração de Padrões é possível avaliar os resultados obtidos, assim então, se são satisfatórios ou não. Essa avaliação é feita por meio de uma matriz de confusão com documentos classificados, viabilizando os cálculos de métricas como acurácia, precisão, sensibilidade, especificidade e *F1-Score*. Também, a rotulação por meio de algoritmos léxicos permite estudar o agrupamento da classificação do sentimento entre positivo, negativo e neutro, durante a pandemia da COVID-19 no Brasil, assim como o seu comportamento nesse período.

3.6 Utilização do conhecimento

Com base nas etapas anteriores e na validação das técnicas aplicadas, o conhecimento enriquecido com os conjuntos de dados criados permitirá diversas análises. Destacam-se abordagens relacionadas às variações de sentimentos no Twitter durante a pandemia da COVID-19,

bem como a análise comparativa entre os classificadores selecionados com base nas métricas estabelecidas para essa análise.

Se os classificadores associados às abordagens léxicas adotadas apresentarem resultados satisfatórios, será viável aplicar essa composição no mundo real para monitorar o sentimento da população sobre diferentes tópicos, como no estudo realizado por Garcia e Berton (2021). Nesse trabalho, os autores avaliaram diferentes técnicas de AM para classificar o sentimento dos textos no Twitter durante a pandemia da COVID-19, tanto em língua inglesa quanto em língua portuguesa. Enquanto para a língua inglesa os autores combinaram algoritmos léxicos para a rotulação do conjunto de dados, para a língua portuguesa eles utilizaram uma modelagem a partir dos emojis presentes nos textos em português, devido à falta das ferramentas necessárias para aquela ocasião. A utilização do conhecimento proposto neste trabalho é similar a realizada pelos autores no que diz respeito à língua inglesa, porém, produzida em língua portuguesa.

CAPÍTULO

4

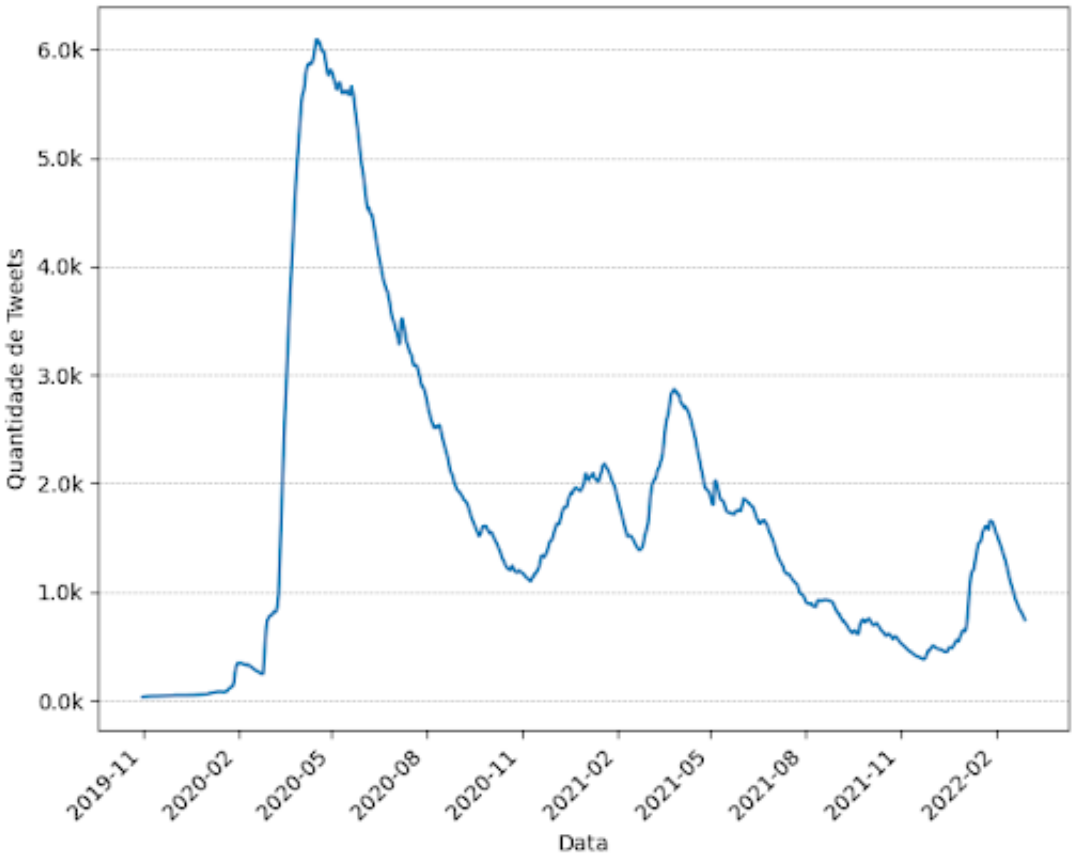
AVALIAÇÃO EXPERIMENTAL

Neste capítulo, é realizada uma análise do processo descrito no Capítulo 3, que segue o fluxo apresentado na Figura 8, no qual a classificação dos sentimentos de tweets é abordada a partir de uma análise léxica. Além disso, é conduzida uma avaliação experimental da comparação entre diferentes classificadores de AM, por meio da aplicação Flamel. Inicialmente, houve a fusão dos dois conjuntos de dados gerados com as aquisições dos dados do twitter com informações entre setembro de 2019 e março de 2022.

O primeiro conjunto de dados é formado por mais de 1.4 milhões de mensagens publicadas em português no Brasil, sua distinção se dá pela aceitação do usuário em conceder acesso à geolocalização dos dispositivos utilizados para publicação das mensagens; a distribuição dos tweets pode ser visualizada na Figura 9. O segundo conjunto de dados formados por mais de 4.7 milhões de mensagens em língua portuguesa pode ser visto na Figura 10.

Na Figura 9, nota-se que em maio de 2020, janeiro de 2021, abril de 2021 e janeiro de 2022 houveram picos de mensagens atreladas às palavras-chave. Esses picos podem estar relacionados respectivamente à primeira onda de COVID-19 no Brasil, o colapso de Manaus, as segunda e terceira ondas do vírus (MOURA *et al.*, 2022; SABINO *et al.*, 2021).

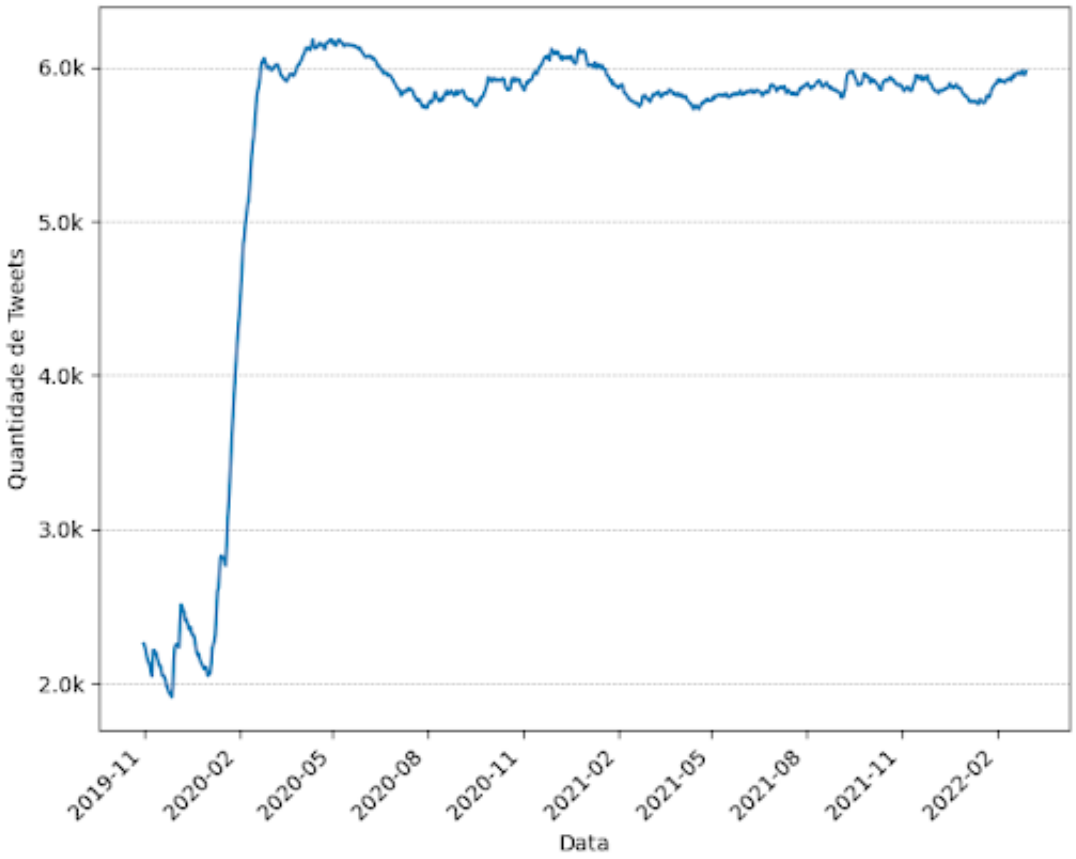
Figura 9 – Distribuição dos tweets publicados no Brasil relacionados com à palavra-chave COVID-19



Fonte: Elaborada pelo autor.

