UNIVERSIDADE ESTADUAL DA PARAÍBA
CAMPUS I - CAMPINA GRANDE
CENTRO DE CIÊNCIAS E TECNOLOGIA
DEPARTAMENTO DE COMPUTAÇÃO
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

DANIEL XAVIER BRITO DE ARAUJO

ANÁLISE DE SENTIMENTOS EM AVALIAÇÕES ONLINE DE PRODUTOS: UM ESTUDO COMPARATIVO ENTRE DIFERENTES MODELOS DE APRENDIZADO DE MÁQUINA

CAMPINA GRANDE
2024

DANIEL XAVIER BRITO DE ARAUJO

ANÁLISE DE SENTIMENTOS EM AVALIAÇÕES ONLINE DE PRODUTOS: UM ESTUDO COMPARATIVO ENTRE DIFERENTES MODELOS DE APRENDIZADO DE MÁQUINA

Trabalho de Conclusão de Curso (Artigo) apresentado ao Departamento do Curso de Computação da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de Bacharel em Ciência da Computação.

Área de concentração: Aprendizado de Máquina.

Orientador: Prof. Dr. Wellington Candeia de Araujo.

CAMPINA GRANDE
2024

A663a Araujo, Daniel Xavier Brito de.

Análise de sentimentos em avaliações online de produtos [manuscrito] : um estudo comparativo entre diferentes modelos de aprendizado de máquina / Daniel Xavier Brito de Araujo. - 2024.

25 f. : il. color.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Ciência da computação) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2024.

"Orientação : Prof. Dr. Wellington Candeia de Araujo, Departamento de Computação - CCT".

1. Aprendizado de máquina. 2. Análise de sentimentos. 3. Algoritmos. I. Título

21. ed. CDD 006.35

Elaborada por Lêda Cristina Diniz Andrade - CRB - 15/1032 BC

DANIEL XAVIER BRITO DE ARAUJO

ANÁLISE DE SENTIMENTOS EM AVALIAÇÕES ONLINE DE PRODUTOS: UM ESTUDO COMPARATIVO ENTRE DIFERENTES MODELOS DE APRENDIZADO DE MÁQUINA

Trabalho de Conclusão de Curso (Artigo) apresentado ao Departamento do Curso de Computação da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de Bacharel em Ciência da Computação.

Área de concentração: Aprendizado de Máquina.

Aprovada em: 12/11/2024.

Documento assinado eletronicamente por:

- **Francisco Anderson Mariano da Silva** (***.120.084-**), em **30/11/2024 10:08:51** com chave **3ebdbeceaf1c11ef89202618257239a1**.
- **Wellington Candeia de Araujo** (***.655.074-**), em **30/11/2024 08:46:10** com chave **b1d3ed40af1011ef86ad1a7cc27eb1f9**.
- **Vinicius Reuteman Feitoza Alves de Andrade** (***.165.784-**), em **01/12/2024 08:38:18** com chave **c2793204afd811ef9f5a06adb0a3afce**.

Documento emitido pelo SUAP. Para comprovar sua autenticidade, faça a leitura do QrCode ao lado ou acesse https://suap.uepb.edu.br/comum/autenticar_documento/ e informe os dados a seguir.
Tipo de Documento: Termo de Aprovação de Projeto Final
Data da Emissão: 02/12/2024
Código de Autenticação: 9c4a2b

LISTA DE ABREVIATURAS E SIGLAS

GPT	Generative Pre-Trained Transformer
HTML	HyperText Markup Language
IA	Inteligência Artificial
NLTK	Natural Language Toolkit
PLN	Processamento de Linguagem Natural
RELU	Rectified Linear Unit
SVM	Support Vector Machines

LISTA DE GRÁFICOS

LISTA DE QUADROS

SUMÁRIO

ANÁLISE DE SENTIMENTOS EM AVALIAÇÕES ONLINE DE PRODUTOS: UM ESTUDO COMPARATIVO ENTRE DIFERENTES MODELOS DE APRENDIZADO DE MÁQUINA

SENTIMENT ANALYSIS IN ONLINE PRODUCT REVIEWS: A COMPARATIVE STUDY BETWEEN DIFFERENT MACHINE LEARNING MODELS

Daniel Xavier Brito de Araujo[1]

RESUMO

O Aprendizado de Máquina está cada vez mais presente no dia a dia da sociedade, possibilitando que várias áreas do mercado possam empregá-la com a finalidade de otimizar seus resultados e encontrar oportunidades. A análise de sentimentos é um recurso crucial que pode viabilizar e facilitar a implementação de tais melhorias. Esse estudo aborda a análise de sentimentos em avaliações que os compradores deixaram nos produtos do *e-commerce* Amazon, sendo feita com o uso de cinco algoritmos de aprendizado de máquina: Naive Bayes (Multinomial), Regressão Logística, Rede Neural Simples, Random Forest e Máquina de Vetores de Suporte Linear. É realizada a coleta dos dados e o seu devido tratamento e com isso é realizado o treinamento dos cinco modelos em cima de tais dados. Finalmente, foi possível a realização de uma avaliação geral sobre os resultados obtidos, bem como uma comparação entre os desempenhos individuais baseada em termos de acurácia, precisão, recall, F1-score.

Palavras-Chave: aprendizado de máquina; análise de sentimentos; algoritmos.

ABSTRACT

Machine Learning is increasingly present in society's daily lives. There are several areas of the market that can employ this technology in order to optimize their results and find opportunities. A crucial point that can enable and facilitate its use is the use of sentiment analysis. This study takes a sentiment analysis approach to reviews that buyers left on Amazon e-commerce products, using five machine learning algorithms: Naive Bayes (Multinomial), Logistic Regression, Simple Neural Network, Random Forest and Vector Machine Linear Support. The concepts and characteristics of each algorithm are discussed and, based on these analyses, data is collected, with the entire data cleaning process carried out and also the training of the five models based on such data. This made it possible to carry out a general assessment of the results obtained, as well as a comparison between individual performances based on accuracy, precision, recall and F1-score.

[1] Graduando em Ciência da Computação - UEPB. E-mail: daniel.araujo@aluno.uepb.edu.br

Keywords: machine learning; sentiment analysis; algorithms.

1 INTRODUÇÃO

Nos últimos anos houve um *boom* no setor de lojas onlines, também conhecidas como *e-commerces*, e hoje muito presentes no dia a dia das pessoas. Não raro, com a finalidade de se comprar um produto, os compradores analisam as avaliações no site para a tomada de decisão de compra, sendo essas opiniões geralmente correspondentes à qualidade e experiências que os usuários compartilharam, podendo funcionar tanto para o cliente quanto para o empreendedor. Existe uma oportunidade de estudo nesse contexto para poder extrair *insights*.

O recurso da análise de sentimentos é um princípio que se utiliza da abordagem computadorizada de textos, considerando as sutilezas de sentimentos e opiniões contidas nas palavras observadas, como discutido por Pang e Lee (2008, p.15), sendo útil em pesquisas que consideram sentimentos individuais, serviços gerais ao cliente e na área de *Marketing*. Neste trabalho, é apresentado o caso em que analisa os sentimentos em avaliações de produtos no *e-commerce* Amazon, possibilitada pelos modelos de aprendizado de máquina Naive Bayes (Multinomial), Regressão Logística, Rede Neural Simples, *Random Forest* e Máquina de Vetores de Suporte Linear.

As avaliações foram coletadas a partir de um banco de dados público com avaliações de usuários da plataforma da Amazon e foram aplicadas técnicas de processamento de linguagem natural que atuam na otimização das palavras contidas nas opiniões dos usuários sobre os produtos e as preparando para a análise. O objetivo é analisar as emoções nas avaliações e determinar qual modelo apresenta o melhor desempenho e quais suas melhores métricas na classificação das emoções.

