Kéthlyn Campos Silva

Relacionamento estatístico entre indicadores obtidos de dados de notícias, Google Trends e bolsa de valores brasileira

Goiânia
2022

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO (TECA) PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES

E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo das Teses e Dissertações disponibilizado na BDTD/UFG é de responsabilidade exclusiva do autor. Ao encaminhar o produto final, o autor(a) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do material bibliográfico

[ ] Dissertação [ ] Tese [x] Outro*:____artigo_____

*No caso de mestrado/doutorado profissional, indique o formato do Trabalho de Conclusão de Curso, permitido no documento de área, correspondente ao programa de pós-graduação, orientado pela legislação vigente da CAPES.

Exemplos: Estudo de caso ou Revisão sistemática ou outros formatos.

2. Nome completo do autor

Kéthlyn Campos Silva

3. Título do trabalho

Relacionamento estatístico entre indicadores obtidos de dados de notícias, Google Trends e bolsa de valores brasileira

4. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador)

Concorda com a liberação total do documento [x] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante:
a) consulta ao(à) autor(a) e ao(à) orientador(a);
b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo da tese ou dissertação.
O documento não será disponibilizado durante o período de embargo.
Casos de embargo:
- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

Obs. Este termo deverá ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **Deborah Silva Alves Fernandes**, **Professor do Magistério Superior**, em 15/09/2022, às 20:30, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **KÉTHLYN CAMPOS SILVA**, **Discente**, em 15/09/2022, às 21:00, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3194172** e o código CRC **033CF1C4**.

Kéthlyn Campos Silva

Relacionamento estatístico entre indicadores obtidos de dados de notícias, Google Trends e bolsa de valores brasileira

Trabalho de conclusão de curso apresentado na Escola de Engenharia Elétrica, Mecânica e de Computação como requisito para a conclusão do curso de Engenharia de Computação e obtenção do título de Engenheiro de Computação.

Universidade Federal de Goiás – UFG
Escola de Engenharia Elétrica, Mecânica e de Computação (EMC)

Orientadora: Prof.ª Dra. Deborah Silva Alves Fernandes

Goiânia
2022

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Silva, Kéthlyn Campos
Relacionamento estatístico entre indicadores obtidos de dados de notícias, Google Trends e bolsa de valores brasileira [manuscrito] / Kéthlyn Campos Silva. - 2022.
XVI, 16 f.: il.

Orientador: Profa. Dra. Deborah Silva Alves Fernandes; co orientadora Márcio Giovane Cunha Fernandes.
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Goiás, Escola de Engenharia Elétrica, Mecânica e de Computação (EMC), Engenharia da Computação, Goiânia, 2022.
Bibliografia.

1. Coleta de dados. 2. Google Trends. 3. Notícias. 4. Análise de Sentimento. 5. Mercado Financeiro. I. Fernandes, Deborah Silva Alves, orient. II. Título.

CDU 519.25

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

DECLARAÇÃO

Aos dias quatorze do mês de setembro de dois mil e vinte e dois, a partir das 14 horas, via sistema de webconferência do Google Meeting, realizou-se a sessão pública de Defesa de Projeto Final de Curso 2 (PFC2) intitulada **"**Relacionamento estatístico entre indicadores obtidos de dados de notícias, Google Trends e bolsa de valores brasileira**"** da aluna do curso de Engenharia de Computação Kéthlyn Campos Silva (matrícula: 201602423). Os trabalhos foram instalados pela Orientadora, Professora Doutora Deborah Silva Alves Fernandes (INF/UFG) com a participação dos demais membros da Banca Examinadora: Professor Doutor Sandrerley R. Pires; Professor Mestre Márcio Giovane C. Fernandes. Durante a arguição os membros da banca fizeram sugestão de alteração do título do trabalho. A Banca Examinadora reuniu-se em sessão secreta a fim de concluir o julgamento do PFC2, tendo sido a candidata **aprovada** pelos seus membros com nota 10,0. Proclamados os resultados pela Professora Doutora Deborah Silva Alves Fernandes, Presidente da Banca Examinadora, foram encerrados os trabalhos e, para constar, lavrou-se a presente ata que é assinada pelos Membros da Banca Examinadora. O documento Ata de Defesa PFC2-KETHELYNCAMPOS (3194163) anexado à esse processo contém as informações da banca.

Documento assinado eletronicamente por **Deborah Silva Alves Fernandes**, **Professor do Magistério Superior**, em 15/09/2022, às 20:21, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Márcio Giovane Cunha Fernandes**, **Usuário Externo**, em 15/09/2022, às 20:26, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Sandrerley Ramos Pires**, **Professor do Magistério Superior**, em 16/09/2022, às 16:51, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **3194154** e o código CRC **48D2591B**.

