COMPARAÇÃO DE MÉTODOS DE PROCESSAMENTO DE LINGUAGEM NATURAL PARA ANÁLISE DE SENTIMENTO APLICADOS A FEEDBACKS DE E-COMMERCE

Crissiano Pires Santana da Silva[1], Marlon de Matos de Oliveira[2]

Resumo: A análise de sentimentos é uma área do Processamento de Linguagem Natural que busca identificar emoções expressas em textos, sendo amplamente aplicada em avaliações de produtos e serviços de *e-commerce*. No contexto do português, essa tarefa ainda apresenta desafios relacionados à ambiguidade linguística e à escassez de recursos computacionais otimizados para o idioma. O objetivo deste trabalho foi realizar uma análise comparativa entre diferentes métodos de Processamento de Linguagem Natural aplicados à análise de sentimentos em comentários de *e-commerce* em português, avaliando métodos léxicos e de aprendizado de máquina. Foram utilizados os léxicos SentiLex-PT, OpLexicon e UniLex, além dos modelos *Support Vector Machine*, *Random Forest* e *Naive Bayes*, em classificações ternária (positivo, negativo e neutro) e binária (positivo e negativo). Os resultados mostraram melhor desempenho dos modelos supervisionados, com destaque para o *Support Vector Machine* com precisão de 70% na classificação ternária e 93% na binária, enquanto o SentiLex-PT foi o léxico mais eficaz com 50% de precisão na ternária e 83% na binária. A exclusão da classe neutra aumentou a precisão e o *F1-Score*, evidenciando que a redução da ambiguidade melhora o desempenho. Conclui-se que a qualidade do corpus e a clareza das categorias influenciam diretamente os resultados.

Palavras-chave: análise de sentimentos; processamento de linguagem natural; aprendizado de máquina; *e-commerce*; língua portuguesa.

[1]Curso de Ciência da Computação, Universidade do Extremo Sul Catarinense (Unesc), crissiano_pires@unesc.net

[2]Orientador. Curso de Ciência da Computação, Universidade do Extremo Sul Catarinense (Unesc), marlon.oliveira@unesc.net

ABSTRACT: Sentiment analysis is a field of Natural Language Processing that aims to identify emotions expressed in text and is widely applied to the evaluation of products and services in e-commerce platforms. In the context of the Portuguese language, this task still faces challenges related to linguistic ambiguity and the scarcity of computational resources optimized for the language. The objective of this study was to perform a comparative analysis of different Natural Language Processing methods applied to sentiment analysis in Portuguese e-commerce reviews, evaluating both lexical and machine learning approaches. The lexicons SentiLex-PT, OpLexicon, and UniLex were used, along with the supervised models Support Vector Machine, Random Forest, and Naive Bayes, in ternary (positive, negative, and neutral) and binary (positive and negative) classifications. The results showed superior performance for the supervised models, with Support Vector Machine achieving 70% accuracy in ternary classification and 93% in binary classification, while SentiLex-PT was the most effective lexicon, with 50% accuracy in the ternary task and 83% in the binary one. The exclusion of the neutral class increased precision and F1-Score, demonstrating that reducing ambiguity improves performance. It is concluded that corpus quality and the clarity of sentiment categories directly influence the results.

Keywords: sentiment analysis; natural language processing; machine learning; e-commerce; Portuguese language.

1 INTRODUÇÃO

A tecnologia avança em ritmo exponencial, resultando na coleta, armazenamento e processamento de grandes volumes de dados. Estimativas indicam que essas informações produzidas no mundo se duplicam a cada vinte meses (Dwivedi; Kasliwal; Soni, 2016). Com isso, a cada dia empresas e organizações sentem a necessidade, em algum momento, de organizar e extrair informações relevantes que possam auxiliar principalmente nas suas tomadas de decisões.

Com essa quantidade de informações, as avaliações sobre uma marca ou serviço, aumentam a visibilidade das empresas nas redes. Por isso, analisar a satisfação do consumidor é essencial para empresas e fabricantes que desejam continuar sendo relevantes no mercado atual (Yang et al., 2020). Nesse sentido, torna-se cada vez mais interessante automatizar o processo de entendimento e o que está sendo falado sobre elas.

Dessa forma, analisar e processar esses dados oferecem vantagens a quem busca obter informações sobre seus produtos e clientes

(Dwivedi; Kasliwal; Soni, 2016). Contudo, extrair informações expressas nessas fontes de dados requer muito esforço quando tratada de forma manual, principalmente devido à quantidade de dados gerados (Alves et al., 2014). Portanto, uma forma de sintetizar essas opiniões de forma automática é por meio do processamento de linguagem humana por máquina, também conhecido por Processamento de Linguagem Natural (Campos; Picchi, 2023).

O Processamento de Linguagem Natural (PLN) é um campo multidisciplinar que integra a Inteligência Artificial e a Mineração de Dados Textuais, desenvolvendo técnicas computacionais para analisar, interpretar e gerar linguagem natural, permitindo que máquinas compreendam e interajam com a comunicação humana de forma automatizada e inteligente (Caseli; Nunes, 2024).

Esse é o campo de estudo em que os pesquisadores se dedicam a entender como os seres humanos produzem, utilizam e compartilham informações por meio da linguagem (Andrade, 2018).

O objetivo do PLN é desenvolver sistemas computacionais capazes de processar e interpretar a linguagem natural de maneira tão eficiente e contextualizada quanto o raciocínio humano. Buscando criar algoritmos e modelos que não apenas reconheçam palavras e estruturas gramaticais, mas também compreendam nuances, intenções e significados implícitos (Andrade, 2018). Dessa forma, é possível realizar muitas tarefas, entre elas, a extração de opiniões e emoções das pessoas em relação a um determinado assunto, utilizando a análise de sentimentos (Titenok, 2022).

A análise de sentimentos tem como objetivo identificar e compreender o sentimento manifestado em um texto (Guide, 2022). Assim, é possível identificar a sua polaridade, que consiste em verificar se o sentimento expresso em um texto é positivo, negativo ou neutro (Cavalcanti et al., 2012). A opinião é o elemento chave na análise de sentimento, que é composta por dois componentes principais: o sentimento e o objeto ao qual ele se direciona. Esse objeto pode ser qualquer entidade, como uma pessoa, uma marca de produto ou qualquer outro assunto relacionado à opinião. O sentimento reflete a emoção ou avaliação expressa em relação a esse objeto (Liu, 2012).

Para identificar e classificar emoções expressas em textos, são utilizados diferentes métodos. Esses métodos são geralmente classificados em duas principais abordagens, aprendizado de máquina e léxicas (Medhat; Hassan; Korashy, 2014).

A abordagem de aprendizado de máquina trata-se de um subcampo da inteligência artificial que estuda algoritmos capazes de aprimorar seu desempenho a partir da experiência (Mullen; Collier, 2004). Simon (1983) também caracteriza o aprendizado de máquina como qualquer modificação em um sistema que resulte na melhoria automática de seu desempenho, seja na repetição de uma mesma tarefa ou na execução de uma tarefa diferente, utilizando a mesma base de dados.