2 REFERENCIAL TEÓRICO

2.1 Coleta de dados de avaliações online

A coleta de dados é a parte mais importante para a análise de sentimentos e qualquer questão relacionada a modelos de aprendizado de máquina é um ponto sensível por todo o histórico de vieses em modelos, mas como tratamos de avaliações online será abordada uma opção menos rígida. Para a coleta, existem técnicas e Hu e Liu (2004) exploraram as mesmas, incluindo abordagens baseadas em *crawling* e técnicas de mineração de opiniões. Eles discutiram os desafios e as melhores práticas para coletar dados de qualidade, como identificação dos produtos por meio de substantivos e frases, mencionaram questões como classificação positiva ou negativa das opiniões que serão abordadas. Trazendo para o contexto do trabalho, a plataforma Kaggle[2] é bastante difundida dentro da comunidade de ciência de dados por ter um amplo número de base de dados e trabalhos relacionados, muitos dos quais incluem avaliações de produtos online de várias plataformas de *e-commerce*, sendo então a opção escolhida para basear este trabalho.

[2] Disponível em https://www.kaggle.com

2.2 Análise de sentimentos

A análise de sentimentos é uma área de pesquisa do processamento de linguagem natural (PLN) que possui o objetivo de identificar e classificar textos escritos na parte das emoções, desenvolvendo uma conclusão sobre o mesmo, seja positiva ou negativa.

Uma visão interessante é discutida por Pang e Lee (2008, p.24-25) sobre principais métodos e técnicas de análise de sentimentos, incluindo abordagens baseadas em léxico e aprendizado de máquina. Essa abordagem foi amplamente pesquisada para várias aplicações como exemplo que é apresentado por Oliveira *et al.* (2019), em seu artigo sobre a análise de *tweets* de usuários com opiniões sobre alguns programas sociais em vigor no Brasil durante o governo de Dilma Rousseff com o objetivo de contribuir para práticas da gestão social.

Outro exemplo de utilização da análise de sentimentos temos no seguinte artigo: “Topics and feelings of entrepreneurs during a crisis period: Analysis of business Twitter hashtags” (Carvache-Franco *et al.*, 2023). Neste estudo, os autores exploram os tópicos discutidos e os sentimentos expressos por empreendedores durante um período de crise, especificamente o da pandemia da Covid de 2019, utilizando *hashtags* do Twitter relacionadas a negócios.

Tais exemplos demonstram algumas aplicações da análise de sentimento e sua relevância em vários contextos, como gerenciamento social e cenários de negócios. Os pesquisadores usaram técnicas de PLN e modelos de aprendizado de máquina para tentar entender a visão que os usuários expressam no texto. Foram gerados *insights* para as tomadas de decisões e para um entendimento aprofundado das necessidades e preferências dos usuários.

Especificamente no campo das avaliações de produtos *online*, essa análise de sentimentos assume um papel central ao organizar as opiniões contidas nos produtos como positivas ou negativas e isso não é só essencial para compreender a percepção dos consumidores em relação aos produtos mas também possibilita às organizações captar *feedbacks* significativos de consumidores para refinar suas estratégias e aperfeiçoar a qualidade dos seus produtos e serviços atendendo as expectativas do mercado.

2.3 Aprendizado de máquina na análise de sentimentos

O aprendizado de máquina é bastante aplicado na análise de sentimentos devido aos diversos algoritmos e técnicas que pesquisadores já utilizaram (Liu, 2012). Ele permite realizar identificações que seriam muito complexas de fazer manualmente, e diversos modelos são aplicados nesta área, cada um com suas próprias características.

Modelos como o Naive Bayes, Máquina de Vetores de Suporte Linear (SVM) e Regressão Logística são reconhecidos no campo das classificações de sentimentos por sua eficiência. Por exemplo, o Naive Bayes é conhecido por sua simplicidade e eficácia como o tratado por Turney (2002). Já o SVM é eficiente em espaços de características de alta dimensão (Cortes; Vapnik, 1998). A Regressão Logística é reconhecida pela sua compreensão a partir de um raciocínio e pela sua rapidez, em que as classificações são mais lineares (Pampel, 2020). Tais modelos são apenas alguns dentre outros muitos modelos que apoiam o aprendizado de máquina.

Utilizando o modelo SVM, Turney (2002) levantou uma abordagem para avaliação de filmes, considerando as sutilezas associadas a sentimentos positivos e negativos no discurso dos usuários no conjunto analisado. Já Lee, Pang e Vaithyanathan (2002) apresentaram avaliações de produtos utilizando Naïve Bayes, demonstrando como o modelo pode ser utilizado para diferenciar as impressões dos consumidores.

Com isso, é possível chegar ao entendimento de que os modelos de aprendizado de máquina proporcionam uma abordagem produtiva quando se trata de análise de sentimentos, visto que uma grande quantidade de informações são processadas e interpretadas e são úteis para servir a diferentes propósitos que vão do campo do *Marketing* até a gestão de reputação de marcas, áreas que são bastante influenciadas pelos sentimentos expressos pelos seus usuários e sua compreensão.

2.4 Classificadores naive bayes (Multinomial)

O modelo Naïve Bayes é probabilístico e se baseia no Teorema de Bayes para a sua funcionalidade. A versão mais notável é a multinomial, que mostra eficácia em tarefas de classificação de texto como exemplo da análise de sentimentos e utilizado muitas vezes na filtragem de *spam* de *e-mails* como foi utilizado por Androutsopoulos *et al.* (2000). Esse modelo possui a premissa *naïve* que implica que a presença ou ausência de uma determinada característica é tratada como um evento independente da presença ou ausência de outras características (Zhang, 2004), conhecidas como a suposição de independência condicional.

Esta suposição de independência é o que faz que o modelo seja rápido e eficiente, que também é observado no seu processo de treinamento relativamente simples. Em uma aplicação prática, como uma classificação de *e-mails* em *spam* ou não *spam*, o Naïve Bayes analisa a probabilidade de um *e-mail* ser *spam* com base na frequência e na combinação de palavras específicas encontradas no texto. A eficácia deste método é demonstrada como exemplo nesse contexto, onde ele pode categorizar e filtrar mensagens indesejadas. E, nessa mesma linha, ele pode ser usado para avaliações de usuários, que com um treinamento correto analisam a probabilidade de ser uma avaliação positiva ou negativa. A representação do Teorema de Bayes é apresentada da seguinte forma:

$$prob(B|A) = \frac{prob(A|B)\ prob(B)}{prob(A)}$$

Onde:

- **P(A)**: A probabilidade de A acontecer.
- **P(B):** A probabilidade de B acontecer.
- **P(A | B)**: A probabilidade de A acontecer dado que B já aconteceu.
- **P(B | A)**: A probabilidade de B acontecer dado que A já aconteceu.
  58429690

Portanto, o modelo é vantajoso devido a sua eficácia em grandes conjuntos de dados, sendo uma escolha segura e confiável para diversos contextos de aplicações. Por outro lado, sua fraqueza pode ser a sua suposição da independência condicional, onde todos os atributos de entrada são independentes

entre si, o que nem sempre é um fato quando se analisa determinados contextos no mundo real, afetando a precisão do modelo quando se analisa situações mais complexas (Stanford NLP Group, 2008).

2.5 Support vector machines (SVM)

Outro modelo adequado para a análise de sentimentos é o SVM. A Máquina de Vetores de Suporte Linear é aplicada em tarefas de classificação e regressão, envolve a identificação do hiperplano ótimo, definido pela superfície de decisão que separa duas classes de dados dentro de um espaço de características, implicando não apenas em encontrar esse hiperplano mas também em maximizar a margem entre elas. (Cortes; Vapnik, 1998). Esta margem é definida como a distância entre o hiperplano e os pontos mais próximos de cada classe, que são os vetores de suporte, que justamente dão o nome ao modelo.