Na Figura 10, nota-se que esse conjunto de dados possui um comportamento de patamar a partir de fevereiro de 2020. Isso é causado por restrições da plataforma, em que limitava a busca diária em até 10 mil tweets. Com o processo de filtragem dos textos indesejados, esse patamar ficou estabelecido por volta de 6 mil mensagens.

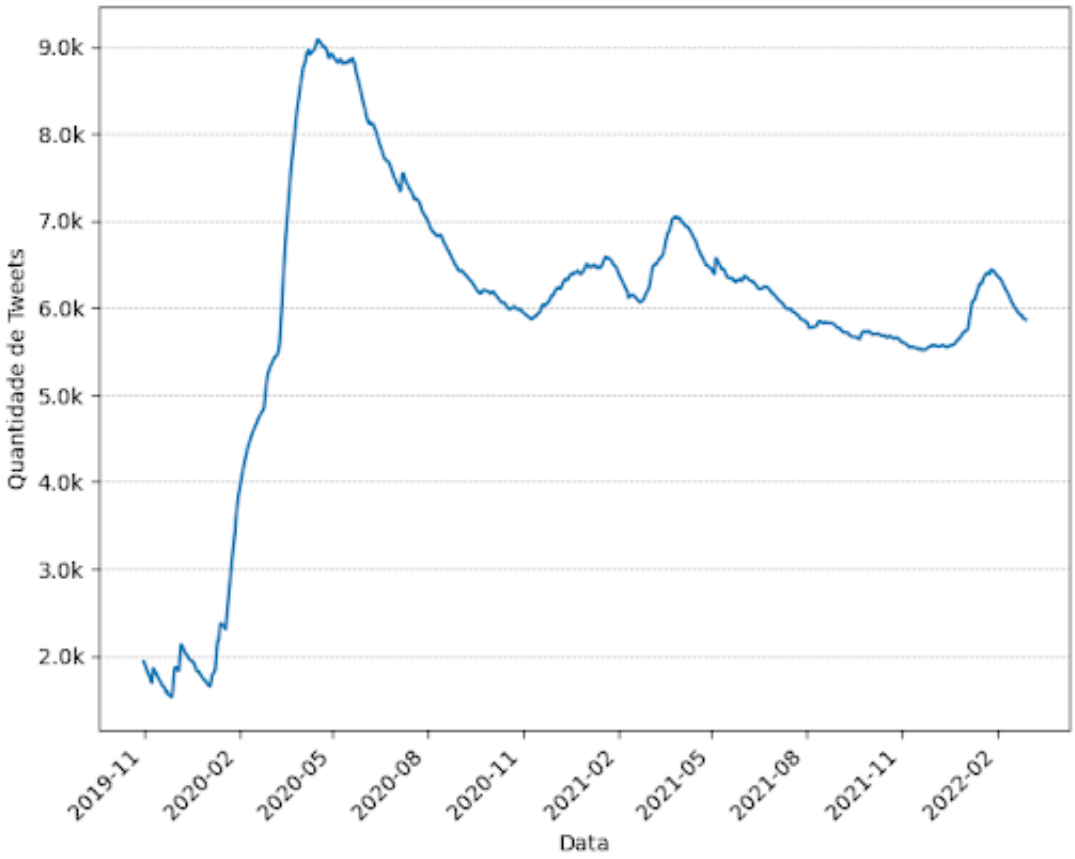
Figura 10 – Distribuição dos *tweets* publicados em língua portuguesa relacionados com à palavra-chave COVID-19



Fonte: Elaborada pelo autor.

A partir da fusão dos dois conjuntos, há um novo conjunto formado por mais de 5.2 milhões de tweets, a distribuição diária dessas mensagens pode ser vista na Figura 11. Nota-se que existem mensagens da interseção dos conjuntos, dessa maneira, sendo incorporadas somente uma vez ao grupo de dados.

Figura 11 – Distribuição dos conjuntos sumarizados de *tweets* relacionados com à palavra-chave COVID-19



Fonte: Elaborada pelo autor.

4.1 Classificação dos sentimentos por meio da abordagem léxica

A partir da formação do conjunto de dados composto por mais de 5.2 milhões de tweets, é possível a aplicação da abordagem léxica para a classificação do sentimento negativo, neutro ou positivo. O algoritmo escolhido para essa tarefa foi a variante do VADER para português intitulada LeIA (ALMEIDA, 2018). Na Tabela 2, tem-se um exemplo da classificação realizada pelo algoritmo.

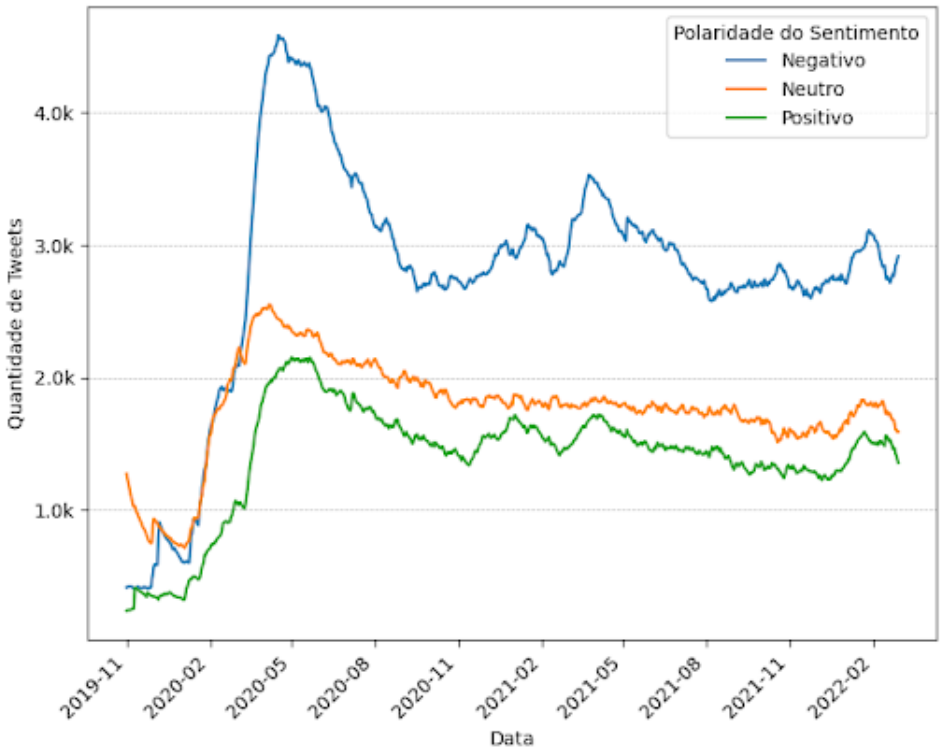
Tabela 2 – Texto e saída do classificador léxico usando o algoritmo LeIA

Texto	Classificação	Polaridade
Medo de estar com Covid	-0,4215	Negativa
Diretora da OMS diz que mundo está entrando em 4ª onda de Covid-19	0,0000	Neutra
Cidades do interior de SP cancelam Carnaval por causa da Covid! Ainda bem!	0,8658	Positiva

Fonte: Elaborada pelo autor.

A presença da palavra medo em “Medo de estar com Covid” enfatizou a polaridade negativa na sentença, entretanto a presença da expressão “Ainda bem!” na sentença “Cidades do interior de SP cancelam Carnaval por causa da Covid! Ainda bem!” acrescentou para a polaridade positiva, por sua vez, o texto “Diretora da OMS diz que mundo está entrando em 4ª onda de Covid-19” apresentou palavras neutras em sua totalidade. A distribuição da polaridade dos tweets no período em análise pode ser vista na Figura 12.

