CENTRO UNIVERSITÁRIO CHRISTUS

SISTEMAS DE INFORMAÇÃO

MAURÍCIO DA SILVA PEREIRA MOURA

UMA ANÁLISE DE SENTIMENTOS DAS MANCHETES SOBRE NOTÍCIAS DO MERCADO FINANCEIRO BRASILEIRO

FORTALEZA

2023

MAURÍCIO DA SILVA PEREIRA MOURA

UMA ANÁLISE DE SENTIMENTOS DAS MANCHETES SOBRE NOTÍCIAS DO MERCADO FINANCEIRO BRASILEIRO

Trabalho de Conclusão de Curso (TCC) apresentado ao curso de Sistemas de Informação do Centro Universitário Christus, como requisito parcial para obtenção do grau de bacharel em Sistemas de Informação.

Orientador: Prof. Dr. Thiago Iachiley Araújo de Souza

FORTALEZA

2023

Dados Internacionais de Catalogação na Publicação
Centro Universitário Christus - Unichristus
Gerada automaticamente pelo Sistema de Elaboração de Ficha Catalográfica do
Centro Universitário Christus - Unichristus, com dados fornecidos pelo(a) autor(a)

M929a Moura, Maurício da Silva Pereira.
Uma análise de sentimentos das manchetes sobre notícias do mercado financeiro brasileiro / Maurício da Silva Pereira Moura. - 2023.
35 f. : il. color.

Trabalho de Conclusão de Curso (Graduação) - Centro Universitário Christus - Unichristus, Curso de Sistemas de Informação, Fortaleza, 2023.
Orientação: Prof. Dr. Thiago Iachiley Araújo de Souza.

1. Análise de Sentimentos. 2. Mercado Financeiro. 3. Manchetes. 4. Processamento de Linguagem Natural. 5. Aprendizagem Supervisionada. I. Título.

CDD 004.07

MAURÍCIO DA SILVA PEREIRA MOURA

UMA ANÁLISE DE SENTIMENTOS DAS MANCHETES SOBRE NOTÍCIAS DO MERCADO FINANCEIRO BRASILEIRO

Trabalho de Conclusão de Curso (TCC) apresentado ao curso de Sistemas de Informação do Centro Universitário Christus, como requisito parcial para obtenção do grau de bacharel em Sistemas de Informação.

Aprovada em:

BANCA EXAMINADORA

Prof. Dr. Thiago Iachiley Araújo de Souza (Orientador)
Centro Universitário Christus (Unichristus)

Prof. Dr. Daniel Nascimento Teixeira
Centro Universitário Christus (Unichristus)

Prof. Me. Felipe Timbó Brito
Centro Universitário Christus (Unichristus)

AGRADECIMENTOS

Agradeço primeiramente a Deus pela força necessária durante minha caminhada, a Ele toda honra e toda glória.

A todos pelo apoio pessoal e emocional, família, minha namorada Marjorie Marques Rodrigues e meu amigo da graduação Carlos Levi da Silva Albuquerque. Gostaria de agradecer especialmente aos voluntários, que disponibilizaram seu tempo e conhecimento para que os dados fossem rotulados, eng. de produção mecânica Francisco Thiago Fernandes de Oliveira, ao eng. eletricista Francisco Aprígio Maciel de Lima e a eng. civil Marjorie Marques Rodrigues.

Ao prof. Dr. Thiago Iachiley Araújo de Souza, meu orientador, pela atenção e apoio que possibilitou a realização deste trabalho. Aos professores participantes da banca examinadora, Prof. Dr. Daniel Nascimento Teixeira e o Prof. Me. Felipe Timbó Brito pelo tempo e pelas imprescindíveis colaborações e sugestões.

Por fim, agradeço aos professores do curso de sistemas de informação da instituição não só pelo conhecimento transferido, mas também pelos momentos na sala de aula, principalmente os conselhos que foram fundamentais para o crescimento acadêmico e profissional e guardarei e levarei esses momentos para fora da academia.

"Devemos acreditar que somos talentosos para algumas coisas, e que essa coisa, a qualquer custo, deve ser alcançada."

(Marie Curie)

RESUMO

A evolução das notícias ao longo dos anos tem sido bastante significativa, impulsionada principalmente pelo avanço da tecnologia e pela disseminação da internet. Destaca-se a importância das manchetes na atração de leitores para as notícias financeiras e como elas podem influenciar a percepção e o sentimento dos leitores. Baseado nisso, este estudo propõe a utilização da análise de sentimentos para identificar automaticamente se as manchetes relacionadas ao mercado financeiro brasileiro são positivas ou negativas. O objetivo geral é criar um modelo de processamento de linguagem natural (NLP) para prever o sentimento das manchetes, enquanto os objetivos específicos envolvem o pré-processamento dos dados, a construção do modelo NLP e a análise de métricas de desempenho. A justificativa do estudo é baseada no impacto das notícias sobre as decisões dos investidores e na dificuldade de prever o comportamento do mercado. O estudo descreve o desenvolvimento do modelo utilizando os classificadores *Support Vector Machine* e *Naive Bayes*. Os resultados dos experimentos mostram a viabilidade da identificação automatizada de emoções e são propostos trabalhos futuros, incluindo o uso de um conjunto de dados mais amplo e a correlação da análise de sentimento com a oscilação dos preços das ações na bolsa de valores.

Palavras-chave: Análise de Sentimentos. Mercado Financeiro. Manchetes. Processamento de Linguagem Natural. Aprendizagem Supervisionada.

ABSTRACT

The evolution of news over the years has been quite significant, driven mainly by the advancement of technology and the dissemination of the internet. The importance of headlines in attracting readers to financial news and how they might influence readers' perception and sentiment is highlighted. Based on this, this study proposes the use of sentiment analysis to automatically identify whether headlines related to the Brazilian financial market are positive or negative. The general objective is to create a natural language processing (NLP) model to predict the sentiment of headlines, while the specific objectives involve the pre-processing of data, the construction of the NLP model and the analysis of performance metrics. The justification of the study is based on the impact of news on investors' decisions and the difficulty of predicting market behaviour. The study describes the development of the model using the Support Vector Machine and Naive Bayes classifiers. The results of the experiments demonstrate the viability of automated emotion identification and future work is proposed, including the use of a wider dataset and the correlation of sentiment analysis with stock market price movements.

Keywords: Sentiment Analysis. Financial Market. Headlines. Natural Language Processing. Supervised Learning.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 Contextualização e delimitação do tema

Devido a democratização da veiculação da informação, investir em ações da bolsa de valores tem tornando-se rotineiro para o brasileiro, segundo a (B3, 2023), empresa de infraestrutura de mercado financeiro, em 2018, o número de investidores pessoas físicas era de cerca de 700 mil e, em dezembro de 2022, foi atingida a marca de 5 milhões, representando um aumento de mais de 700%. O que significa que a população está pesquisando e aprendendo como esse ambiente funciona, e uma fonte valiosa para essa pesquisa são as notícias.

A principal forma de atrair a atenção do público para as notícias veiculadas é por meio das manchetes. Efetivamente, a manchete é um resumo da notícia, indispensável por ter a função de informar clara e objetivamente, de modo que o leitor consiga entender e ter uma ideia do que abordará o texto (GUIRALDELLI; SÁ, 2014).