De modo geral, as técnicas de aprendizado de máquina dividem-se em dois tipos: métodos supervisionados e não supervisionados (Tan; Steinbach; Kumar, 2006). Os supervisionados utilizam dados rotulados para treinar o modelo, permitindo que ele aprenda a realizar previsões com base em exemplos conhecidos, empregando técnicas como árvores de decisão, modelos probabilísticos, lineares e baseados em regras (Medhat; Hassan; Korashy, 2014). Já os não supervisionados buscam identificar padrões e agrupamentos em conjuntos de dados sem rótulos, por meio de algoritmos de clusterização e redução de dimensionalidade, que revelam estruturas e relações ocultas nos dados (Yang; Chen, 2017).

A abordagem léxica, também chamada de abordagem de dicionário, identifica emoções em textos a partir da polaridade atribuída às palavras (Medhat; Hassan; Korashy, 2014). Essa técnica pode ser estruturada de duas formas: baseada em dicionário, que associa termos a valores numéricos ou categorias qualitativas, ou baseada em corpus, que define polaridades a partir de padrões linguísticos extraídos de grandes conjuntos de textos, embora exija maior processamento e, por vezes, apresente desempenho inferior (Dang; Zhang; Chen, 2010).

Portanto, esta pesquisa tem como objetivo realizar uma análise comparativa entre técnicas de Processamento de Linguagem Natural aplicadas à análise de sentimentos, considerando duas abordagens distintas: o uso de métodos supervisionados de aprendizado de máquina e a utilização de dicionários léxicos. Utilizando como base de dados feedbacks de *e-commerce*, visando a comparação de performance das técnicas aplicadas.

Os objetivos específicos desta pesquisa consistem em compreender o conceito de Processamento de Linguagem Natural, explorar os métodos aplicados na análise de sentimentos, implementar diferentes técnicas voltadas a esse tipo de análise, avaliar a eficácia dessas abordagens em avaliações de *e-commerce* no contexto do português brasileiro.

Este trabalho está organizado em cinco partes: a Introdução,

que apresenta o tema, os objetivos e a motivação; os Trabalhos Correlatos, que revisam pesquisas anteriores; os Materiais e Métodos, onde são descritos o dataset, o pré-processamento, os léxicos, os classificadores e as métricas; o Resultados e Discussão, que compara o desempenho dos métodos léxicos e supervisionados; e a Conclusão, que resume os achados e sugere melhorias futuras.

2 TRABALHOS CORRELATOS

Diversas pesquisas no campo de PLN têm explorado abordagens baseadas em léxicos para a análise de sentimentos. Os trabalhos a seguir destacam o uso e a eficácia dessas técnicas.

No estudo desenvolvido por Junqueira e Fernandes (2018), os autores fizeram uma comparação empregando técnicas léxicas e de aprendizado de máquina para análise de sentimentos em tweets em português sobre as Olimpíadas de 2016. Para a abordagem baseada em léxicos, foram utilizados os dicionários SentiLex, OpLexicon e LIWC, enquanto nas abordagens de aprendizado de máquina testaram os classificadores Naïve Bayes, Support Vector Machines (SVM), Máxima Entropia, Random Forest e Árvore de Decisão. Entre os métodos avaliados, o algoritmo SVM destacou-se como o mais eficiente com 89,5% de acurácia, assim como o léxico SentiLex superou os demais recursos lexicais com 76%. Os autores utilizaram diversas métricas para comparar as abordagens e os algoritmos, como, precisão, acurácia, *recall*, abrangência e medida de concordância.

O trabalho de Penczkoski e Penteado (2019) investiga a eficácia de ferramentas de PLN, *Natural Language Toolkit* e *spaCy*, combinadas com quatro léxicos em português Sentilex, Unilex, WordNet Affect BR e OpLexicon, para classificar sentimentos em avaliações de hotéis. A metodologia incluiu coleta de dados via *web scraping*, com mais de 5000 avaliações, pré-processamento, limpeza, tokenização e avaliação por métricas como acurácia e F1-score. Os resultados mostraram que a combinação com o léxico OpLexicon e a ferramente NLTK obteve o melhor desempenho 83,8% de acurácia, enquanto o WordNet Affect BR teve o pior 27%, revelando desafios na análise de textos neutros/negativos e limitações na cobertura lexical.

A pesquisa de Araújo, Gonçalves e Benevenuto (2013) compara oito métodos de análise de sentimentos aplicados ao Twitter, avaliando sua eficácia em classificar a polaridade de mensagens. Os autores analisaram dois conjuntos de dados: um extenso corpus de tweets e outro rotulado ma-

nualmente. Os resultados mostram que métodos léxicos como SentiWordNet e SenticNet alcançaram alta cobertura de 90%, enquanto técnicas baseadas em emoticons, embora precisas, tiveram abrangência limitada de 4% a 13%. A ferramenta iFeel, criada para comparar os diferentes métodos em tempo real, complementa o trabalho.

No artigo de Ribeiro Alison e da Silva (2018) é apresentado um estudo comparativo sobre métodos de análise de sentimentos aplicados a tweets, com o objetivo de identificar a abordagem mais eficaz para classificar textos curtos como positivos, negativos ou neutros. Os métodos analisados incluem técnicas baseadas em aprendizado de máquina, dicionários léxicos, *emoticons, part-of-speech, ensembles e word embeddings*. A pesquisa busca responder as questões como qual método possui a melhor performance, qual modelo de *ensemble* é mais eficiente e como os diferentes léxicos se comparam em termos de precisão e contradições. Foram utilizadas duas bases de dados principais: Sanders e HCR (Health Care Reform). Os resultados mostraram que, na base Sanders, o método baseado em dicionário léxico, Opinion Lexicon, alcançou a maior acurácia de 79,09%, seguido por word embeddings com 79,36%. Na base HCR, o Opinion Lexicon também se destacou, com 69,11% de acurácia.

3 MATERIAIS E MÉTODOS

Esta pesquisa caracteriza-se como quantitativa e descritiva, uma vez que busca mensurar e comparar o desempenho de diferentes métodos de PLN aplicados à análise de sentimentos.

A base de dados utilizada foi um dataset público já existente, composto por comentários reais de consumidores em plataformas de *e-commerce* escritos em língua portuguesa.

Foram aplicados métodos léxicos e algoritmos de aprendizado de maquina supervisionado à análise de sentimentos, com o objetivo de avaliar o desempenho de cada abordagem de forma independente e possibilitar uma comparação entre os resultados obtidos. Antes da aplicação dos métodos, foi realizado o pré-processamento dos textos do dataset, etapa fundamental para garantir a padronização dos dados e eliminar ruídos que poderiam comprometer a qualidade da análise.

Após o tratamento e aplicação dos métodos, foi feita uma avaliação comparativa de desempenho. Para isso, foram utilizadas métricas quantitativas. Essas métricas permitiram medir a qualidade das classificações realizadas por cada método em relação à polaridade dos textos

Essa pesquisa foi realizada em um notebook com 16Gb de RAM, 250Gb de HD e sistema operacional Windows. O desenvolvimento prático desse estudo foi realizado em Python, utilizando o Google Colab, um serviço gratuito em nuvem que possibilita a escrita e execução de código diretamente no navegador, sem a necessidade de instalação local de software.

3.1 BASE DE DADOS

A base de dados utilizada nesta pesquisa tem como origem o corpus B2W-Reviews01. Esse corpus reúne avaliações de produtos de clientes de comércio eletrônico, coletadas do site Americanas.com entre janeiro e maio de 2018, estando disponível publicamente em um repositório no GitHub (Real; Oshiro; Mafra, 2019).