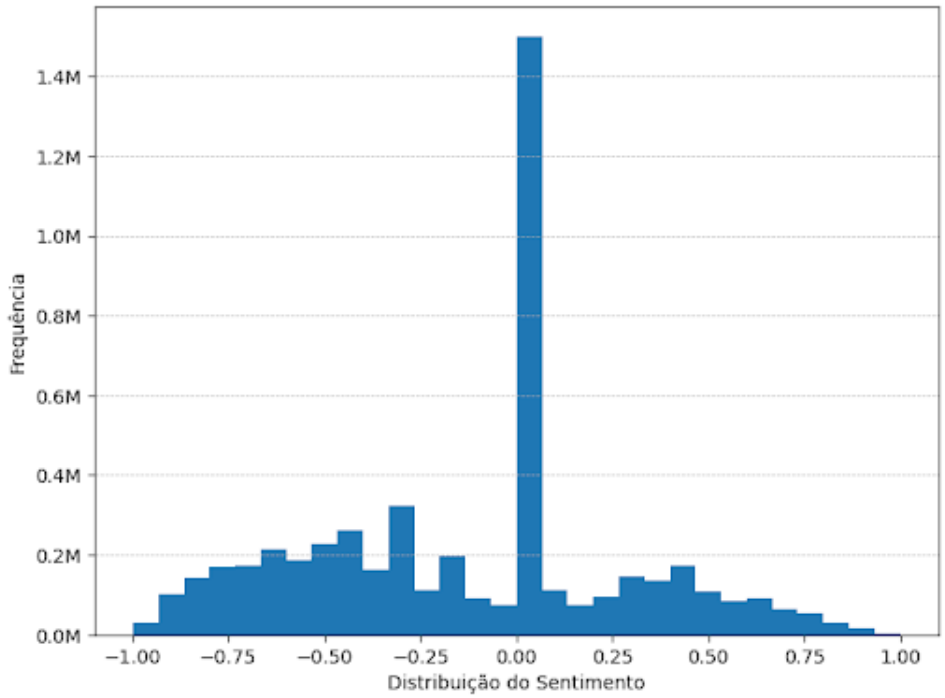
Figura 12 – Quantidade de *tweets* relacionados com à palavra-chave COVID-19 ao longo do tempo e suas polaridades



Fonte: Elaborada pelo autor.

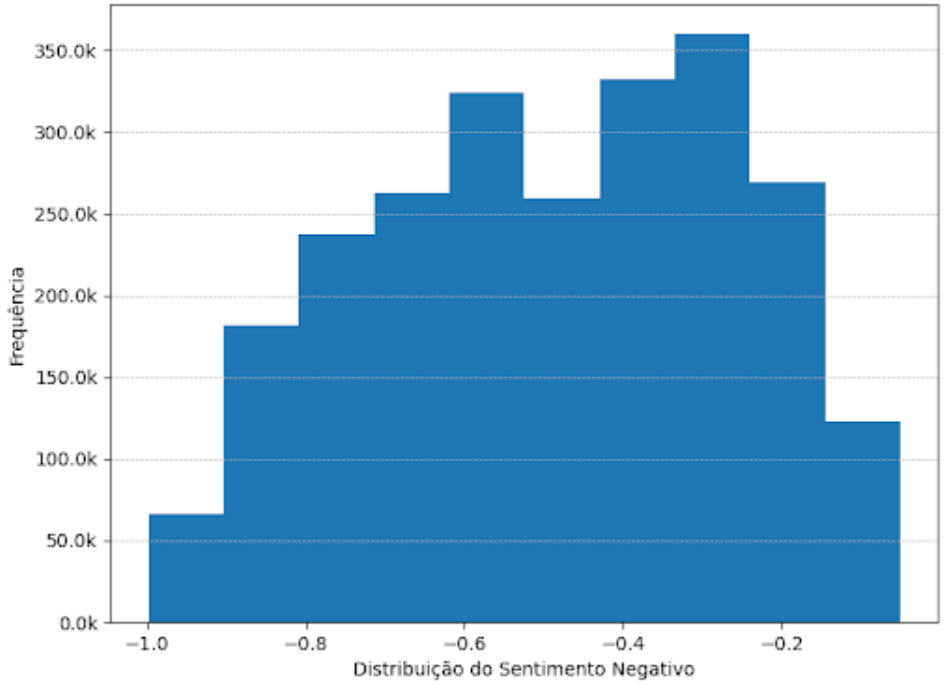
Os picos relacionados ao sentimento negativo da Figura 13 estão correlacionados com os presentes nas Figuras 9 e 11, consequentemente aos episódios já descritos anteriormente. A partir da pontuação da classificação podemos ter a distribuição geral desses valores, como pode ser observado na Figura 13; e das distribuições segregadas conforme as polaridades negativa, neutra e positiva, respectivamente exibidas nas Figuras 14, 15 e 16.

Figura 13 – Distribuição da pontuação da polaridade dos textos



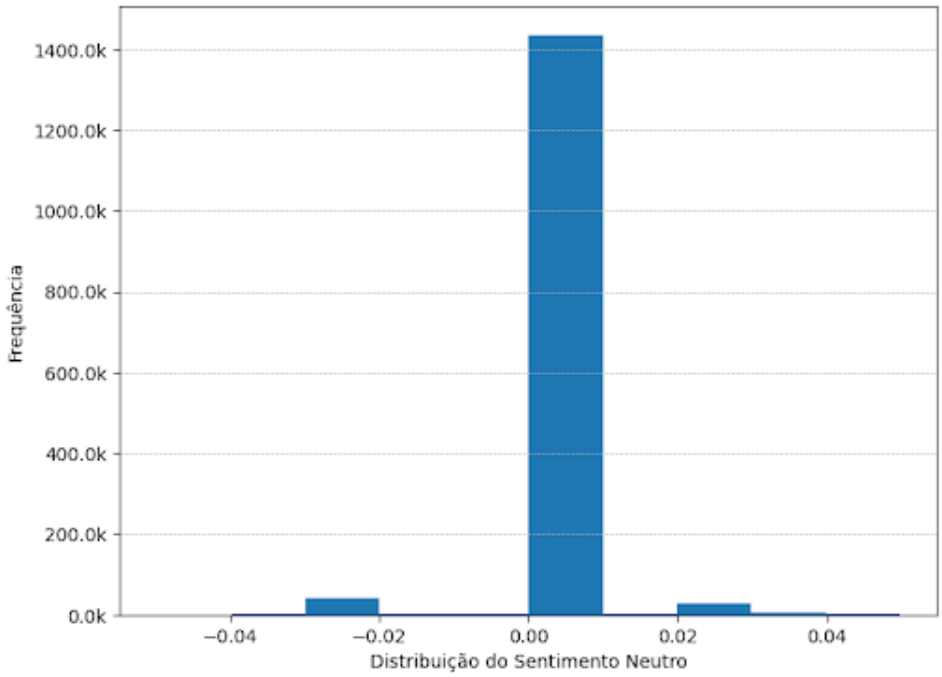
Fonte: Elaborada pelo autor.

Figura 14 – Distribuição do sentimento para a classe negativa



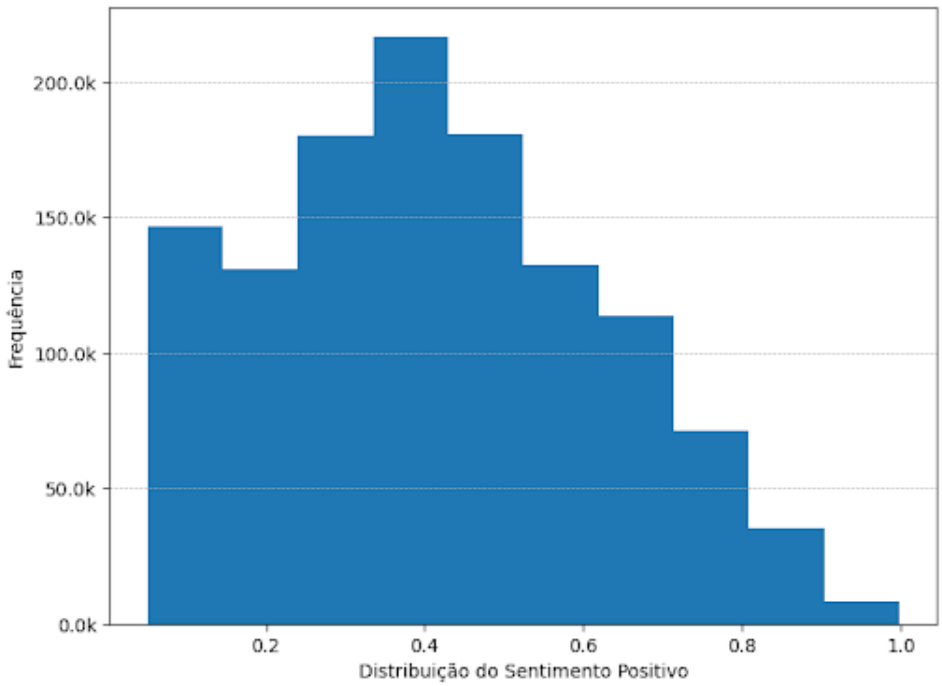
Fonte: Elaborada pelo autor.