Potencialmente as manchetes podem determinar a forma com que muitas pessoas leem as notícias (KONNIKOVA, 2016). As manchetes podem determinar a forma de percepção dos leitores em relação ao restante do conteúdo, afetando inclusive a maneira como as pessoas vão se lembrar destas notícias (ECKER *et al.*, 2014).

Para analisar essas manchetes, pode-se utilizar a técnica de análise de sentimentos, processo de análise textual capaz de categorizar a emoção escrita no texto em positivo ou negativo. Essa análise tem o intuito de identificar como os leitores se sentem em relação às manchetes da mídia não especializada sobre o mercado financeiro do Brasil.

1.2 Problematização

Como visto na contextualização, as notícias sobre finanças circulam nos meios de comunicação a todo instante e os leitores tentam informar-se o mais rápido possível, a fim de inteirar-se dos últimos acontecimentos tornando as manchetes a principal fonte de pesquisa na busca por informações, pois elas resumem de forma rápida todo o assunto que a notícia aborda. Dessa forma, como analisar os sentimentos que as manchetes causam nos investidores brasileiros? Portanto, faz-se necessário identificar se a opinião do leitor é positiva ou negativa de forma automática e imparcial sobre o mercado financeiro brasileiro, com o objetivo de entender a dinâmica desses sentimentos e seus efeitos através das manchetes.

1.3 Pressupostos

A Análise de Sentimentos e a Mineração de Opinião tem recebido muita atenção com o crescimento das redes sociais na Internet e uma aplicação comum para a classificação de textos tem sido classificar as opiniões em positivas, negativas e neutras (BALAGE *et al.*, 2013), para essa finalidade utiliza-se aprendizado de máquina.

> O aprendizado de máquina geralmente se distingue em três casos: aprendizado supervisionado, não supervisionado e por reforço. Mais especificamente, o problema de aprendizado supervisionado envolve aprender uma função a partir de exemplos de suas entradas e saídas. Além disso, os dados de treinamento devem estar rotulados de alguma forma (ALBUQUERQUE, 2022).

Assim, o pressuposto deste trabalho é a utilização da técnica de aprendizado supervisionado, usando como entrada manchetes sobre o mercado financeiro brasileiro classificadas em positivo e negativo.

1.4 Objetivos

1.4.1 Objetivo geral

Este trabalho tem como objetivo principal criar um modelo utilizando NLP (*Natural Language Processing*) para predizer o sentimento de manchetes relacionadas ao mercado financeiro brasileiro.

1.4.2 Objetivos específicos

1. Realizar o pré-processamento de manchetes de um jornal, com o propósito de preparar os dados como entrada para o modelo que será elaborado.
2. Construir um modelo NLP (*Natural language processing*) a partir de um aprendizado supervisionado para identificar sentimentos positivos ou negativos nas manchetes de um jornal.
3. Analisar, experimentalmente, métricas de desempenho obtidas após o treinamento do modelo com base nos dados analisados.

1.5 Justificativa

Os investidores recebem uma variedade de notícias positivas ou negativas sobre várias empresas no cotidiano. As informações veiculadas nessas notícias afetam as suas decisões. Estas decisões refletem os seus padrões de negociação através de alterações no preço das ações nos mercados financeiros (VIJAY *et al.*, 2018). O impacto dos textos das notícias financeiras depende tanto do comportamento individual como do comportamento global do mercado (mentalidade de rebanho), tornando-o difícil de prever (VIJAY *et al.*, 2018). Dessa forma, o presente trabalho realizou uma análise de sentimento sobre as manchetes das notícias relacionadas ao mercado financeiro brasileiro, a fim de verificar a polaridade desses dados.

1.6 Estrutura do trabalho

Os capítulos a seguir estão divididos de forma em que o segundo capítulo apresenta o referencial teórico e abrange as principais definições e conceitos à compreensão do tema. O terceiro capítulo reúne os trabalhos correlatos, em que os temas abordados possuem similaridade com esta pesquisa. No quarto capítulo é apresentada a metodologia adotada para a análise de sentimento em notícias sobre o mercado financeiro brasileiro. O quinto capítulo apresenta os resultados obtidos com a base de dados utilizada. Por fim, no sexto capítulo será apresentada a conclusão do trabalho proposto e possíveis trabalhos futuros.

2 REFERENCIAL TEÓRICO

2.1 Aprendizado de máquina supervisionado

O aprendizado supervisionado é uma técnica de aprendizado de máquina em que um algoritmo é treinado com um conjunto de dados rotulados para aprender a fazer previsões ou classificações precisas em novos dados. Nesse processo, o algoritmo é fornecido com exemplos de entrada (conhecidos como "dados de treinamento") juntamente com as saídas desejadas correspondentes (conhecidas como "rótulos"ou "alvos").

Com base nesses dados rotulados, o algoritmo é capaz de aprender um modelo matemático que pode mapear as entradas para as saídas desejadas com uma alta precisão. O objetivo final é usar esse modelo para fazer previsões precisas em dados não vistos anteriormente.

O aprendizado supervisionado é amplamente utilizado em aplicações de reconhecimento de padrões, classificação, regressão, previsão e outras tarefas de análise de dados. O conceito de aprendizado supervisionado, do inglês *Supervised Learning*, está vinculado a relação entre um conjunto de variáveis de entrada que gera uma variável de saída, e que gera um mecanismo que torna possível a predição de novos valores a partir de dados que não foram ainda vistos (CUNNINGHAM *et al.*, 2008).

O aprendizado supervisionado é o que a maioria das pessoas pensam quando se fala sobre aprendizado de máquina, no qual o modelo só pode melhorar se puder medir a diferença entre a saída esperada (o rótulo) e suas previsões (HAPKE *et al.*, 2019).

2.2 Linguagem de processamento natural

O processamento da linguagem natural (NLP, do inglês *Natural Language Processing*) é a disciplina das ciências da computação, da inteligência artificial e da linguística, que procura dar às máquinas a capacidade de compreender a linguagem das pessoas.

Linguagem de processamento natural (NLP, do inglês *Natural Language Processing*) é uma subárea da inteligência artificial e da linguística que se concentra na interação entre a linguagem humana e os computadores. A técnica permite que as máquinas compreendam e interpretem a linguagem humana, incluindo a fala e o texto, de maneira semelhante à forma como os seres humanos a utilizam.

O Processamento de Linguagem Natural (PLN) é uma área da computação que tem

como objetivo extrair representações e significados mais completos de textos livres escritos em linguagem natural (INDURKHYA; DAMERAU, 2010).

Desde o surgimento das técnicas de Processamento de Linguagem Natural - PLN, muitos avanços foram obtidos, mas a compreensão plena de linguagem natural por métodos computacionais está ainda longe de ser resolvida (VIEIRA; LOPES, 2010).

Embora os humanos ainda sejam melhores em reconhecer sentimentos dentro de um diálogo permanente, devido à nossa capacidade em reter informações sobre o contexto de uma declaração, as máquinas estão ficando cada vez melhores em manter o contexto (HAPKE *et al.*, 2019).