O funcionamento da SVM na matemática pode ser descrito, o hiperplano é definido por uma equação:

$$w0 \cdot z + b0 = 0$$

Onde:

- w0 é o vetor de pesos, equivalente a w.
- z é o vetor de características.
- b0 é o termo de bias, equivalente a b.

A solução procura maximizar a margem, onde ||w|| é a norma do vetor w e o treinamento da SVM envolve a resolução de um problema de otimização quadrática para encontrar os valores de w e b que maximizam essa margem.

As SVMs são conhecidas por seu desempenho excepcional em espaços de alta dimensão e em situações onde a margem de separação entre as classes é clara, tornando-as particularmente úteis em aplicações de reconhecimento de padrões como tratado por Osuna *et al.* (1997). Um exemplo que reflete isso está no contexto de reconhecimento facial dentro da classificação de imagens e textos, podendo ser treinada para diferenciar imagens que representam um rosto das que não representam.

A vantagem do modelo está na capacidade de lidar com espaços de alta dimensão de forma eficaz, além de sua robustez contra *overfitting*, definido pelos modelos que aprendem os dados de treinamento, mas por outro lado, tem um desempenho ruim com novos dados, principalmente em cenários onde o número de características supera o número de amostras (Joachims, 1998). A escolha do *kernel* apropriado (por exemplo, polinomial, linear) e a sintonia dos parâmetros podem ser processos complexos e cruciais para o desempenho do modelo.

2.6 Regressão logística

A Regressão Logística é um modelo aplicado em tarefas de classificação binária, onde se pode categorizar um resultado em uma de duas possíveis categorias (Wiley, 2013).

A probabilidade é dada por P(Y=1), onde Y é a variável dependente binária, é modelada como uma função dos preditores ou variáveis independentes (X). A equação da regressão logística que incorpora essa transformação é:

$$Pi = \frac{1}{1 + exp(-\alpha - \beta 1 xi1 - \beta 2 xi2 - \ldots - \beta k xik)}$$

Em que:

- pi representa a probabilidade do evento Y ocorrer.
- α é o intercepto da regressão.
- β1,β2,…,βk são os coeficientes das variáveis independentes.
- xi1,xi2,…,xik são os valores das variáveis independentes para a iii-ésima observação.
- exp é a base do logaritmo natural (aproximadamente 2.71828).

O modelo tem algumas vantagens, são elas: simplicidade de implementação e interpretação clara dos coeficientes, onde tais coeficientes são indicadores de força e a direção da relação entre cada preditor e a probabilidade do evento. É eficaz em problemas de classificação binária, proporcionando um bom equilíbrio entre facilidade de uso e desempenho. Por outro lado, tem suas desvantagens. O método está restrito a problemas binários (0's e 1's) e portanto não atende a contextos onde as classificações possuem mais de duas categorias. Para contornar a tal binariedade podem ser feitas adaptações, a exemplo da Regressão Logística Multinomial. Um exemplo prático disso tudo é fazendo uma previsão de comportamentos de compra de clientes. Em que ele pode ser treinado para prever a probabilidade de um cliente comprar um produto com base em alguns dados como renda e histórico de compras, esta análise pode ser valiosa para empresas como os e-commerces.

2.7 Random forest

O modelo *Random Forest* pode ser compreendido através de sua abordagem de construção de árvores. É conhecido pela construção de árvores de decisão durante a fase do treinamento do modelo, onde cada árvore fornece uma predição independente. Pertence à categoria de métodos ensemble, que são definidos por:

> Um algoritmo que utiliza vários modelos em conjunto com o objetivo de obter um modelo final com melhores resultados [...]. O processo envolve a criação de um conjunto de modelos, que pode ser composto por diferentes algoritmos ou variações do mesmo algoritmo (Cruz, 2023, p. 01).

Na fase de classificação, o resultado final é determinado pelo modo das classes previstas pelas árvores individuais, enquanto na regressão o que prevalece é a média das predições. Já na fase do treinamento, várias árvores de decisão são construídas e cada árvore é treinada em um subconjunto aleatório dos dados, com a seleção de características também feita aleatoriamente, garantindo a diversidade entre as árvores, uma característica principal que confere ao modelo sua robustez. Tal processo é uma aplicação do princípio de *bagging* ou *bootstrap aggregating*, onde a agregação de múltiplas modelos (árvore, neste caso) aumenta a precisão geral do sistema (Breiman, 2001).

Tem a vantagem da ajuda na diminuição do risco do modelo se ajustar demais aos dados de treinamento devido a sua diversidade entre as árvores e tem destaque na capacidade de processar grandes conjuntos de dados com muitas variáveis e sua robustez contra o *overfitting*. Apesar de suas vantagens, possui

como desvantagem a complexidade e a dificuldade em interpretar os modelos individuais, devido ao grande número de árvores e à natureza aleatória de sua construção, dificultando o entendimento de como uma determinada previsão foi feita.

2.8 Redes neurais simples

Redes Neurais, de acordo com Krose e Van Der Smagt (1996), é um modelo inspirado na Biologia no qual uma rede de neurônios artificiais simulam os neurônios de um cérebro biológico, projetado para reconhecer padrões complexos. É utilizado em contextos de aprendizado de máquina e inteligência artificial. Sua estrutura é composta por múltiplas conexões entre as diversas camadas de neurônios que são essenciais para a capacidade da rede de aprender representações não lineares dos dados, cada neurônio em uma camada está conectado a vários neurônios na próxima camada através de um conjunto de pesos (Bishop, 2006). Tais conexões possibilitam a modelagem de relações complexas entre as variáveis de entrada e saída.

Cada neurônio de uma camada calcula uma soma ponderada de suas entradas e aplica uma função de ativação não linear, como a função *sigmóide* ou ReLU (*Rectified Linear Unit*). A saída de um neurônio é dada por algumas funções de ativação, como exemplo 3 funções:

Sigmóide:

$$\alpha(x) = \frac{1}{1+e^{x}} \qquad \alpha'(x) = \alpha(x)(1-\alpha(x))$$

Fonte: Deep Learning Book, 2022.

ReLU (Rectified Linear Unit):

$$ReLU(x) = max\{0, x\} \qquad ReLU'(x) = \begin{cases} 1, se\ x \geq 0 \\ 0, c.c. \end{cases}$$

Fonte: Deep Learning Book, 2022.

Tangente Hiperbólica (tanh):

$$tanh(x) = 2\alpha(2x) - 1 \qquad tanh'(x) = 1 - tanh^{2}(x)$$

Fonte: Deep Learning Book, 2022.

As vantagens das redes neurais simples incluem o fato de ser possível ajustá-las para uma ampla variedade de tarefas, desde classificação até regressão, além de possuírem flexibilidade e a capacidade de capturar representações não lineares complexas. Nas possíveis desvantagens, podem ter propensão ao *overfitting*, especialmente em redes com muitas camadas e neurônios. Além disso, elas exigem uma quantidade significativa de dados de treinamento para generalizar efetivamente.

3 METODOLOGIA

3.1 Coleta de dados

Este trabalho possui como foco a Amazon, um dos maiores *e-commerces* que existe. Conhecida pela oferta de milhões de produtos, tem demonstrado uma *performance* financeira impressionante. A empresa obteve resultados em lucro líquido de US$30,4 bilhões (Globo, 2024). Possui alta variedade de produtos, uma outra forte característica da Amazon são as avaliações dos usuários que variam entre positivas e negativas e são acompanhadas de notas atribuídas pelos consumidores, sendo uma parte essencial para a análise do desempenho e da satisfação dos consumidores da plataforma.

Para esta pesquisa, foi indispensável utilizar um banco de dados público com uma ampla gama de avaliações. A base de dados escolhida está localizada na plataforma Kaggle[3], amplamente utilizada por cientistas de dados e profissionais da área, não oferecendo somente uma extensa variedade de bases de dados, mas também um *hub* para competições de ciência de dados, possui *notebooks* de código interativos e fóruns de discussão, facilitando a troca de conhecimentos e experiências na área.