Figura 15 – Distribuição do sentimento para a classe neutra



Fonte: Elaborada pelo autor.

Figura 16 – Distribuição do sentimento para a classe positiva



Fonte: Elaborada pelo autor.

As distribuições das classes positivas e negativas, respectivamente nos intervalos (0,05; 1,00] e [-1,00; -0,05), são parcialmente assimétricas, deslocadas respectivamente para direita e esquerda do eixo, porém para a classe neutra há uma assimetria e concentração dos valores

muito próximos de zero, visto que o classificador léxico assume como neutro valores no intervalo [-0,05; 0,05]. Na Tabela 3, podemos ver as quantidades e proporções de tweets por polaridade do sentimento, há de se notar que existe um desbalanceamento entre as classes.

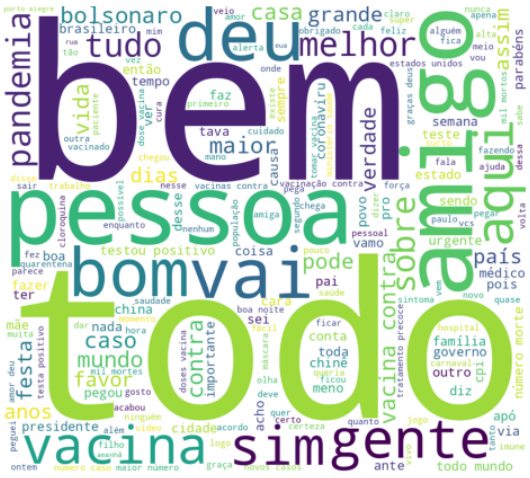
Tabela 3 – Quantidade e proporção de *tweets* classificados

Polaridade	Quantidade (em milhões)	Proporção (%)
Negativa	2,18	46,62
Neutra	1,46	30,72
Positiva	1,11	23,41

Fonte: Elaborada pelo autor.

A partir da frequência de palavras em cada uma das classes, foram geradas as nuvens de palavras. Nessa ferramenta de visualização é possível constatar a diferença de concentração das palavras conforme o seu tamanho para o conjunto. Essas nuvens podem ser vistas nas Figuras 17a e 17b.

Figura 17 – Nuvens de palavras geradas a partir dos *tweets* classificados



(a) Nuvem de palavras da classe positiva (b) Nuvem de palavras da classe negativa

Fonte: Elaborada pelo autor.

4.2 Construção de classificadores a partir dos dados rotulados

A partir da classificação com a abordagem léxica do conjunto de tweets rotulados com sentimentos negativo, neutro ou positivo, foram aplicadas distintas técnicas de AM para identificar a abordagem que melhor performa no domínio construído para uma tarefa de classificação multinomial. Para a geração dos classificadores foram utilizados 80% do conjunto de dados para treino (quando necessário esse dividido, 65% para treino e 15% para validação) e 20% para teste, com reamostragem para balanceamento das classes, sendo desbalanceadas conforme Tabela 3.

4.2.1 Classificadores baseados em árvores

O pré-processamento para as técnicas de AM por árvores aleatórias e de decisão é praticamente o mesmo. São realizados processos de redução das palavras para os seus radicais (stemming em inglês), remoção de *stopwords*, tags, URL e demais símbolos não alfanuméricos, também a filtragem e exclusão de palavras pouco presentes no conjunto de dados, dessa maneira, reduzindo a quantidade de atributos e tornando os modelos mais genéricos com relação ao domínio, evitando o *overfitting*. Para árvores de decisão foi utilizado o algoritmo CART, porque pode lidar com dados de diferentes tipos, como categóricos e numéricos, sem a necessidade de pré-processamento extensivo. Nas Tabelas 4 e 5, é apresentada a performance dos modelos baseados em árvores.

Tabela 4 – Performance da abordagem com árvores de decisão

Classe	Precisão	Revocação	F1-Score
Negativa	0.78	0.77	0.77
Neutra	0.71	0.74	0.73
Positiva	0.72	0.70	0.71

Fonte: Elaborada pelo autor.

Tabela 5 – Performance da abordagem com árvores aleatórias

Classe	Precisão	Revocação	F1-Score
Negativa	0.83	0.82	0.83
Neutra	0.76	0.80	0.78
Positiva	0.80	0.76	0.78

Fonte: Elaborada pelo autor.

4.2.2 Abordagem com Naive Bayes

A abordagem com Naive Bayes é importante, pois a técnica estabelece uma *baseline* para comparação da performance dos outros modelos. A etapa de pré-processamento é a mesma estabelecida para os modelos baseados em árvores. Inicialmente, são realizados processos de redução das palavras para os seus radicais, remoção de *stopwords*, tags, URLs e demais símbolos não alfanuméricos, também a filtragem e exclusão de palavras pouco frequentes no conjunto de dados. A Tabela 6 apresenta a performance da abordagem com Naive Bayes para classificação multinomial.

Tabela 6 – Performance da abordagem com Naive Bayes

Polaridade	Precisão	Revocação	F1-Score
Negativa	0,67	0,79	0,73
Neutra	0,61	0,53	0,57
Positiva	0,72	0,60	0,65

Fonte: Elaborada pelo autor.

4.2.3 Abordagem com regressão logística multinomial

Na aplicação do algoritmo para classificação multinomial com regressão logística, é realizado pré-processamento idêntico aos modelos baseados em árvore, dessa forma, são realizados *stemming*, remoção de *stopwords*, tags, URLs e símbolos não alfanuméricos, também, foram desconsideradas palavras com baixa ocorrência. Essa abordagem, assim como a Naive Bayes, é elencada com um bom *baseline* comparativo entre os algoritmos. A Tabela 7 apresenta a performance do classificador gerado.

Tabela 7 – Performance da abordagem com regressão logística multinomial

Polaridade	Precisão	Revocação	F1-Score
Negativa	0,67	0,79	0,73
Neutra	0,61	0,53	0,57
Positiva	0,72	0,60	0,65

Fonte: Elaborada pelo autor.

4.2.4 Abordagem com SVM

O algoritmo de SVM por algum tempo foi considerado o estado da arte para aplicações em processamento de linguagem natural quando se trata de AS. O seu pré-processamento foi o mesmo utilizado para as abordagens com regressão logística multinomial, árvores de decisão e naive Bayes. A performance pode ser visto na Tabela 8.

Tabela 8 – Performance da abordagem com SVM

Polaridade	Precisão	Revocação	F1-Score
Negativa	0,81	0,85	0,83
Neutra	0,78	0,78	0,78
Positiva	0,81	0,74	0,77

Fonte: Elaborada pelo autor.