Portanto, o NLP usa técnicas de aprendizado de máquina, processamento de linguagem natural e análise estatística para extrair significado de textos e fala, incluindo a identificação de palavras-chave, análise de sentimento, tradução automática, reconhecimento de fala e resposta a perguntas, entre outras aplicações.

O objetivo final do NLP é permitir que as máquinas entendam a linguagem humana de maneira natural e possam interagir com os seres humanos de forma inteligente e significativa. O NLP é amplamente utilizado em aplicações de processamento de texto e fala, como *chatbots*, assistentes virtuais, análise de sentimentos, correção gramatical, resumo de texto, entre outros.

2.3 Análise de sentimento

A Análise de sentimento é um processo de mineração de dados que envolve a identificação e extração de informações subjetivas e emocionais de um texto, como opiniões, atitudes, emoções e sentimentos. Esse processo é realizado por meio de algoritmos de aprendizado de máquina que analisam o texto e classificam as palavras ou frases como positivas, negativas ou neutras.

A análise de sentimentos que tem por objetivo identificar e extrair de forma automática, as opiniões, sentimentos e emoções, expressados em um texto (NARAYANAN *et al.*, 2009).

> A área de Análise de Sentimentos ou Mineração de Opinião visa descobrir, quantificar e qualificar computacionalmente opiniões e seus conceitos relacionados, com objetivo de avaliar o sentimento sobre um determinado produto, analisar empresas na bolsa de valores, analisar sentimento sobre determinadas pessoas, entre outros (CASTRO, 2022).

A análise de sentimento é uma técnica de descoberta de conhecimento por meio

da mineração de dados, sua finalidade é revelar a opinião das pessoas sobre temas específicos (OLIVEIRA; BERMEJO, 2017). Dessa forma, a análise de sentimento pode ser aplicada a diferentes tipos de dados de texto, como avaliações de produtos, comentários de redes sociais, pesquisas de opinião e outros tipos de conteúdo gerado pelo usuário na internet.

Essa técnica é bastante útil para empresas e organizações que desejam entender a percepção do público em relação a seus produtos, serviços ou marcas, e para governos e organizações que desejam monitorar a opinião pública em relação a questões políticas ou sociais.

2.4 Métricas de avaliação

Essa etapa consiste em demonstrar quais métricas foram utilizadas para verificação do modelo, elas possibilitam avaliar, quantitativamente, os algoritmos de classificação utilizados nos experimentos. Dessa forma possibilitam identificar se as técnicas de pré-processamento aplicadas resultam numa melhoria significativa para o projeto ou não.

Segundo, (SHALEV-SHWARTZ; BEN-DAVID, 2014) a métrica de acurácia, é descrita como a fração de previsões corretas do modelo, a métrica de sensibilidade, e caracterizada como o percentual de instâncias classificadas de modo correto como positivas em meio às outras instâncias da base que são, efetivamente, positivas.

Também segundo, (SHALEV-SHWARTZ; BEN-DAVID, 2014) a métrica de precisão o percentual de instâncias classificadas corretamente como positivas em meio a todas as instâncias também classificadas como positivas.

De acordo com, (AGARWAL, 2020), o índice F1 ou F1-score é um número localizado entre 0 e 1, que representa a média harmônica entre precisão e recall. Segue na figura 1 uma tabela resumo dos parâmetros apresentados.

Figura 1 – Resumo das métricas.

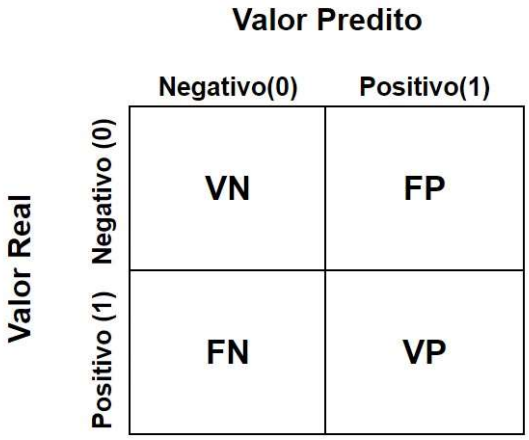
Acurácia	Precisão	Sensibilidade	F1 - score
$\frac{VP+VN}{VP+FN+VN+FP}$	$\frac{VP}{VP+FP}$	$\frac{VP}{VP+FN}$	$\frac{precisao * revocacao}{precisao + revocacao}$

Fonte: Elaborada pelo autor.

2.4.1 Matriz de confusão

Uma matriz de confusão, permite visualizar o desempenho de um modelo através de uma tabela, onde cada linha da matriz representa as instâncias de uma classe real enquanto cada coluna representa as instâncias de uma classe predita, ou vice-versa. A matriz de confusão é uma tabela comparativa dos valores que um algoritmo trouxe como predição em relação aos valores reais. A figura 2 apresenta um tipo de matriz confusão.

Figura 2 – Matriz de confusão.



Fonte: Elaborada pelo autor.

Onde na diagonal principal da matriz os valores: Verdadeiro positivo (VP) e Verdadeiro Negativo (VN), ocorre quando, no conjunto real, a classe que estamos buscando foi prevista corretamente.

Por exemplo na posição (VP), quando o sentimento é positivo e o modelo previu corretamente que o sentimento do texto analisado é positivo.

Por exemplo, na posição (VN), o sentimento não é positivo, e o modelo previu corretamente que o sentimento do texto analisado não é positivo.

Onde na diagonal secundária da matriz os valores Falso Positivo (FP) e Falso Negativo (FN): ocorrem quando, no conjunto real, a classe que não estamos buscando prever foi prevista incorretamente.

Por exemplo, na posição (FP) o sentimento é positivo, mas o modelo previu que o texto analisado contém sentimento negativo.

Por exemplo na posição (FN): Por exemplo, o sentimento é negativo, mas o modelo previu que o texto analisado contém sentimento positivo.

3 TRABALHOS CORRELATOS

Na última década o estudo da polarização das manchetes tem investigado e gerado pesquisas correlatas acerca da análise de sentimentos e sua importância para o mercado de ações. Esses trabalhos são citados a seguir.

3.1 Análise de sentimento nas manchetes de notícias brasileiras

Em (RAMOS *et al.*, 2016) é apresentada uma análise da polaridade das notícias produzidas pelos jornais online brasileiros, usando uma metodologia experimental baseada na análise de sentimentos como forma de capturar a força das polaridades expressas nas manchetes de notícias, a fim de determinar se elas são exageradamente negativas ou positivas, ou se elas são neutras.

O trabalho coleta e analisa o conteúdo de 59.510 notícias produzidas por 8 diferentes jornais brasileiros. A coleta foi realizada por meio de uma *web crawler* executado durante o período de 06 de setembro de 2015 até 30 de novembro de 2015. Para cada uma das páginas web recuperadas, é realizado um *parsing* do conteúdo retornado e recuperadas informações relativas a cada uma das notícias.

Na fase de rotulação é criado um subconjunto de 100 notícias aleatórias com um número proporcional de notícias de cada uma das fontes coletadas. As manchetes são rotuladas manualmente por 3 voluntários. Para cada manchete, é solicitado aos voluntários que classifiquem o seu conteúdo como negativo, neutro ou positivo.