A base de dados de avaliações da Amazon escolhida no Kaggle para este trabalho abrange um total de 249.354 avaliações. Esta base tem colunas chave para a análise, incluindo: *ProductId*, que representa o identificador único de cada produto; *UserId*, o identificador do usuário que realizou a avaliação; *Score*, que mostra a nota atribuída ao produto pelo usuário; e *Sentiment*, uma classificação que é impactada de acordo com o grau de sentimento positivo ou negativo do comentário, baseada na pontuação *Score* atribuída. Esta estrutura facilita uma análise detalhada do comportamento do consumidor e das percepções em relação aos produtos listados na Amazon.

3.2 Descrição dos dados

O conjunto de dados utilizado neste estudo consiste em 249.354 avaliações de produtos coletadas da Amazon. Incluem alguns atributos, que possuem identificadores de produto e usuário, nomes de perfil, numeradores e denominadores de utilidade, pontuações de avaliação, sumários e tempo, sendo este último baseado nas datas das avaliações que abrangem um amplo intervalo, o que pode indicar variações nas opiniões dos consumidores ao longo do tempo.

No início, o atributo mais importante é o *Score* porque é a nota que o comprador deu ao produto e com base nessa nota cada avaliação foi categorizada em uma de duas classes de sentimento: Positiva ou Negativa. O *Score* é definido pelas avaliações que estão associadas a pontuações, que variam de 1 a 5. O resultado indica uma tendência geral para avaliações moderadamente positivas, tendo a média das pontuações de aproximadamente 3.37.

Uma nova classificação é feita, adicionando uma nova coluna na tabela considerando o seguinte: se o *Score* for de valor (1, 2 ou 3) consideramos o sentimento como negativo, se o *Score* for 4 e 5 consideramos como positivo. Dessa forma a nova coluna *Sentiment* foi determinada com *Positive* ou *Negative*, e a contagem geral ficou *Positive* com 124677 linhas e *Negative* também com 124677 linhas.

Com 50% das avaliações classificadas como Positivas e 50% classificadas como Negativas, a distribuição das classes de sentimento no conjunto de dados é perfeitamente equilibrada. Esta distribuição equilibrada facilita a avaliação

[3] Disponível em https://www.kaggle.com/datasets/arhamrumi/amazon-product-reviews?resource=download

comparativa dos modelos de aprendizado de máquina sem o viés introduzido por classes desproporcionais.

3.3 Pré-processamento dos dados

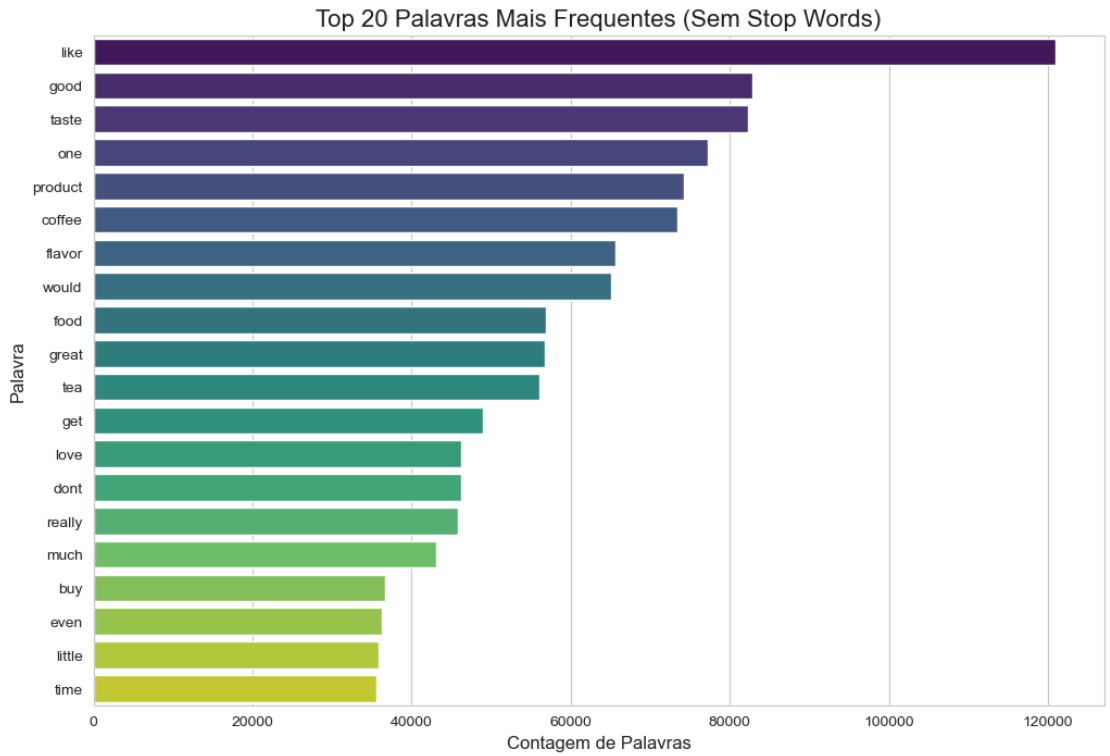
O pré-processamento dos dados antes de proceder com quaisquer comparações entre modelos de aprendizado de máquina é uma etapa essencial, envolvendo a limpeza e a padronização dos dados, preparando-os adequadamente para as técnicas de aprendizado de máquina. Para a limpeza dos dados foi usada uma abordagem que incluiu a remoção de pontuações, números, caracteres especiais, conversão de todo o texto para letras minúsculas para manter a uniformidade dos dados, eliminação de tags HTML a também como a remoção de emojis e outros caracteres não convencionais. Para execução dessas tarefas, utilizou-se a linguagem de programação Python em conjunto com bibliotecas especializadas como *regular expression* (RE). Em seguida é criada uma nova coluna na base de dados, denominada *Cleaned_Text*, que contém os textos processados.

A próxima etapa é fundamental para a análise textual. A tokenização dos dados consiste na quebra do texto em unidades básicas, ou *tokens* como palavras individuais, neste trabalho realizado utilizando a biblioteca *Natural Language Toolkit* (NLTK) de Python, conforme a documentação oficial disponível em: https://www.nltk.org/.

Transforma frases complexas, como "*as far as the ingredients are concerned*", em listas de *tokens* simplificados, por exemplo, [*as, far, as, the, ingredients, are, concerned*]. Dessa forma, é possível criar outra coluna na base de dados, chamada *Tokenized_Text*.

Além disso, é realizada a remoção de *stopwords*, que são palavras comuns que, por não portarem significado relevante isoladamente, são excluídas da análise. Foi utilizado uma lista abrangente de *stopwords* disponibilizada pela biblioteca NLTK, o que permitiu uma limpeza textual mais efetiva. Em que é possível observar um exemplo de resultado no Gráfico 1:

Gráfico 1 - Top 20 Palavras Mais Frequentes (Sem Stop Words)



Fonte: Elaborado pelo autor, 2024.

A análise prosseguiu com a aplicação do processo de *stemming*, que simplifica as palavras até suas raízes, diminuindo assim a dimensionalidade do conjunto de dados e acelerando significativamente a análise textual. Por exemplo, palavras como [*far, ingredients, concerned, chips, seem*] são reduzidas a suas formas radicais [*far, ingredi, concern, chip, seem*].

Por último, é feita a vetorização dos textos, uma etapa bem importante que faz a conversão dos textos para um formato compreensível pelos algoritmos. A técnica de TF-IDF (Frequência do Termo-Inverso da Frequência nos Documentos) é empregada para destacar a importância de termos específicos dentro de um documento em relação a uma coleção de documentos. Uma das formas para se realizar o cálculo é utilizando a seguinte fórmula:

$$TFIDF = log\left(\frac{N}{DF}\right) * ln\left(1 + \frac{M}{T}\right)$$

Em que:

- N é o número de vezes que a palavra ocorre no corpus.
- DF é o número total de documentos na coleção.
- M representa o número de vezes que a palavra ocorre neste documento específico.
- T é o número total de palavras neste documento específico.