A partir desse corpus, Ehrenfried e Todt (2022) desenvolveram três subconjuntos balanceados de dados[3]: *B2W-Polarity-Balanced, B2W-Rating-Balanced* e *B2W-Recommend-Balanced*. Para esta pesquisa, optou-se pela utilização do subconjunto *B2W-Polarity-Balanced*, que organiza previamente os comentários em três classes de sentimento, positiva, negativa e neutra. Os dados são balanceados, composto por 48.945 avaliações, sendo 16.315 em cada classe.

Durante o desenvolvimento do estudo, também foi conduzida uma análise binária, considerando apenas as classes positiva e negativa, com o objetivo de avaliar o impacto da exclusão da classe neutra nos resultados de classificação.

3.2 PRÉ-PROCESSAMENTO

O pré-processamento textual é uma etapa fundamental na análise de sentimentos (Campos; Picchi, 2023). O seu principal objetivo consiste na filtragem e limpeza dos dados, eliminando redundâncias e informações desnecessárias que não contribuem para classificação dos sentimentos (Gonçalves et al., 2006).

Para o pré-processamento nesse trabalho foram utilizadas bibiliotecas disponiveis na linguagem Python: Pandas, para criação e manipulação dos dados, (*Natural Language Toolkit*) NLTK, spaCy para processamento avançado de linguagem natural. A seguir, apresentam-se as técnicas de pré-processamento adotadas.

Durante o pré-processamento, inicialmente realizou-se a tokenização, segmentando cada comentário em unidades linguísticas menores.

[3]Disponível em https://github.com/HenriqueVarellaEhrenfried/B2W-Datasets.git

Em seguida, registros contendo valores ausentes foram removidos, garantindo a consistência do conjunto de dados. A pontuação foi eliminada com o uso do *SpaCy*, reduzindo ruídos textuais. A remoção de *stopwords* foi realizada com o NLTK, mantendo-se a palavra "não" devido ao seu papel determinante na polaridade de sentimento; números também foram descartados por não contribuírem semanticamente. Todo o texto foi convertido para letras minúsculas, evitando distinções indevidas entre termos equivalentes. Por fim, a lematização foi aplicada com o *SpaCy*, reduzindo as palavras às suas formas básicas e promovendo padronização do corpus para melhor representação semântica na etapa de análise.

O comentário original "Atendeu as expectativas e SUPER recomendo!!!" foi pré-processado, resultando nos tokens ['atender', 'expectativa', 'super', 'recomendar'], evidenciando a padronização do texto para análise.

Com a etapa de pré-processamento finalizada, é possível aplicar os textos processados nos métodos para avaliação.

3.3 DICIONÁRIOS

Para a análise com dicionarios foram selecionados três léxicos em português frequentemente usados em trabalhos correlatos: SentiLex-PT[4], Unilex[5] e OpLexicon[6].

O SentiLex-PT é um léxico de sentimentos para a língua portuguesa composto por 7.014 lemas e 82.347 formas flexionadas, abrangendo adjetivos, substantivos, verbos e expressões idiomáticas. Seu foco é a análise de opiniões e sentimentos voltados a entidades humanas, permitindo identificar a polaridade (positiva, negativa ou neutra) associada a diferentes predicados. (Silva; Carvalho; Sarmento, 2012).

O UniLex é um dicionário léxico desenvolvido para análise de sentimentos em textos escritos em português brasileiro. Ele foi criado com o objetivo de classificar palavras segundo sua polaridade emocional. O método propõe uma abordagem baseada na linguagem nativa, evitando perdas de significado causadas por traduções para o inglês, comuns em outros modelos. O UniLex utiliza técnicas como TF-IDF e contagem de frequência de termos para atribuir pesos às palavras e calcular a polaridade geral dos textos (Souza; Pereira; Dalip, 2017).

[4]Disponível em https://b2find.eudat.eu/dataset/b6bd16c2-a8ab-598f-be41-1e7aeecd60d3.
[5]Disponível em https://dicionariounilex.wixsite.com/unilex.
[6]Disponível em https://www.inf.pucrs.br/linatural/wordpress/recursos-e-ferramentas/oplexicon/.

O OpLexicon é outro léxico de opinião para o português construído a partir da combinação de diferentes recursos linguísticos e métodos, com o objetivo de suportar tarefas de mineração de opinião e análise de sentimentos. O léxico combina três abordagens complementares: *corpus-based*, utilizando coocorrências de termos em textos para determinar polaridade; *thesaurus-based*, explorando relações semânticas de sinônimos e antônimos; e *translation-based*, traduzindo lexicons de outros idiomas, especialmente o inglês, para o português. O resultado final é um recurso com 7.077 palavras e expressões polares, cujas polaridades foram definidas a partir de heurísticas para resolver conflitos entre fontes (Souza et al., 2011).

Para todos esses dicionários a polaridade numérica atribuída segue a convenção: 1 para sentimentos positivos, -1 para negativos e 0 para neutros.

Na literatura, existem diversos métodos para extrair o sentimento de um texto com base em dicionários léxicos. Neste trabalho, foi adotado o método adaptado de soma das pontuações dos termos proposto por Hamouda e Rohaim (2011). Nesse método, cada palavra da frase é buscada em um léxico de sentimentos, e o valor de polaridade associado a ela é somado ao longo da análise. O resultado dessa soma indica o sentimento global da frase: um valor positivo representa polaridade positiva, zero neutralidade e um valor negativo representa polaridade negativa, e assim como no método SO-CAL[7] de Taboada et al. (2011), também é considerado o efeito das negações. Quando o algoritmo identifica uma palavra de negação, a polaridade da palavra seguinte é invertida, garantindo uma interpretação mais fiel do contexto.

3.4 CLASSIFICADORES

Para a análise com os classificadores, foram selecionados três algoritmos de aprendizado de máquina supervisionado mais utilizados em trabalhos correlatos: SVM, *Naive Bayes* e *Random Forest*.

O Support Vector Machine (SVM), no contexto da classificação, tem como objetivo separar os dados em diferentes grupos por meio da construção de um hiperplano no espaço de características. Esse hiperplano é definido de modo que a margem entre as classes, isto é, a distância entre o hiperplano e os pontos mais próximos de cada classe, seja maximizada, garantindo uma separação mais precisa e robusta entre os grupos

[7] *Semantic Orientation Calculator*, calcula a polaridade de um texto utilizando dicionários de palavras com suas respectivas orientações semânticas (polaridade e força).

de dados (Vilarinho, 2019).

O algoritmo *Naive Bayes* (NB) utiliza as características dos dados de treinamento para atribuir rótulos de classe a novos exemplos. Ele faz isso calculando probabilidades baseadas nos atributos de cada instância, que são usadas como parâmetros para classificar novos elementos (Freire, 2023). Para esse cálculo, assume-se que os atributos são independentes entre si, o que justifica o termo *naive* (ingênuo). As probabilidades estimadas para cada classe são então combinadas com as probabilidades prévias das classes, obtidas durante o treinamento, resultando na classificação final do elemento (Vilarinho, 2019).

O *Random Forest* é um conjunto de árvores de decisão construídas a partir de amostras aleatórias do conjunto de dados, seguindo o princípio de "dividir para conquistar". A classificação final é definida por votação majoritária entre todas as árvores. Cada nó de uma árvore representa uma característica dos dados e, ao se expandir, gera novos ramos de decisão. Diferentemente de outros algoritmos, ele escolhe a melhor característica apenas entre um subconjunto aleatório de atributos, aumentando a diversidade do modelo (Vilarinho, 2019).