Com esta base rotulada é possível analisar a capacidade de predição de cada um dos métodos para o contexto proposto. O conjunto de dados rotulados é utilizado como baseline para comparar o desempenho de duas estratégias de análise de sentimento.

Os autores utilizam o método *SentiStrength* e *iFeel*. O método *SentiStrength*, específico para o idioma português, é o método que apresenta o melhor desempenho, com uma acurácia de 64%, considerando as 3 classes (negativa, neutra e positiva).

O trabalho realiza uma série de análises comparativas das oito fontes e verifica a tendência do número de sentimentos das manchetes, em seguida realiza uma análise temporal com o objetivo de verificar se a polaridade das notícias publicadas é afetada por uma determinada época ou devido a ocorrência de algum evento específico.

Dessa forma, o trabalho mostra que existe no cenário brasileiro grande concentração

de notícias classificadas como neutras, e poucas classificadas no extremo negativo ou positivo. Já a análise temporal revela que o número de manchetes neutras e negativas é constantemente superior ao número de notícias positivas.

O trabalho citado acima está relacionado com o tema deste trabalho, pois busca analisar a polaridade das manchetes, a fim de determinar se elas são negativas, positivas ou neutras, além disso, os autores utilizam o conjunto de dados rotulados por voluntários como *baseline* para comparar o desempenho de duas estratégias de análise de sentimento e realiza uma série de análises quantitativas e temporais.

A diferença entre os trabalhos é que as manchetes deste trabalho são específicas do mercado financeiro e foram utilizados dois classificadores para automatizar a identificação apenas da polaridade positiva ou negativa, e a análise temporal poderá ser realizada em um trabalho futuro.

3.2 Identificando polaridade em manchetes de notícias utilizando *Naive Bayes*

O trabalho realizado por (OLIVEIRA, 2016) objetiva identificar as emoções em manchetes de textos jornalísticos utilizando uma abordagem baseada em aprendizado de máquina supervisionado. As emoções utilizadas no trabalho são chamadas de Emoções Básicas (ou puras) e consistem em seis emoções: tristeza, raiva, medo, alegria, desgosto e surpresa.

O autor utiliza 1.750 notícias como conjunto de dados rotulados, onde, cada notícia rotulada é uma notícia na qual foi atribuída a uma das seis emoções básicas. Caso o texto não tenha uma emoção predominante, a notícia é rotulada como “neutra”. Das 1.750 notícias, têm-se 250 notícias rotuladas para cada uma das seis emoções e 250 notícias rotuladas como “neutra”.

Cada manchete de notícia é submetida com sua emoção ao final do texto processado. Deste modo, é possível para o modelo de predição do classificador associar as palavras do texto processado a uma emoção (previamente rotulada) e construir seu aprendizado. Este aprendizado é usado na fase de teste para o classificador predizer a emoção de determinada manchete de notícia.

No trabalho, foram utilizadas duas técnicas para a avaliação do modelo, a divisão fixa e a validação cruzada, com os experimentos observa-se que o classificador *Naive Bayes* obteve uma acurácia média de 59,55% utilizando a técnica de divisão fixa e 58,46% utilizando a técnica de validação cruzada. Assim, de acordo com os resultados, vimos que é possível identificar emoções de forma automatizada através do classificador *Naive Bayes*.

O trabalho citado acima está relacionado com este trabalho por se propor a identificar emoções em manchetes. Entretanto, o autor utilizou apenas o classificador *Naive Bayes* para automatizar a identificação dessas emoções, enquanto a abordagem proposta neste trabalho é usar dois classificadores nos dados e determinar qual deles performa melhor na identificação das emoções.

3.3 Análise de sentimentos e SVM na classificação de *tweets* depressivos

O trabalho de (CORTES; MELO, 2021), apresenta a análise de sentimentos baseada em aprendizagem de máquina que pode ser utilizada para auxiliar na classificação de *tweets* depressivos e não depressivos em redes sociais.

O processo de construção do *dataset* se inicia com o cadastro e solicitação de acesso à API do *Tweeter*. Após a liberação são definidos os termos das *strings* de busca, cujo objetivo é extrair dados tanto para a classe depressiva quanto para a classe não depressiva. Definiu-se o *MongoDB Atlas* como ferramenta de banco de dados *NoSql* para o armazenamento dos *tweets* extraídos.

O processo de extração e armazenamentos de *tweets* teve início no mês de maio de 2020 e se estendeu até o mês de novembro de 2020. Nesse período foram coletados 15.747 *tweets* da classe depressiva e 15.430 *tweets* da classe não depressiva, somando um total de 31.177 *tweets*.

Em seguida são criados dois vetores chamados *tweets* e classes, onde o vetor *tweets* recebe somente as informações dos textos e o vetor classes apenas as classificações correspondentes a cada *tweet*, por conseguinte foi realizado todo o processo de tratamento de texto este processo são passos provenientes da área de Processamento de Linguagem Natural (PLN) e foram utilizados através da biblioteca NLTK.

O último passo antes da utilização dos algoritmos de aprendizagem de máquina é realizar uma etapa de separação dos dados para que o modelo possa ser testado e validado, utilizou-se a técnica de re-amostragem conhecida como validação cruzada através da função *cross_val_predict*, presente na biblioteca *sklearn*.

O trabalho executa dois experimentos, sendo o primeiro um subconjunto de 5 mil *tweets* e o segundo com a base completa contendo 31.177 *tweets*. Em seguida verifica-se o desempenho dos algoritmos *Naive Bayes* e SVM de aprendizagem de máquina, sendo que o modelo SVM mostra-se mais eficaz que o modelo de *Naive Bayes*, trazendo como resultado uma

acurácia de 94% ao experimento realizado com a quantidade máxima de *tweets* do dataset.

O trabalho citado acima está relacionado com este por utilizar os dois classificadores, SVM e o *Naive Bayes*. Os autores propõem verificar o desempenho em ambos os classificadores, onde o Support Vector Machine apresenta uma melhor performance em relação ao *Naive Bayes*.

A distinção entre os trabalhos está na utilização do conjunto de dados para os dois classificadores, no trabalho mencionado são utilizados um subconjunto e o conjunto de dados completo, enquanto neste trabalho foi utilizado todo o conjunto de dados rotulado.

4 METODOLOGIA

Neste capítulo é apresentada a metodologia experimental proposta para este trabalho, que inclui o processo de coleta dos dados e a estratégia adotada para a inferência dos sentimentos expressa nas manchetes.

4.1 Dados utilizados

Para esse estudo foi utilizado um conjunto de dados do (KAGGLE, 2023), plataforma online de Ciência de dados que permite a estudantes e profissionais encontrarem e participarem de competições de aprendizado de máquina, colaborar em projetos de código aberto e acessar conjuntos de dados e recursos educacionais.

O conjunto de dados consiste em 167.053 exemplos, subdivididos em: Título, Url do Artigo, Artigo Completo e Categoria. No conjunto a subdivisão título refere-se às manchetes. Os dados utilizados são da página do Jornal Brasileiro Folha De São Paulo - http://www.folha.uol.com.br/ e compreende o período de janeiro de 2015 a setembro de 2017.

