UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

VICTOR AUGUSTUS LOPES COSTA

COMPARAÇÃO DO DESEMPENHO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA COM A INTEGRAÇÃO DE UM LÉXICO DE OPINIÃO PARA ANÁLISE DE SENTIMENTOS NA LÍNGUA PORTUGUESA

MEDIANEIRA

2024

VICTOR AUGUSTUS LOPES COSTA

COMPARAÇÃO DO DESEMPENHO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA COM A INTEGRAÇÃO DE UM LÉXICO DE OPINIÃO PARA ANÁLISE DE SENTIMENTOS NA LÍNGUA PORTUGUESA

Comparison of the performance of machine learning algorithms integrated with an opinion lexicon for sentiment analysis in the Portuguese language

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Ciência da Computação do Curso de Bacharelado em Ciência da Computação da Universidade Tecnológica Federal do Paraná.

Orientador: Prof. Dr. Alan Gavioli

Coorientador: Prof. Dr. Davi Pereira dos Santos

MEDIANEIRA

2024

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

Siglas

ABSA	Análises de Sentimentos Baseadas em Aspectos
IA	Inteligência Artificial
MLP	*Multilayer Perceptron*
PLN	Processamento de Linguagem Natural
PMI	*Pointwise Mutual Information*
SVM	*Support Vector Machine*

SUMÁRIO

1 INTRODUÇÃO

A análise de sentimentos é uma área de estudo que busca extrair e analisar as opiniões e emoções de textos, utilizando técnicas de aprendizado de máquina, mineração de dados e Processamento de Linguagem Natural (PLN) (PEREIRA, 2021). Ela classifica os dados em classes positivas, negativas e neutras, de acordo com a opinião expressa no texto, que pode ser uma crítica de filme, uma avaliação de uma pessoa ou de um produto, por exemplo. A análise de sentimentos utiliza o PLN para mapear os sentimentos e emoções referentes a determinado tópico. Essa abordagem vem se tornando atrativa em diversas áreas por proporcionar às empresas, por exemplo, uma visão de como um produto está sendo percebido pelo público (PANDYA; MEHTA, 2020).

O uso tradicional de algoritmos de aprendizado de máquina isolados é uma técnica comumente utilizada para a análise de sentimentos, apresentando resultados satisfatórios em diversas aplicações (WU *et al.*, 2021). O SVM e *Naive Bayes* são dois algoritmos comumente usados em análises de sentimentos, juntamente de outros, e em grande maioria, acabam se destacando por apresentarem resultados mais satisfatórios, como em estudos de Januário *et al.* (2022), Souza *et al.* (2023), Silva *et al.* (2023), Yogi *et al.* (2024).

No entanto, a combinação de técnicas de aprendizado de máquina com léxicos de sentimentos podem apresentar uma melhora significativa na precisão do resultado da análise. Essa integração possibilita a detecção do conhecimento emocional associado às palavras, além da semântica delas, o que resulta em um desempenho superior na detecção de sentimentos em textos que seguem regras gramaticais fracas, como os encontrados em redes sociais (WU *et al.*, 2021).

Em paralelo a isso, as abordagens de análise de sentimentos no contexto da língua portuguesa ainda são pouco exploradas. Um dos principais desafios está na falta de recursos e de ferramentas específicas de PLN neste idioma, o que limita o desenvolvimento de técnicas eficazes. Por ser um idioma com vocabulário extenso e flexibilidade gramatical, seu processamento automatizado apresenta dificuldades (SOUZA *et al.*, 2023).

Com o propósito de aumentar e melhorar os recursos de PLN e na detecção de polaridade em textos para a língua portuguesa, surgiu a criação do léxico de opinião *OpLexicon*. Uma ferramenta composta por uma combinação de três métodos diferentes, com um conjunto diversificado de palavras e expressões, sendo abrangente para domínios gerais (SOUZA; VIEIRA, 2011).

Conforme apresentado por Wu *et al.* (2021), a utilização de uma combinação de léxico de sentimentos juntamente com algoritmos de aprendizado de máquina entrega resultados mais precisos. Estudos apresentados de análise de sentimentos em bases de dados em português brasileiro não utilizam essa técnica, e muitos pesquisadores relatam as dificuldades de realizar as análises no idioma.

A combinação de algoritmos de aprendizado de máquina com léxico de sentimentos ou de opinião para a análise de uma base de dados em português brasileiro é uma pesquisa que visa observar a diferença entre a técnica de algoritmos de aprendizado de máquina isolados da técnica de combinação. Conforme apresentado em estudos realizados por Cardoso *et al.* (2021), Wu *et al.* (2021), Souza *et al.* (2023), Silva *et al.* (2023), Ramanathan, Hajri e Ruth (2024), Muthukrishnan M. (2024), Yogi *et al.* (2024), a análise de sentimentos é realizada, em sua grande maioria, em base de dados compostas por textos extraídos de redes sociais.

Contudo, podem ser utilizadas bases de dados de outras plataformas, como a *Steam*. A *Steam* é uma plataforma de distribuição de jogos digitais, que contém uma comunidade ativa onde os usuários compartilham suas experiências com seus jogos favoritos (PACHECO, 2023). Este trabalho propõe um estudo comparativo entre as duas abordagens: o uso dos algoritmos SVM e *Naive Bayes* isolados e o uso dos mesmos com a combinação do léxico de opinião *OpLexicon*, com o objetivo de tentar alcançar resultados mais aprimorados.

1.1 Objetivos

Analisar os resultados dos algoritmos SVM e *Naive Bayes* com e sem a aplicação da combinação do léxico de opinião *OpLexicon*, para a identificação eficaz da polaridade em textos de avaliações de jogos digitais, na plataforma *Steam*. Este objetivo geral divide-se nos seguintes objetivos específicos:

- Ajustar um modelo SVM e um modelo *Naive Bayes* para a tarefa de detecção de polaridade nos textos das avaliações;
- Avaliar o desempenho dos algoritmos SVM e *Naive Bayes*;
- Ajustar um modelo SVM e um modelo *Naive Bayes* combinados com o modelo *OpLexicon* para a tarefa de detecção de polaridade nos mesmos textos;
- Avaliar o desempenho dos algoritmos SVM e *Naive Bayes* combinados com o léxico *OpLexicon*;
- Comparar o desempenho de ambos resultados obtidos pelas duas abordagens avaliadas.

1.2 Justificativa

A análise de sentimentos é uma área que tem maior proximidade com dados em inglês em suas análises, enquanto dados provenientes de outros idiomas são pouco explorados, gerando um déficit de recursos na área para idiomas diferentes do inglês (FREITAS; VIEIRA,

2015). Segundo Coutinho (2022), cada língua contém seus próprios níveis de dificuldade, e poder realizar o processamento de informações de forma automática nas demais línguas é muito importante. Para a língua portuguesa, esses desafios em técnicas de PLN estão muito presentes, mesmo com trabalhos que contribuem para a criação de novas abordagens dessas técnicas para o idioma.

Diante desses dados, a combinação de novas técnicas de análise de sentimentos para dados na língua portuguesa brasileira pode ser de grande importância para a comunidade científica, a fim de melhorar os recursos e contribuir para a obtenção de melhores resultados nas análises feitas com bases de dados em português. A escolha dos algoritmos SVM e *Naive Bayes* é dada a partir dos trabalhos de Januário *et al.* (2022), Souza *et al.* (2023), Silva *et al.* (2023) e Yogi *et al.* (2024), que utilizaram diversos algoritmos em suas pesquisas e os que mais obtiveram resultados satisfatórios foram o SVM e *Naive Bayes*. Outro fator decisivo na escolha foi o baixo custo computacional e um tempo de treinamento menor, diferente de Redes Neurais, como o BERT (*Bidirectional Encoder Representations from Transformers)* , que necessita de um poder computacional significativamente maior e sua eficiência se torna superior apenas em bases maiores, em bases menores, a diferença entre o BERT e os algoritmos tradicionais não é muito expressiva (BORGES, 2024).

Espera-se que esse trabalho possa analisar e comprovar que a combinação de um léxico de opinião juntamente com algoritmos de aprendizado de máquina possa obter resultados mais aprimorados em comparação com a utilização dos algoritmos isolados.

1.3 Estrutura do trabalho

Este documento está estruturado em capítulos e seções da seguinte forma: O Capítulo 1 aborda uma introdução que contextualiza e apresenta a análise de sentimentos, sua problemática e importância, define os objetivos gerais e específicos, e, por fim, a justificativa. No Capítulo 2, é apresentado o “Referencial Teórico”, que detalha a análise de sentimentos e traz as definições de aprendizado de máquina, com subseções dedicadas aos algoritmos SVM e *Naive Bayes*, também é abordado o léxico de opinião, com uma subseção dedicada ao método *OpLexicon*, além do método PLN. Ainda neste capítulo, por fim, são apresentados, de forma resumida, os trabalhos relacionados, que trazem uma revisão de pesquisas anteriores sobre o tema que será tratado neste trabalho. No Capítulo 3, são apresentados os “Materiais e Métodos”, descrevendo as ferramentas, abordagens e técnicas utilizadas para a resolução da proposta.

2 REFERENCIAL TEÓRICO

Neste capítulo é apresentada uma revisão de trabalhos que abordam diferentes métodos aplicados à análise de sentimentos, alguns com o foco específico em textos em português. São discutidas abordagens baseadas em aprendizado de máquina, como *Naive Bayes* e SVM, combinadas com técnicas lexicais e redes neurais, avaliando seu desempenho em postagens de mídias sociais, avaliações de produtos e notícias financeiras. Esses estudos proporcionam uma base sólida para entender os desafios e avanços na área, em especial sobre a complexidade gramatical do português brasileiro e as limitações de recursos de processamento de linguagem natural no idioma.

2.1 Processamento de Linguagem Natural (PLN)

O PLN consiste em uma área da Inteligência Artificial (IA), onde ajuda os computadores a realizar tarefas de entendimento da linguagem humana de maneira eficaz, como interpretar, manipular e entender (ANCHIÊTA *et al.*, 2021). A linguagem natural, por definição, é uma linguagem que foi desenvolvida através do uso humano na comunicação e do uso natural, ao contrário da linguagem artificial, que são construídas e criadas de maneira deliberada (SARKAR, 2019).

A Comissão Especial de Processamento de Linguagem Natural (CE - PLN) da Sociedade Brasileira de Computação (SBC) define que o objetivo do PLN é investigar, propor e desenvolver modelos, métodos, técnicas e sistemas computacionais voltados para a automação da interpretação e geração da língua humana. Dentre as principais aplicações da PLN estão: análise de sentimentos, análise morfo-sintática, análise semântica, categorização textual, ferramentas de auxílio à escrita, perguntas e respostas, recuperação e extração de informação, sumarização automática e tradução automática de textos (ANCHIÊTA *et al.*, 2021).

A base da área de PLN é construída por alguns termos, conceitos e definições, que auxiliam no entendimento de diversas aplicações da IA. Entre esses termos, os principais são: *corpus*, *tokens*, normalização, *stopwords*, *stemming*, lematização, etiquetador e analisador sintático (ANCHIÊTA *et al.*, 2021). Todos esses termos serão detalhados a seguir.

Corpus ou corpora: a definição de *corpus* é dada como um conjunto de dados linguísticos. É utilizado em estudos linguísticos e lexicográficos há muitos anos, quando a popularização dos computadores nos anos 1980 gerou mais força para a técnica. O material dentro dele é coletado a partir de um propósito e produzido de forma “natural”. Um corpus é composto por materiais já classificados, em que essa classificação pode ser de diferentes segmentos de texto, como palavras, expressões, frases, parágrafos ou o texto inteiro, por exemplo (FREITAS, 2023).

Tokens e *tokenização*: um *token* consiste em uma sequência de caracteres que juntos podem formar uma palavra, como “escada”, ou um símbolo de pontuação, como “;” por exemplo. Desse modo, a separação e indexação dessas sequências de caracteres é chamada de *tokenização* (NIEVES; MENDONÇA; FERREIRA, 2021).

Normalização: é uma etapa do pré-processamento dentro do PLN que realiza a padronização dos dados. Essa etapa consiste na remoção ou substituição de caracteres ou palavras, *Uniform Resource Locator (URL)*, itens duplicados, datas e horas, valores monetários, entre outros. Além disso, inclui uma correção ortográfica simples em palavras e abreviações. Essa etapa garante uma uniformidade e limpeza nos dados, facilitando as próximas análises (BRITO; GOMES, 2019).

Stopwords: esse termo é utilizado para classificar palavras ou termos que aparecem com uma alta frequência nos textos, mas são irrelevantes para a análise de dados, como artigos, preposições e pronomes. Essas palavras são agrupadas em uma lista focada apenas nessa função, para realizar a remoção das mesmas dos textos. Alguns exemplos são as palavras “as, os, um, uma, com, de, para, etc.” (JESUS; VIEIRA, 2023).

Stemming: é um processo que consiste na conversão de uma palavra em sua raiz. Ele pode melhorar o desempenho da análise pois ele unifica todas as palavras que contêm a mesma raiz, sendo consideradas com significados semelhantes (RIZKI; TJAHYANTO; TRIALIH, 2019). A conversão para a raiz se dá pela remoção de afixos, por exemplo, usando a palavra “desconectado”, ao remover o prefixo “des-” e o sufixo “-ado”, obtém-se a raiz “conect-”. As palavras “reconectado”, “reconectar” e “ conectou” compartilham dessa mesma raiz (VILELA; CUNHA, 2024).

Lematização: o processo dele é semelhante ao *stemming*. Na lematização, agrupa-se as variações flexionadas de uma palavra para que sejam identificadas como um único elemento, chamado de lema da palavra ou sua forma de vocabulário. Em resumo, ela relaciona textos com significados semelhantes a uma única palavra. Por exemplo, as palavras “melhor” e “bom” são diferentes, mas possuem o mesmo significado. Com isso, a lematização define que essas palavras são variações uma da outra (KHYANI *et al.*, 2021).