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE GOIÁS
ESCOLA DE ENGENHARIA ELÉTRICA, MECÂNICA E DE COMPUTAÇÃO
COORDENAÇÃO DE ESTÁGIO E PROJETO FINAL

ATA DE AVALIAÇÃO DE PROJETO FINAL

CURSO

( ) Eng. Elétrica ( ) Eng. Mecânica ( ) Eng. de Computação
( ) Projeto Final 1 (x) Projeto Final II

AVALIAÇÃO DE PROJETO FINAL

Título do projeto: Relacionamento estatístico entre indicadores obtidos de dados de notícias, Google Trends e bolsa de valores brasileira

BANCA AVALIADORA

Membro 1: Profa. Deborah Silva Alves Fernandes (INF-UFG)

Membro 2: Prof. Sandrerley Ramos Pires (EMC-UFG)

Membro 3: Prof. Márcio Giovane Cunha Fernandes (UEG)

ESTUDANTES

Matrícula	Nome
201602423	**Kéthlyn Campos Silva**
---	---

NOTAS

Matrícula	Membro 1				Membro 2				Membro 3				Média
	NPT	NTE	NAA	NF	NPT	NTE	NAA	NF	NPT	NTE	NAA	NF	
201602423	10	10	10	10	10	10	10	10	10	10	10	10	10,0
------	-	-	-	-	-	-	-	-	-	-	-	-	-

NPT – Nota plano de trabalho; NTE – Nota do trabalho escrito; NAA – Nota de apresentação e arguição

Para Eng. Elétrica, Mecânica e PFC2 da Eng. Da Computação: NF = 0,1 x NPT + 0,45xNTE + 0,45 x NAA

Para PFC1 da Eng. Da Computação: NF = 0,3 x NPT + 0,7 x NAA

Goiânia, 14 de setembro de 2022.

Profa. Dra. Deborah Silva Alves Fernandes

Membro 1

Prof. Dr. Sandreley Ramos Pires

Membro 2

Prof. Msc. Márcio Giovane Cunha Fernandes

Membro 3

FREQUÊNCIA – a ser preenchido pelo orientador(a)	
Nome do(a) estudante	Frequência (%)
Kéthlyn Campos Silva	100
-----	-----

Profa. Dra. Deborah Silva Alves Fernandes
Professor(a) Orientador(a)

ATA DE AVALIAÇÃO DE PROJETO FINAL - Observações

Preencher com modificações solicitadas, caso existam. Em caso de reprovação, informar a justificativa.

Relacionamento estatístico entre indicadores obtidos de dados de notícias, Google Trends e bolsa de valores brasileira

Kéthlyn C. Silva*, graduanda em Engenharia de Computação. Deborah S. A. Fernandes†, Professora Associada. Márcio Giovane C. Fernandes‡, Professor Associado. *EMC/UFG. †INF/UFG. ‡SI/UEG. E-mails: kethlyncampos@discente.ufg.br*, deborah.fernandes@ufg.br†, marcio.giovane@ueg.br‡

***Resumo*—Atividades de coleta, processamento e análise automática de dados não estruturados são amplamente utilizadas na academia e fora desta para a descoberta de conhecimento em grandes bases de dados. Neste artigo é detalhado o processo de coleta, pré-processamento e obtenção de indicadores de notícias, de dados do Google Trends e da bolsa de valores brasileira. A partir desses dados foram obtidos indicadores de sentimento de notícias para o quais foi aplicado o modelo CNN (*Convolutional Neural Networks*), com obtenção de F1-score de 96%. Quanto aos resultados, algumas correlações interessantes foram obtidas, das quais destacam-se uma correlação inversa entre o sentimento das notícias e o preço de fechamento ajustado classificada como "moderada" (de acordo com a escala de Cohen) em 2019; a relação entre volume de buscas em 2020 com o preço de fechamento e com o volume de negociações obteve um coeficiente negativo caracterizado como "muito grande" e um positivo definido como "grande", respectivamente. Além disso, foram notadas correlações inversas de caráter "grande" entre o preço de fechamento e o volume de negociações tanto em 2020 quanto em 2021.**

***Palavras-chave*—Coleta de dados, Google Trends, Notícias, Análise de Sentimento, Mercado Financeiro**

***Abstract*—Activities for the collection, processing and automatic analysis of unstructured data are widely used in academia and beyond for the discovery of knowledge in large databases. This article details the process of collecting, preprocessing and obtaining news indicators, data from Google Trends and the Brazilian stock exchange. From these data, indicators of news sentiment were obtained for which the CNN model (Convolutional Neural Networks) was applied, obtaining an F1-score of 96%. As for the results, some interesting correlations were obtained, of which we highlight an inverse correlation between news sentiment and the adjusted closing price classified as "moderate" (according to the Cohen scale) in 2019; the relationship between search volume in 2020 with the closing price and with the trading volume obtained a negative coefficient characterized as "very large" and a positive coefficient defined as "large", respectively. In addition, "large" inverse correlations were noted between the closing price and trading volume in both 2020 and 2021.**