Para o trabalho foi filtrado um período de dois anos, compreendido entre janeiro de 2015 a dezembro de 2016, reduzindo o conjunto para 132.873 exemplos. Em seguida, na coluna categoria, seletou-se apenas a categoria mercado, restringindo a quantidade de exemplos para 16.385 dados.

Devido ao tempo limitado de produção deste trabalho e para tornar viável a leitura analítica e a categorização dos dados pelos voluntários, na fase de rotulação, optou-se pela utilização de 5.000 exemplos dentre os 16.385 selecionados anteriormente.

A figura 3 apresenta uma amostra do conjunto de dados utilizados na etapa de rotulação.

Figura 3 – Amostra do conjunto de dados.

Tensão politica assusta investidor estrangeiro e preocupa o Planalto	O governo Michel Temer entrou em alerta depois de ser informado por investidores estrangeiros sobre ...	2016-11-12	mercado		http://www1.folha.uol.com.br/mercado/2016/12/1840321-tensao-politica-assusta-investidor-estrangeiro-...
Crise financeira nos Estados trava Parcerias Público-Privadas	Além dos problemas com as próprias concessões, o governo federal ainda vai precisar lidar com uma cr...	2016-11-12	mercado		http://www1.folha.uol.com.br/mercado/2016/12/1840326-crise-financeira-nos-estados-trava-parcerias-pu...

Fonte: Elaborada pelo (KAGGLE, 2023).

Os exemplos foram rotulados manualmente por 3 voluntários, investidores com mais de 1 ano de experiência na bolsa. Para a rotulação, buscou-se identificar a natureza da notícia, seu impacto no mercado financeiro e os setores ou empresas específicas envolvidas.

Dentre os critérios de avaliação das manchetes foram levadas em consideração:

1. Familiarize-se com os conceitos financeiros: entendimento dos termos e conceitos financeiros, como ações, *commodities*, índices de mercado, moedas, fusões e aquisições, lucros e perdas, entre outros.
2. Identificação do contexto: entender o contexto em que a manchete está sendo apresentada, ou seja, manchetes sobre aumento das taxas de juros pode ser positiva para os investidores que buscam rendimentos fixos, mas pode ter um impacto negativo no mercado de ações.
3. Determinar o impacto no mercado: Avaliar como a notícia pode afetar o mercado financeiro Brasileiro. Identifique se a manchete é de natureza positiva ou negativa e seu potencial impacto nos preços das ações, taxas de câmbio, *commodities* etc.
4. Analisar as empresas ou setores envolvidos: Muitas manchetes financeiras estão relacionadas a empresas ou setores específicos. Identifique as empresas mencionadas na notícia e considere o impacto direto que a notícia pode ter sobre elas. Isso pode envolver rotular a notícia com o nome da empresa ou setor afetado.

Portanto a rotulação de notícias financeiras pode variar dependendo do contexto e dos requisitos específicos. É importante adaptar as diretrizes às suas necessidades e usar a razoabilidade ao aplicá-las.

Nesse contexto verificou-se o sentimento que a manchete transmitiu em relação ao mercado de ações do Brasil e classificou-as em positivas, negativas e neutras. Podemos verificar que o Exemplo 1 exprime um sentimento positivo, o Exemplo 2 um sentimento negativo e o Exemplo 3 um sentimento neutro.

Exemplo 1: *Bolsa sobe 0,6% sustentada por Vale e Petrobras; dólar encosta em R$3,15*

Exemplo 2: *Siderúrgicas e Vale pressionam e Bolsa acumula sexta baixa; dólar cai 0,3%*

Exemplo 3: *Livros destacam mulheres inovadoras; veja lançamentos*

4.2 Pré-processamento de dados

O pré-processamento de dados é uma fase crucial, visto que permite escolher quais dados possuem nexo para estruturarem o conjunto de dados. É um processo que compreende a preparação, organização e estruturação de dados.

Para isso é realizado um conjunto de atividades que envolvem converter dados brutos em dados preparados, ou seja, em formatos úteis e eficientes. As principais causas de baixa qualidade de dados incluem a ocorrência de atributos irrelevantes, valores ausentes ou redundantes (PADILHA; CARVALHO, 2017).

Dessa forma um correto pré-processamento é fundamental para garantir a qualidade e confiabilidade das nossas conclusões. A figura 4 mostra o fluxo de tratamento dos dados através das colunas, categorias e rótulos na qual o conjunto de dados deve seguir para ser rotulado.

Figura 4 – Fluxo de tratamento dos dados.

Fonte: Elaborada pelo autor.

A tabela 1 mostra a distribuição dos sentimentos classificados pelos voluntários e os percentuais em relação a quantidade de manchetes, onde obteve-se, 1.826 positivos, 1.666 negativos e 1.508 neutros:

Tabela 1 – Distribuição dos sentimentos classificados pelos voluntários.

Positivos	1.826	36,52%
Negativos	1.666	33,32%

Fonte:Elaborada pelo autor.

4.2.1 Pandas

Pandas é uma biblioteca desenvolvida sobre a linguagem *Python* que disponibiliza estruturas de dados e ferramentas para tratar dados rotulados, usada para limpar, formatar e padronizar os dados.

4.2.1.1 Remover valores duplicados

Uma parte importante da análise de dados é analisar valores duplicados e removê-los. O método *drop_duplicates()* do Pandas ajuda a remover duplicatas do conjunto de dados. Foram retiradas linhas que acabam por se repetir no conjunto de dados e que alterariam a análise.

A seguir verifica-se na tabela 2 o comparativo dos valores após a remoção dos valores duplicados.

Tabela 2 – Comparativo dos valores após a remoção dos valores duplicados.

	Positivo	**Negativo**	**Total**
Antes	1.666	1.665	3.492
Depois	1.824	1.665	3.489

Fonte: Elaborada pelo autor.

4.2.1.2 Remover expressões regulares

Expressões Regulares são padrões que associam sequências de caracteres no texto. Podem ser usadas para extrair ou substituir porções de texto, bem como, endereço ou link de imagens em uma página HTML, remover caracteres inválidos e remover vírgulas.

Expressões que não possuem significado para o texto, portanto devem ser removidas para não atrapalhar a análise.

Exemplo 1:

Antes: Inflação sobe 0,16% em setembro puxada por alta do preço da gasolina

Depois: Inflação sobe 016% em setembro puxada por alta do preço da gasolina

4.2.1.3 Remover caractere especial

Ao analisar dados, principalmente para treinar modelos de inteligência artificial, alguns dados podem prejudicar a análise. Por isso, temos que utilizar apenas os dados coerentes para o modelo.

Exemplo 2:

Antes: Inflação sobe 016% em setembro puxada por alta do preço da gasolina

Depois: Inflação sobe 016 em setembro puxada por alta do preço da gasolina

4.2.1.4 Converter strings maiúsculas em minúsculas

No Python pode-se facilmente converter uma *string* em maiúscula ou minúscula através das funções *upper()* e *lower()*, respectivamente. Nessa etapa transformou-se todos os caracteres do texto para o seu formato minúsculo, evitando que palavras sejam diferenciadas apenas por estarem maiúsculas.