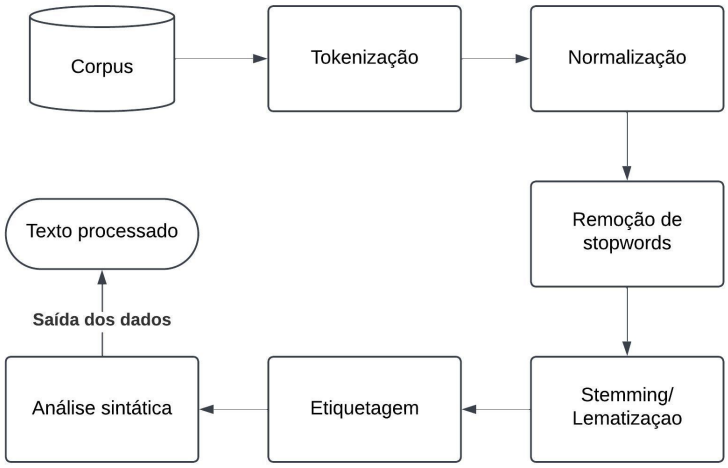
Etiquetador: definido como um sistema de marcação de classes gramaticais, o etiquetador é responsável por identificar itens lexicais e suas classes, como substantivos ou verbos, que uma palavra, em uma frase, pode representar, criando uma estrutura sintática ou árvore de derivação. Sua principal função é a resolução de ambiguidades, atribuindo categorias gramaticais corretas a palavras com múltiplos significados (SANTOS; PAIVA; BITTENCOURT, 2016).

Analisador sintático: um analisador sintático (*parser*) é uma ferramenta que realiza a verificação da escrita de um texto ou código de acordo com as regras de linguagem impostas. Sua entrada consiste em uma descrição formal da gramática da linguagem e sua saída é um código-fonte que pode reconhecer se uma determinada cadeia (sequência de palavras ou símbolos) é válida de acordo com as regras dessa gramática (BOSS; VENSKE, 2008).

A Figura 1 representa o processo de PLN, detalhando as etapas principais que os dados passam para serem transformados em informações processadas. A entrada dos dados é representada pelo "corpus", que é um conjunto de textos a ser analisado. O primeiro passo no processo é a *tokenização*, onde o texto é dividido em unidades menores, chamadas de *tokens*. Em seguida, o texto passa por uma etapa de normalização, que visa uniformizar os dados re-

movendo caracteres indesejados, como pontuação e símbolos especiais. A etapa seguinte, de remoção de *stopwords*, elimina palavras comuns, como artigos e preposições, que não agregam valor significativo ao contexto, ajudando a reduzir o ruído nos dados. O texto é então submetido ao *stemming* ou lematização, processos que buscam reduzir as palavras à sua raiz ou forma canônica. Nas últimas etapas ocorrem a etiquetagem e a análise sintática, que atribuem categorias semânticas às palavras, como substantivos e verbos, e organizam o texto em uma estrutura gramatical, respectivamente. Ao final do processo, a saída dos dados é o "texto processado", pronto para ser utilizado em tarefas como análise de sentimentos, classificação de texto ou outros tipos de análise linguística.

Figura 1 – Representação da ordem dos processos de PLN



Fonte: Autoria própria (2024).

Nos anos de 1990, uma ferramenta foi introduzida na comunidade de PLN por Church e Hanks (1990), a *Pointwise Mutual Information* (PMI). Ela tem sido utilizada para abordar ou melhorar problemas de PLN, como as co-ocorrências bidirecionais (CRUYS, 2011). A PMI também fornece uma quantificação do grau de relação entre uma palavra e uma classe de sentimento, ajudando a definir se uma sentença é positiva, negativa ou neutra (SHAPIRO; SUDHOF; WILSON, 2022). O último dado apresentado refere-se à definição da polaridade de uma frase, uma tarefa da análise de sentimentos.

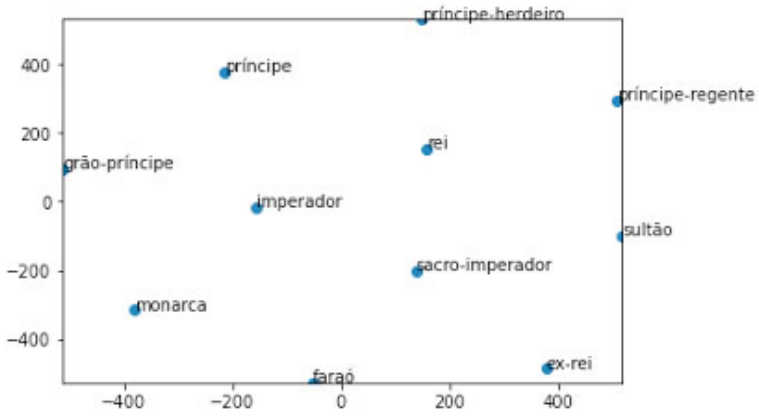
No PLN, por utilizar palavras como unidades básicas de entrada, é importante que elas sejam representadas de maneira significativa. A técnica de *word embeddings* é amplamente utilizada em sistemas modernos de PLN, e é considerada eficiente para fornecer esse tipo de representação em palavras. *Embeddings* são representações de palavras em vetores dentro de um espaço n-dimensional. Esses vetores são aprendidos a partir de grandes corpora não anotados e são capazes de capturar conhecimentos morfológicos, sintáticos e semânticos (HARTMANN *et al.*, 2017).

No trabalho de Hartmann *et al.* (2017), é realizada uma avaliação de diferentes modelos de *embeddings* para o idioma português treinados a partir de quatro métodos. Um dos métodos

utilizados foi o *Word2Vec*, amplamente empregado em PLN para gerar *embeddings* de palavras. Ele realiza o treinamento a partir de dois modelos diferentes: O *Continuos Bag-of-Words* (CBOW), que tenta prever a palavra que foi omitida dentro da sequência de palavras que ele recebe, e o *Skip-Gram*, que tenta prever as palavras vizinhas de uma palavra recebida.

A Figura 2 representa de forma abstrata o espaço vetorial, sendo possível a visualização da proximidade de palavras que contêm uma semelhança de semântica. Cada palavra no espaço contém um vetor n-dimensional contínuo de números reais que a representa. Através desse vetor, é possível capturar relações sintáticas e semânticas das palavras. A derivação da posição das palavras no espaço vetorial entrega essas relações. A relação de similaridade entres as palavras pode ser encontrada através da distância entre os vetores que estão em regiões próximas. É mostrada a proximidade entre palavras como "príncipe", "monarca", "rei" e "imperador", comprovando a característica de similaridade pois são termos similares.

Figura 2 – Representação abstrata de espaço vetorial: algoritmo Word2Vec



Fonte: Anchiêta *et al.* (2021).

Uma das aplicações do PLN, como citado acima, é a análise de sentimentos. Os recursos do PLN auxiliam na análise de texto de opinião, que é uma tarefa difícil, dado o fato de esses textos serem, em sua grande maioria, escritos de maneira informal, contendo gírias, ironia, sarcasmo, abreviações e *emoticons* (PEREIRA, 2021).

2.2 Análise de Sentimentos no PLN

A análise de sentimentos tem como o principal objetivo a extração de emoções, avaliações, atitudes e opiniões de textos escritos por pessoas sobre diversos assuntos, como produtos e eventos. Ela ajuda na interpretação da polaridade dos textos, ou seja, se são positivos, negativos ou neutros, podendo até calcular a intensidade dessas emoções (CUI *et al.*, 2023). A análise de sentimentos é amplamente utilizada em campos como o comércio, negócios, política e serviços, ajudando profissionais de diversos setores na tomada de decisões (KHATUA; KHATUA; CAMBRIA, 2020). Por serem textos escritos por pessoas, a grande maioria deles são textos não estruturados, o que é um dos maiores desafios na análise de sentimentos, pois as

ferramentas de PLN apresentam limitações que dificultam a realização de uma análise precisa em textos nesse formato (CUI *et al.*, 2023).

A análise de sentimentos utiliza o PLN junto com outros processos avançados, como a classificação de sentimentos (supervisionada ou não supervisionada), análise subjetiva ou objetiva, e extração de opiniões. Umas das principais tarefas é a classificação de sentenças subjetivas e objetivas, onde é feita uma análise dos textos para identificar aspectos que expressam emoções ou sentimentos, classificando-os como sentenças subjetivas. As sentenças subjetivas são compostas por perspectivas, pensamentos, comentários e opiniões de pessoas. Já as sentenças objetivas, por serem compostas por fatos, são descartadas, pois não expressam opinião ou sentimento (PANDYA; MEHTA, 2020).

Para classificar a subjetividade, uma das formas de fazê-lo é utilizar o chamado "classificador básico", que usa algoritmos como *Naive Bayes* ou SVM. No caso do *Naive Bayes*, Pang e Lee (2004) propuseram um modelo em que cada sentença de um documento é avaliada quanto à probabilidade de ser subjetiva ou objetiva. No contexto da análise de subjetividade, o *Naive Bayes* usa a fórmula de probabilidade condicional para calcular a probabilidade de uma sentença s_i pertencer à classe de subjetividade (subjetivo). Essa probabilidade é calculada com base em um conjunto de características, como palavras presentes na sentença. Formalmente, isso é representado como $P(\text{subjetivo}|s_i)$, sendo essa probabilidade baseada na frequência das palavras da sentença em textos subjetivos e objetivos previamente rotulados.

O estudo de Pang e Lee (2004) mostra que a probabilidade de uma sentença s_i ser subjetiva ou objetiva é determinada pela fórmula de *Bayes*, representada por $P_{\text{NB}}^{\text{sub}}$. Para cada sentença, a pontuação de subjetividade, representada por $\text{ind}_1(s_i)$, é calculada por meio da Equação 1:

$$P_{\text{NB}}^{\text{sub}}(s_i) = \frac{P(s_i|\text{subjetivo}) \cdot P(\text{subjetivo})}{P(s_i)} \tag{1}$$

onde $P(s_i|\text{subjetivo})$ representa a probabilidade de observar as palavras da sentença s_i em textos subjetivos, e $P(\text{subjetivo})$ é a probabilidade a priori de que uma sentença seja subjetiva. De forma complementar, a pontuação de objetividade, representada por $\text{ind}_2(s_i)$, é calculada como $\text{ind}_2(s_i) = 1 - \text{ind}_1(s_i)$, garantindo que a soma das probabilidades de subjetividade e objetividade de uma sentença seja igual a 1. Esse processo permite ao classificador *Naive Bayes* atribuir uma etiqueta de subjetividade ou objetividade a cada sentença com base na análise probabilística das palavras presentes na sentença. No trabalho de Pang e Lee (2004), essas pontuações individuais são utilizadas em um sistema mais complexo de detecção de subjetividade que utiliza corte mínimo de grafos. Mesmo em um modelo básico, o uso de *Naive Bayes* oferece uma maneira eficaz de prever a subjetividade, explorando a frequência das palavras em diferentes contextos de polaridade, permitindo a classificação com base em dados textuais previamente observados.

A definição da polaridade de uma palavra também é dada pela da orientação semântica (HU; LIU, 2004). Nesse processo, faz-se uso do PMI, uma ferramenta que mede a frequência de

co-ocorrência entre pares de palavras, avaliando a intensidade de sua relação. É útil para determinar o grau de relação entre duas variáveis aleatórias. Muitos estudos combinam o PMI com outras medidas para a extração de palavras, a fim de superar suas limitações (SRIVASTAVA; PANDEY; AGGARWAL, 2022).

No estudo de Srivastava, Pandey e Aggarwal (2022), é mostrado que o PMI é uma medida ideal da teoria da informação, no que diz respeito às normas de associação de palavras. O PMI pode ser utilizado para determinar se há realmente uma conexão entre duas coisas ou se é apenas coincidência. Conforme mostrado na Equação 2, o PMI(palavra 1, palavra 2) é definido como a informação mútua entre duas palavras, levando em consideração as probabilidades de P(palavra 1) e P(palavra 2).

$$PMI(palavra\ 1,palavra\ 2) = log\left(\frac{P(palavra_1 \& palavra_2)}{P(palavra_1)P(palavra_2)}\right) \quad (2)$$

A quantidade de dados utilizados no treinamento afeta diretamente o desempenho do PMI. Se as palavras 1 e 2 não ocorrerem no mesmo corpus, a precisão do algoritmo será prejudicada. O uso de sistemas de *Information Retrieval* (IR) supera essa limitação, pois esses sistemas avaliam consultas por meio da análise de dados estatísticos extraídos de coleções de documentos. Assim, o *PMI-IR* utiliza a coleção disponibilizada pelo sistema IR para realizar suas análises (SRIVASTAVA; PANDEY; AGGARWAL, 2022).

A análise de sentimentos é classificada por Bhuiyan, xu e Josang (2009) em três abordagens principais: nível de documento, nível de sentença e nível de característica (ou aspecto), como ilustrado na Figura 3. Quando as avaliações são classificadas no nível de documento ou de sentença, não é possível determinar exatamente o que o autor da opinião gosta ou não. Por outro lado, o nível de características investiga maneiras de classificar cada aspecto, mas é mais difícil de executar (WESTERSKI, 2007). A seguir, é detalhado cada um desses níveis.

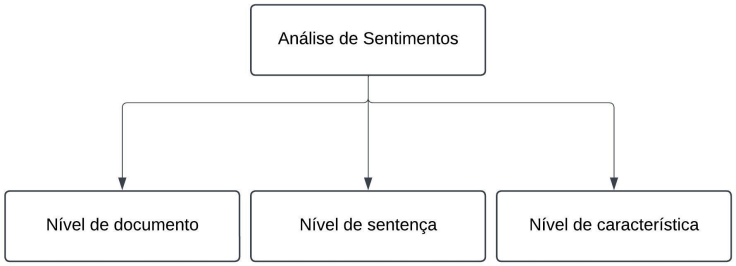
Nível de Documento: este é o primeiro nível da análise de sentimentos, que se baseia exclusivamente no conteúdo do documento. A tarefa neste nível é determinar se o documento inteiro expressa um sentimento positivo ou negativo. Este nível de análise pressupõe que o documento expresse opiniões sobre um único tópico, por exemplo, uma resenha de filme (FREITAS, 2015).