Para que um modelo de aprendizado de máquina consiga compreender e processar a linguagem natural, é necessário realizar a extração de atributos (*features*), variáveis presentes no conjunto de dados que permitem ao algoritmo identificar padrões. Essas *features* determinam quais características do texto serão utilizadas como entrada para o classificador (Malta; Kuroiva, 2020). Como os algoritmos de aprendizado de máquina lidam exclusivamente com valores numéricos, tais características precisam ser transformadas em vetores numéricos, possibilitando sua interpretação matemática (Freire, 2023).

Neste estudo foi aplicada a técnica de vetorização *Term Frequency–Inverse Document Frequency* (TF-IDF), que combina a frequência do termo no documento (TF) com a frequência inversa do termo nos documentos (IDF) (Malta; Kuroiva, 2020). Essa estratégia atribui pesos aos termos, de modo a reduzir a importância de palavras que aparecem com muita frequência e, portanto, têm menor relevância para a análise (Evangelista, 2022). Configurado com bigrams (1,2), um limite máximo de 5.000 atributos e remoção automática de termos pouco informativos.

A performance de um classificador é avaliada pelo seu desempenho na classificação do conjunto de teste. Como essa amostra é menor que a de treinamento, existe a possibilidade de não representar adequa-

damente toda a base de dados, o que pode comprometer o desempenho do algoritmo. A fim de evitar esse problema, foi utilizada a estratégia de validação cruzada *K-Fold* (Vilarinho, 2019).

Esse tipo de treinamento divide a base de dados em k partes, chamadas *folds*. Em cada etapa, uma parte é usada para testar o modelo, enquanto as demais servem para o treinamento. O processo se repete até que todo modelo seja treinado e testado com todas as partes. Por fim, todas as métricas de desempenho são calculadas para cada *fold* e as médias dos resultados são usadas para avaliação.

Inicialmente, foram utilizados 10 *folds* neste estudo. No entanto, optou-se por reduzir para 5 *folds*, uma vez que o tempo de processamento, devido ao grande volume de dados do *dataset*, mostrou-se muito elevado.

3.5 MÉTRICAS PARA AVALIAÇÃO DE DESEMPENHO

Um fator essencial na avaliação dos métodos aplicados à análise de sentimentos está relacionado às métricas adotadas para avaliar seu desempenho (Gonçalves et al., 2013). Essas métricas são adotados na literatura para verificar a eficácia de técnicas e algoritmos na identificação da polaridade dos sentimentos expressos em textos (Sokolova; Lapalme, 2009).

Neste estudo, foram utilizadas cinco métricas[8] para avaliar o desempenho dos métodos léxicos e dos algoritmos de aprendizado de máquina: matriz de confusão, precisão, abrangência, F1-*score* e acurácia.

A matriz de confusão é composta por quatro conceitos principais: **Verdadeiros Positivos (VP)** correspondem às instâncias positivas corretamente classificadas, **Verdadeiros Negativos (VN)** referem-se às instâncias negativas corretamente classificadas, **Falsos Positivos (FP)** representam instâncias negativas classificadas como positivas, **Falsos Negativos (FN)** dizem respeito a instâncias positivas classificadas como negativas (Sousa, 2016).

Em problemas de classificação multiclasse, como no caso da análise de sentimentos com adição do rótulo neutro, a matriz de confusão se expande para uma estrutura 3×3. Adicionando novos conceitos: **Verdadeiros neutros (VE)** indicam as instâncias neutras corretamente classificadas e **Falsos neutros (FE)** correspondem a instâncias positivas ou negativas incorretamente atribuídas à neutras (Penczkoski; Penteado, 2019), como pode ser visto na Tabela 1 .

[8] https://scikit-learn.org/stable/api/sklearn.metrics.html

Tabela 1 - Exemplo de matriz de confusão 3×3

	Verdadeiro Positivo (P)	**Verdadeiro Negativo (N)**	**Verdadeiro Neutro (E)**
Predito Positivo (P)	VP	FP	FP
Predito Negativo (N)	FN	VN	FN
Predito Neutro (E)	FE	FE	VE

Fonte: Adaptado de Penczkoski e Penteado (2019).

Esses valores são utilizados para calcular indicadores como precisão, abrangência, F1 e acurácia.

A precisão, ou *precision*, é uma métrica que indica a proporção de previsões positivas que realmente pertencem à classe prevista. Para o calculo é dividido a quantidade de exemplos que foram corretamente classificados como pertencentes a uma classe pelo total de exemplos que o método classificou como pertencentes a essa classe, incluindo acertos e erros (Sokolova; Lapalme, 2009).

A abrangência, também conhecida como revocação (*recall* em inglês), quantifica a capacidade de um método em identificar corretamente os casos de uma determinada classe, sendo calculada pela razão entre o número de casos corretamente classificados e o total de ocorrências que realmente pertencem a ela (Sokolova; Lapalme, 2009).

O F1, ou *F1-score*, é uma métrica que combina a precisão e a revocação em um único valor, usando a média harmônica entre elas (Araújo; Gonçalves; Benevenuto, 2013).

A acurácia, ou *accuracy*, representa a proporção total de classificações corretas realizadas pelo método, considerando todas as classes, em relação ao número total de previsões analisadas (Sokolova; Lapalme, 2009).

4 RESULTADOS E DISCUSSÃO

Durante as análises, observou-se que o desempenho dos métodos apresentou valores mais baixos com a classe neutra incluída. Isso levantou a hipótese de que a base de dados utilizada não apresentava comentários realmente neutros em quantidade significativa, o que possivelmente influenciou negativamente os resultados nas análises que incluíram essa classe.

Diante disso, optou-se por realizar duas abordagens distintas. A primeira considerou as três classes (ternário): positiva, negativa e neutra e a segunda abordagem restringiu-se a duas classes (binário): positiva e negativa.

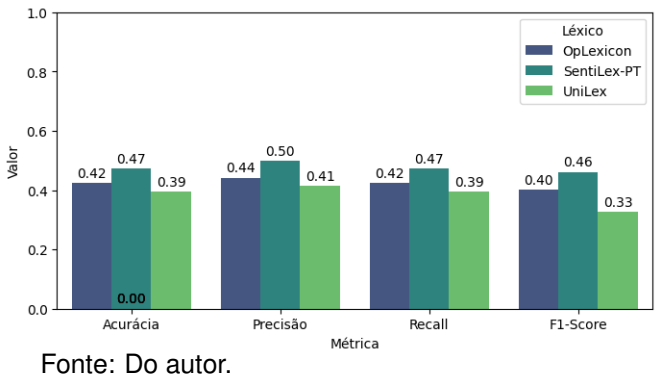
4.1 DESEMPENHO DOS MÉTODOS LÉXICOS

A seguir, são apresentados os resultados das análises dos métodos léxicos considerando os cenários de classificação ternária e binária de sentimentos.

4.1.1 Análise léxica com sentimentos de classificação ternária.

Os resultados da Figura 1 mostram que o SentiLex-PT apresentou o melhor desempenho geral entre os léxicos avaliados, alcançando acurácia de 47,19%, precisão de 49,75%, *recall* de 47,19% e F1-Score de 46,15%. Esses valores indicam que o léxico foi o mais eficiente na identificação correta das classes de sentimento, mantendo um bom equilíbrio entre precisão e abrangência, o que reflete sua maior adequação ao domínio linguístico analisado.

Figura 1 - Métricas dos léxicos com classe ternária.