A classe *TfidfVectorizer* da biblioteca *sklearn* foi utilizada conforme a sua documentação[4] e realizou o cálculo da equação, onde se obteve o resultado ((249354, 5000), 31.933051003793803), que revela uma matriz gerada que possui 249.354 documentos e considera 5.000 termos mais frequentes. A densidade média de elementos não-nulos é de aproximadamente 31,93, o que significa que, em média, cada documento contém cerca de 32 termos significativos com valores de TF-IDF diferentes de zero.

Dessa forma foi feita a limpeza da base para que seja realizada as análises de sentimentos e comparações dos modelos.

3.4 Treinamento e resultado dos modelos

Com a base preparada, se realiza o treinamento dos modelos citados neste trabalho, em seguida será apresentado os resultados para comparação. Após, é feita uma divisão do conjunto de dados, em que se divide 70% do conjunto para o treinamento dos modelos, esta etapa foi possibilitada por causa da biblioteca *sklearn.model_selection* do Python, onde se obtém as dimensões dos três conjuntos: ((174547, 5000), (37403, 5000), (37404, 5000)), 15% para validações e 15% para os testes. Todos os modelos, exceto a rede neural simples (que foi utilizado o *tensorflow*) foram feitos usando a biblioteca *sklearn* e avaliados seus respectivos desempenhos com algumas métricas como acurácia, precisão, *recall* e *f1-score*.

4 MÉTRICAS

Para obter resultados mais explicativos sobre cada modelo utilizado, é importante o uso de métricas, e é possível identificar algumas que estão muito bem definidas na área. A biblioteca do *sklearn*, que é bem completa para o assunto, possui o método *classification_report* que gera automaticamente as métricas em um relatório, facilitando a análise e visualização. Essas métricas são: Acurácia, Precisão, *Recall* e *F1-Score*.

Cada uma dessas métricas é descrita por um modelo matemático específico, de acordo com (Duarte, 2019):

$$Acurácia = \frac{Verdadeiros\ positivos + Verdadeiros\ negativos}{Total}$$

$$Precisão = \frac{Verdadeiros\ positivos}{Verdadeiros\ positivos + Falso\ positivos}$$

$$Recall = \frac{Verdadeiros\ positivos}{Verdadeiros\ positivos + Falsos\ Negativos}$$

$$F1 = \frac{2 * precisão * recall}{precisão + recall}$$

Com elas, é possível ter uma base para a comparação dos resultados e identificar determinados comportamentos que esses modelos possam apresentar.

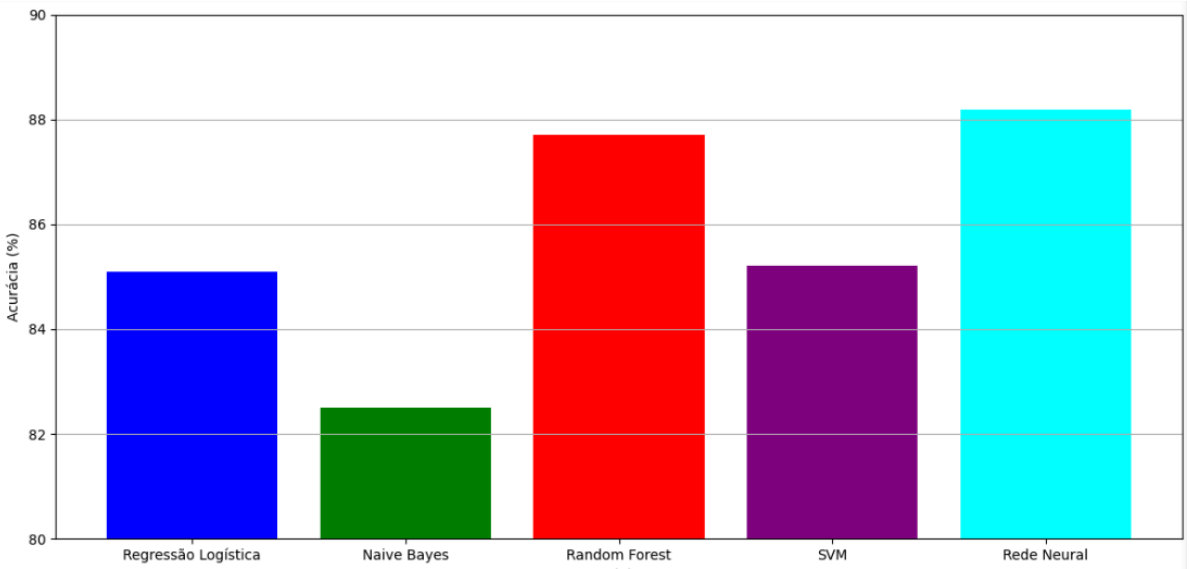
[4] Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

Cada métrica possui um objetivo, e a partir disso foram criados gráficos para as comparações de cada tipo de métrica.

4.1 Acurácia

A acurácia é uma métrica que mede a proporção das previsões corretas com relação ao total de previsões feitas, ou seja, o quão frequente o modelo faz previsões corretas, e na classificação binária onde temos dois resultados (Positivos ou Negativos).

Gráfico 2 - Acurácia dos Modelos



Fonte: Elaborado pelo autor, 2024.

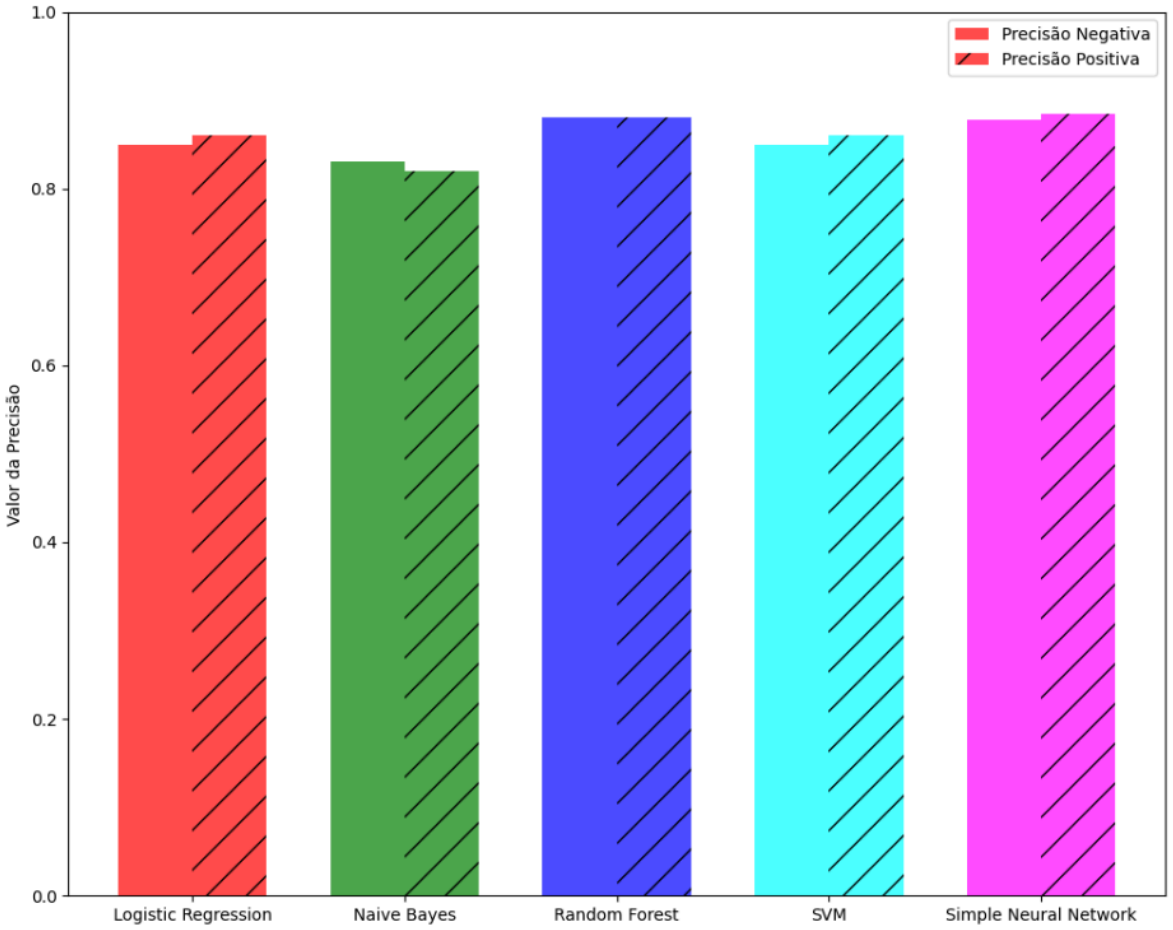
O Gráfico 2 demonstra a porcentagem que cada modelo obteve com a acurácia. Com esses resultados, é possível afirmar que a rede Neural Simples obteve o melhor desempenho, devido à capacidade das redes neurais de modelar relações não lineares e complexas dos dados. No entanto, podem ser propensas a *overfitting*, situação no qual o modelo se ajusta muito bem ao conjunto de dados e passa a trazer resultados não tão interessantes ao verificar novos dados. O uso de *Dropout*, que é uma técnica conhecida no código ajudou a regularizar, pois desativa aleatoriamente um subconjunto de neurônios em uma camada específica em cada época. *Random Forest* é um modelo poderoso e ficou em segundo lugar, pode capturar relações não lineares e é menos propenso ao *overfitting* devido ao *ensemble* de árvores como a técnica de *bagging* conforme citado neste trabalho. Regressão Logística e SVM tiveram acurácia parecidas e se demonstraram competentes, eles são particularmente úteis quando a relação entre as características e as classes é mais linear. Naive Bayes (Multinomial) teve a menor acurácia dos cinco, mas ainda assim apresentou desempenho razoável. Pode ser adequado em cenários onde a velocidade de treinamento e predição são críticos.