Nível de sentença: neste nível, a tarefa principal é analisar cada sentença no documento, com o objetivo de determinar se cada sentença expressa uma opinião de polaridade positiva, negativa ou neutra. Esse processo está relacionado à classificação de subjetividade, distinguindo sentenças objetivas, que apresentam fatos, das subjetivas que refletem opiniões ou emoções (FREITAS, 2015).

Nível de características: por fim, o último nível de análise realiza um estudo mais aprofundado. Em vez de se concentrar em construções linguísticas maiores, como documentos, parágrafos, sentenças ou frases, este nível de análise foca diretamente no objeto da opinião. Isso

significa que, em vez de observar apenas a estrutura do texto, ele examina aspectos específicos (FREITAS, 2015).

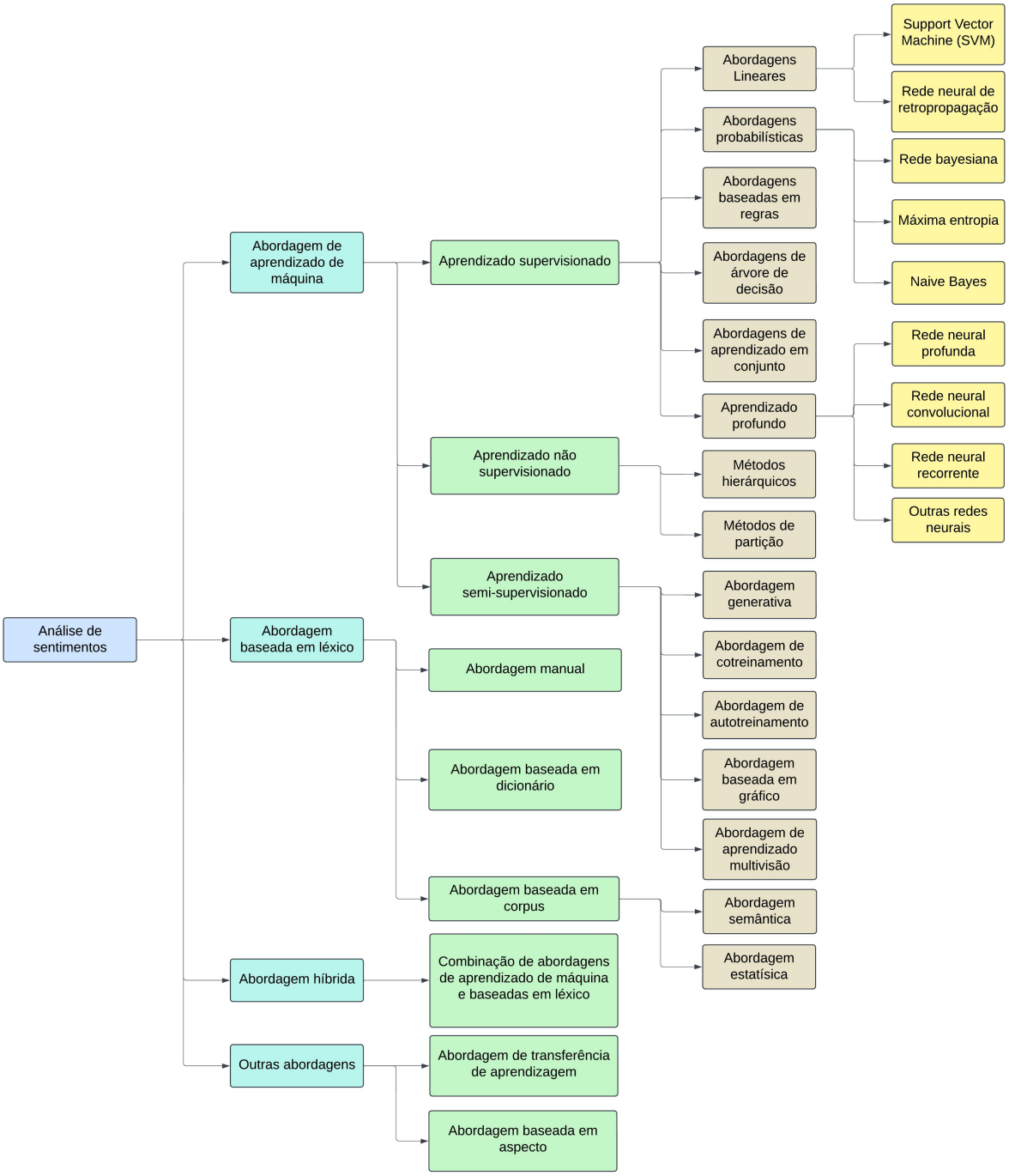
Figura 3 – Classificação da pesquisa em mineração de opiniões



Fonte: Adaptado de Bhuiyan, xu e Josang (2009).

Abordagens de aprendizado de máquina, baseadas em léxico e híbridas são as três principais abordagens existentes dentro da análise de sentimentos. Todas elas contêm técnicas e abordagens específicas. Há também outras abordagens, como a transferência de aprendizado e a abordagem baseada em aspecto, por exemplo (BIRJALI; KASRI; BENI-HSSANE, 2021). Na Figura 4, são ilustradas todas essas abordagens dentro da análise de sentimentos em um fluxograma.

Figura 4 – Abordagens de análise de sentimentos



Fonte: Adaptado de Birjali, Kasri e Beni-Hssane (2021).

Para realizar uma análise de sentimentos, é necessária a coleta de dados, sendo seu objetivo principal. Nesse caso, canais de comunicação social, como o *X*, *Facebook* ou outros meios pré-existentes, são as principais fontes desses dados. Base de dados de blogs, fóruns, resenhas, artigos de notícias e redes sociais são os principais meios de comunicação dos quais os pesquisadores extraem dados para realizar estudos (PANDYA; MEHTA, 2020).

A análise de sentimentos enfrenta uma escassez de recursos para muitas línguas, incluindo o português, em comparação com o inglês. Alguns estudos foram realizados para desenvolver novas soluções para a classificação de sentimentos em outras línguas. Muitas das ferramentas desenvolvidas utilizam um método de tradução automática para traduzir essas línguas para o inglês. Em seguida, aplicam as técnicas de *word embeddings* em inglês, as polaridades de um léxico de sentimentos e um modelo de rede neural convolucional para a classificação (ZHANG; WANG; LIU, 2018).

Por outro lado, as línguas possuem suas próprias particularidades, expressões e gírias que variam conforme a cultura local. A tradução automática não resolve bem essas questões. Isso resulta na dificuldade dessas estratégias de análise de sentimentos multilíngue em extrair o verdadeiro sentimento de determinada palavra, visto que sua eficiência se aplica apenas às expressões gerais de sentimentos compartilhadas entre os idiomas (CHEN *et al.*, 2019).

Entre os idiomas com recursos linguísticos limitados disponíveis, o português, mesmo que seja uma das cinco línguas mais faladas na web e tenha aproximadamente 290 milhões de falantes no mundo, se enquadra nessa classificação. Estudos já foram realizados para identificar o estado da arte da mineração de textos para o idioma, e sobre a extração automática de termos para o português brasileiro (PEREIRA, 2021).

Essas limitações reafirmam a problemática desta pesquisa, que é a falta de ferramentas e recursos de PLN adequados para a análise de textos no idioma português, o que dificulta o progresso de estudos e aplicações voltadas para essa língua.

2.3 Aprendizado de Máquina

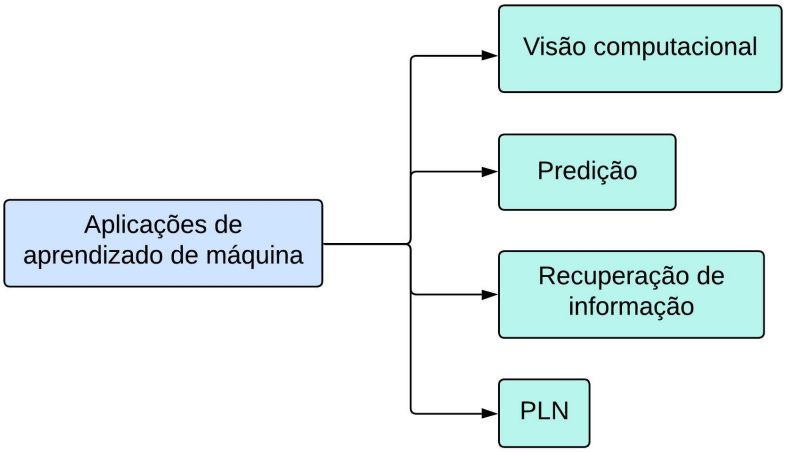
Durant (1953) refletiu sobre o aprendizado em seu livro *"The Pleasures of Philosophy"*. Ele afirma que aprender é um fenômeno muito pessoal para nós. No entanto, o aprendizado não se limita apenas aos humanos. Desde as espécies mais simples, como um paramécio, até as plantas, apresentam comportamentos inteligentes. Apenas as coisas materiais naturais não vivas não estão envolvidas no aprendizado. Isso cria uma relação entre viver e aprender. Criações não vivas geradas pela natureza dificilmente tem algo para aprender. Porém é questionado se pode-se introduzir aprendizado em criações humanas não vivas, chamadas de máquinas (MOHAMMED; KHAN; BASHIER, 2016).

A diferença entre a execução de tarefas por um ser humano e por uma máquina reside na sua inteligência. Por meio do sistema neural, os seres humanos realizam um processo de percepção de dados, onde são organizados, reconhecidos, comparados com experiências armazenadas na memória e interpretados, permitindo que o cérebro tome decisões e direcione partes do corpo para reagir adequadamente. Uma máquina não é capaz de lidar com dados coletados da mesma maneira. Ela não possui a capacidade de interpretar dados da mesma forma que um ser humano, relacioná-los a experiências anteriores, aprender com elas e armazenar

novas experiências na memória. Ou seja, as máquinas não aprendem com a experiência da mesma forma que os humanos (MOHAMMED; KHAN; BASHIER, 2016).

O aprendizado de máquina é um subcampo da ciência da computação desenvolvido a partir da combinação das áreas de reconhecimento de padrões IA (SIMON, 2013). O uso do aprendizado de máquina se popularizou rapidamente dentro da ciência da computação. É amplamente utilizado em buscas na web, filtros de spam, sistemas de recomendação, colocação de anúncios e muitas outras aplicações (DOMINGOS, 2012). Tecnicamente, o aprendizado de máquina abrange diversos domínios de aplicação. Entre seus domínios estão a visão computacional, predição, PLN e recuperação de informações. Essa classificação é ilustrada na Figura 5 (SHINDE; SHAH, 2018).

Figura 5 – Aplicações de aprendizado de máquina

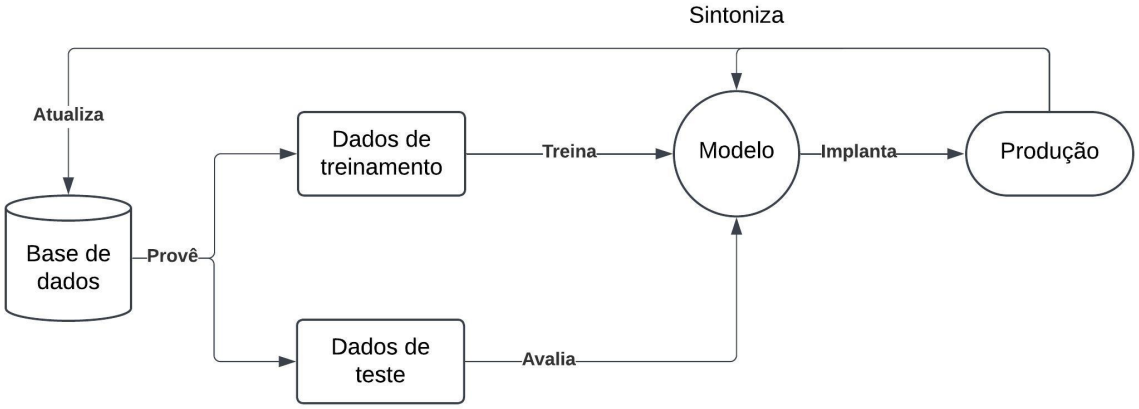


Fonte: Adaptado de Shinde e Shah (2018).

A abordagem supervisionada é uma técnica de aprendizado de máquina que utiliza dados rotulados para aprender uma função que mapeia entradas para suas respectivas saídas. Por outro lado, na abordagem não supervisionada, os algoritmos identificam e apresentam padrões nos dados de forma autônoma. Quando recebem novos dados, utilizam as características aprendidas para classificar as novas entradas (MAHESH, 2019).

Os algoritmos de aprendizado supervisionado são aqueles que requerem supervisão externa, onde o conjunto de dados é dividido em dois: o conjunto de treinamento e o conjunto de teste. O conjunto de treinamento inclui variáveis de saída que o algoritmo deve prever ou classificar, permitindo-lhe aprender padrões a partir dos dados (MAHESH, 2019). Durante o processo de aprendizado, é criada uma função que pode prever os valores de saída quando novos dados são introduzidos. Após o treinamento, o sistema gera resultados para dados de entrada, compara-os com os resultados reais e esperados, e ajusta o modelo conforme necessário para melhorar as previsões ou classificações (SARAVANAN; SUJATHA, 2018). Esse funcionamento do aprendizado supervisionado pode ser visto pela representação da Figura 6.

Figura 6 – Fluxo de trabalho da aprendizagem supervisionada



Fonte: Adaptado de Mahesh (2019).

Os métodos supervisionados podem ser divididos em duas categorias principais: regressão e classificação. Os algoritmos de regressão têm como objetivo modelar dependências e relações entre a saída alvo e as características de entrada, com o intuito de prever novos dados que ainda não foram observados. Já os algoritmos de classificação atribuem rótulos de classe para exemplos de um determinado domínio. Alguns dos principais algoritmos de classificação incluem a árvore de decisão, *Naive Bayes*, SVM e Redes Neurais Artificiais (MRABET; MAKKAOUI; FAIZE, 2021).