Exemplo 3:

Antes: Inflação sobe 016 em setembro puxada por alta do preço da gasolina

Depois: inflação sobe 016 em setembro puxada por alta do preço da gasolina

4.2.1.5 Remover número

Os números foram retirados com a finalidade de simplificar e focar nas palavras e padrões linguísticos relevantes e dessa forma padronizar o texto evitando que eles influenciem indevidamente o processo de classificação.

Exemplo 4:

Antes: inflação sobe 016 em setembro puxada por alta do preço da gasolina

Depois: inflação sobe em setembro puxada por alta do preço da gasolina

4.2.2 NLTK

NLTK significa *Natural Language Toolkit*, é uma biblioteca em *Python* utilizada para processamento de linguagem natural (NLP), amplamente utilizada em pesquisas acadêmicas e na indústria de processamento de linguagem natural, de código aberto e gratuito para uso, distribuição e modificação.

Essa biblioteca fornece uma série de ferramentas para trabalhar com tarefas comuns de NLP, como *tokenização*, etiquetagem de partes do discurso, análise sintática, extração de informações, classificação de texto, entre outras.

4.2.2.1 Stopwords

O NLTK possui uma lista de *stopwords* para o Português. *Stopwords* são palavras que podem ser consideradas irrelevantes para o entendimento do sentido de um texto, ou seja, palavras semanticamente irrelevantes. Exemplos: as, e, os, de, para, com, sem, foi. Essas palavras são geralmente removidas de um texto durante a fase de pré-processamento.

Exemplo 5:

Antes: inflação sobe 016 em setembro puxada por alta do preço da gasolina

Depois: inflação sobe 016 setembro puxada alta preço gasolina

4.2.2.2 Tokenização

A *tokenização*, também conhecida como segmentação de palavras, quebra a sequência de caracteres em um texto localizando o limite de cada palavra, ou seja, os pontos onde uma Palavra termina e outra começa (PALMER, 2010).

Dessa forma, a *tokenização* pode normalizar um texto, por exemplo mapeando suas palavras para versões apenas com letra minúscula, expandindo contrações e até mesmo extraindo o radical de cada palavra, processo este conhecido por *"stemming"*.

A ideia da *tokenização* é dividir algum texto existente em pedaços menores. Por exemplo, um parágrafo pode ser convertido em frases e posteriormente em palavras. Para aplicar a *tokenização* foi usado *word_tokenize* do pacote *nltk.tokenize.*

Exemplo 6:

Antes: inflação sobe setembro puxada alta preço gasolina

Depois: [’inflação’, ’sobe’, ’setembro’, ’puxada’, ’alta’, ’preço’, ’gasolina’]

4.2.2.3 Stemização

Ao usar um verbo que é conjugado em tempos verbais diferentes que está localizado em uma coluna de uma data frame ou uma lista, a lematização encurtará todos esses verbos conjugados para o menor comprimento possível de caractere, preservando assim a raiz do verbo.

Entretanto a *stemização* não se limita apenas aos verbos e pode ser utilizada em todas as classes gramaticais que possuam os mesmos radicais da palavra original como os adjetivos, substantivos e outros verbos.

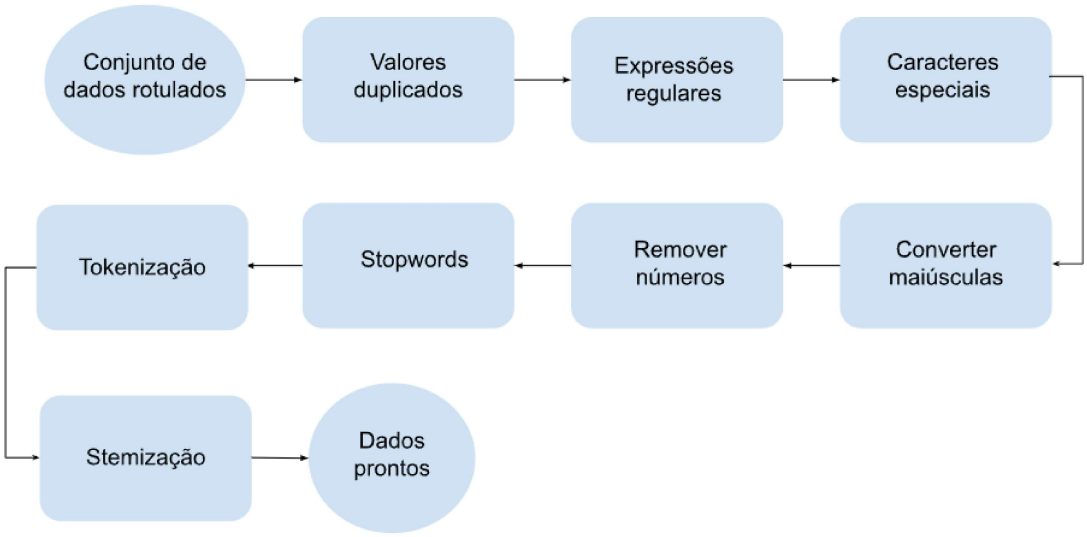
Exemplo 7:

Antes: [’inflação’, ’sobe’, ’setembro’, ’puxada’, ’alta’, ’preço’, ’gasolina’]

Depois: inflaçã sob setembr pux alta prec gasolin

A figura 5 apresenta o fluxo de processamento das manchetes através da retirada dos valores duplicados, expressões regulares, caracteres especiais, converter maiúsculas, remover números, *stopwords*, *tokenização*, *stemização* que o conjunto de dados rotulados segue para que esteja apto a ser usado.

Figura 5 – Fluxo de processamento das manchetes.



Fonte: Elaborada pelo autor.

4.3 Construção do modelo

Para este trabalho foi desenvolvido um modelo de aprendizado de máquina supervisionado utilizando um total de 3.489 dados rotulados como positivo ou negativo.

Em seguida, foi aplicada a técnica de linguagem natural conhecida como *bag of words*, a ideia principal dessa técnica é simplificar o texto, ignorando a estrutura gramatical e a ordem das palavras, considerando apenas a frequência das palavras no documento.

Ao final desse processo, os dados foram inseridos na função *train_test_split*, fornecida pela biblioteca de aprendizado de máquina *scikitlearn* em *Python*. Essa técnica é comumente usada para dividir um conjunto de dados em dois subconjuntos: um conjunto de treinamento e um conjunto de teste.

A divisão dos dados ficou 20% para teste e 80% para treino. Essa divisão é feita para avaliar o desempenho do modelo em dados não vistos durante o treinamento e, assim, verificar a capacidade de generalização do modelo.

Após a divisão do conjunto de dados foram utilizados dois algoritmos de aprendizado de máquina para realização da classificação do modelo, o primeiro deles é o *Support Vector Machine* e o segundo o *Naive Bayes*.