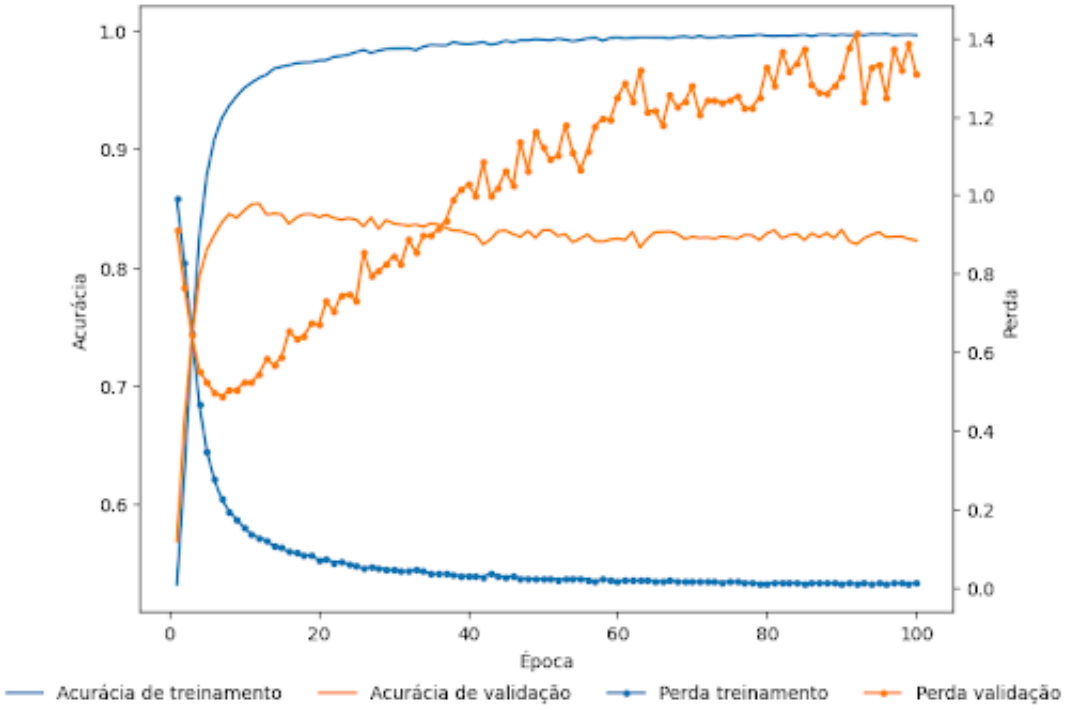
4.2.5 Classificadores baseados em redes neurais

Na abordagem da RNN com LSTM, que foi selecionada devido à sua capacidade de lidar com dados sequenciais, dependências de longo prazo e entradas de comprimento variável;

o pré-processamento envolve *stemming*, reduções de *stopwords*, símbolos não alfanuméricos, menções e URLs. Após a etapa de filtragem, a vetorização das palavras é realizada com seleção das 8000 mais comuns para a abordagem. Na camada de saída deve ter 3 valores, um para cada classe; com função de ativação *softmax*, por ser um problema de classificação multi classe; e função de perda definida pela entropia cruzada categórica, por ser um problema de classificação multi classe.

Na Figura 18, tem-se o gráfico que exibe a acurácia e perda para os dados de treinamento e validação em cada uma das épocas, em que se faz presente a estabilidade para a métrica de acurácia a partir da 30ª época. A Tabela 9 apresenta a performance do modelo no conjunto dos dados de teste.

Figura 18 – Curvas de perda e acurácia para os dados de treinamento e validação para LSTM



Fonte: Elaborada pelo autor.

Tabela 9 – Performance da RNN com LSTM

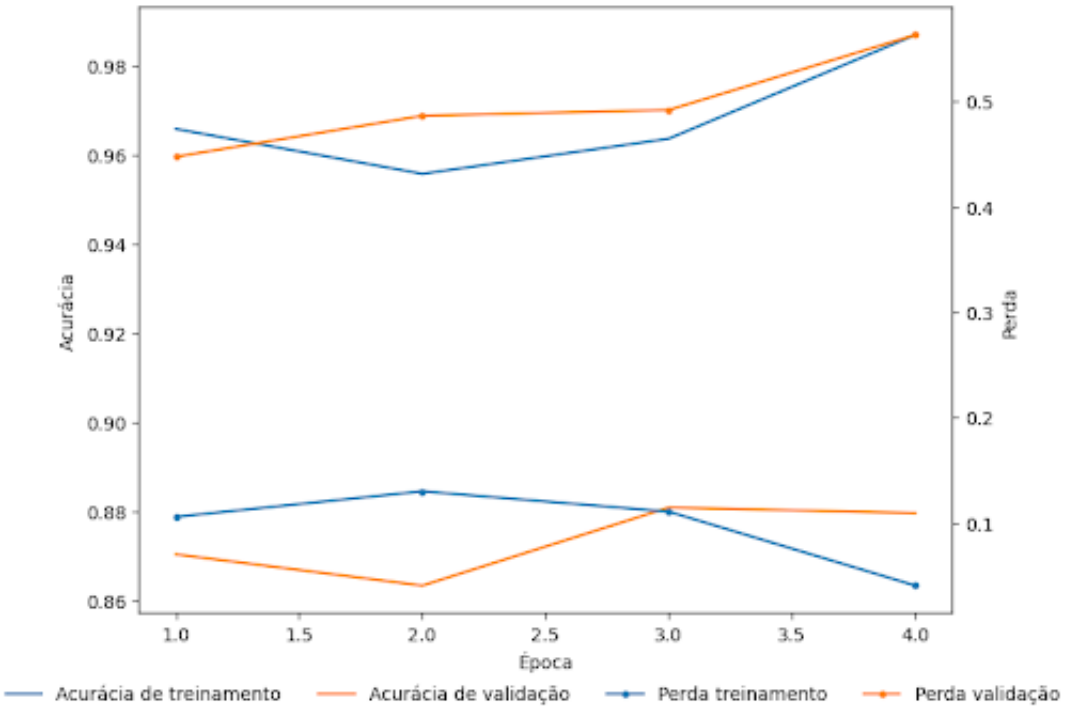
Polaridade	Precisão	Revocação	F1-Score
Negativa	0,85	0,89	0,87
Neutra	0,85	0,74	0,79
Positiva	0,71	0,76	0,73

Fonte: Elaborada pelo autor.

Para a abordagem com o BERT foi realizado o pré-processamento com a remoção de *retweets*, URLs, símbolos alfanuméricos e menções. O modelo pré-treinado foi BERT-uncased, que assume a utilização de palavras em minúsculo antes do processo de tokenização. É composto

por aproximadamente 110 milhões de parâmetros, onde no processo de ajuste fino ocorre adaptação com uma camada *softmax* na saída com um número de neurônios iguais a 3, devido às nossas classes negativa, neutra e positiva. Também, para o ajuste fino foram utilizadas 4 épocas, as métricas de cada época do ajuste são exibidas na Figura 19 com a performance no conjunto de dados de teste na Tabela 10.

Figura 19 – Curvas de perda e acurácia para os dados de treinamento e validação para o BERT



Fonte: Elaborada pelo autor.

Tabela 10 – Performance do BERT

Polaridade	Precisão	Revocação	F1-Score
Negativa	0,91	0,92	0,91
Neutra	0,89	0,86	0,87
Positiva	0,82	0,83	0,83

Fonte: Elaborada pelo autor.

4.3 Considerações finais

O resultado consolidado dos experimentos apresentados neste capítulo são obtidos a partir do *framework* adaptado proposto na Figura 8, que podem ser vistos na Tabela 11, com destaque para os modelos baseados em redes neurais. A abordagem a partir do BERT, demonstrou-se superior com relação a todas as métricas de avaliação, da Figura 10, pode-se verificar que para uma única época de treinamento, a abordagem se demonstrou promissora. Também, para a

abordagem com LSTM um desempenho similar foi obtido, na Figura 9 é possível analisar a estabilização da acurácia de treinamento a partir da época de número vinte.

Nas técnicas estabelecidas tradicionalmente como bons *baselines* de avaliação, os resultados foram satisfatórios com a melhor performance para SVM, seguida pela regressão logística multinomial e naive Bayes. A comparação com os modelos baseados em redes neurais se faz necessária, pois embora a performance das técnicas de base comparativa foram inferiores, requerem uma capacidade computacional de treinamento menor do que as consideradas o estado da arte para AS.

As abordagens baseadas em árvores de decisão performaram de maneira intermediária entre as técnicas baseadas em redes neurais e as *baselines*, entretanto, com superioridade para a técnica de árvores aleatórias, que apresentou melhores resultados quando comparada com árvores de decisão. Embora não tenha sido realizado, a importância dessa abordagem se dá pela interpretabilidade do modelo obtido após a etapa de treinamento.

Em todas as abordagens a classe neutra obteve os piores valores na avaliação, o que já era esperado, ao estar contida em um intervalo mais restrito pela abordagem léxica e com sentenças que podem apresentar palavras de polaridades opostas.