2.3.1 *Support Vector Machine* (SVM)

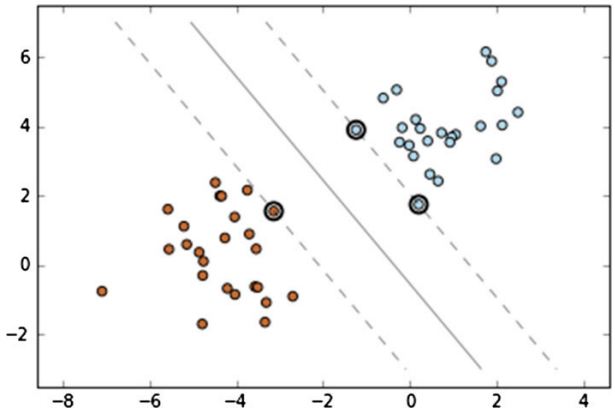
Baseada no princípio de minimização do risco estrutural, a SVM é uma das ferramentas mais sofisticadas de aprendizado de máquina. Originalmente, o algoritmo foi desenvolvido para classificação binária, onde um hiperplano é construído para maximizar a margem entre duas classes. Seu objetivo é determinar uma hipótese h que minimize o erro verdadeiro, entendido como a probabilidade de h cometer um erro ao classificar um exemplo de teste não visto. Para isso, pode-se utilizar um limite superior que relacione o erro verdadeiro da hipótese h ao seu erro no conjunto de treinamento, bem como à complexidade de H, medida pela dimensão de VC (Vapnik-Chervonenkis), que representa o espaço de hipóteses contendo h. Assim, a SVM procura encontrar a hipótese h que minimize esse limite do erro verdadeiro, controlando de maneira eficaz e eficiente a dimensão de VC de H (VAPNIK, 1999).

A SVM é um algoritmo altamente versátil. Ela aprende tanto funções lineares de limiar quanto classificadores mais complexos, como classificadores polinomiais ou redes neurais de múltiplas camadas. Uma característica importante é o aprendizado independente do número de características no conjunto de dados, pois a SVM avalia a distância entre os dados das

diferentes classes e maximiza a margem entre elas, sem contar o número de características (JOACHIMS, 1998).

A Figura 7 ilustra a separação de duas classes em um espaço bidimensional, onde o hiperplano é posicionado de modo a maximizar a margem entre as classes. Os círculos representam os vetores de suporte, que são os pontos mais próximos do hiperplano e que determinam sua posição.

Figura 7 – Separação de classes e maximização da margem em SVM



Fonte: Chauhan, Dahiya e Sharma (2019).

Zhang, Su e Xu (2006) explicam o processo de aprendizado da SVM para a classificação de texto seguindo algumas etapas. Primeiro, antes de usar a SVM é necessário representar o texto numericamente. A representação vetorial pode ser expressa da seguinte forma: $\mathbf{x} = [wd_1, wd_2, \ldots, wd_n]$, onde o texto d é transformado em um vetor de características, sendo que cada elemento do vetor correspondente ao valor de uma característica do texto. Essas características são ponderadas por um valor wd_i, que representa a importância ou o peso da característica i no texto d. Para o aprendizado com margem suave, utiliza-se a formulação dual de Lagrange (Equação 3). A "margem suave" é utilizada quando os dados não são linearmente separáveis, permitindo alguns erros, como dados que ficam do lado errado do hiperplano. A formulação dual de Lagrange é utilizada na SVM para transformar o problema de maximização da margem em uma abordagem mais eficiente e viável computacionalmente. Ela permite incorporar as restrições diretamente na função objetivo, garantindo que o problema seja resolvido considerando a separação das classes. Através dos multiplicadores de Lagrange, o foco é dado nos vetores de suporte, reduzindo a complexidade computacional e concentrando o cálculo nos pontos mais relevantes para definir a margem máxima.Tornando o método eficiente e adequado para problemas de classificação.

$$\max_{\alpha} \sum_{i} a_i - \frac{1}{2} \sum_{i,j} a_i a_j y_i y_j K(\mathbf{x}_i \cdot \mathbf{x}_j) \tag{3}$$

$$\text{s.t. } 0 \leq a_i \leq C \quad (\text{para } i = 1, \ldots, M), \quad \sum_{i=1}^{M} a_i y_i = 0$$

$\sum_i a_i$ representa a maximização da soma dos multiplicadores a_i, que está diretamente ligada a separação dos dados. $\frac{1}{2}\sum_{i,j} a_i a_j y_i y_j K(\mathbf{x}_i \cdot \mathbf{x}_j)$ é o termo de penalização que representa a interação entre os vetores de suporte (pontos próximos ao hiperplano). Esse termo ajuda a determinar a margem máxima. y_i e y_j são as classes de i e j que contém valores típicos +1 ou -1 para problemas binários. $K(\mathbf{x}_i \cdot \mathbf{x}_j)$ é a função de kernel que calcula a similaridade entre os exemplos $\mathbf{x}_i$ e $\mathbf{x}_j$. Pode ser linear, polinomial, etc. O kernel transforma os dados para um espaço dimensional superior onde eles podem ser linearmente separáveis.

Após o treinamento do modelo, a função de decisão (Equação 4) de um SVM é usada para classificar novos dados. A função calcula a soma ponderada das similaridades entre o ponto de teste x e os pontos de treinamento x_i, ponderadas pelos multiplicadores a_i e pelos rótulos y_i (ZHANG; SU; XU, 2006).

$$f(\mathbf{x}) = \text{sgn}(\sum_{i=1}^{M} a_i y_i K(\mathbf{x}, \mathbf{x}_i) + b) \tag{4}$$

Em resumo, sendo considerado o classificador de texto com mais precisão, a SVM encontra um hiperplano que separa os dados em duas categorias, positiva ou negativa. O nome desse algoritmo é dado por conta da utilização de um vetor de suporte, que é um conjunto de pontos de dados, usados para determinar o limite de cada plano. O classificador prevê a qual lado da margem o dado pertence (JARDIM; MORA; SANTANA, 2021).

2.3.2 *Naive Bayes*

O *Naive Bayes* pertence a uma família de "classificadores probabilísticos" que é baseada no teorema de Bayes com fortes suposições de independência entre as características. Ele pertence ao grupo dos modelos de redes Bayesianas mais simples, porém ele pode alcançar níveis de precisão altos quando se é combinado com a estimativa da densidade do núcleo (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). A técnica de classificação do *Naive Bayes* baseada no teorema de Bayes pode ser vista através da Equação 5.

$$P(Y|X) = \frac{P(X|Y) \cdot P(Y)}{P(X)} \tag{5}$$

No *Naive Bayes*, assume-se que a probabilidade de uma palavra em um documento pertencer a uma determinada categoria não está relacionada à probabilidade de outras palavras estarem na mesma categoria. A classificação de um documento, dentro dessa técnica, é dividida em quatro etapas. A primeira é transformar o conjunto de dados em uma tabela de frequência das palavras. A segunda etapa consiste no cálculo da probabilidade a priori usando a Equação 6

$$P(C) = \frac{N_c}{N} \tag{6}$$

em que $P(C)$ representa a probabilidade da classe, N_c é o número total de ocorrências de uma classe específica no conjunto de treinamento, e N indica o número total de ocorrências de todas as classes no conjunto de treinamento. Na terceira etapa, é realizado o cálculo da probabilidade condicional de cada atributo (palavra) em relação à classe, chamada de verossimilhança. Por último, é calculada a probabilidade a posteriori utilizando a Equação 7 (JARDIM; MORA; SANTANA, 2021).

$$C_{\text{map}} = argmax P(X1, X2, X3, ..., Xn) \cdot P(C) \tag{7}$$

Na Equação 7, C_{map} representa a classe mais provável (*Maximum a Posteriori – map*). É indicado por $argmax$ que a classe escolhida é aquela que maximiza a expressão $P(X1, X2, X3, ..., Xn) \cdot P(C)$, onde $P(X1, X2, X3, ..., Xn)$ representa a probabilidade conjunta dos atributos $(X1, X2, X3, ..., Xn)$ ocorrerem.

2.3.3 Desempenho dos algoritmos

O desempenho dos algoritmos de aprendizado supervisionado é medido a partir de métricas. Avaliar a eficácia de modelos antes de implementá-los na prática é essencial. Na área de aprendizado de máquina, existem várias medidas para medir o desempenho do aprendizado supervisionado, entre as principais, há a acurácia, precisão, *recall* e *f1-score*, as quais podem ser derivadas da matriz de confusão (MRABET; MAKKAOUI; FAIZE, 2021).

A acurácia é uma métrica de desempenho que indica a proporção de previsões corretas em relação ao total de registros do conjunto de dados utilizados para avaliação (ASTHANA; HAZELA, 2020). A acurácia nem sempre reflete a eficiência do modelo de aprendizado de forma precisa, pois depende do equilíbrio entre os dados utilizados para a avaliação. Portanto, em cenários de dados desbalanceados, a acurácia pode ser uma métrica enganosa. A representação da acurácia pode ser vista na Equação 8, onde N é o tamanho da amostra de dados de teste usados na avaliação, CC o conjunto de classificação correta e Card(CC) a quantidade de elementos dentro do conjunto de CC (MRABET; MAKKAOUI; FAIZE, 2021).

$$\text{Acurácia} = \frac{1}{N} Card(CC) \tag{8}$$

A matriz de confusão, também conhecida como matriz de erro ou contingência, é usada para avaliar o desempenho de modelos de classificação (ASTHANA; HAZELA, 2020). Ela conta o número de vezes que cada rótulo de classe ocorre, tanto nos casos em que a classificação foi correta quanto nos casos em que houve erro. Em um contexto de classificação, os valores de saída costumam ser binários, como 1 ou 0, ou classificados como verdadeiro ou falso, ou ainda em formato de texto, como “spam” e “não spam”. Por padrão, ela aplica o valor binário para contar as ocorrências de cada classe para cada estado (YETURU, 2020). É chamado de verdadeiro-positivo (TP) e verdadeiro-negativo (TN) o número de instancias positivas e negati-

vas previstas corretamente. Já as instancias positivas e negativas previstas incorretamente são chamadas de falso-positivo (FP) e falso-negativo (FN) (MRABET; MAKKAOUI; FAIZE, 2021). A matriz de confusão é representada pelo Quadro 1.

Quadro 1 – Representação da Matriz de Confusão

	Positivo	**Negativo**
Positivo	TP	FP
Negativo	FN	TN

Fonte: Adaptado de Mrabet, Makkaoui e Faize (2021).

A precisão é uma métrica utilizada para medir a proporção de previsões corretas realizadas para a classe verdadeiro-positivo (TP) em relação ao total de instâncias previstas como classe positiva. No caso são incluídas as instâncias de verdadeiro-positivo (TP) e falso-positivo (FP). Assim, a precisão avalia o quão assertivo é o modelo ao prever a classe positiva, considerando apenas as instâncias classificadas como tal (YETURU, 2020). Mrabet, Makkaoui e Faize (2021) representam a métrica de precisão pela Equação 9.

$$\text{Precisão} = \frac{TP}{TP + FP} \tag{9}$$

A métrica *recall* se concentra principalmente na proporção de todas as previsões corretas realizadas para a classe verdadeiro-positivo (TP) em relação ao total de classes positivas reais presentes no conjunto de dados. Ele mede a capacidade do modelo identificar corretamente as instâncias da classe positiva (YETURU, 2020). Mrabet, Makkaoui e Faize (2021) apresentam a Equação 10 como a equação da métrica de *recall*.

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

Quando se há um contexto de problemas de classificação com dados desbalanceados, a métrica *F-Score* é a opção que se é usada para superar as limitações da métrica de acurácia. Ela representa a média harmônica entre a precisão e o *recall*. Ele é usado com frequência no campo de recuperação de informações para medir o desempenho em buscas, classificação de documentos e classificação de consultas (SHOBHA; RANGASWAMY, 2018). A Equação 11 ilustra o cálculo do *F1-Score* (MRABET; MAKKAOUI; FAIZE, 2021).

$$\textit{F1-Score} = \frac{2}{\frac{1}{recall} + \frac{1}{\text{precisão}}} \tag{11}$$

2.4 Léxico de Opinião

O léxico de opinião é uma peça fundamental na área de PLN, especialmente para a análise de sentimentos. Esse léxico é uma coleção de palavras, frases e expressões idiomáticas que carregam consigo uma carga semântica de opinião, que geralmente são usadas para

expressar sentimentos positivos ou negativos. Palavras como “lindo”, “maravilhoso”, “bom” e “incrível” são palavras de opinião positivas, e “ruim”, “pobre”, e “terrível” são palavras de opinião negativas, por exemplo. Ele permite que máquinas identifiquem e classifiquem a polaridade de um texto, se ele expressa uma opinião positiva, negativa ou neutra (LIU *et al.*, 2010).

A definição de polaridade das palavras dentro de um léxico de opinião vem do processo de identificação da orientação semântica, que visa determinar se uma palavra possui uma conotação positiva ou negativa. Essa orientação indica como a palavra desvia do padrão para seu grupo semântico. Palavras que expressam um estado desejável, como “incrível” ou “eficiente”, tem uma orientação positiva, enquanto palavras que representam estados indesejáveis, como “decepcionante” ou “problemático”, têm uma orientação negativa (HU; LIU, 2004).

A identificação da orientação semântica (*SO*) de um texto vem através de um cálculo (Equação 12) utilizando o algoritmo *PMI-IR* (*Pointwise Mutual Information with Inverse Ratio*) para estimar a orientação semântica. Esse cálculo é feito de forma que se é comparada a similaridade de uma frase com uma palavra de referência positiva, “excelente” por exemplo, e a similaridade com uma palavra de referência negativa, “ruim” por exemplo. A representação da formula desse cálculo é dada da seguinte forma (TURNEY, 2002):

$$\textit{SO}(\text{frase}) = \textit{PMI}(\text{frase}, "excelente") - \textit{PMI}(\text{frase}, "ruim") \tag{12}$$

Na Equação, a orientação semântica de uma frase é a diferença entre o PMI da frase em relação à palavra de referência positiva (“excelente”) e o PMI da frase em relação à palavra de referência negativa (“ruim”) (TURNEY, 2002).