A escolha desses dois classificadores se deu a fim de verificar a performance aplicada ao conjunto de manchetes, pois segundo (PARDO; NUNES, 2002) afirmam que *Naive Bayes* tem desempenho melhor quando há um grande conjunto de treinamento, no entanto as Máquinas

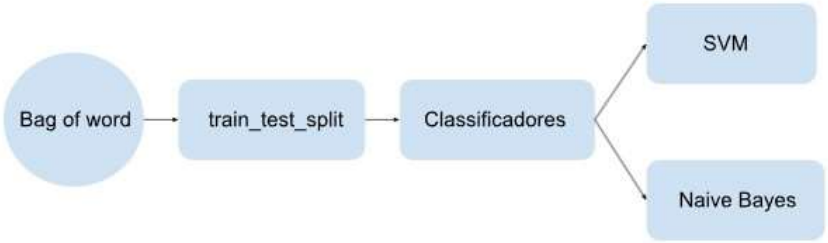
de Vetores de Suporte, por sua vez, apresentam grande potencial para classificação quando se dispõe de um conjunto de treinamento limitado (LORENA; CARVALHO, 2007).

Após a classificação dos modelos foi comparado a acurácia de ambos e escolhido o modelo com melhor desempenho. A acurácia é uma métrica comum utilizada para avaliar o desempenho de um modelo de classificação.

No entanto, esse trabalho não se limitou apenas a essa métrica para verificar o desempenho dos modelos, devido a casos, onde as classes estão desbalanceadas, ou seja, quando há uma grande diferença no número de exemplos de cada classe, a acurácia pode não ser uma métrica adequada para avaliar o desempenho do modelo.

Nesses casos, é recomendado utilizar outras métricas, como precisão, *recall*, *F1-score* ou matriz de confusão, a figura 6 apresenta o fluxo do modelo criado para a análise de sentimentos das manchetes, que fornecem uma visão mais detalhada do desempenho do modelo para cada classe, essas métricas serão especificadas na parte de resultados desse trabalho.

Figura 6 – Fluxo do modelo.



Fonte: Elaborada pelo autor.

5 RESULTADOS

Este capítulo detalha as especificações do ambiente de desenvolvimento e execução do modelo, além de apresentar e analisar os resultados obtidos a partir da base de dados que os modelos executaram.

5.1 Especificação do ambiente

Para desenvolvimento do modelo foi usado *Google Colaboratory*, também conhecido como *Google Colab*, que é um ambiente de notebook baseado em nuvem que permite escrever, executar e colaborar em código *Python*. Essa plataforma é gratuita e fornecida pela Google que oferece recursos de computação em nuvem e acesso às unidades de processamento gráfico.

A vantagem do *Google Colab* é a sua capacidade de compartilhar e colaborar em tempo real com outras pessoas. Os *notebooks* podem ser compartilhados com colegas ou equipes, permitindo a edição e execução colaborativa do código. Além disso, o *Colab* permite o uso de recursos de armazenamento em nuvem, como o *Google Drive*, para importar e exportar dados de forma conveniente.

O *Google Colaboratory* permitiu que a criação do modelo fosse utilizada em várias máquinas, pois sua manipulação foi feita por meio de um navegador da web e não requer nenhuma configuração complexa, uma vez que todo o ambiente de desenvolvimento e computação está disponível na nuvem. A praticidade da plataforma permitiu o acompanhamento do progresso do trabalho.

5.2 Apresentação das métricas e análise de resultados

Para a apresentação dos resultados e avaliação da qualidade dos modelos propostos foram utilizadas as seguintes métricas: acurácia, precisão, sensibilidade e *F1-score*. E para medir o desempenho do modelo foi utilizada a matriz de confusão.

5.3 Resultados obtidos

A seguir são apresentados os resultados obtidos referentes aos modelos utilizados. Na Tabela 3 é possível visualizar as acurácias obtidas após a análise dos classificadores SVM e *Naive Bayes*.

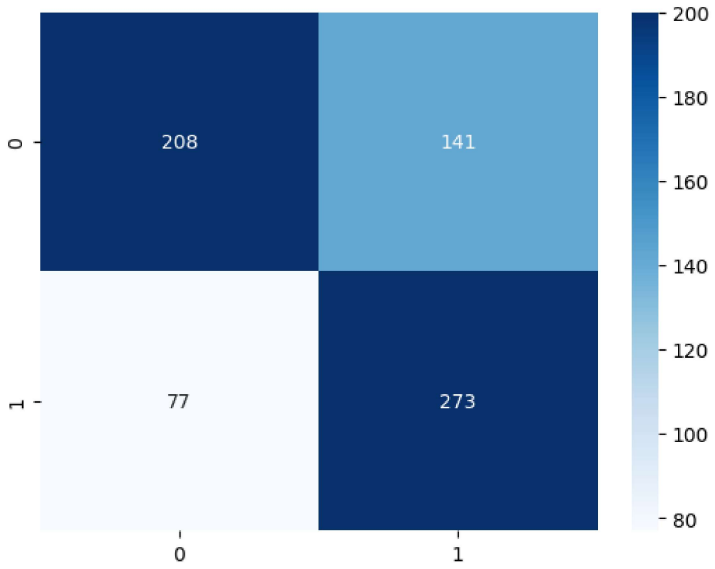
Tabela 3 – Acurácias obtidas dos classificadores.

	SVM	**Naive Bayes**
Acurácia	0.68	0.61

Fonte: Elaborada pelo autor.

A figura 7 apresenta a matriz de confusão referentes ao SVM. Como pode ser observado na matriz de confusão, referente ao SVM, o número de acertos para a classe 0 é de 208 em contradição com 141 predições erradas pelo modelo. Já na classe 1, obtivemos 273 predições acertadas pelo modelo e 77 predições erradas.

Figura 7 – Matriz confusão SVM.



Fonte: Elaborada pelo autor.

A tabela 4 apresenta os valores obtidos pelo classificador SVM para as métricas de precisão, revocação e f1-score para os pólos negativo e positivo.

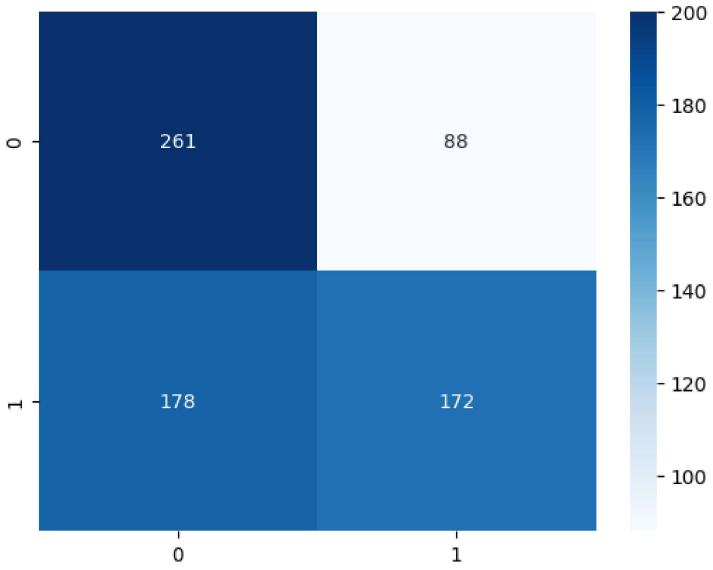
Tabela 4 – Valores obtidos pelo SVM.

	Precisão	**Revocação**	**F1 - Score**
Negativo	0.73	0.60	0.66
Positivo	0.66	0.78	0.71

Fonte: Elaborada pelo autor.