Como foi apresentado na Figura 12, é possível observar a correlação existente entre a quantidade de tweets com alguns eventos marcantes já descritos. Destacam-se pelos picos de sentimentos majoritariamente negativos, por estarem associados a palavras e fatos característicos dessa polaridade, tais como morte e contaminação de pessoas pela COVID-19. Ainda nas Figuras 17a e 17b, pode-se ver nas nuvens de palavras, as sentenças mais comuns nas polaridades positiva, bem, melhor, amigo, dentre outras; e negativa, morte, pandemia, doença, dentre outras.

Tabela 11 – Consolidação dos resultados

Modelo	Classe	Precisão	Revocação	F1-Score
Árvores de Decisão	Negativa	0.78	0.77	0.77
	Neutra	0.71	0.74	0.73
	Positiva	0.72	0.70	0.71
Árvores Aleatórias	Negativa	0.83	0.82	0.83
	Neutra	0.76	0.80	0.78
	Positiva	0.80	0.76	0.78
Naive Bayes	Negativa	0.67	0.79	0.73
	Neutra	0.61	0.53	0.57
	Positiva	0.72	0.60	0.65
Regressão Logística	Negativa	0.71	0.78	0.79
	Neutra	0.70	0.74	0.72
	Positiva	0.75	0.72	0.73
SVM	Negativa	0.81	0.85	0.83
	Neutra	0.78	0.78	0.78
	Positiva	0.81	0.74	0.77
LSTM	Negativa	0.85	0.89	0.87
	Neutra	0.85	0.74	0.79
	Positiva	0.71	0.76	0.73
BERT	Negativa	**0.91**	**0.92**	**0.91**
	Neutra	**0.89**	**0.86**	**0.87**
	Positiva	**0.82**	**0.83**	**0.83**

Fonte: Elaborada pelo autor.

CAPÍTULO

5

CONCLUSÕES

Neste capítulo são apresentadas as conclusões obtidas acerca do problema de aplicação de análise de sentimento e comparação de classificadores em dados textuais provenientes do Twitter relacionados a COVID-19 no Brasil e em língua portuguesa entre novembro de 2019 e março de 2022. A partir do *framework* adaptado proposto na Figura 8, foram coletadas informações da RS, totalizando mais de nove milhões de mensagens; sendo posteriormente pré-processadas, classificadas com relação à polarização da sentença por um algoritmo léxico e utilizadas como base comparativa entre diferentes abordagens com AM supervisionado. Por fim, são apresentados e discutidos os próximos trabalhos pretendidos.

5.1 Principais resultados

Este trabalho tinha como objetivo a investigação analítica da composição entre algoritmos léxicos de PLN e classificadores de AM, que se demonstrou eficaz, como pode ser visto na Figura 12, que apresenta a variação do sentimento presente nas sentenças no decorrer do tempo, podendo ser correlacionada com eventos ocorridos durante a pandemia da COVID-19; e nas Figuras 17a e 17b, que respectivamente apresentam as nuvens de palavras para as classes positiva e negativa. Também, tinha por objetivo a comparação entre diferentes classificadores de AM, sendo atingido através dos resultados consolidados pela Tabela 11, a qual apresenta o desempenho para os classificadores de AM. Dessa maneira, todos os propósitos iniciais deste trabalho foram realizados com sucesso.

Com o aumento do uso das RS, transformando a maneira pela qual os indivíduos e empresas se relacionam e se comunicam, houve também um acréscimo do volume de informações geradas pelas pessoas. Com isso, técnicas para AS e MO desempenham um papel preponderante para estudar o comportamento e impacto dos indivíduos com relação a fenômenos locais e globais, visto que, possuem a capacidade da escalabilidade de soluções.

Inicialmente, com o desenvolvimento da aplicação Alexandria, responsável pela coleta dos dados brutos oriundos do Twitter, que resultaram em mais de nove milhões de mensagens, sendo depois pré-processadas para um conjunto formado por pouco menos de cinco milhões de Tweets. Sem a associação das abordagens léxicas com AM, desenvolvida na aplicação Flamel; seria humanamente inviável fazer a rotulação desse grande conjunto de dados. Dessa forma, a associação entre as técnicas léxicas e AM demonstrou ser eficiente conforme o *framework* proposto por Melville, Gryc e Lawrence (2009).

Conforme descrito na seção 4.1, em que temos as curvas dos sentimentos presentes nas sentenças com picos para a classe negativa em momentos onde houveram catarses relacionadas à saúde pública. Na Figura 12 há sinais de crescimento atrelado a sentença negativa durante os períodos relacionados a maio de 2020, janeiro de 2021, abril de 2021 e janeiro de 2022, onde houveram respectivamente à primeira onda de COVID-19 no Brasil, o colapso de Manaus, as segunda e terceira ondas do vírus.

Com os textos já rotulados em sentimentos positivos e negativos, deu-se início a um trabalho relacionado a aplicação de algoritmos de AM, buscando a otimização e o melhoramento das abordagens segundo as técnicas selecionadas. A técnica considerada o estado da arte, BERT; se sobressai com relação às demais, como pode ser visto na Tabela 11; entretanto, demandando uma alta capacidade de processamento, pois se trata de um modelo que utiliza elevado número de parâmetros para treinamento. Também com destaque para as abordagens SVM e LSTM que performaram relativamente próxima ao BERT e demandam de uma capacidade computacional inferior à utilizada pela abordagem campeã.

5.2 Trabalhos futuros

Como trabalhos futuros pretende-se realizar o treinamento dos modelos utilizando *k-fold* para validação cruzada e considerar outras abordagens com AM, sobretudo as relacionadas a aprendizado não supervisionado para a classificação de sentimentos, acredita-se no potencial dessa abordagem devido ao tamanho do conjunto de dados formulado para os experimentos deste trabalho.

REFERÊNCIAS

ACHEAMPONG, F. A.; NUNOO-MENSAH, H.; CHEN, W. Transformer models for text-based emotion detection: A review of bert-based approaches. **Artificial Intelligence Review**, Springer, p. 1–41, 2021. Citado na página 44.

AGGARWAL, C. C. **An Introduction to Social Network Data Analytics**. [S.l.]: Springer, 2011. Citado na página 23.

AGGARWAL, C. C.; AGGARWAL, C. C. **Mining Text Data**. [S.l.]: Springer, 2015. Citado na página 27.

____. **Machine Learning for Text: An Introduction**. [S.l.]: Springer, 2018. Citado nas páginas 40, 41, 42, 43 e 49.

AGRAWAL, R.; IMIELIŃSKI, T.; SWAMI, A. Mining association rules between sets of items in large databases. In: **Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data**. [S.l.: s.n.], 1993. p. 207–216. Citado na página 32.

ALMEIDA, R. J. A. **LeIA - Léxico para Inferência Adaptada**. [S.l.]: GitHub, 2018. <https://github.com/rafjaa/LeIA>. Acesso em: 10/07/2023. Citado nas páginas 53 e 58.

ASLAM, S. **Twitter Statistics**. 2023. Disponível em: <https://www.omnicoreagency.com/twitter-statistics>. Acesso em: 10/07/2023. Citado na página 38.

BAI, Y.; YAO, L.; WEI, T.; TIAN, F.; JIN, D.-Y.; CHEN, L.; WANG, M. Presumed asymptomatic carrier transmission of covid-19. **Jama**, American Medical Association, v. 323, n. 14, p. 1406–1407, 2020. Citado na página 24.

BOJANOWSKI, P.; GRAVE, E.; JOULIN, A.; MIKOLOV, T. Enriching word vectors with subword information. **Transactions of the Association for Computational Linguistics**, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA Journals-info . . . , v. 5, p. 135–146, 2017. Citado na página 43.

CAMBRIA, E.; SCHULLER, B.; XIA, Y.; HAVASI, C. New avenues in opinion mining and sentiment analysis. **IEEE Intelligent Systems**, IEEE, v. 28, n. 2, p. 15–21, 2013. Citado na página 34.

CARVALHO, J.; PLASTINO, A. On the evaluation and combination of state-of-the-art features in twitter sentiment analysis. **Artificial Intelligence Review**, Springer, v. 54, p. 1887–1936, 2021. Citado na página 47.

CARVALHO, V. O. d. **Generalização de Regras de Associação Utilizando Conhecimento de Domínio e Avaliação do Conhecimento Generalizado**. Tese (Doutorado) — Universidade de São Paulo, 2007. Citado na página 32.