A criação de um léxico de opiniões exige que seja feita uma combinação de diferentes métodos, tornando o léxico robusto e abrangente. As três abordagens principais consistem na abordagem manual, baseada em dicionário e baseada em corpus. A manual é uma abordagem precisa, porém cara e trabalhosa, exigindo muito tempo e esforço, por ser realizada através de uma curadoria humana, onde é realizada de forma manual a anotação e validação de palavras com suas polaridades, e a inclusão de termos específicos de domínio. A abordagem baseada em dicionário é uma técnica que utiliza um conjunto inicial de palavras de opinião, junto de recursos lexicais e dicionários online, como o *WordNet* por exemplo. É realizada a rotulação automática dos termos com sua polaridade sentimental correspondente. Também é feita a expansão do léxico, buscando sinônimos, antônimos e relações semânticas. Por fim, a abordagem baseada em Corpus consiste na análise de grandes conjuntos de dados textuais para identificar padrões linguísticos e sintáticos entre palavras de dentro do corpus para verificar a presença de opiniões. Também há técnicas que exploram as estatísticas de co-ocorrência das palavras, como a técnica de propagação de polaridade, para rotular as palavras com suas determinadas polaridades (DARWICH *et al.*, 2019).

Outra característica presente no léxico de opinião é a definição de domínio, se ele será um léxico específico em domínio ou livre de domínio. Essa também é uma parte de muita im-

portância no léxico, pois define se o léxico será focado em análise em textos de diferentes áreas ou não. O léxico específico em domínio contém palavras e expressões que possuem uma polaridade positiva ou negativa em uma área em específico. Uma palavra pode conter duas polaridades, positiva e negativa, e o contexto que ela está inserida define qual é sua polaridade. Por exemplo, a palavra “longo” pode ter uma polaridade positiva quando seu domínio está inserido na saúde (“o método especificado para usar este dispositivo é confuso, mas o dispositivo é de longa duração”), já no âmbito de livros ela pode conter a polaridade negativa (“este livro contém uma história muito longa e confusa”) (VISHNU; APOORVA; GUPTA, 2014).

O léxico que for livre de domínio irá conter palavras e expressões que possuem a mesma polaridade, positiva ou negativa, em diferentes domínios. Por exemplo, a palavra “confuso”, em sua maioria das vezes, contém a polaridade negativa em diversas áreas em que ela é inserida (VISHNU; APOORVA; GUPTA, 2014).

No estudo de Machado (2023) é relatado a existência de um atraso nos métodos de PLN para a língua portuguesa em comparação com a língua inglesa por conta do baixo número de recursos disponíveis para a língua, o que dificulta a realização de pesquisas. É apontado alguns avanços, como o primeiro corpus em português com aspectos anotados, que foi criado em 2014, ou em 2018, que teve a criação de um corpus com anotação de aspectos implícitos. O léxico de opinião é uma ferramenta que se enquadra dentro dessa escassez de recursos. Foi com essa necessidade dentro da comunidade científica que a criação do *OpLexicon* surgiu.

O *OpLexicon* é um léxico de opinião desenvolvido para a língua portuguesa pelos autores Souza e Vieira (2011) na Pontifícia Universidade Católica do Rio Grande do Sul (PUC-RS), com o objetivo de oferecer um novo recurso de PLN para a análise de sentimentos e detecção de polaridade de textos. O léxico foi criado a partir da junção de 3 métodos já consolidados na literatura: o método baseado em corpus fundamentado no estudo de Turney (2002), uma variação do método baseado em tradução de Mihalcea, Banea e Wiebe (2007), e um método baseado em tesauro semelhante ao proposto por Kamps *et al.* (2004). Um tesauro consiste em um conjunto de vocabulários que são relacionados entre si de maneira semântica. Ele atua como uma ferramenta para manter um controle terminológico, padronizando o uso de palavras. Pode ser organizado de forma hierárquica, como categorias maiores e subcategorias menores, ou por aproximação semântica. Seu objetivo é organizar e facilitar a busca de informações de maneira eficiente (SALES; CAFÉ, 2009). A aplicação dos três métodos resultou na criação de três léxicos distintos que foram combinados para formar um único léxico abrangente para o português, contendo na época 7077 palavras e expressões polares.

Diferente do método baseado em corpus aplicado por Turney (2002) que utiliza a técnica PMI-IR, a metodologia aplicada no *OpLexicon* consiste na utilização de um corpus específico para bi-gramas – uma análise de ocorrências de pares de palavras consecutivas no corpus – extraídos de um corpus em português, sem aplicar a análise de proximidade semântica, como Turney faz.

O método baseado em tesauro proposto por Kamps *et al.* (2004) utiliza o *WordNet*, que é uma base de dados lexical para a língua inglesa, que organiza as palavras em conjuntos de sinônimos chamados de *sysnsets*. Esse método calcula a distância entre as palavras no *WordNet* para identificar a polaridade. No *OpLexicon* o método utiliza o TEP (*TeP Thesaurus*), que é um tesauro para a língua portuguesa. Ele contém anotações de sinônimos e antônimos de palavras, que são usadas diretamente para determinar a polaridade. Nele, foi calculada a distância (Equação 13) entre palavras positivas e negativas no TeP para selecionar apenas as palavras cuja a distância para algum termo semente fosse menor ou igual a 3.

$$EVA(x) = \frac{\min(d(x,p)) - \min(d(x,n))}{\min(d(p,n))}, \quad \text{para cada } p \text{ em } \textit{Seed}_{pos} \text{ e } n \text{ em } \textit{Seed}_{neg} \qquad (13)$$

Na Equação 13, x representa a palavra que está sendo analisada no texto para definir a polaridade. $d(x,p)$ mede a distância semântica entre a palavra x e uma palavra positiva p pertencente ao conjunto *Seed*$_{pos}$ (semente de palavras positivas). Essa distância pode ser baseada em uma métrica específica no tesauro. A medida da distância semântica entre a palavra x e uma palavra negativa n pertencente ao conjunto *Seed*$_{neg}$ (semente de palavras negativas) é dado por $d(x,n)$. $d(p,n)$ é a distância entre todas as combinações de palavras dos conjuntos *Seed*$_{pos}$ e *Seed*$_{neg}$.

Por fim, o método baseado em tradução escolhido para o *OpLexicon* se baseia no uso do *Liu's English Opinion Lexicon*, que consiste em um léxico da língua inglesa com 6800 entradas, para gerar um léxico de opinião em português, para isso é usado o *Google Translate*, que realiza a tradução das palavras de forma automática em vez de utilizar um dicionário bilíngue manual. Ao contrário do método imposto por Mihalcea, Banea e Wiebe (2007), que usa o dicionário, gerando um trabalho manual para as traduções.

Por ser um léxico mais abrangente, ele é considerado um léxico não específico de domínio, pois é composto por diferentes classes de palavras (substantivos, verbos e adjetivos), podendo ser utilizado para diversas aplicações, como análise de resenhas de filmes, por exemplo (SOUZA; VIEIRA, 2011).

2.5 Trabalhos Relacionados

A análise de sentimentos tem sido amplamente explorada na literatura, com diferentes abordagens para a extração e classificação da polaridade textual. Métodos baseados em aprendizado de máquina, como SVM e *Naive Bayes*, continuam sendo amplamente utilizados. Estratégias híbridas que combinam aprendizado de máquina com léxicos de opinião também vêm sendo propostas para aprimorar a eficiência dos resultados da análise de sentimentos. Nesta seção, são discutidos estudos que aplicam essas abordagens, destacando seus principais resultados e desafios, bem como sua relevância para este trabalho.

É apresentado por Cardoso *et al.* (2021) uma análise comparativa de diferentes abordagens para a análise de sentimentos, com o foco específico na língua portuguesa. Um conjunto de 2.330 dados de postagens da plataforma X com abordagem em vôlei foram utilizados, onde foram classificados em 1.298 dados positivos e 1.032 negativos. Aprendizado de máquina, métodos lexicais e um modelo de comitê compõem as abordagens analisadas. Dos resultados obtidos, a abordagem de comitê, utilizando um modelo de *Stacking* que combina os algoritmos *Decision Tree*, *Naive Bayes*, SVM e Regressão Logística, apresentou o melhor desempenho, chegando a 82% de acurácia, 81% de *recall* e 82% de *f1-score*. O estudo destaca as dificuldades que existem na análise de sentimentos em textos curtos e coloquiais, tal como a variação de linguagem, ambiguidade e a presença de sarcasmo. É discutido a importância de conter uma base de dados bem rotulada e a necessidade de abordagens que considerem a natureza específica dos dados em português.

O estudo de Souza e Filho (2021) apresenta uma proposta de utilização de diferentes estratégias de incorporação de documentos e algoritmos de classificação para a análise de sentimentos de textos em português brasileiro. O trabalho unifica cinco conjuntos de dados públicos de avaliações de usuários brasileiros referente a filmes, produtos e aplicativos, totalizando mais de 2,2 milhões de amostras. Foram utilizados os algoritmos *Random Forest*, Regressão Logística e *LightGBM* para os testes, com destaque para o *LightGBM*, que alcançou o melhor desempenho nos testes, com *ROC-AUC* variando entre 88,5% e 96,1% dependendo do conjunto de dados. Também é discutido os desafios de generalização dos modelos entre diferentes conjuntos de dados e observa que um modelo treinado com todos os dados combinados apresenta boa generalização, com uma redução de desempenho de apenas 0,1% a 0,5% em relação aos modelos específicos para cada conjunto de dados.

A proposta de Wu *et al.* (2021) consiste em um modelo que combina léxicos de sentimentos com redes neurais para analisar textos de mídias sociais que apresentam características de gramática fraca. É destacado na pesquisa a limitação que existe nas abordagens tradicionais de análise de sentimentos, que muitas vezes dependem de regras linguísticas rígidas, as quais não se aplicam de forma adequada em textos informais e aleatórios encontrados em plataformas como o *Weibo*. O modelo proposto utiliza a pontuação de sentimentos definida para cada palavra no léxico, convertendo o texto em vetores de pontuação de sentimentos, e, em seguida, emprega uma rede neural para aprender representações emocionais ocultas. Os resultados apresentados demonstram que essa abordagem não apenas melhora a precisão geral em dados positivos e negativos, mas também a eficácia da combinação de conhecimento emocional tradicional com técnicas de aprendizado profundo, superando métodos que utilizam apenas redes neurais sem a integração dos léxicos de sentimentos.

É apresentado por Januário *et al.* (2022) um estudo comparativo entre abordagens lexicais e de aprendizado de máquina para análise de sentimentos em notícias financeiras do mercado brasileiro extraídas dos sites Folha de São Paulo, Estadão, InfoMoney, G1 e outros, relatando notícias de empresas como Bradesco, Petrobras, Vale, Magazine Luiza e Gol. A análise

foi realizada a partir de um conjunto de dados com 828 notícias dessas empresas brasileiras, onde foram rotuladas como positivas ou negativas de forma manual, e os algoritmos aplicados foram o *Naive Bayes* e a rede neural *Multilayer Perceptron* (MLP). Os resultados apresentam que as abordagens lexicais se tornam menos eficazes que os métodos de aprendizado de máquina, onde é alcançado uma acurácia de 58% nas abordagens lexicais, 75,36% com o *Naive Bayes* e 78,98% na MLP. Os resultados destacaram a superioridade dos algoritmos de aprendizado de máquina para a análise de sentimentos no contexto financeiro.

A pesquisa de Li *et al.* (2022) retrata a importância da identificação de sentimentos conflitantes em Análises de Sentimentos Baseadas em Aspectos (ABSA). É dissertado a relevância do uso para entender as opiniões do público, em destaque a de consumidores, especialmente em comentários que apresentam emoções opostas. Em vista disso, os autores propõem a implementação de um novo modelo denominado *D-MA-EGCN*, que combina a representação de texto aprimorada pelo modelo *BERT* com uma rede neural de convolução de arestas para identificar as relações de dependência entre palavras. A avaliação do modelo ocorreu com a utilização do conjunto de dados *SemEval*, que contém 3.841 frases de comentários de restaurantes, cada uma etiquetada com categorias de aspectos e polaridades sentimentais correspondentes. As análises mostram que o *D-MA-EGCN* alcançou uma acurácia de 61,5% no reconhecimento de emoções conflitantes, representando um aumento de aproximadamente 20% em relação a abordagens anteriores.

Souza *et al.* (2023), apresentaram uma pesquisa sobre a eficácia de diferentes algoritmos de classificação de sentimentos em textos em português, extraídos das plataformas X e *Steam*. Os autores realizaram uma comparação entre abordagens de aprendizado de máquina, onde na plataforma X o algoritmo em destaque foi o *Stacking* com meta-classificador SVM com um alcance de 81,5% de acurácia, enquanto na plataforma *Steam* o destaque foi o algoritmo *Stacking* com *Random Forest* que obteve 82,8% de resultado. A pesquisa ressalta que a utilização de dados de plataformas digitais se torna importante para compreender as opiniões dos usuários sobre jogos digitais, além de abordar as dificuldades enfrentadas na análise de sentimentos em língua portuguesa, por conta da escassez de recursos nas ferramentas de PLN e a complexidade gramatical que o idioma apresenta.

O artigo de Silva *et al.* (2023) propõe uma abordagem de análise de sentimentos com foco em textos em português brasileiro, onde foram extraídos da plataforma X, com o objetivo de identificar sentimentos positivos e negativos como potenciais indicadores de quadros depressivos. No estudo, são utilizados dois algoritmos de aprendizado de máquina, *Naive Bayes* e SVM, aplicados a um conjunto de dados de 4.797 postagens, sendo 2.019 positivas e 2.778 negativas. Após a realização do pré-processamento e o uso do modelo *Bag of Words* para a extração de características, os resultados das análises indicaram que ambos os algoritmos tiveram desempenhos semelhantes. O SVM obteve uma leve vantagem e comparação com o *Naive Bayes*, com uma precisão de 81% para os textos negativos, enquanto o *Naive Bayes* resultou

em 80%. No geral, a acurácia dos modelos foi de 79%, com diferenças mínimas em termos de *f1-score* e *recall*.