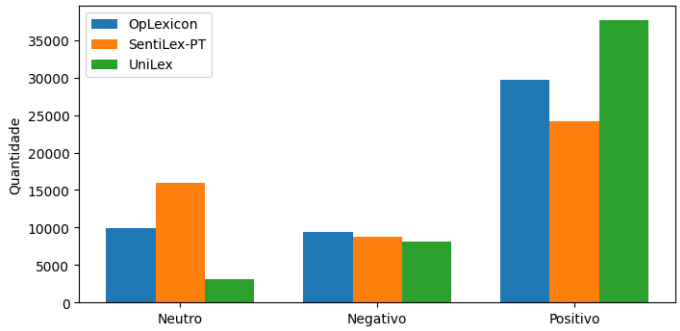


Fonte: Do autor.

O OpLexicon, por sua vez, obteve desempenho intermediário, com acurácia de 42,42% e F1-Score de 40,02%, demonstrando consistência na classificação, embora inferior ao SentiLex-PT. Já o UniLex apresentou as menores métricas, com acurácia de 39,43% e F1-Score de 32,69%, sugerindo dificuldades em capturar corretamente as polaridades dos textos, possivelmente devido a limitações em sua cobertura lexical e representação dos termos afetivos.

A distribuição das quantidades conforme a Figura 2 de classificações por polaridade evidencia diferenças marcantes entre os léxicos utilizados. O UniLex apresentou forte viés para a classe positiva, com 37.673 textos classificados como positivos, contra apenas 3.118 neutros e 8.154 negativos, indicando baixa cobertura de termos neutros e negativos.

Figura 2 - Distribuição de sentimentos por léxico.

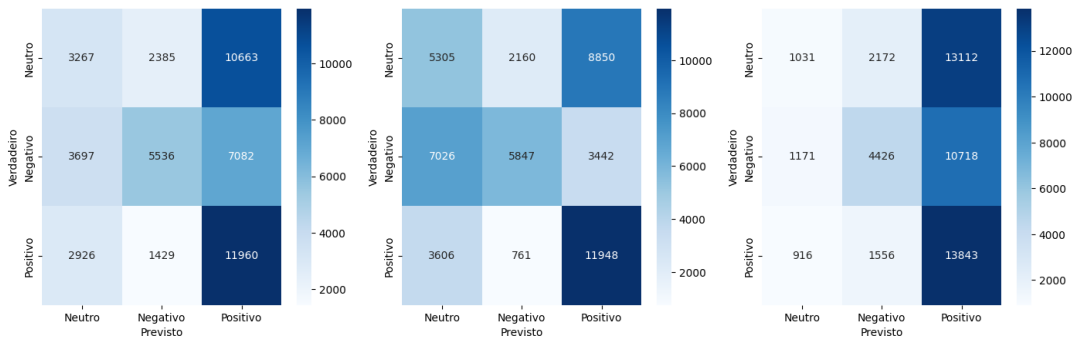


Fonte: Do autor.

O OpLexicon também mostrou predominância de positivos (29.705 textos), mas com distribuição mais equilibrada em relação às outras classes (9.890 neutros e 9.350 negativos), sugerindo maior diversidade léxica. Já o SentiLex-PT apresentou a distribuição mais balanceada, com 24.240 positivos, 15.937 neutros e 8.768 negativos, demonstrando melhor capacidade de diferenciar sentimentos no corpus.

A análise das matrizes de confusão, conforme a Figura 3, evidencia diferenças claras entre os léxicos avaliados. O OpLexicon mostrou forte viés positivo, com 10.663 neutros e 7.082 negativos incorretamente classificados como positivos, obtendo 11.960 acertos positivos, mas precisão moderada devido ao alto número de falsos positivos. O SentiLex-PT apresentou o desempenho mais equilibrado, com 11.948 positivos e 5.847 negativos corretamente classificados, embora ainda tenha registrado 7.026 falsos positivos e 3.442 falsos negativos, mantendo boa separação entre as polaridades. Já o UniLex demonstrou forte viés para a classe positiva, com 13.843 acertos, porém 10.718 negativos e 13.112 neutros incorretamente atribuídos a essa classe. De modo geral, a classe positiva foi a mais facilmente identificada, enquanto a neutra apresentou maior confusão.

Figura 3 - Matrizes de confusão dos léxicos terciário.



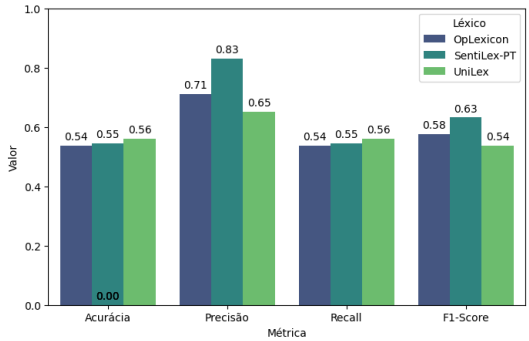
(a) OpLexicon (b) SentiLex-PT (c) UniLex

Fonte: Do autor.

4.1.2 Análise léxica com sentimentos de classificação binária.

Conforme Figura 4, o SentiLex-PT apresentou o melhor desempenho entre os léxicos avaliados, com precisão de 83,05% e F1-Score de 63,19%, demonstrando maior capacidade de identificar corretamente textos positivos e negativos. O OpLexicon obteve desempenho intermediário, com precisão de 71,14% e F1-Score de 57,60%, enquanto o UniLex apresentou os valores mais baixos, com precisão de 65,17% e F1-Score de 53,71%, evidenciando menor consistência nas classificações.

Figura 4 - Métricas dos léxicos com sentimentos positivo e negativo.

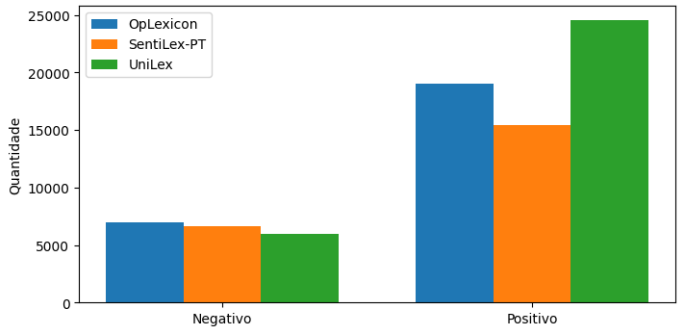


Fonte: Do autor.

De modo geral, a remoção da classe neutra resultou em maior precisão para todos os léxicos, pois eliminou a ambiguidade de casos menos polarizados. No entanto, a acurácia e o *recall* permaneceram em níveis moderados, entre 53% e 56%, o que indica que ainda há dificuldades na separação clara entre sentimentos positivos e negativos, mesmo em cenários binários.

A análise da distribuição, apresentada na Figura 5, evidencia diferenças marcantes entre os três léxicos. O OpLexicon classificou 19.042 textos como positivos e 6.965 como negativos, indicando uma tendência à polaridade positiva. O SentiLex-PT, por sua vez, apresentou uma distribuição mais equilibrada, com 15.390 positivos e 10.632 negativos, demonstrando melhor distinção entre as classes e justificando seu maior F1-Score. Já o UniLex mostrou forte viés positivo, com 24.561 positivos e 5.982 negativos, o que explica seu desempenho inferior em precisão e F1-Score.

Figura 5 - Distribuição de sentimentos por léxico.