CHOWDHARY, K.; CHOWDHARY, K. Natural language processing. **Fundamentals of Artificial Intelligence**, Springer, p. 603–649, 2020. Citado nas páginas 24 e 45.

CUTLER, A.; CUTLER, D. R.; STEVENS, J. R. Random forests. **Ensemble Machine Learning: Methods and Applications**, Springer, p. 157–175, 2012. Citado na página 29.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018. Citado nas páginas 44 e 45.

DHARMA, E. M.; GAOL, F. L.; WARNARS, H.; SOEWITO, B. The accuracy comparison among word2vec, glove, and fasttext towards convolution neural network (cnn) text classification. **Journal of Theoretical and Applied Information Technology**, v. 100, n. 2, p. 31, 2022. Citado na página 43.

DUBEY, A. D. Twitter sentiment analysis during covid-19 outbreak. **Available at SSRN 3572023**, 2020. Disponível em: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3572023>. Citado na página 24.

ENDSUY, R. D. Sentiment analysis between vader and eda for the us presidential election 2020 on twitter datasets. **Journal of Applied Data Sciences**, v. 2, n. 1, p. 08–18, 2021. Citado na página 47.

FILHO, P. B.; PARDO, T. A. S.; ALUÍSIO, S. An evaluation of the brazilian portuguese liwc dictionary for sentiment analysis. In: **Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology**. [S.l.: s.n.], 2013. Citado nas páginas 35 e 53.

FRIEDMAN, N.; GEIGER, D.; GOLDSZMIDT, M. Bayesian network classifiers. **Machine Learning**, Springer, v. 29, p. 131–163, 1997. Citado na página 33.

GALLAGHER, K. **The Social Media Demographics Report**. 2017. Disponível em: <https://www.businessinsider.com/the-social-media-demographics-report-2017-8>. Acesso em: 10/07/2023. Citado na página 37.

GARCIA, K.; BERTON, L. Topic detection and sentiment analysis in twitter content related to covid-19 from brazil and the usa. **Applied Soft Computing**, Elsevier, v. 101, p. 107057, 2021. Citado na página 54.

GIACHANOU, A.; CRESTANI, F. Like it or not: A survey of twitter sentiment analysis methods. **ACM Computing Surveys (CSUR)**, ACM New York, NY, USA, v. 49, n. 2, p. 1–41, 2016. Citado na página 46.

GOLDBERG, Y. **Neural Network Methods for Natural Language Processing**. [S.l.]: Springer Nature, 2022. Citado na página 30.

GUPTA, B.; RAWAT, A.; JAIN, A.; ARORA, A.; DHAMI, N. Analysis of various decision tree algorithms for classification in data mining. **International Journal of Computer Applications**, Foundation of Computer Science, v. 163, n. 8, p. 15–19, 2017. Citado na página 29.

HABIMANA, O.; LI, Y.; LI, R.; GU, X.; YU, G. Sentiment analysis using deep learning approaches: An overview. **Science China Information Sciences**, Springer, v. 63, p. 1–36, 2020. Citado na página 45.

HEARST, M. A.; DUMAIS, S. T.; OSUNA, E.; PLATT, J.; SCHOLKOPF, B. Support vector machines. **IEEE Intelligent Systems and Their Applications**, IEEE, v. 13, n. 4, p. 18–28, 1998. Citado na página 30.

HUTTO, C.; GILBERT, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: **Proceedings of the International AAAI Conference on Web and Social Media**. [S.l.: s.n.], 2014. v. 8, n. 1, p. 216–225. Citado na página 35.

JANG, B.; KIM, I.; KIM, J. W. Word2vec convolutional neural networks for classification of news articles and tweets. **Plos One**, Public Library of Science San Francisco, CA USA, v. 14, n. 8, p. e0220976, 2019. Citado na página 43.

JIANQIANG, Z.; XIAOLIN, G. Comparison research on text pre-processing methods on twitter sentiment analysis. **IEEE Access**, IEEE, v. 5, p. 2870–2879, 2017. Citado nas páginas 47 e 52.

JOACHIMS, T. *et al.* A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. In: CITESEER. **ICML**. [S.l.], 1997. v. 97, p. 143–151. Citado na página 30.

LEE, F. L. Social media, political information cycle, and the evolution of news: The 2017 chief executive election in hong kong. **Communication and the Public**, SAGE Publications Sage UK: London, England, v. 3, n. 1, p. 62–76, 2018. Citado nas páginas 23 e 37.

LIU, B. **Sentiment Analysis and Opinion Mining**. [S.l.]: Springer Nature, 2022. Citado nas páginas 25, 26, 27, 34, 36, 37 e 52.

MACHADO, M. T.; PARDO, T. A.; RUIZ, E. E. S. Creating a portuguese context sensitive lexicon for sentiment analysis. p. 335–344, 2018. Citado nas páginas 35 e 53.

MAIMON, O.; ROKACH, L. **Data Mining and Knowledge Discovery Handbook**. [S.l.]: Springer, 2005. v. 2. Citado nas páginas 28 e 29.

MAKS, I.; VOSSEN, P. A lexicon model for deep sentiment analysis and opinion mining applications. **Decision Support Systems**, Elsevier, v. 53, n. 4, p. 680–688, 2012. Citado na página 26.

MAMMONE, A.; TURCHI, M.; CRISTIANINI, N. Support vector machines. **Wiley Interdisciplinary Reviews: Computational Statistics**, Wiley Online Library, v. 1, n. 3, p. 283–289, 2009. Citado na página 30.

MARCACINI, R. M.; ROSSI, R. G.; MATSUNO, I. P.; REZENDE, S. O. Cross-domain aspect extraction for sentiment analysis: A transductive learning approach. **Decision Support Systems**, Elsevier, v. 114, p. 70–80, 2018. Citado na página 44.

MEDHAT, W.; HASSAN, A.; KORASHY, H. Sentiment analysis algorithms and applications: A survey. **Ain Shams Engineering Journal**, Elsevier, v. 5, n. 4, p. 1093–1113, 2014. Citado nas páginas 26, 27, 28, 32 e 33.

MELVILLE, P.; GRYC, W.; LAWRENCE, R. D. Sentiment analysis of blogs by combining lexical knowledge with text classification. In: **Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. [S.l.: s.n.], 2009. p. 1275–1284. Citado na página 70.

MICHALSKI, R. S.; CARBONELL, J. G.; MITCHELL, T. M. **Machine Learning: An Artificial Intelligence Approach**. [S.l.]: Springer Science & Business Media, 2013. Citado nas páginas 28, 30 e 34.

MILLER, G. A. Wordnet: A lexical database for english. **Communications of the ACM**, ACM New York, NY, USA, v. 38, n. 11, p. 39–41, 1995. Citado na página 35.

MISHRA, P.; RAJNISH, R.; KUMAR, P. Sentiment analysis of twitter data: Case study on digital india. In: IEEE. **2016 International Conference on Information Technology (InCITe) - The Next Generation IT Summit on the Theme-Internet of Things: Connect Your Worlds**. [S.l.], 2016. p. 148–153. Citado nas páginas 24 e 37.

MOURA, E. C.; CORTEZ-ESCALANTE, J.; CAVALCANTE, F. V.; BARRETO, I. C. d. H. C.; SANCHEZ, M. N.; SANTOS, L. M. P. Covid-19: Evolução temporal e imunização nas três ondas epidemiológicas, brasil, 2020–2022. **Revista de Saúde Pública**, SciELO Brasil, v. 56, 2022. Citado na página 55.

MURTHY, G.; ALLU, S. R.; ANDHAVARAPU, B.; BAGADI, M.; BELUSONTI, M. Text based sentiment analysis using lstm. **International Journal of Engineering Research Technology**, v. 9, n. 05, 2020. Citado na página 47.

NIELSEN, M. A. **Neural Networks and Deep Learning**. [S.l.]: Determination Press San Francisco, CA, USA, 2015. v. 25. Citado na página 30.

NIGAM, K.; LAFFERTY, J.; MCCALLUM, A. Using maximum entropy for text classification. In: STOCKHOLOM, SWEDEN. **IJCAI-99 Workshop on Machine Learning for Information Filtering**. [S.l.], 1999. v. 1, n. 1, p. 61–67. Citado na página 34.

ORTIZ-OSPINA, E.; ROSER, M. The rise of social media. **Our World in Data**, 2023. Citado nas páginas 23 e 26.