A abordagem de Ramanathan, Hajri e Ruth (2024) para a análise de sentimentos foca na integração de características semânticas para aprimorar a classificação de sentimentos em dados extraídos da plataforma X. O estudo apresenta a complexidade da identificação de sentimentos em postagens que frequentemente carecem de expressões sentimentais explícitas, propondo um método que difere os textos subjetivos e objetivos, seguido por uma categorização de sentimentos positivos e negativos. Foram introduzidos conceitos semânticos, como entidades relacionadas a temas específicos, que ajudam a contextualizar e interpretar as emoções expressas nas mensagens. A metodologia mostrou que a incorporação de características semânticas não apenas melhora a precisão de detecção de sentimentos, mas também oferece uma análise mais aprofundada das opiniões dos usuários, destacando a relevância do conhecimento de fundo na análise de sentimentos.

Muthukrishnan M. (2024) discorrem sobre a crescente demanda por técnicas que otimizem a análise de sentimentos diante do volume exponencial de dados gerados em plataformas de mídia social. Na pesquisa é introduzida uma técnica de seleção de características chamada de *Chi-Vec*, que visa melhorar a eficiência e a precisão dos modelos de aprendizado profundo ao selecionar características relevantes de forma eficaz. É destacada a importância de superar as limitações de escalabilidade dos modelos de aprendizado profundo, trazendo uma abordagem que se adapta a diferentes conjuntos de dados e aplicações, juntamente com a melhoria no desempenho em espaços de características de alta dimensão. Nos resultados, uma taxa de precisão de 97,96%, 98,41% e 94,45% são obtidos, para os conjuntos *CBET*, *ATIS* e *AWARE*, respectivamente.

O artigo de Yogi *et al.* (2024) apresenta um estudo sobre a análise de sentimentos em mídias sociais utilizando um conjunto de dados com 1,6 milhões de postagens de natureza diversa da plataforma X. Cada postagem foi rotulada com a polaridade de sentimento: -1 para negativa, 0 para neutra e +1 para positiva. Foram aplicados algoritmos de aprendizado de máquina, como o SVM, *Naive Bayes*, Redes Neurais e *Random Forest*, para avaliar a precisão e eficácia na classificação de sentimentos. Os resultados chegaram a diferenças significativas entre os modelos. O *Naive Bayes* chegou à uma precisão de 78%, enquanto o SVM alcançou 85%, resultando no algoritmo mais preciso. O estudo também inclui a análise de *f1-scores*, porém com modelos testados em diferentes plataformas, como X, *Facebook* e *Instagram*, apresentando variações de desempenho conforme a rede social.

Os estudos analisados demonstram que abordagens supervisionadas, como SVM e *Naive Bayes*, apresentam bons resultados na análise de sentimentos, especialmente em dados textuais curtos e estruturados. Além disso, pesquisas indicam que a combinação dessas técnicas com léxicos pode aprimorar a extração de sentimentos ao incorporar conhecimento semântico adicional. No entanto, a necessidade de soluções mais eficientes para o idioma português ainda se mostra um desafio.

3 MATERIAIS E MÉTODOS

Neste capítulo são apresentados todos os materiais usados para a realização da pesquisa, incluindo as informações sobre os softwares e base de dados, bem como a metodologia utilizada para o procedimento para que os objetivos propostos sejam alcançados.

3.1 Materiais

Nesta seção são apresentados os materiais que se pretendem utilizar para a realização deste trabalho.

3.1.1 Ambientes de desenvolvimento

Para o desenvolvimento desta pesquisa, será utilizado o *Google Collaboratory*, um serviço do *Jupyter* Notebook hospedado que não requer configuração para uso. Ele oferece um ambiente de treinamento que permite escrever e executar códigos em *Python* diretamente no navegador. Seu acesso é gratuito e inclui recursos de computação, como *GPUs* e *TPUs*. Na área de aprendizado de máquina, é possível importar conjuntos de dados, treiná-los, classificá-los e avaliar a eficácia dos modelos. Os documentos interativos do *Google Collaboratory* são hospedados nos servidores em nuvem do Google, permitindo o uso do hardware da empresa, independentemente da capacidade do computador local. Além disso, o serviço possibilita a hospedagem simultânea de diferentes modelos de aprendizado e o armazenamento em nuvem, podendo também ser vinculado ao Google Drive.

A linguagem de programação escolhida para o desenvolvimento deste trabalho foi o *Python*. Ela é conhecida por ser intuitiva e fácil de aprender, tanto para programadores iniciantes quanto para aqueles experientes em outras linguagens. Desenvolvida sob uma licença de código aberto, é livre para uso e distribuição, inclusive para fins comerciais. A linguagem é amplamente utilizada em diversas áreas, como ciência de dados, computação científica e numérica. Com sua ampla variedade de bibliotecas, tanto padrão quanto externas, é possível desenvolver projetos mais complexos de forma eficiente.

3.1.2 Bibliotecas

Será utilizada a biblioteca Pandas, que possui código aberto em *Python* e é usada para a manipulação e análise de dados. Ela oferece alto desempenho e diversas ferramentas especializadas, como o *DataFrame* e Series, que facilitam o trabalho com dados tabulares e rotulados. Além disso, suporta a leitura de diversos formatos de arquivos, como *CSV* e *Excel*.

Outra biblioteca escolhida para o trabalho foi o *Scikit-learn*. Trata-se de uma biblioteca de aprendizado de máquina, também em *Python*, que fornece ferramentas para mineração e análise de dados. Ela é utilizada para treinar e avaliar modelos de aprendizado supervisionado e não supervisionado, incluindo algoritmos para classificação, regressão, agrupamento, redução de dimensionalidade e validação cruzada.

O *Gensim* é uma biblioteca com foco em modelagem de tópicos e PLN. Ela implementa modelos de *embeddings* como o *Word2Vec*, *Doc2Vec* e *Latent Dirichlet Allocation* (LDA), permitindo a realização de análises semânticas e o aprendizado de representações vetoriais de palavras. Junto com ela, será usado o *NLTK*, que também é uma biblioteca de código aberto em *Python* para PLN. O *NLTK* fornece ferramentas para a análise e manipulação de linguagem natural, incluindo processos como *tokenização*, *stemming*, lematização, análise sintática, extração de informações e remoção de *stopwords*. Ele poderá ser importante para a etapa do pré-processamento dos dados.

O *Matplotlib* será utilizado para produzir gráficos estatísticos, como linhas, barras, dispersão, histogramas, entre outros. Junto com ele, será usada a biblioteca *Seaborn*, que é baseada no próprio *Matplotlib* e simplifica a criação de visualizações estatísticas em *Python*. Ela também possui integração direta com o Pandas.

3.1.3 Léxico de opinião

O *OpLexicon* é um léxico de opinião para a língua portuguesa, que foi descrito com mais detalhes na seção 2.4, que está na sua versão 3.0. Essa versão foi criada de maneira automática e revisada por linguistas em relação à correção de alguns erros de polaridade de alguns adjetivos presentes no *OpLexicon* V2.1 e foi adicionado um novo campo, classificação de polaridade, a qual foi revisada de forma manual ou atribuída automaticamente. O recurso é composto por uma lista de palavras classificadas com sua categoria morfológica e anotadas com a polaridade positiva, negativa ou neutra.

3.1.4 Base de dados

Para que seja possível realizar o treinamento dos modelos de aprendizado de máquina nas tarefas de análise de sentimentos, é necessário utilizar um conjunto de dados textuais considerado grande. A base de dados empregada na pesquisa será uma base desenvolvida por Jorge e Pardo (2023), contendo um corpus com mais de 2 milhões de comentários em português brasileiro, extraídos de 10 mil jogos na plataforma *Steam*[1].

O conjunto inclui tanto os dados puros quanto os dados já filtrados e vetorizados. Os dados puros tiveram seu nome e gênero anotados manualmente. Os dados filtrados contam com

[1] https://store.steampowered.com/about/

mais de 230 mil comentários em português brasileiro, retirados da plataforma *Steam* após um processo de filtragem que considerou apenas aqueles com 3 votos ou mais. Os comentários foram classificados e agrupados em 10 gêneros diferentes, sendo eles *indie*, ação, aventura, *rpg*, estratégia, simulação, terror, corrida, *fps* e esporte. Posteriormente, foram divididos em duas partes: uma metade (part_50) foi destinada ao treinamento dos vetores de documento (*doc2vec*), enquanto a outra (rest_part_50) foi utilizada para o treino e teste do modelo GBM (*Gradient Boost Machine*), uma técnica proposta por Friedman (2001), que se baseia na aproximação de funções, uma extensão do paradigma de descida de gradiente para espaços de funções, em vez de espaços de parâmetros, aplicando incrementos aditivos em cada etapa do treinamento. A Tabela 1 representa como os dados da base estão estruturados após a realização do pré-processamento, transformação, filtragem e separação. Os atributos estão organizados em colunas, como ilustrado na tabela. A primeira coluna, que não contém um nome, indica a identificação de cada dado por um número. A coluna *"Text(dirty)"* representa os dados brutos, sem passar pelo pré-processamento. A coluna *"text"* contém os dados normalizados, colocados em letras minúsculas, excluindo números, pontuações e caracteres especiais. *"Helpful"* é a coluna que representa de forma binária se um dado contém uma pontuação maior que 0.5. As colunas "wv.1" e "wv.2" são os vetores dos dados, como os dados foram transformados em vetores de 1000 dimensões, há uma sequência até o "wv.1000". A *"uppercase.ratio"* é a representação vetorial da proporção de letras maiúsculas nos dados. Também há outras colunas como "n.sentences", "n.words", n.question", entre outras, mas não foram exibidas na tabela.

Tabela 1 – Representação dos dados estruturados.

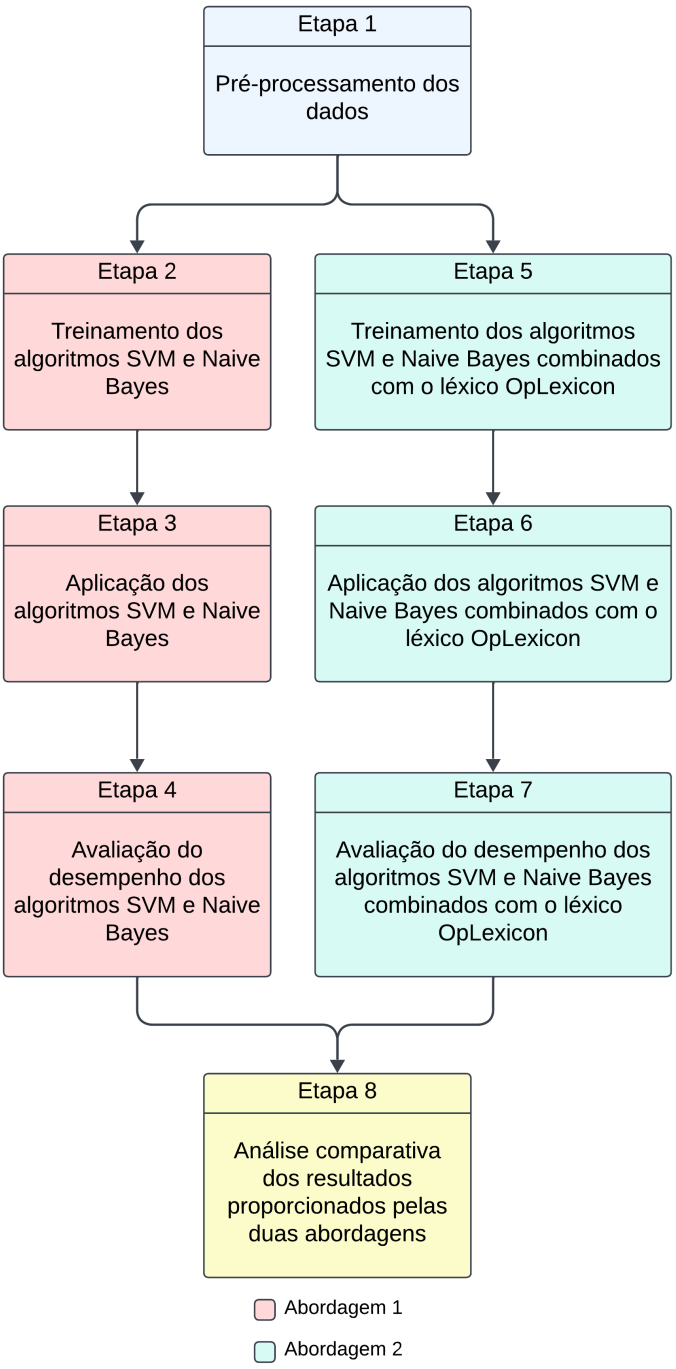
	...	Text(dirty)	Text	Helpful	wv.1	wv.2	...	uppercase.ratio
0	...	Lembro que foi um dos...	lembro primeiros jogos ps joguei...	1	-0.013191	0.107437	...	0.056561
1	...	QUEEE NOSTALGIA JOGAR ESSE...	queee nostalgia jogar jogooooo...	1	-0.000177	0.000166	...	0.846154
2	...	Conheci esse jogo somente agora...	conheci jogo somente agora...	1	-0.001358	0.006560	...	0.048673
3	...	Da pra Esmerilhar o carro...	pra esmerilhar carro completamente...	1	0.000192	0.000072	...	0.030000
4	...	OLD, um dos melhores jogos de...	old melhores jogos corrida demolição...	1	0.000302	-0.000011	...	0.035714
...	...	...	...	...	...	...	...	...
3578	...	Grande jogo	grande jogo	0	-0.000036	-0.000403	...	0.090909
3579	...	que jogo bom!\n\nAguardando...	jogo bom aguardando skin cazé...	1	0.000151	-0.000144	...	0.054545
3580	...	Jogo muito legal pra se divertir...	jogo legal pra divertir sozinho amigos...	1	0.000309	-0.000126	...	0.007246
3581	...	Game god god, muito daora...	game god god daora jogar	0	-0.000041	-0.000192	...	0.029412
3582	...	SAI DAQUI CROCODILO SAFADO...	sai daqui crocodilo safado sa fa	1	-0.000166	-0.000393	...	0.202797

Fonte: Adaptado de Jorge (2022).

3.2 Métodos

Nesta seção são apresentados os métodos que serão aplicados para o desenvolvimento dos objetivos propostos para este trabalho. Na Figura 8 estão representadas, a partir de um fluxograma, as etapas de execução deste trabalho.

Figura 8 – Fluxograma da metodologia do trabalho



Fonte: Autoria própria (2024).