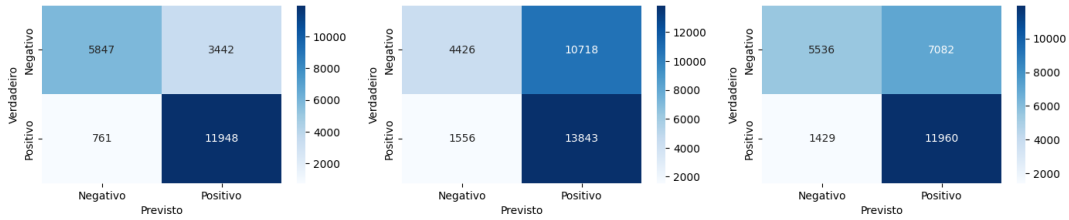


Fonte: Do autor.

No geral, distribuições mais balanceadas resultam em melhores métricas de classificação, enquanto viés forte em uma única polaridade, como no UniLex, reduz a qualidade da predição.

Na análise binária, conforme a Figura 6 observou-se melhor equilíbrio entre as polaridades. O SentiLex-PT apresentou o melhor desempenho, com 11.948 acertos positivos e 5.847 negativos, além de apenas 3.442 falsos positivos e 761 falsos negativos, evidenciando alta precisão e F1-Score. O OpLexicon obteve 11.960 acertos positivos e 5.536 negativos, porém apresentou 7.082 falsos positivos e 1.429 falsos negativos, indicando tendência a superestimar sentimentos positivos. Já o UniLex registrou 13.843 acertos positivos, mas 10.718 falsos positivos, revelando forte viés positivo e menor sensibilidade para sentimentos negativos. No geral, o SentiLex-PT mostrou-se mais equilibrado e consistente, seguido pelo OpLexicon, enquanto o UniLex apresentou maior desequilíbrio entre as classes.

Figura 6 - Matrizes de confusão dos léxicos binário.



(a) SentiLex-PT (b) UniLex (c) OpLexicon

Fonte: Do autor.

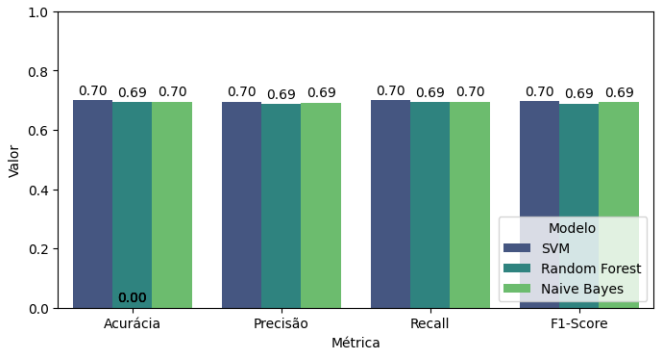
4.2 DESEMPENHO DOS MÉTODOS DE APRENDIZADO DE MÁQUINA

A seguir, são apresentados os resultados das análises de aprendizado de máquina também considerando os cenários de classificação ternária e binária de sentimentos.

4.2.1 Análise de aprendizado de máquina com sentimentos de classificação ternária.

Os resultados, conforme a Figura 7, indicam que o SVM apresentou melhor acurácia de 70% e maior F1-Score também com 70%. Esse comportamento é coerente com a literatura, já que o SVM costuma se destacar em tarefas de classificação de texto com alta dimensionalidade, como em Junqueira e Fernandes (2018) onde algoritmo SVM destacou-se como o mais eficiente com 89,5% de acurácia.

Figura 7 Comparação de métricas por modelo.

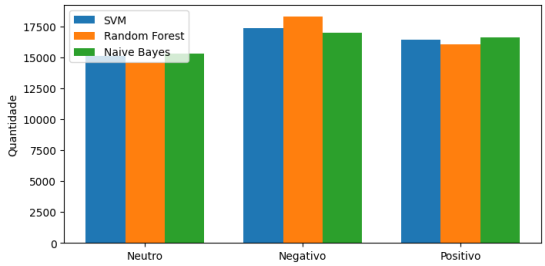


Fonte: Do autor.

O Random Forest apresentou desempenho muito próximo, confirmando sua robustez e estabilidade mesmo sem ajustes complexos de hiperparâmetros. O Naive Bayes, embora mais simples, também mostrou resultados competitivos, com valores de F1-Score semelhantes aos demais modelos. De forma geral, as diferenças entre os três algoritmos foram pequenas, inferiores a 1%, o que evidencia que todos conseguiram capturar padrões relevantes nos dados.

A análise da distribuição de sentimentos na Figura 8 revelou que os três modelos apresentaram maior concentração de predições na classe negativa, seguida pela positiva e, por último, pela neutra. Essa tendência foi mais evidente no Random Forest, que classificou 18.324 textos como negativos, indicando um leve viés para essa polaridade. O SVM apresentou a distribuição mais equilibrada entre as classes (15.097 neutros, 17.463 negativos e 16.385 positivos), o que se refletiu em seu melhor desempenho global, com maior F1-Score. Já o Naive Bayes manteve resultados próximos aos do SVM, com leve vantagem na identificação de textos positivos.

Figura 8 - Distribuição dos sentimentos por modelo.

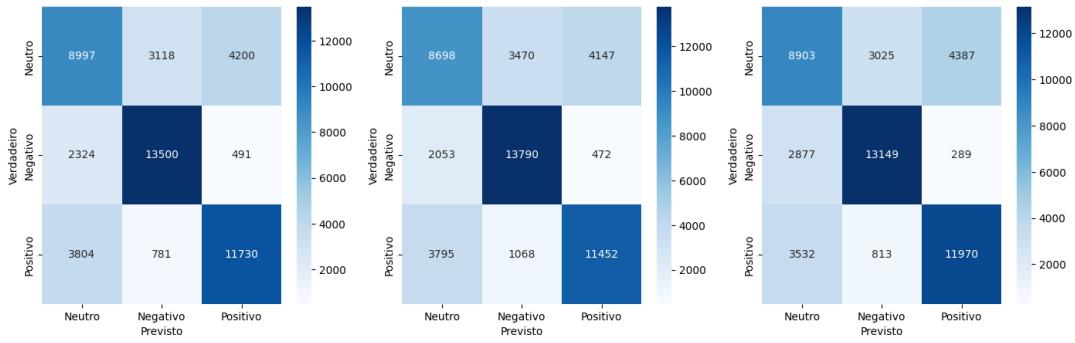


Fonte: Do autor.

De forma geral, observou-se que sentimentos negativos e positivos são mais facilmente reconhecidos pelos modelos, enquanto a classe neutra apresentou maior taxa de erro, o que é esperado devido à natureza ambígua desse tipo de texto.

Analisando as matrizes de confusão na Figura 9 , o modelo SVM apresentou 11.730 acertos na classe positiva, 13.500 na negativa e 8.997 na neutra, com 4.200 positivos classificados como neutros e 3.118 neutros como negativos, demonstrando bom equilíbrio geral. O Random Forest teve 11.452 acertos positivos, 13.790 negativos e 8.698 neutros, mas com 4.147 positivos confundidos como neutros e 3.470 neutros como negativos, indicando viés para a neutralidade. Já o Naive Bayes obteve 11.970 acertos positivos, 13.149 negativos e 8.903 neutros, com 2.877 negativos incorretamente classificados como neutros. De forma geral, os três modelos identificaram melhor os sentimentos positivos e negativos, sendo o SVM e o Naive Bayes mais equilibrados, enquanto o Random Forest mostrou tendência à classificação neutra.