***Index Terms*—Data Mining, Google Trends, News, Sentiment Analysis, Financial Market**

I. Introdução

COM a facilidade de acesso à internet proporcionada por uma época em que a tecnologia avança cada vez mais rapidamente, mais de 5 bilhões de usuários ([1]) geram quantidades progressivamente maiores de dados por meio diversas fontes como pesquisas em buscadores, redes sociais, aplicativos, entre outros. Por conseguinte, tanto empresas como orgãos governamentais e pessoas físicas despertaram para a importância da utilização dessas informações de um modo estratégico.

No contexto do mercado financeiro, o uso de indicadores, obtidos de diversas fontes, é uma ferramenta relevante no auxílio à tomada de decisão por parte dos investidores. Dessa forma, diversos estudos são realizados com o intuito de formular novos indicadores a partir de dados disponibilizados na internet, como feito por [2] ao utilizar o sentimento obtido através da análise de notícias publicadas na rede. Ademais, devido à baixa liquidez e alta volatilidade do mercado financeiro brasileiro, os preços não refletem todas as informações disponíveis, o que diverge da Hipótese de Mercados Eficientes apresentada por [3]. Esta divergência de preços abre margem para oportunidades de lucro, tornando-se atrativo para investidores do mundo todo, como descrito em [4].

Sendo o Brasil um país emergente, há um crescente interesse das pessoas pelo mercado financeiro refletido em buscas por notícias e conteúdo a respeito deste assunto. De acordo com [5], há mais de 5 milhões de pessoas físicas investidores na custódia da B3 (bolsa brasileira). O estudo e análise de dados coletados da internet, em especial textos em língua portuguesa é pertinente pois, trata-se de um problema de processamento de linguagem natural (PLN) com vários temas de pesquisa em aberto. O conceito de processamento de linguagem natural se mostra cada vez mais presente na vida das pessoas, seja através de uma pesquisa em um buscador da internet, na utilização de um tradutor, ou um comando de voz a uma assistente virtual. Esta tecnologia fornece uma melhor experiência para o usuário, uma vez que busca entender a linguagem do mesmo, sendo uma ferramenta fundamental para as empresas oferecerem um serviço agradável e eficiente aos seus clientes, uma vez que através da análise de sentimento, um subcampo do PLN, pode-se obter as opiniões dos consumidores.

Portanto, este trabalho apresentará: o processo de coleta de dados da internet (notícias em língua portuguesa, informações sobre motores de busca do Google, e da bolsa de valores brasileira B3); uma metodologia para obtenção de indicadores do mercado financeiro com base nos dados adquiridos; uma análise estatística e discussão sobre a relação entre os indicadores levantados. Nas seções seguintes estão descritas uma revisão bibliográfica, o método proposto, os resultados e discussões e, por fim, a conclusão é exibida na seção 5.

II. Revisão Bibliográfica

A opinião e o sentimento público têm ganhado cada vez mais relevância tanto no âmbito privado quanto público, vez que as pessoas tendem a definir suas opiniões baseadas nas opiniões de outras. Isso pode ser percebido no campo das redes sociais pelo poder de influência dos chamados agentes influenciadores, que interferem na formação de opinião de seus seguidores a partir de suas postagens. Os dados que são comumente coletados e utilizados para resgatar essa opinião, no contexto do mercado financeiro são: notícias, dados de redes sociais e dados de busca na web. Em [6] apresenta o uso da rede social Twitter como fonte de coleta. Uma grande dificuldade enfrentada na manipulação dos tweets foi o Processamento de Linguagem Natural, em função da falta de padronização na escrita dos tweets e da presença de erros ortográficos e truncamento de palavras. No pré-processamento, foram feitas filtragens de retweets, de links, de pontuação e de tweets contendo palavras e expressões selecionadas. Nesses dados aplicaram a análise de sentimento para classificação de sua polaridade em positiva e negativa. Com os indicadores obtidos através da polarização automática de tweets, construíram um simulador de compra e venda de ações que conseguiu demonstrar a efetividade destes no processo decisório para a aquisição desses ativos. Em [7], há o uso de dados do Twitter e notícias sobre o mercado financeiro brasileiro para uma análise estatítica do relacionamento entre indicadores de sentimentos obtidos através desses dados.

O trabalho de [8] elabora uma comparação de oito abordagens para extração de dados da web, tendo o algoritmo *Trinity* obtido melhor performance. Por meio desta técnica, foram extraídos textos contidos em sites de notícias, e ainda dados de todos os tipos de sites por meio de um *web crawler*, a fim de extrair links, e de árvore ternária para processar os dados extraídos.