A seguir a figura 8 mostra a matriz de confusão referentes ao classificador *Naive Bayes*. Como pode ser observado na matriz de confusão, referente ao *Naive Bayes* (NB), o número de acertos para a classe 0 é de 261 em contradição com 88 predições erradas pelo modelo. Já na classe 1, o número de predições acertadas pelo modelo foi 172 e 178 foram as predições erradas.

Figura 8 – Matriz confusão NB.



Fonte: Elaborada pelo autor.

A tabela 5 apresenta os valores obtidos de precisão, revocação e f1-score para os pólos negativo e positivo no NB.

Tabela 5 – Valores obtidos pelo NB.

	Precisão	**Revocação**	**F1 - Score**
Negativo	0.59	0.75	0.66
Positivo	0.66	0.49	0.56

Fonte: Elaborada pelo autor.

O modelo *Naive Bayes* apresentou um resultado menos satisfatório, em relação ao SVM que apresentou métricas melhores desde a acurácia, após a verificação das outras métricas, comprovou-se que para o conjunto de dados do trabalho ele foi mais eficiente em sua classificação.

6 CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho apresenta a criação de um modelo para a análise de sentimentos da polarização das manchetes sobre notícias do mercado brasileiro. Para a realização do trabalho, a rotulação dos dados requereu muita atenção acerca da dinâmica das palavras, termos como baixa, queda, alta e aumento se repetiram com frequência.

Logo após, foram aplicadas técnicas de processamento de linguagem natural e aprendizado de máquina no desenvolvimento do método e na realização dos experimentos. O modelo foi desenvolvido na linguagem de programação *Python* utilizando a ferramenta *Natural Language ToolKit* (NLTK) para o auxílio na fase de pré-processamento das manchetes de notícias. A NLTK foi escolhida devido a fácil utilização e integração com a linguagem de programação *Python*.

Utilizou-se os classificadores *Support Vector Machine* e *Naive Bayes* para verificar qual dos modelos apresenta melhor resultado para os dados propostos (descritas na Seção 2.4). Dessa forma, a abordagem utilizando o classificador *Support Vector Machine* obteve os melhores resultados em relação ao classificador *Naive Bayes*.

Os resultados dos experimentos mostram a viabilidade da identificação automatizada de emoções através dos classificadores. São propostos como trabalhos futuros, incluir o uso de um conjunto de dados mais amplo e recente, adicionar a classe neutra, reaplicar os classificadores, e correlacionar a análise de sentimento com a oscilação dos preços das ações na bolsa de valores.

REFERÊNCIAS

AGARWAL, R. ***The 5 Classification Evaluation metrics every Data Scientist must know.*****(2019)**. 2020.

ALBUQUERQUE, C. L. D. S. **Análise de sentimento sobre comentários em sites de e-commerce no idioma português/br.** 2022.

B3. **Perfil pessoas físicas**. 2023. Disponível em: <https://www.b3.com.br/pt_br/market-data-e-indices/servicos-de-dados/market-data/consultas/mercado-a-vista/perfil-pessoas-fisicas/perfil-pessoa-fisica/>. Acesso em: 15 de maio de 2023.

BALAGE, P.; PARDO, T. A. S.; ALUÍSIO, S. ***an evaluation of the brazilian portuguese liwc dictionary for sentiment analysis.*** In: **Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology**. [S.l.: s.n.], 2013.

CASTRO, C. P. N. d. **Extração de dados e análise de sentimento:** com diferentes dicionários léxicos. Universidade Federal de São Carlos, 2022.

CORTES, O. A. C.; MELO, W. E. de O. **Utilizando Análise de Sentimentos e SVM na Classificação de Tweets Depressivos**. **Anais do Computer on the Beach**, v. 12, p. 102–110, 2021.

CUNNINGHAM, P.; CORD, M.; DELANY, S. J. ***Supervised learning machine learning techniques for multimedia.*** [S.l.]: Springer, 2008.

ECKER, U. K.; LEWANDOWSKY, S.; CHANG, E. P.; PILLAI, R. ***The effects of subtle misinformation in news headlines.*** **Journal of experimental psychology: applied**, American Psychological Association, v. 20, n. 4, p. 323, 2014.

GUIRALDELLI, L. A.; SÁ, M. C. P. de. **Estudando os efeitos da ambiguidade no discurso jornalístico manchete. Entrepalavras**, v. 4, n. 1, p. 82–98, 2014.

HAPKE, H.; HOWARD, C.; LANE, H. **Natural Language Processing in Action:** ***Understanding, analyzing, and generating text with Python.*** [S.l.]: Simon and Schuster, 2019.

INDURKHYA, N.; DAMERAU, F. J. ***Handbook of natural language processing.*** [S.l.]: Chapman and Hall/CRC, 2010.

KAGGLE. ***News of the Brazilian Newspaper***. 2023. Disponível em: <https://www.kaggle.com/datasets/marlesson/news-of-the-site-folhauol/>. Acesso em: 04 de junho de 2023.

KONNIKOVA, M. **What makes people feel upbeat at work. The New Yorker**, 2016.

LORENA, A. C.; CARVALHO, A. C. D. **Uma introdução às *support vector machines*. Revista de Informática Teórica e Aplicada**, v. 14, n. 2, p. 43–67, 2007.

NARAYANAN, R.; LIU, B.; CHOUDHARY, A. ***sentiment analysis of conditional sentences***. In: **Proceedings of the 2009 conference on empirical methods in natural language processing**. [S.l.: s.n.], 2009. p. 180–189.

OLIVEIRA, A. D. C. M. d. **Identificando emoções em manchetes de notícias escritas em português do Brasil utilizando *Naïve Bayes.*** 2016.

OLIVEIRA, D. J. S.; BERMEJO, P. H. d. S. **Mídias sociais e administração pública:** análise do sentimento social perante a atuação do governo federal brasileiro. **Organizações & Sociedade**, SciELO Brasil, v. 24, p. 491–508, 2017.

PADILHA, V. A.; CARVALHO, A. **Mineraçao de dados em python. Instituto de Ciências Matemáticas e de Computaçao da Universidade de Sao Paulo**, 2017.

PALMER, D. D. **Text Preprocessing.** ***Handbook of natural language processing.***, v. 2, p. 9–30, 2010.

PARDO, T. A. S.; NUNES, M. d. G. V. **Aprendizado Bayesiano Aplicado ao Processamento de Línguas Naturais.** 2002.

RAMOS, P.; REIS, J.; BENEVENUTO, F. **Uma Análise da Polaridade Expressa nas Manchetes de Notícias Brasileiras.** In: SBC. **Anais do V Brazilian Workshop on Social Network Analysis and Mining**. [S.l.], 2016. p. 187–198.

SHALEV-SHWARTZ, S.; BEN-DAVID, S. ***Understanding machine learning: From theory to algorithms.*** [S.l.]: Cambridge university press, 2014.

VIEIRA, R.; LOPES, L. **Processamento de linguagem natural e o tratamento computacional de linguagens científicas. Em corpora**, p. 183, 2010.

VIJAY, N.; SINGH, S.; MALHOTRA, G. ***sentiment analysis: gauging the effect of news on stock prices in indian stock market.*** **International Journal of Trade, Economics and Finance**, v. 9, n. 4, p. 148–152, 2018.