Figura 9 - Matrizes de confusão dos modelos de aprendizado de máquina.



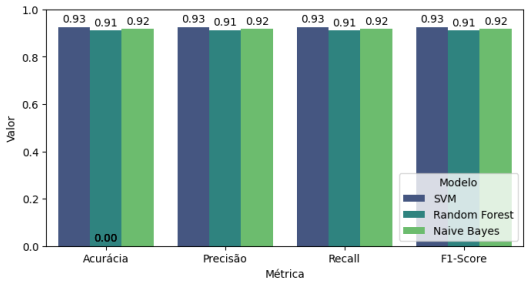
(a) SVM (b) Random Forest (c) Naive Bayes

Fonte: Do autor.

4.2.2 Análise de aprendizado de máquina com sentimentos de classificação binária.

De forma geral, a remoção da classe neutra resultou em um aumento expressivo no desempenho dos três modelos, como pode ser visto na Figura 10, refletido no crescimento das métricas de acurácia e F1-Score. Os três classificadores demonstraram boa capacidade de generalização no cenário binário.

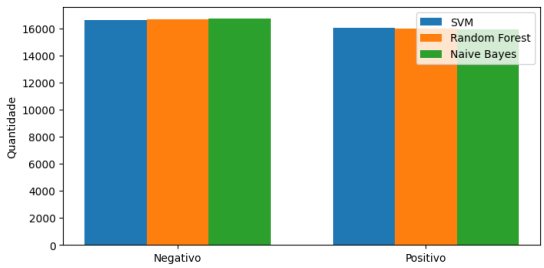
Figura 10 - Comparação de métricas por modelo.



Fonte: Do autor.

A distribuição de sentimentos revela um equilíbrio consistente entre as classes negativa e positiva nos três modelos avaliados, com ligeira predominância da classe negativa como pode ser visto na Figura 11. O SVM apresentou o melhor desempenho geral, com 16.604 negativos e 16.026 positivos, resultando em acurácia de 92,59%. O Random Forest e o Naive Bayes mantiveram resultados próximos. O número de acertos foi alto em todas as classes, indicando que a remoção da classe neutra reduziu ambiguidades e tornou a classificação mais precisa e equilibrada.

Figura 11 - Exemplo de distribuição de sentimentos.

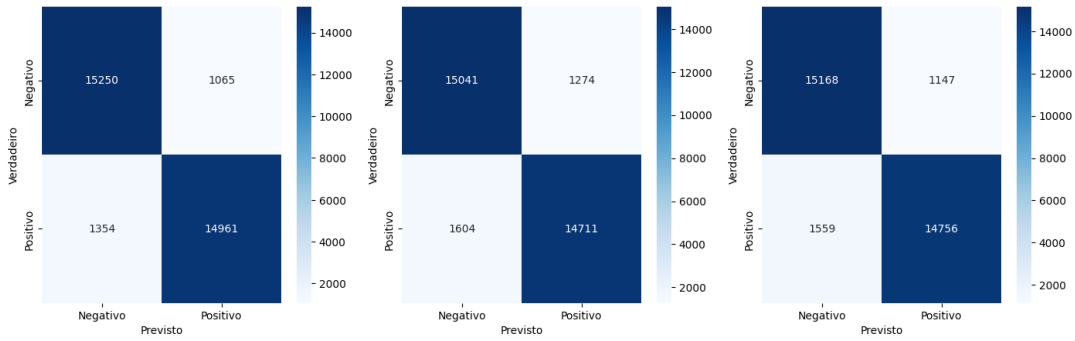


Fonte: Do autor.

A análise das matrizes de confusão mostra que todos os modelos apresentaram alto número de acertos nas duas classes, com pequenas variações entre eles. O SVM, apresentado na Figura 12a, obteve bom desempenho, com 15.250 verdadeiros negativos e 14.961 verdadeiros positivos, e poucos erros (1.065 falsos positivos e 1.354 falsos negativos), refletindo sua acurácia. O Random Forest, na Figura 12b, registrou 15.041

acertos negativos e 14.711 positivos, com um leve aumento nos erros, especialmente nos falsos negativos (1.604). Já o Naive Bayes, Figura 12c, apresentou 15.168 acertos negativos e 14.756 positivos, com desempenho intermediário. De forma geral, as três matrizes indicam equilíbrio entre as classes.

Figura 12 - Matrizes de confusão modelos de aprendizado de máquina.



(a) SVM (b) Random Forest (c) Naive Bayes

Fonte: Do autor.

5 CONCLUSÃO

Esse trabalho teve como objetivo realizar uma análise comparativa das técnicas de PLN aplicadas à análise de sentimentos em avaliações de e-commerce em português, com foco na comparação entre métodos léxicos e algoritmos de aprendizado de máquina. De modo geral, os resultados alcançados demonstraram que ambos os grupos de métodos apresentam desempenhos consistentes, sendo que os modelos supervisionados superaram os léxicos em termos de acurácia e equilíbrio entre as classes de sentimento. Entre os léxicos, o SentiLex-PT destacou-se por seu desempenho mais estável, enquanto o SVM se sobressaiu entre os modelos supervisionados.

No cenário de classificação ternária, observou-se maior dificuldade dos métodos em distinguir corretamente a classe neutra, resultando em taxas de erro mais elevadas. Já na classificação binária, ao eliminar a neutralidade e focar nas polaridades opostas, houve um aumento expressivo na precisão e no F1-Score, evidenciando que a remoção de ambiguidades melhora o desempenho geral. Esses resultados reforçam que a eficácia das abordagens depende diretamente da clareza das categorias de sentimento e da qualidade da base de dados utilizada.

Como trabalhos futuros, sugere-se a aplicação de modelos baseados em aprendizado profundo. Além disso, recomenda-se ampliar o conjunto de dados, e explorar métodos híbridos, combinando abordagens

léxicas e supervisionadas para aprimorar a precisão e a robustez da análise de sentimentos em língua portuguesa.

REFERÊNCIAS

ALVES, A. L. F. et al. **A comparison of svm versus naive-bayes techniques for sentiment analysis in tweets: A case study with the 2013 fifa confederations cup**. Association for Computing Machinery, New York, NY, USA, p. 123–130. Acesso em 10 mar. 2025. Disponível em: <https://doi.org/10.1145/2664551.2664561>.

ANDRADE, F. W. G. **Mineração de texto: previsão de tendências no mercado de ações por meio de notícias publicadas na internet**. Dissertação (Dissertação (Mestrado em Administração)) — Fundação Mineira de Educação e Cultura – FUMEC, 2018.

ARAúJO, M.; GONçALVES, P.; BENEVENUTO, F. **Métodos para Análise de Sentimentos no Twitter**. [S.l.]: UFMG, 2013.

CAMPOS, E.; PICCHI, M. de L. P. **Do processamento de linguagem natural à análise de sentimento**. Interface Tecnológica, Faculdade de Tecnologia de Taquaritinga (Fatec), v. 20, n. 1, p. 170–179. Acesso em: 22 de maio 2025. Disponível em: <https://doi.org/10.31510/infa.v20i1.1667>.

CASELI, H. M.; NUNES, M. G. V. (Ed.). **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. 2. ed. BPLN, 2024. Acesso em: 07 de abr. 2025. ISBN 978-65-00-95750-1. Disponível em: <https://brasileiraspln.com/livro-pln/2a-edicao/>.