4.2 Precisão

A precisão é a métrica que mostra a porcentagem de acertos de determinados tipos de avaliação, ou seja, dentre as classificações positivas, quantas estão corretas e da mesma forma as avaliações negativas, sendo importante em

situações onde os falsos positivos são mais críticos. O Gráfico 3 apresenta dados relacionados à métrica.

Gráfico 3 - Precisão dos Modelos



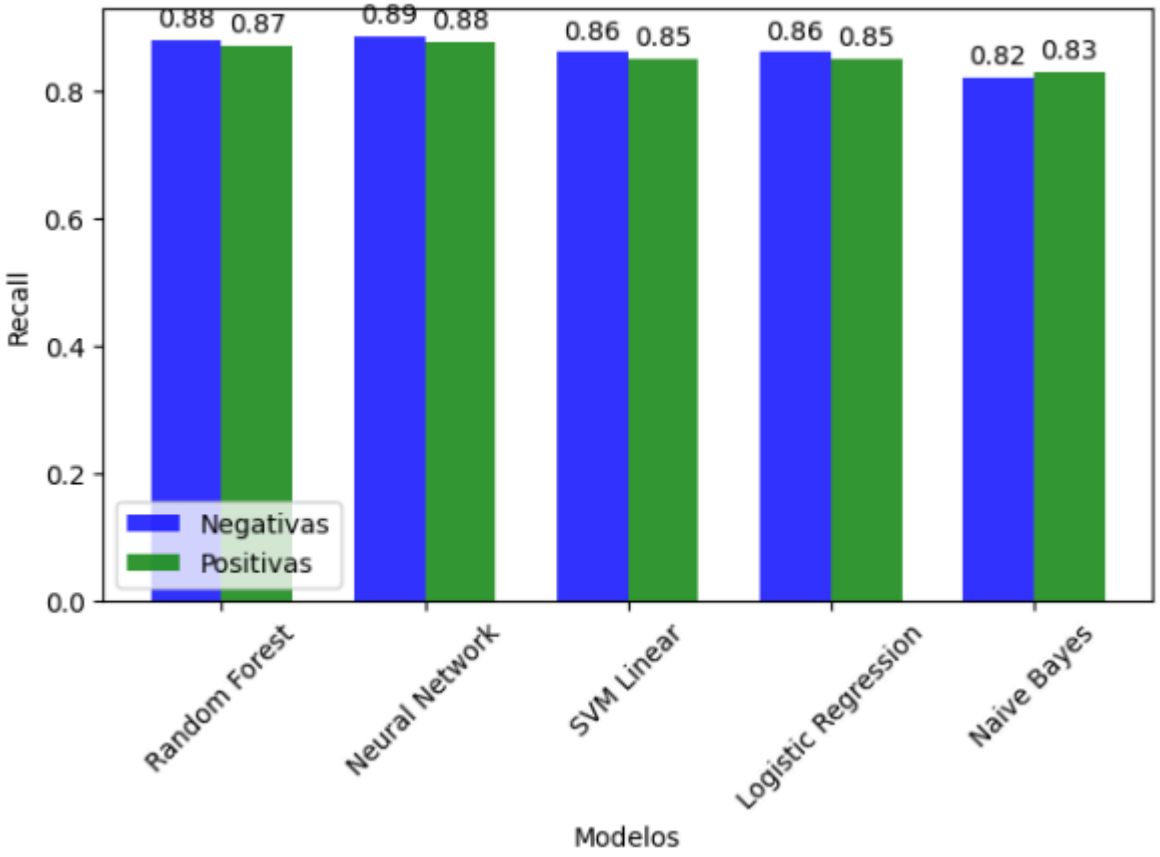
Fonte: Elaborado pelo autor, 2024.

Com esses resultados conseguimos observar que a precisão para positivo e negativo é bem equilibrada nos modelos, ou seja, não há um viés forte para uma classe específica. Rede Neural Simples e *Random Forest* mostraram altas precisões para as avaliações de ambas as classes, com rede neural sendo ligeiramente superior para as avaliações positivas. Regressão logística e SVM apresentaram precisões muito próximas, que é consistente pelo fato que se baseiam em limites de decisão lineares. São eficazes porém menos precisos que *Random* e Rede Neural. Naive Bayes (Multinomial) teve a menor precisão entre os cinco modelos que pode ser devido às suas suposições de independência entre as características, o que raramente é verdadeiro em dados de texto onde o contexto e a sequência das palavras são importantes. O desempenho dos modelos em termos de precisão está em uma faixa entre (82% e 88%), ou seja, todos são relativamente eficazes para o conjunto de dados passado e os modelos mais complexos como *Random Forest* e Rede Neural demonstram ser os modelos mais promissores.