PEIXOTO, L. H. R. **Aprendizado de Máquina Aplicado no Atendimento de Reclamações de Clientes**. Dissertação (Mestrado) — Universidade de São Paulo, 2021. Citado nas páginas 49 e 50.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. [S.l.: s.n.], 2014. p. 1532–1543. Citado na página 43.

PETRESCU, A.; TRUICĂ, C.-O.; APOSTOL, E.-S. Sentiment analysis of events in social media. In: IEEE. **2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP)**. [S.l.], 2019. p. 143–149. Citado na página 51.

PHILANDER, K.; ZHONG, Y. Twitter sentiment analysis: Capturing sentiment from integrated resort tweets. **International Journal of Hospitality Management**, Elsevier, v. 55, p. 16–24, 2016. Citado na página 24.

PISNER, D. A.; SCHNYER, D. M. Support vector machine. In: **Machine Learning**. [S.l.]: Elsevier, 2020. p. 101–121. Citado na página 30.

QUINLAN, J. R. Induction of decision trees. **Machine Learning**, Springer, v. 1, p. 81–106, 1986. Citado nas páginas 28 e 29.

RAHMAN, S. A. E.; ALOTAIBI, F. A.; ALSHEHRI, W. A. Sentiment analysis of twitter data. In: IEEE. **2019 International Conference on Computer and Information Sciences (ICCIS)**. [S.l.], 2019. p. 1–4. Citado na página 47.

RAMOS, J. *et al.* Using tf-idf to determine word relevance in document queries. In: CITESEER. **Proceedings of the First Instructional Conference on Machine Learning**. [S.l.], 2003. v. 242, n. 1, p. 29–48. Citado na página 36.

REZENDE, S. O. **Sistemas Inteligentes: Fundamentos e Aplicações**. [S.l.]: Editora Manole Ltda, 2003. Citado nas páginas 40, 41, 49, 52 e 53.

RISH, I. *et al.* An empirical study of the naive bayes classifier. In: **IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence**. [S.l.: s.n.], 2001. v. 3, n. 22, p. 41–46. Citado na página 33.

ROSSI, R. G. **Classificaçao Automática de Textos por Meio de Aprendizado de Máquina Baseado em Redes**. Tese (Doutorado) — Universidade de São Paulo, 2016. Citado nas páginas 42 e 43.

RUEDEN, L. V.; MAYER, S.; BECKH, K.; GEORGIEV, B.; GIESSELBACH, S.; HEESE, R.; KIRSCH, B.; PFROMMER, J.; PICK, A.; RAMAMURTHY, R. *et al.* Informed machine learning - a taxonomy and survey of integrating prior knowledge into learning systems. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, v. 35, n. 1, p. 614–633, 2021. Citado na página 41.

SABINO, E. C.; BUSS, L. F.; CARVALHO, M. P.; PRETE, C. A.; CRISPIM, M. A.; FRAIJI, N. A.; PEREIRA, R. H.; PARAG, K. V.; PEIXOTO, P. da S.; KRAEMER, M. U. *et al.* Resurgence of covid-19 in manaus, brazil, despite high seroprevalence. **The Lancet**, Elsevier, v. 397, n. 10273, p. 452–455, 2021. Citado na página 55.

SANTOS, F. L. D.; LADEIRA, M. The role of text pre-processing in opinion mining on a social media language dataset. In: IEEE. **2014 Brazilian Conference on Intelligent Systems**. [S.l.], 2014. p. 50–54. Citado na página 34.

SEBASTIANI, F.; ESULI, A. Sentiwordnet: A publicly available lexical resource for opinion mining. In: EUROPEAN LANGUAGE RESOURCES ASSOCIATION (ELRA) GENOA, ITALY. **Proceedings of the 5th International Conference on Language Resources and Evaluation**. [S.l.], 2006. p. 417–422. Citado na página 34.

SHERSTINSKY, A. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. **Physica D: Nonlinear Phenomena**, Elsevier, v. 404, p. 132306, 2020. Citado na página 31.

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. Citado na página 26.

SINOARA, R. A.; CAMACHO-COLLADOS, J.; ROSSI, R. G.; NAVIGLI, R.; REZENDE, S. O. Knowledge-enhanced document embeddings for text classification. **Knowledge-Based Systems**, Elsevier, v. 163, p. 955–971, 2019. Citado nas páginas 42 e 43.

SOHRABI, C.; ALSAFI, Z.; O'NEILL, N.; KHAN, M.; KERWAN, A.; AL-JABIR, A.; IOSIFIDIS, C.; AGHA, R. World health organization declares global emergency: A review of the 2019 novel coronavirus (covid-19). **International Journal of Surgery**, Elsevier, v. 76, p. 71–76, 2020. Citado na página 24.

SOLOMON, M. R.; TUTEN, T. L. Social media marketing. **Social Media Marketing**, SAGE Publications Ltd, p. 1–448, 2017. Citado nas páginas 26 e 37.

SUN, S.; LUO, C.; CHEN, J. A review of natural language processing techniques for opinion mining systems. **Information Fusion**, Elsevier, v. 36, p. 10–25, 2017. Citado na página 37.

TWITTER. **Twitter Statistics**. 2020. Disponível em: <https://help.twitter.com/en/using-twitter/retweet-faqs>. Acesso em: 10/07/2023. Citado na página 38.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. Attention is all you need. **Advances in Neural Information Processing Systems**, v. 30, 2017. Citado na página 44.

VERMA, J. P.; AGRAWAL, S.; PATEL, B.; PATEL, A. Big data analytics: Challenges and applications for text, audio, video, and social media data. **International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI)**, v. 5, n. 1, p. 41–51, 2016. Citado na página 23.

VILLAVICENCIO, C.; MACROHON, J. J.; INBARAJ, X. A.; JENG, J.-H.; HSIEH, J.-G. Twitter sentiment analysis towards covid-19 vaccines in the philippines using naïve bayes. **Information**, MDPI, v. 12, n. 5, p. 204, 2021. Citado na página 47.

WAGH, R.; PUNDE, P. Survey on sentiment analysis using twitter dataset. In: IEEE. **2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)**. [S.l.], 2018. p. 208–211. Citado na página 47.

WIEBE, J.; BRUCE, R.; O'HARA, T. P. Development and use of a gold-standard data set for subjectivity classifications. In: **Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics**. [S.l.: s.n.], 1999. p. 246–253. Citado na página 27.

WIEBE, J. *et al.* Learning subjective adjectives from corpora. **Aaai/iaai**, Austin, TX, v. 20, n. 0, p. 0, 2000. Citado na página 27.

WILSON, T.; WIEBE, J.; HOFFMANN, P. Recognizing contextual polarity in phrase-level sentiment analysis. In: **Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing**. [S.l.: s.n.], 2005. p. 347–354. Citado na página 25.

XU, T.; PENG, Q.; CHENG, Y. Identifying the semantic orientation of terms using s-hal for sentiment analysis. **Knowledge-Based Systems**, Elsevier, v. 35, p. 279–289, 2012. Citado na página 26.

YU, Y.; SI, X.; HU, C.; ZHANG, J. A review of recurrent neural networks: Lstm cells and network architectures. **Neural Computation**, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA Journals-info . . . , v. 31, n. 7, p. 1235–1270, 2019. Citado na página 31.

YUAN, X.; LI, L.; WANG, Y. Nonlinear dynamic soft sensor modeling with supervised long short-term memory network. **IEEE Transactions on Industrial Informatics**, IEEE, v. 16, n. 5, p. 3168–3176, 2019. Citado na página 31.

ZHANG, C.; ZHANG, S. **Association Rule Mining: Models and Algorithms**. [S.l.]: Springer, 2002. Citado na página 32.

ZHANG, Y.; JIN, R.; ZHOU, Z.-H. Understanding bag-of-words model: A statistical framework. **International Journal of Machine Learning and Cybernetics**, Springer, v. 1, p. 43–52, 2010. Citado na página 36.

ZIMBRA, D.; ABBASI, A.; ZENG, D.; CHEN, H. The state-of-the-art in twitter sentiment analysis: A review and benchmark evaluation. **ACM Transactions on Management Information Systems (TMIS)**, ACM New York, NY, USA, v. 9, n. 2, p. 1–29, 2018. Citado na página 47.

ZURADA, J. **Introduction to Artificial Neural Systems**. [S.l.]: West Publishing Company, 1992. Citado na página 30.


ICMC
USP
SÃO CARLOS