3.2.1 Pré-processamento dos dados

Neste trabalho será utilizada a base de dados de Jorge e Pardo (2023), que já consta com os dados separados em um modelo de vetores de documentos *Doc2Vec* com 1000 dimensões já treinado utilizando metade do conjunto de dados. Os dados estão transformados

em uma representação das sentenças dos comentários através de vetores que permitem que sentenças com significados semelhantes possuam representações semelhantes.

Será realizada uma verificação nos dados já transformados em vetores para averiguar a necessidade de um novo pré-processamento, onde poderá ser feito o processo de *tokenização*, normalização, remoção de *stopwords*, *stemming*, lematização e etiquetagem. Todo esse processo poderá ser feito através da biblioteca *NLTK*. Após esse processo, será necessária a transformação dos dados para *embeddings*, que é planejado ser feita por meio da aplicação do *Gensim*. A escolha da biblioteca *Gensim* foi dada a partir da utilização da mesma na transformação dos dados na base de dados de Jorge e Pardo (2023).

3.2.2 Treinamento dos algoritmos SVM e *Naive Bayes*

Na segunda etapa, os algoritmos SVM e *Naive Bayes* serão treinados utilizando o conjunto de dados pré-processados que será separado para essa finalidade. O treinamento será realizado com os vetores previamente gerados pela técnica *Doc2Vec*, que representam os comentários textuais em um espaço vetorial de 1000 dimensões.

Como o tipo de *kernel* da SVM e o modelo específico de *Naive Bayes* ainda não foram definidos, será explorado um processo inicial de validação cruzada para testar diferentes configurações e identificar os melhores parâmetros para cada algoritmo. A biblioteca *Scikit-learn* se propõe a ser utilizada para implementar os dois algoritmos e conduzir esse processo.

3.2.3 Aplicação dos algoritmos SVM e *Naive Bayes*

Após o treinamento, os modelos SVM e *Naive Bayes* serão aplicados à outra metade do conjunto de dados, que foi reservada para testes. Nesta etapa, os modelos serão utilizados para extrair a polaridade das avaliações textuais, identificando se cada vetor de entrada é classificado como positivo, negativo ou neutro.

3.2.4 Avaliação do desempenho dos algoritmos SVM e *Naive Bayes*

Para avaliar o desempenho inicial dos algoritmos SVM e *Naive Bayes*, serão utilizadas métricas específicas, que incluem a análise dos dados extraídos e comparações qualitativas. Pretende-se utilizar as métricas *F1-Score*, precisão e *recall*, pois elas permitem avaliar se o modelo está identificando corretamente as polaridades, enquanto minimiza os erros. Também será avaliado o uso da acurácia como métrica complementar, caso os dados estejam balanceados, pois podem ocorrer resultados enganosos. Pretende-se aplicar a biblioteca *Scikit-learn* para o cálculo das métricas e a geração de visualizações descritivas.

3.2.5 Treinamento dos algoritmos SVM e *Naive Bayes* combinados com o léxico *OpLexicon*

Nesta etapa, será introduzido o léxico *OpLexicon*, que fornece polaridades para palavras em português. Cada avaliação no conjunto de dados será analisada para verificar a ocorrência de palavras presentes no léxico, e suas polaridades serão integradas aos vetores representativos das sentenças.

Os algoritmos SVM e *Naive Bayes* serão novamente treinados com esses vetores enriquecidos pelas informações do léxico, permitindo que os modelos incorporem nuances adicionais de polaridade durante o aprendizado. Esse processo será realizado utilizando os mesmos dados e parâmetros ajustados na etapa anterior.

3.2.6 Aplicação dos algoritmos SVM e *Naive Bayes* combinados com o léxico *OpLexicon*

Os modelos treinados com a integração do léxico *OpLexicon* serão aplicados ao conjunto de teste. Nesta etapa, será realizada novamente a extração de polaridades para cada comentário, agora com o auxílio das informações adicionais fornecidas pelo léxico.

3.2.7 Avaliação do desempenho dos algoritmos SVM e *Naive Bayes* combinados com o léxico *OpLexicon*

Os resultados obtidos nesta etapa serão avaliados com as mesmas métricas descritas na subseção 3.2.4.

3.2.8 Análise comparativa dos resultados proporcionados pelas duas abordagens

Por fim, será realizada uma análise comparativa entre os resultados das duas abordagens: (1) o treinamento e aplicação dos algoritmos SVM e *Naive Bayes* sem o léxico *OpLexicon*; e (2) o treinamento e aplicação desses algoritmos com a integração do léxico. Além disso, pretende-se gerar gráficos comparativos para visualizar os ganhos ou perdas de desempenho em diferentes aspectos avaliados. Para essa geração, deverá ser utilizada a biblioteca *Matplotlib* combinada com a *Seaborn*.

Essa análise será guiada pelas métricas de desempenho calculadas nas etapas anteriores, destacando vantagens e limitações de cada abordagem. A discussão incluirá os impactos observados no contexto específico da língua portuguesa e da base de dados utilizada, além de recomendações para trabalhos futuros.

REFERÊNCIAS

ANCHIÊTA, R. *et al.* PIn: Das técnicas tradicionais aos modelos de deep learning. **Sociedade Brasileira de Computação**, 2021. Disponível em: https://doi.org/10.5753/sbc.7973.3.1. Acesso em: 9 nov. 2024.

ASTHANA, P.; HAZELA, B. Applications of machine learning in improving learning environment. *In*: ____. **Multimedia Big Data Computing for IoT Applications: Concepts, Paradigms and Solutions**. Singapore: Springer Singapore, 2020. p. 417–433. Disponível em: https://doi.org/10.1007/978-981-13-8759-3_16. Acesso em: 19 nov. 2024.

BHUIYAN, T.; XU, Y.; JOSANG, A. State-of-the-art review on opinion mining from online customers' feedback. *In*: . [*s.n.*], 2009. p. 385–390. Disponível em: https://www.researchgate.net/publication/40635992_State-of-the-Art_Review_on_Opinion_Mining_from_Online_Customers'_Feedback. Acesso em: 14 nov. 2024.

BIRJALI, M.; KASRI, M.; BENI-HSSANE, A. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. **Knowledge-Based Systems**, v. 226, p. 107134, 2021. ISSN 0950-7051. Disponível em: https://www.sciencedirect.com/science/article/pii/S095070512100397X. Acesso em: 06 dez. 2024.

BORGES, B. R. **Análise comparativa de algoritmos de classificação de texto**. Outubro 2024. Trabalho de Conclusão de Curso — Universidade Federal de Uberlândia, Uberlândia, Brasil, Outubro 2024. Acesso Aberto. Disponível em: https://repositorio.ufu.br/handle/123456789/43701. Acesso em: 18 fev. 2025.

BOSS, S. L. B.; VENSKE, S. M. G. S. Analisadores sintáticos: conflitos e ambiguidades. **Revista Científica da FAI**, Santa Rita do Sapucaí, MG, v. 8, n. 1, p. 9–18, 2008. Disponível em: https://www.fai-mg.br/portal/pesquisa/revistas/revista-cientifica. Acesso em: 11 nov. 2024.

BRITO, P. F.; GOMES, L. Desenvolvimento do módulo de pre-processamento da ferramenta sentimentall. **Singular. Engenharia, Tecnologia e Gestão**, v. 1, n. 1, p. 27–35, 2019. Disponível em: https://doi.org/10.33911/singular-etg.v1i1.22. Acesso em: 9 nov. 2024.

CARDOSO, M. H. *et al.* Comparison between different approaches to sentiment analysis in the context of the portuguese language. *In*: **2021 16th Iberian Conference on Information Systems and Technologies (CISTI)**. [*s.n.*], 2021. p. 1–6. Disponível em: https://doi.org/10.23919/CISTI52073.2021.9476501. Acesso em: 14 out. 2024.

CHAUHAN, V. K.; DAHIYA, K.; SHARMA, A. Problem formulations and solvers in linear svm: a review. **Artificial Intelligence Review**, v. 52, n. 2, p. 803–855, 2019. ISSN 1573-7462. Disponível em: https://doi.org/10.1007/s10462-018-9614-6. Acesso em: 22 nov. 2024.

CHEN, Z. *et al.* Emoji-powered representation learning for cross-lingual sentiment classification. *In*: **The World Wide Web Conference**. New York, NY, USA: Association for Computing Machinery, 2019. (WWW '19), p. 251–262. ISBN 9781450366748. Disponível em: https://doi.org/10.1145/3308558.3313600. Acesso em: 14 nov. 2024.

CHURCH, K.; HANKS, P. Word association norms, mutual information, and lexicography. **Computational linguistics**, v. 16, n. 1, p. 22–29, 1990. Disponível em: https://aclanthology.org/J90-1003. Acesso em: 12 nov. 2024.

COUTINHO, J. F. **Avaliação de Técnicas de Combinação de Embeddings para a Análise de Sentimentos de Produtos Escritos em Português-BR**. 2022. Dissertação (Mestrado) — Universidade Federal de Pernambuco, Recife, 2022. Disponível em: https://repositorio.ufpe.br/handle/123456789/48620. Acesso em: 1 nov. 2024.

CRUYS, T. Van de. Two multivariate generalizations of pointwise mutual information. *In*: **Proceedings of the Workshop on Distributional Semantics and Compositionality**. [*s.n.*], 2011. p. 16–20. Disponível em: https://aclanthology.org/W11-1303. Acesso em: 12 nov. 2024.

CUI, J. *et al.* Survey on sentiment analysis: evolution of research methods and topics. **Artificial Intelligence Review**, v. 56, n. 8, p. 8469–8510, aug 2023. ISSN 1573-7462. Disponível em: https://doi.org/10.1007/s10462-022-10386-z. Acesso em: 3 out. 2024.

DARWICH, M. *et al.* Corpus-based techniques for sentiment lexicon generation: A review. **Journal of Digital Information Management**, v. 17, p. 296–305, 10 2019. Disponível em: https://doi.org/10.6025/jdim/2019/17/5/296-305. Acesso em: 5 nov. 2024.

DOMINGOS, P. A few useful things to know about machine learning. **Commun. ACM**, Association for Computing Machinery, New York, NY, USA, v. 55, n. 10, p. 78–87, out. 2012. ISSN 0001-0782. Disponível em: https://doi.org/10.1145/2347736.2347755. Acesso em: 16 nov. 2024.

DURANT, W. **The Pleasures of Philosophy: A Survey of Human Life and Destiny**. Simon and Schuster, 1953. (A Touchstone book). ISBN 9780671581107. Disponível em: https://books.google.com.br/books?id=zQ01AAAAIAAJ. Acesso em: 16 nov. 2024.

FREITAS, C. Dataset e corpus. *In*: **Processamento de linguagem natural: conceitos, técnicas e aplicações em português**. [*s.n.*], 2023. Disponível em: https://brasileiraspln.com/livro-pln/1a-edicao/. Acesso em: 9 nov. 2024.

FREITAS, L. A. d. **Feature-level sentiment analysis applied to brazilian portuguese reviews**. 2015. Tese de Doutorado — Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, Brasil, 2015. Programa de Pós-Graduação em Ciência da Computação. Disponível em: http://hdl.handle.net/10923/7247. Acesso em: 14 nov. 2024.

FREITAS, L. A. D.; VIEIRA, R. Exploring resources for sentiment analysis in portuguese language. *In*: **2015 Brazilian Conference on Intelligent Systems (BRACIS)**. [*s.n.*], 2015. p. 152–156. Disponível em: https://doi.org/10.1109/BRACIS.2015.52. Acesso em: 1 nov. 2024.

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 29, n. 5, p. 1189–1232, 2001. ISSN 00905364, 21688966. Disponível em: http://www.jstor.org/stable/2699986. Acesso em: 07 dez. 2024.

HARTMANN, N. *et al.* **Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks**. 2017. Disponível em: https://arxiv.org/abs/1708.06025. Acesso em: 20 nov. 2024.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**: Second edition. 2. ed. Springer New York, NY, 2009. 465-576 p. (Springer Series in Statistics). ISBN 978-0-387-84857-0. Disponível em: https://doi.org/10.1007/978-0-387-84858-7. Acesso em: 19 nov. 2024.

HU, M.; LIU, B. Mining and summarizing customer reviews. *In*: **Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: Association for Computing Machinery, 2004. (KDD '04), p. 168–177. ISBN

1581138881. Disponível em: https://doi.org/10.1145/1014052.1014073. Acesso em: 8 nov. 2024.

JANUÁRIO, B. A. *et al.* Sentiment analysis applied to news from the brazilian stock market. **IEEE Latin America Transactions**, v. 20, n. 3, p. 512–518, 2022. Disponível em: https://doi.org/10.1109/TLA.2022.9667151. Acesso em: 18 out. 2024.

JARDIM, S.; MORA, C.; SANTANA, T. A multilingual lexicon-based approach for sentiment analysis in social and cultural information system data. *In*: **2021 16th Iberian Conference on Information Systems and Technologies (CISTI)**. [*s.n.*], 2021. p. 1–6. Disponível em: https://doi.org/10.23919/CISTI52073.2021.9476631. Acesso em: 8 nov. 2024.

JESUS, E. L. de; VIEIRA, L. K. **Aplicando PLN para Análise de Sentimentos do Twitter**. 2023 — Trabalho de Conclusão de Curso II, PUC Minas, 2023. Disponível em: https://bib.pucminas.br/acervo/536073. Acesso em: 10 nov. 2024.

JOACHIMS, T. Text categorization with support vector machines: Learning with many relevant features. *In*: NÉDELLEC, C.; ROUVEIROL, C. (Ed.). **Machine Learning: ECML-98**. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998. p. 137–142. ISBN 978-3-540-69781-7. Disponível em: https://doi.org/10.1007/BFb0026683. Acesso em: 17 nov. 2024.

JORGE, G. A. Z. **SteamBR**. 2022. Disponível em: https://github.com/germanojorge/SteamBR. Acesso em: 23 out. 2024.

JORGE, G. A. Z.; PARDO, T. A. S. Steambr: a dataset for game reviews and evaluation of a state-of-the-art method for helpfulness prediction. **Anais**, 2023. Disponível em: https://doi.org/10.5753/brasnam.2023.230132. Acesso em: 23 out. 2024.