4.3 Recall

O *recall* é a sensibilidade ou taxa de verdadeiros positivos, utilizado na avaliação de modelos de classificação com o objetivo de medir a proporção de verdadeiros positivos identificados corretamente pelo modelo em relação ao número total de casos que são de fato positivos. É a medida de quão bem um modelo de classificação identifica os resultados relevantes. Com base nos modelos, tivemos os seguintes dados de *Recall*, vistos no Gráfico 4:

Gráfico 4 - Recall por Modelo e Sentimento



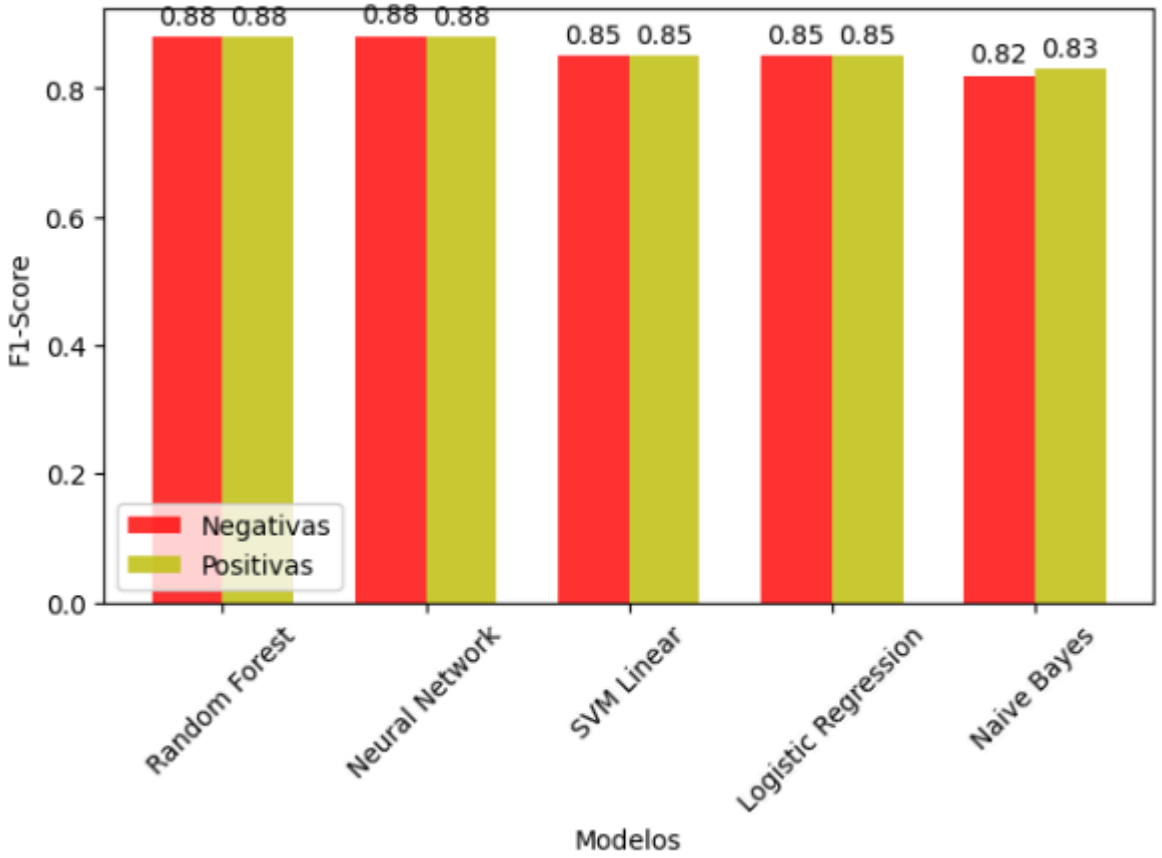
Fonte: Elaborado pelo autor, 2024.

Random Forest e Rede Neural Simples apresentam os melhores *recalls*, com a Rede Neural tendo um pequeno aumento e sendo o de melhor *recall*. SVM e Regressão tem um *recall* inferior, o que significa que estão perdendo mais casos positivos do que os citados acima. Naive Bayes (Multinomial) tem o *recall* mais baixo, o que pode ser um indicativo de que está falhando em identificar um número maior de casos positivos reais.

4.4 F1-score

Essa métrica combina precisão com a sensibilidade (*recall*) de um modelo. Ele é particularmente útil quando se deseja buscar um equilíbrio entre os dois, especialmente em situações onde uma taxa alta de falsos positivos ou falsos negativos é crítica. *F1-Score* é a média harmônica de precisão e *recall*. Com isso, a métrica recuperada dos nossos modelos foi a seguinte, de acordo com o Gráfico 5:

Gráfico 5 - F1-Score por Modelo e Sentimento



Fonte: Elaborado pelo autor, 2024.

Como já observado que as últimas métricas do *Random Forest* e Rede Neural Simples foram maiores, aqui não seria diferente. Os dois mantêm um bom equilíbrio entre precisão e *recall* e isso indica que são modelos muito eficazes para esse conjunto de dados, conseguindo tanto identificar corretamente os verdadeiros positivos quanto evitar os falsos positivos. SVM e Regressão Logística possuem *F1-Scores* iguais para avaliações negativas e positivas, refletindo um equilíbrio. No entanto são valores menores que os citados acima. Naive Bayes (Multinomial) possui menos *F1-Scores*, o que é consistente com a sua menor precisão e *recall*.

5 RESULTADOS

A Rede Neural Simples demonstrou a melhor acurácia e altos valores em todas as outras métricas. Isso sugere que esse modelo é capaz de capturar bem a complexidade dos dados e tem um bom desempenho na classificação das avaliações negativas e positivas.

Random Forest teve um desempenho muito próximo ao da rede neural simples, com as métricas sendo ligeiramente inferiores. Esse modelo também é

muito robusto e eficaz para a tarefa, sendo uma alternativa viável se a velocidade de treinamento e predição foram fatores importantes.

SVM e Regressão Logística apresentaram desempenhos similares entre si e foram efetivos, mas o desempenho se mostra menor em comparação com rede neural simples e *random forest*. Podem ser considerados para uso quando a simplicidade do modelo e velocidade são mais críticas do que alcançar melhor desempenho.

Naive Bayes (Multinomial) mostrou as métricas mais baixas entre os modelos testados. Apesar de ser um modelo rápido e fácil, pode não ser a melhor escolha para esse conjunto de dados específico. É possível demonstrar os resultados por meio do Quadro 1:

Quadro 1 - Resultado dos Modelos

	Random Forest	Regressão Logística	Rede Neural	Naive Bayes	SVM
Melhor Desempenho Geral					
Simplicidade e Eficiência					
Modelos Avançados					

Fonte: Elaborado pelo autor, 2024.

No ponto de melhor desempenho geral, o *Random Forest* foi o campeão em termos de acurácia e métricas de classificação, com isso se sugere que o modelo é o mais adequado entre os demais testados na análise de sentimentos. No quesito simplicidade e eficiência, o modelo de Regressão Logística se destacou dentre os demais, pois oferece a combinação ideal de performance e interpretabilidade. Como ponto de modelos avançados a Rede Neural mostrou um potencial interessante, especialmente considerando a sua capacidade de melhorar as métricas com mais treinamento. Um contraponto no entanto é sua complexidade e a necessidade de mais dados e ajustes cuidadosos a tornam uma escolha mais avançada, adequada para cenários onde há disponibilidade de recursos computacionais, podendo considerar um possível *Overfitting*, onde a rede neural apresentou melhorias contínuas no desempenho durante o treinamento, mas existindo esse risco. A introdução de mais regularizações (como *early stopping*) poderia ajudar a mitigar isso.

Dependendo dos recursos disponíveis e do objetivo final do estudo (por exemplo, necessidade de interpretabilidade vs. desempenho máximo), a escolha do modelo pode variar. Para aplicações práticas, *Random Forest* ou Regressão Logística podem ser preferíveis, enquanto redes neurais oferecem um caminho para explorar se houver interesse em modelos mais complexos.

6 CONSIDERAÇÕES FINAIS

A análise de sentimentos é muito importante pois fornece um norte para impulsionar vários setores, incluindo o da Tecnologia, sendo a ponta do *iceberg* tecnológico que continua gerando novos conhecimentos e trabalhos, é muito

provável que nos próximos anos sejam feitas novas descobertas, principalmente com o fortalecimento das IAs. A sociedade como um todo se beneficia com tais avanços na área.

Neste trabalho foram feitas comparações de cinco modelos de aprendizado de máquina para análise de sentimentos. É possível compreender qual algoritmo possui mais atributos positivos, mas também é possível entender que nem sempre o mais completo é o mais adequado, para cada trabalho pode existir algum modelo que pode se encaixar melhor, por ser menos complexo ou possuir alguma outra característica que seja mais apropriada, por exemplo.