CAVALCANTI, D. C. et al. **Análise de Sentimento em Citações Científicas para Definição de Fatores de Impacto Positivo**. [S.l.: s.n.], 2012. 1–10 p.

DANG, Y.; ZHANG, Y.; CHEN, H. **A Lexicon-Enhanced Method for Sentiment Classification: An Experiment on Online Product Reviews**. [S.l.]: IEEE Intelligent Systems, 2010. v. 25. 46-53 p.

DWIVEDI, S.; KASLIWAL, P.; SONI, S. **Comprehensive study of data analytics tools (RapidMiner, Weka, R tool, Knime)**. [S.l.]: 2016 Symposium on Colossal Data Analysis and Networking (CDAN), 2016. 1-8 p.

EHRENFRIED, H. V.; TODT, E. Should i buy or should i pass: E-commerce datasets in portuguese. In: PINHEIRO, V. et al. (Ed.). **Computational Processing of the Portuguese Language**. Cham: Springer International Publishing, 2022. p. 420–425. ISBN 978-3-030-98305-5.

EVANGELISTA, J. L. **Aprendizado de máquina aplicado à análise de sentimento em avaliações de usuários de aplicativos do google play**. Biblioteca Digital: Trabalhos de Especialização, Ciências Exatas e da Terra,

Inteligência Artificial Aplicada. Acesso em: 8 out. 2025. Disponível em: <https://acervodigital.ufpr.br/xmlui/handle/1884/82085>.

FREIRE, F. E. d. S.

Aprendizado de máquina para classificação de sentimentos: uma análise comparativa — Universidade Estadual de Campinas, Campinas, SP, 2023. Orientação de Guilherme Palermo Coelho. 1 recurso on-line (41 p.), il., digital, arquivo PDF. Disponível em: <https://hdl.handle.net/20.500.12733/17674>. Acesso em: 7 out. 2025.

GONÇALVES, T. et al. Analysing part-of-speech for portuguese text classification. In: **Computational Linguistics and Intelligent Text Processing**. Berlin, Heidelberg: Springer, 2006. p. 551–562.

GONçALVES, P. et al. **Comparing and combining sentiment analysis methods**. Association for Computing Machinery, New York, NY, USA, p. 27–38. Acesso em 10 maio 2025. Disponível em: <https://doi.org/10.1145/2512938.2512951>.

GUIDE, B. F. **Detecção automática de discurso de ódio punitivista em redes sociais**. Tese (Tese de Doutorado) — Universidade de São Paulo, Faculdade de Filosofia, Letras e Ciências Humanas, São Paulo, 2022. Acesso em 13 mar. 2025. Disponível em: <https://doi.org/10.31510/infa.v20i1.1667>.

HAMOUDA, A.; ROHAIM, M. Reviews classification using sentiwordnet lexicon. In: **World Congress on Computer Science and Information Technology**. [S.l.]: s.n., 2011.

JUNQUEIRA, K. T. C.; FERNANDES, A. M. d. R. **Análise de sentimento em redes sociais no idioma português com base em mensagens do twitter**. [S.l.]: Anais do Computer on the Beach, 2018. 681–690 p. 8, 9, 10, 21.

LIU, B. **Sentiment analysis and opinion mining**. [S.l.]: Morgan & Claypool Publishers, 2012. v. 5. 1–167 p.

MALTA, H. A. L.; KUROIVA, M. A. R. L. **Aprendizado de Máquina e Processamento de Linguagem Natural aplicados à identificação de discurso de ódio**. 2020. [S.l.: s.n.].

MEDHAT, W.; HASSAN, A.; KORASHY, H. **Sentiment analysis algorithms and applications: A survey**. Ain Shams Engineering Journal, v. 5, n. 4, p. 1093–1113. ISSN 2090-4479. Acesso em 20 abr. 2025. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2090447914000550>.

MULLEN, T.; COLLIER, N. **Sentiment Analysis using Support Vector Machines with Diverse Information Sources**. Barcelona, Spain: Association for Computational Linguistics, 2004. 412–418 p. Acesso em 15 out. 2025. Disponível em: <https://aclanthology.org/W04-3253/>.

PENCZKOSKI, R.; PENTEADO, R. J. **Comparação de Ferramentas de Processamento de Linguagem Natural para Análise de Sentimento em Português: Um Estudo de Caso em Avaliações Online de Hotéis**. Trabalho de Conclusão de Curso — Universidade Tecnológica Federal do Paraná, Ponta Grossa, 2019.

REAL, L.; OSHIRO, M.; MAFRA, A. B2w-reviews01: An open product reviews corpus. In: **Proceedings of STIL - Symposium in Information and Human Language Technology**. [S.l.: s.n.], 2019.

RIBEIRO ALISON E DA SILVA, N. **Um estudo comparativo sobre métodos de análise de sentimentos em tweets**. [S.l.]: Universidade Federal de Goiás, 2018.

SILVA, M. J.; CARVALHO, P.; SARMENTO, L. Building a sentiment lexicon for social judgement mining. In: **Computational Processing of the Portuguese Language, 10th International Conference, PROPOR 2012**. Coimbra, Portugal: Springer, 2012. (Lecture Notes in Computer Science / Lecture Notes in Artificial Intelligence, v. 7243), p. 218–228.

SIMON, H. A. **Machine Learning: An Artificial Intelligence Approach**. [S.l.]: Springer Berlin Heidelberg, 1983. 25–37 p.

SOKOLOVA, M.; LAPALME, G. **A systematic analysis of performance measures for classification tasks**. Information Processing Management, v. 45, n. 4, p. 427–437. ISSN 0306-4573. Acesso em 2 jun. 2025. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0306457309000259>.

SOUSA, R. C. C. d. **Identificando sentimentos de texto em português com o SentiWordNet traduzido**. 2016. Trabalho de Conclusão de Curso (Graduação em Ciência da Computação) – Universidade Federal do Ceará, Campus de Quixadá.

SOUZA, K. F. de; PEREIRA, M. H. R.; DALIP, D. H. **Unilex: Método léxico para análise de sentimentos textuais sobre conteúdo de tweets em português brasileiro**. Abakós, CEFET-MG, Belo Horizonte, Brasil, v. 5, n. 2, p. 79–96. ISSN 2316-9451.

SOUZA, M. et al. Construction of a portuguese opinion lexicon from multiple resources. In: **In 8th Brazilian Symposium in Information and Human Language Technology - STIL, Mato Grosso**. [S.l.: s.n.], 2011.

TABOADA, M. et al. **Lexicon-Based Methods for Sentiment Analysis**. [S.l.]: Computational Linguistics, 2011. v. 37. 267-307 p.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. [S.l.]: Addison Wesley, 2006.

TITENOK, Y. **Natural Language Processing vs Text Mining**. 2022. Acesso em: 30 set. 2024. Disponível em: <https://sloboda-studio.com/blog/natural-language-processing-vs-text-mining>.

VILARINHO, G. N. **SentiElection: análise de sentimento no Twitter baseada em centralidade de palavras**. Dissertação (Dissertação (Mestrado em Computação Aplicada)) — Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, 2019. Acesso em: 2025-10-07.

YANG, L. et al. **Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning**. [S.l.]: IEEE Access, 2020. v. 8. 23522-23530 p.

YANG, P.; CHEN, Y. **A survey on sentiment analysis by using machine learning methods**. [S.l.]: 2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), 2017. 117-121 p.