KAMPS, J. *et al.* Using WordNet to measure semantic orientations of adjectives. *In*: LINO, M. T. *et al.* (Ed.). **Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)**. Lisbon, Portugal: European Language Resources Association (ELRA), 2004. Disponível em: https://aclanthology.org/L04-1473/. Acesso em: 5 nov. 2024.

KHATUA, A.; KHATUA, A.; CAMBRIA, E. Predicting political sentiments of voters from twitter in multi-party contexts. **Applied Soft Computing**, v. 97, p. 106743, 2020. ISSN 1568-4946. Disponível em: https://www.sciencedirect.com/science/article/pii/S1568494620306815. Acesso em: 13 nov. 2024.

KHYANI, D. *et al.* An interpretation of lemmatization and stemming in natural language processing. **Journal of University of Shanghai for Science and Technology**, v. 22, n. 10, p. 350–357, 2021. Disponível em: https://www.researchgate.net/publication/348306833_An_Interpretation_of_Lemmatization_and_Stemming_in_Natural_Language_Processing. Acesso em: 11 nov. 2024.

LI, P. *et al.* A conflict opinion recognition method based on graph neural network in aspect-based sentiment analysis. *In*: **2022 5th International Conference on Data Science and Information Technology (DSIT)**. [*s.n.*], 2022. p. 1–6. Disponível em: https://doi.org/10.1109/DSIT55514.2022.9943870. Acesso em: 9 out. 2024.

LIU, B. *et al.* Sentiment analysis and subjectivity. **Handbook of natural language processing**, Oxfordshire, v. 2, n. 2010, p. 627–666, 2010. Disponível em: https://www.taylorfrancis.com/chapters/mono/10.1201/9781420085938-36/sentiment-analysis-subjectivity-bing-liu-nitin-indurkhya-fred-damerau. Acesso em: 8 nov. 2024.

MACHADO, M. T. **Tese de Doutorado: Avaliação de Técnicas X na Ciência da Computação**. oct 2023. Tese (Tese de Doutorado) — Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação, São Carlos, Brasil, oct 2023. Data de Defesa: 16 de outubro de 2023. Disponível em: https://doi.org/10.11606/T.55.2023.tde-16012024-151720. Acesso em: 8 nov. 2024.

MAHESH, B. Machine learning algorithms -a review. **International Journal of Science and Research (IJSR)**, v. 9, 01 2019. Disponível em: https://doi.org/10.21275/ART20203995. Acesso em: 16 nov. 2024.

MIHALCEA, R.; BANEA, C.; WIEBE, J. Learning multilingual subjective language via cross-lingual projections. *In*: ZAENEN, A.; BOSCH, A. van den (Ed.). **Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics**. Prague, Czech Republic: Association for Computational Linguistics, 2007. p. 976–983. Disponível em: https://aclanthology.org/P07-1123. Acesso em: 5 nov. 2024.

MOHAMMED, M.; KHAN, M.; BASHIER, E. **Machine Learning: Algorithms and Applications**. [*s.n.*], 2016. ISBN 9781498705387. Disponível em: https://doi.org/10.1201/9781315371658. Acesso em: 16 nov. 2024.

MRABET, M. A. E.; MAKKAOUI, K. E.; FAIZE, A. Supervised machine learning: A survey. *In*: **2021 4th International Conference on Advanced Communication Technologies and Networking (CommNet)**. [*s.n.*], 2021. p. 1–10. Disponível em: https://doi.org/10.1109/CommNet52204.2021.9641998. Acesso em: 17 nov.2024.

MUTHUKRISHNAN M., A. S. . R. R. S. P. A fused feature selection technique for enhanced sentiment analysis using deep learning. **Brazilian Archives of Biology and Technology**, Instituto de Tecnologia do Paraná - Tecpar, v. 67, p. e24240183, 2024. ISSN 1516-8913. Disponível em: https://doi.org/10.1590/1678-4324-2024240183. Acesso em: 9 out. 2024.

NIEVES, T.; MENDONÇA, E. A. d.; FERREIRA, S. L. Processamento de linguagem natural na indústria aec: uma abordagem para tradução de regulamentos edilícios brasileiros para o domínio bim. **SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E COMUNICAÇÃO NA CONSTRUÇÃO**, v. 3, n. 00, p. 1–14, ago. 2021. Disponível em: https://eventos.antac.org.br/index.php/sbtic/article/view/613. Acesso em: 10 nov. 2024.

PACHECO, M. **O que é Steam? Entenda para que serve e como usar a plataforma**. 2023. https://www.terra.com.br/gameon/plataformas-e-consoles/o-que-e-steam-entenda-para-que-serve-e-como-usar-a-plataforma,87822de5f0355917aacb8530fdb4d559a5x9nayi.html. Acesso em: 30 out. 2024.

PANDYA, S.; MEHTA, P. A review on sentiment analysis methodologies, practices and applications. **International Journal of Scientific and Technology Research**, p. 601–609, 10 2020. Disponível em: https://www.ijstr.org/research-paper-publishing.php?month=feb2020. Acesso em: 3 out. 2024.

PANG, B.; LEE, L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *In*: **Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)**. Barcelona, Spain: [*s.n.*], 2004. p. 271–278. Disponível em: https://aclanthology.org/P04-1035. Acesso em: 15 nov. 2024.

PEREIRA, D. A. A survey of sentiment analysis in the portuguese language. **Artificial Intelligence Review**, v. 54, n. 2, p. 1087–1115, 2021. ISSN 1573-7462. Disponível em: https://doi.org/10.1007/s10462-020-09870-1. Acesso em: 7 out. 2024.

RAMANATHAN, V.; HAJRI, H. A.; RUTH, A. Conceptual level semantic sentiment analysis using twitter data. *In*: **2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)**. [*s.n.*], 2024. p. 1–8. Disponível em: https://doi.org/10.1109/ADICS58448.2024.10533498. Acesso em: 10 out. 2024.

RIZKI, A. S.; TJAHYANTO, A.; TRIALIH, R. Comparison of stemming algorithms on indonesian text processing. **TELKOMNIKA (Telecommunication Computing Electronics and Control)**, v. 17, n. 1, p. 95–102, 2019. Disponível em: http://doi.org/10.12928/telkomnika.v17i1.10183. Acesso em: 11 nov. 2024.

SALES, R. d.; CAFÉ, L. Diferenças entre tesauros e ontologias. **Perspectivas em Ciência da Informação**, Escola de Ciência da Informação da UFMG, v. 14, n. 1, p. 99–116, Jan 2009. ISSN 1413-9936. Disponível em: https://doi.org/10.1590/S1413-99362009000100008. Acesso em: 09 dez. 2024.

SANTOS, J. J.; PAIVA, R.; BITTENCOURT, I. I. Avaliaçao léxico-sintática de atividades escritas em algoritmo genético e processamento de linguagem natural: Um experimento no enem. **Revista Brasileira de Informática na Educação**, v. 24, n. 02, p. 92–107, 2016. Disponível em: https://doi.org/10.5753/rbie.2016.24.02.92. Acesso em: 11 nov. 2024.

SARAVANAN, R.; SUJATHA, P. A state of art techniques on machine learning algorithms: A perspective of supervised learning approaches in data classification. *In*: **2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)**. [*s.n.*], 2018. p. 945–949. Disponível em: https://doi.org/10.1109/ICCONS.2018.8663155. Acesso em: 17 nov. 2024.

SARKAR, D. **Text Analytics with Python: A Practitioner's Guide to Natural Language Processing**. 2. ed. Apress, Berkeley, CA, 2019. XXIV, 674 p. Published: 21 May 2019 (eBook), 22 May 2019 (Softcover). ISBN 978-1-4842-4353-4. Disponível em: https://doi.org/10.1007/978-1-4842-4354-1. Acesso em: 20 nov. 2024.

SHAPIRO, A. H.; SUDHOF, M.; WILSON, D. J. Measuring news sentiment. **Journal of Econometrics**, v. 228, n. 2, p. 221–243, 2022. ISSN 0304-4076. Disponível em: https://www.sciencedirect.com/science/article/pii/S0304407620303535. Acesso em: 12 nov. 2024.

SHINDE, P. P.; SHAH, S. A review of machine learning and deep learning applications. *In*: **2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)**. [*s.n.*], 2018. p. 1–6. Disponível em: https://doi.org/10.1109/ICCUBEA.2018.8697857. Acesso em: 16 nov. 2024.

SHOBHA, G.; RANGASWAMY, S. Chapter 8 - machine learning. *In*: GUDIVADA, V. N.; RAO, C. (Ed.). **Computational Analysis and Understanding of Natural Languages: Principles, Methods and Applications**. Elsevier, 2018, (Handbook of Statistics, v. 38). p. 197–228. Disponível em: https://www.sciencedirect.com/science/article/pii/S0169716118300191. Acesso em: 19 nov. 2024.

SILVA, M. S. D. *et al.* Machine learning strategies to analyze positive or negative sentiments in twitter texts. *In*: **2023 18th Iberian Conference on Information Systems and Technologies (CISTI)**. [*s.n.*], 2023. p. 1–5. Disponível em: https://doi.org/10.23919/CISTI58278.2023.10211836. Acesso em: 18 out. 2024.

SIMON, P. **Too Big to Ignore: The Business Case for Big Data**. Wiley, 2013. (Wiley and SAS Business Series). ISBN 9781118641866. Disponível em: https://books.google.com.br/books?id=1ekYIAoEBrEC. Acesso em: 21 nov. 2024.

SOUZA, F. D.; FILHO, J. Baptista de Oliveira e S. Sentiment analysis on brazilian portuguese user reviews. *In*: **2021 IEEE Latin American Conference on Computational Intelligence (LA-CCI)**. [*s.n.*], 2021. p. 1–6. Disponível em: https://doi.org/10.1109/LA-CCI48322.2021.9769838. Acesso em: 18 out. 2024.

SOUZA, G. A. D. *et al.* Comparison between sentiment analysis approaches applied to digital games. *In*: **2023 18th Iberian Conference on Information Systems and Technologies (CISTI)**. [*s.n.*], 2023. p. 1–6. Disponível em: https://doi.org/10.23919/CISTI58278.2023.10211536. Acesso em: 14 out. 2024.

SOUZA, M.; VIEIRA, R. Construction of a portuguese opinion lexicon from multiple resources. **Anais do Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana, 2011, Brasil.**, p. 59–66, 2011. Disponível em: https://repositorio.pucrs.br/dspace/handle/10923/14064. Acesso em: 4 nov. 2024.

SRIVASTAVA, A. K.; PANDEY, D.; AGGARWAL, A. Summarization of medical document using pointwise mutual information (pmi)-based for web document summarization. *In*: **2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)**. [*s.n.*], 2022. p. 1–6. Disponível em: https://doi.org/10.1109/ICRITO56286.2022.9964523. Acesso em: 15 nov. 2024.

TURNEY, P. D. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. *In*: **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**. Morristown, US: Association for Computational Linguistics, 2002. (ACL '02), p. 417–424. Disponível em: https://doi.org/10.3115/1073083.1073153. Acesso em: 5 nov. 2024.

VAPNIK, V. N. **The Nature of Statistical Learning Theory**. 2. ed. Springer, New York, NY, 1999. XX, 314 p. (Information Science and Statistics). Originally published as a monograph. ISBN 978-0-387-98780-4. Disponível em: https://doi.org/10.1007/978-1-4757-3264-1. Acesso em: 22 nov. 2024.

VILELA, C. D.; CUNHA, G. R. **Coleta e armazenamento de dados morfológicos na língua portuguesa**. ago. 2024. Trabalho de Conclusão de Graduação — Universidade Federal do Rio de Janeiro, Brasil, ago. 2024. Disponível em: http://hdl.handle.net/11422/23630. Acesso em: 12 nov. 2024.

VISHNU, K.; APOORVA, T.; GUPTA, D. Learning domain-specific and domain-independent opinion oriented lexicons using multiple domain knowledge. *In*: **2014 Seventh International Conference on Contemporary Computing (IC3)**. [*s.n.*], 2014. p. 318–323. Disponível em: https://doi.org/10.1109/IC3.2014.6897193. Acesso em: 8 nov. 2024.

WESTERSKI, A. Sentiment analysis: Introduction and the state of the art overview. **Universidad Politecnica de Madrid, Spain**, p. 211–218, 2007. Disponível em: https://www.adamwesterski.com/wp-content/files/docsCursos/sentimentA_doc_TLAW.pdf. Acesso em: 14 nov. 2024.

WU, X. *et al.* Sentiment analysis of weak-ruletext based on the combination of sentiment lexicon and neural network. *In*: **2021 IEEE 6th International Conference on Cloud**

Computing and Big Data Analytics (ICCCBDA). [*s.n.*], 2021. p. 205–209. Disponível em: https://doi.org/10.1109/ICCCBDA51879.2021.9442593. Acesso em: 9 out. 2024.

YETURU, K. Chapter 3 - machine learning algorithms, applications, and practices in data science. *In*: Srinivasa Rao, A. S.; RAO, C. (Ed.). **Principles and Methods for Data Science**. Elsevier, 2020, (Handbook of Statistics, v. 43). p. 81–206. Disponível em: https://www.sciencedirect.com/science/article/pii/S0169716120300225. Acesso em: 19 nov. 2024.

YOGI, K. S. *et al.* Enhancing accuracy in social media sentiment analysis through comparative studies using machine learning techniques. *In*: **2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)**. [*s.n.*], 2024. v. 1, p. 1–6. Disponível em: https://doi.org/10.1109/ICKECS61492.2024.10616441. Acesso em: 10 out. 2024.

ZHANG, B.-f.; SU, J.-s.; XU, X. A class-incremental learning method for multi-class support vector machines in text classification. *In*: **2006 International Conference on Machine Learning and Cybernetics**. [*s.n.*], 2006. p. 2581–2585. Disponível em: https://doi.org/10.1109/ICMLC.2006.258853. Acesso em: 17 nov. 2024.

ZHANG, L.; WANG, S.; LIU, B. Deep learning for sentiment analysis: A survey. **WIREs Data Mining and Knowledge Discovery**, v. 8, n. 4, p. e1253, 2018. Disponível em: https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1253. Acesso em: 14 nov. 2024.