Para análises futuras, devem ser levadas em consideração o contexto específico da aplicação, priorizando o alinhamento entre o objetivo do modelo e os recursos disponíveis. Por exemplo, o Random Forest é indicado para situações que exigem equilíbrio entre robustez e simplicidade operacional, enquanto redes neurais são mais adequadas em problemas que demandam alta precisão e tem disponibilidade de recursos computacionais avançados.

O uso de modelos generativos como o GPT pode servir de base para futuros trabalhos a partir da análise de sentimentos em que ele usa os modelos de forma especializada e já retorna resultados interessantes. Um outro tópico é o estudo dos vieses que podem ser trabalhos em todos os pontos das análises, verificando no discurso ou também nos modelos trabalhados indícios de enviesamento.

REFERÊNCIAS

ALLISON, Paul D. **Logistic regression using SAS: theory and application**. 2a Edição. Cary: SAS Institute, 2012.

AMAZON supera expectativa de receita, que vai a US$ 170 bilhões no quarto trimestre. **O Globo**, Rio de Janeiro, 01 fev. 2024. Disponível em: https://oglobo.globo.com/economia/negocios/noticia/2024/02/01/amazon-supera-expectativa-de-receita-que-vai-a-us-170-bilhoes-no-quarto-trimestre.ghtml. Acesso em: 12 ago. 2024.

ANDROUTSOPOULOS, Ion; KOUTSIAS, John; CHANDRINOS, Konstantinos V.; SPYROPOULOS, Constantine D. An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages. **Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval**, New York, p. 160 - 167, jul. 2000. Disponível em: https://arxiv.org/pdf/cs/0008019. Acesso em: 1 mar. 2024.

BISHOP, Christopher M. **Pattern recognition and machine learning**. New York: Springer, 1996.

BREIMAN, Leo. Random forests. **Machine Learning**, v. 45, p. 5 - 32, out. 2001. Disponível em: https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf. Acesso em: 9 jul. 2024.

CARVACHE-FRANCO, Orly; VÍQUEZ-PANIAGUA, Ana Gabriela; CARVACHE-FRANCO, Mauricio; CARVACHE-FRANCO, Wilmer; PÉREZ-OROZCO, Allan. Topics and feelings of entrepreneurs during a crisis period: Analysis of business Twitter hashtags. **TEC Empresarial 2023**, Costa Rica, v. 17, n. 2, p. 33-47,

2023. Disponível em: https://www.scielo.sa.cr/pdf/tec/v17n2/1659-3359-tec-17-02-33.pdf. Acesso em: 29 jan. 2024.

CORTES, Corinna; VAPNIK, Vladimir. Support-vector networks. **Machine Learning**, Boston, p. 273 - 297, 1995. Disponível em: https://link.springer.com/article/10.1007/BF00994018. Acesso em: 21 fev. 2024.

CRUZ, Marcelo. Machine learning: conhecendo as técnicas de bagging e boosting. **Alura**, 2023. Disponível em: https://www.alura.com.br/artigos/machine-learning-tecnicas-bagging-boosting. Acesso em: 9 jul. 2024.

CYPRIANO, Eduardo. **Análise de Dados em Astronomia I**. São Paulo, 2015. Slide. Disponível em: https://edisciplinas.usp.br/pluginfile.php/799829/mod_resource/content/1/aula2.pdf. Acesso em: 7 mar. 2024.

DATA science academy. **Deep Learning Book.** 2022. Disponível em: https://www.deeplearningbook.com.br/.

DUARTE, Mateus P. Machine Learning: métricas de avaliação (Acurácia, Precisão e Recall). **Medium**, 9 nov. 2020. Disponível em: <https://medium.com/@mateuspdua/machine-learning-m%C3%A9tricas-de-avalia%C3%A7%C3%A3o-acur%C3%A1cia-precis%C3%A3o-e-recall-d44c72307959/>. Acesso em: 23 set. 2024.

HU, Minqing; LIU, Bing. Mining and summarizing customer reviews. **Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, Chicago, 2004. Disponível em: https://www.cs.uic.edu/~liub/publications/kdd04-revSummary.pdf. Acesso em: 8 jan. 2024.

JOACHIMS, Thorsten. **Text categorization with support vector machines: learning with many relevant features**. Machine Learning: ECML-98. Berlin, Heidelberg: Springer, 1998. p. 137-142. Disponível em: https://link.springer.com/chapter/10.1007/BFb0026683. Acesso em: 20 jun. 2024.

JOHN, Jobin. TF-IDF Vectorizer Explained. **Egochi**, 6 jul. 2022. Disponível em: https://www.egochi.com/tfidfvectorizer/. Acesso em: 25 set. 2024.

KROSE, Ben; VAN DER SMAGT, Patrick. **An introduction to neural networks**. 8a Edição. Amsterdam: The University of Amsterdam, 1996.

LIU, Bing. **Sentiment analysis and opinion mining**. 1a Edição. San Rafael: Morgan&Claypool Publishers, 2012.

OLIVEIRA, Daniel José Silva; BERMEJO, Paulo Henrique de Souza; PEREIRA, José Roberto; BARBOSA, Daniely Aparecida. A aplicação da técnica de análise de sentimento em mídias sociais como instrumento para as práticas da gestão social

em nível governamental. **Revista de Administração Pública**, Rio de Janeiro, v. 53, n. 1, p. 235-251, jan./fev. 2019. Disponível em: https://www.scielo.br/j/rap/a/GD3F8HdkQKGSHy8zzV8w9Ys. Acesso em: 13 jan. 2024.

OSUNA, E.; FREUND, R.; GIROSIT, F. Training support vector machines: an application to face detection. **Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition**, San Juan, p. 130-136, jun. 1997. Disponível em: https://ieeexplore.ieee.org/document/609310. Acesso em: 17 jun. 2024.
PAMPEL, Fred C. Logistic regression: a primer. Thousand Oaks: Sage Publications, 2000. Disponível em: https://doi.org/10.4135/9781412984805. Acesso em: 26 fev. 2024.

PANG, Bo; LEE, Lillian. Opinion mining and sentiment analysis. **Foundations and Trends in Information Retrieval**, v. 2, n.1-2, p.1-135, 2008. Disponível em: https://www.cs.cornell.edu/home/llee/omsa/omsa.pdf. Acesso em: 10 jul. 2024.

PANG, Bo; LEE, Lillian; VAITHYANATHAN, Shivakumar. Thumbs up? Sentiment classification using machine learning techniques. **Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing**, Philadelphia, p. 79 - 86, jul. 2002. Disponível em: https://www.cs.cornell.edu/home/llee/papers/sentiment.pdf. Acesso em: 23 fev. 2024.

TURNEY, Peter D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, Philadelphia, p. 417-424, jul. 2002. Disponível em: https://aclanthology.org/P02-1053.pdf. Acesso em: 19 fev. 2024.

ZHANG, Harry. The optimality of naive Bayes. 2004. **Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference**, California, p. 562-567, mai. 2004. Disponível em: https://cdn.aaai.org/FLAIRS/2004/Flairs04-097.pdf. Acesso em: 5 mar. 2024.

AGRADECIMENTOS

À Deus por conceder toda força e sabedoria durante toda a caminhada.

Aos meus familiares, principalmente meus pais Sebastião e Paula por todo esforço e apoio incondicional.

À minha namorada Ludmila por toda ajuda e companheirismo em todos os momentos.

Aos meus colegas, em especial Lucas pela colaboração e troca de conhecimento durante todo o percurso.

Ao meu orientador Wellington pela sua prestatividade e disposição em me apoiar neste momento final de graduação.

E aos demais professores, pela dedicação e por guiarem com o seu ensino e inspiração.