Uma análise de diferentes técnicas de extração de dados é realizada em [9], havendo o Algoritmo Genético atingido um melhor desempenho na execução de *web crawling*, enquanto o sistema *WHISK* se revelou mais vantajoso como *web wraper* dentre os sistemas de *machine learning* analisados. Além disso, é destacada a importância da fase de manutenção das ferramentas de coleta, em que há uma verificação para atestar que a mesma está funcionando e não foi afetada por mudanças na estrutura das páginas.

Em [10] é proposta a realização de web scraping utilizando expressões regulares para extração do título, data de publicação, autor, artigo da notícia e URL das notícias publicadas nos sites Detik, Tribunnews e Liputan6, uma vez que estes são os sites de notícias mais acessados pelos usuários na Indonésia, de acordo com o *Alexa Traffic Rank* (ATR). Para os sites Detik e Tribunnews, a precisão e recall atingida pela estratégia foi igual a 1 e o F-Measure de 100%, enquanto que para o site Liputan6 a precisão e recall foi igual a 0.95 e F-Measure de 95%.

Os dados provindos da internet são de grande valia para diversas aplicações, porém especialmente para empresas, pois com a análise destes dados é possível a criação de estratégias para aceleração do crescimento das empresas. Deste modo, a linguagem Python tem sido utilizada como ferramenta na tarefa de extração destes dados, uma vez que possui uma grande quantidade de bibliotecas e comunidade empenhada. Em [11] os pesquisadores propõem uma abordagem para extração de dados não estruturados da web utilizando a linguagem Python 3.6 e o software de *web scraping* Scrapy, com o intuito de fazer uma análise da informação extraída. Esse processo é realizado em três etapas: o *web crawler* navega pelos links desejados da web, a informação é extraída dos links das fontes e, então, armazenada em um arquivo csv. Na criação do script de *web crawling* foi usado o método XPath para encontrar detalhes dos elementos das Pesquisas Frequentes da rede social Reddit. Utilizando o script escrito, a extração dos dados foi feita com grande facilidade e o resultado da análise dos dados foi apresentado na forma de porcentagem, contendo os assuntos mais procurados no site.

Os autores de [12] propõem um estudo de técnicas de web scraping com intuito de extrair declarações de líderes de governo, sobre o COVID-19, de sites de notícias para a realização de uma análise de sentimentos e emoções, e para verificar o efeito destas declarações no número diário de casos. Como os *web scrapers* existentes não atendiam aos requisitos dos autores, foi construída uma nova metodologia para criação de um web scraper com analisadores (parsers) existentes. Uma combinação das bibliotecas Rvest e RSelenium foi escolhida para análise de HTML, uma vez que Rvest é de fácil uso e o RSelenium trabalha bem com páginas feitas com Javascript. O pacote stringr foi usado para manipular e limpar os dados. Os dados foram armazenados no formato csv. O site de notícias para extração utilizado foi o site da CNN. Uma vez que todas as notícias relacionadas ao COVID-19 para cada dia foram obtidas, a acurácia da metodologia apresentada foi de 100%.

Uma quantidade massiva de dados, de cerca de 2.5 quintilhões de bytes, é gerada diariamente pelas pessoas ao utilizarem a internet. A análise destes dados se faz necessária para sua aplicação em campanhas mais direcionadas, melhoria de produtos, etc. Para extração destes dados das páginas web, uma ferramenta muito utilizada é o web scraping, contudo a maioria das técnicas de web scraping exige um bom conhecimento da estrutura do DOM. [13] apresenta formas alternativas de realizar o web scraping utilizando Processamento de Linguagem Natural (NLP) e Machine Learning (ML) para contornar a necessidade de se conhecer a estrutura do DOM. A linguagem Python e suas bibliotecas, como SpaCy e Natural Language Toolkit, são utilizadas para o processamento do HTML das páginas desejadas. A primeira abordagem apontada é a sumarização do texto, na qual as palavras recebem pesos, é criado corpus e as sentenças são classificadas de acordo com seus scores, sendo assim as sentenças que possuem os maiores scores são escolhidas para compor o sumário do texto. O sumário é filtrado utilizando palavras-chave, que neste caso são relacionadas à epidemia do corona. Informações como país, estado e cidade foram extraídas utilizando a biblioteca pycountry. A segunda abordagem apresentada é utilizando Named Entity Recognition (NER) que classifica as palavras e segmentos de sentenças em entidades pré-definidas. O texto passa pelo modelo NER en_core_web_sm da biblioteca SpaCy, que retorna as palavras classificadas em diferentes

entidades. Os autores utilizaram a sumarização do texto e o Named Entity Recognition para extração de dados de notícias para a construção de um preditor de epidemias. A sumarização do texto não é tão eficiente quanto os métodos convencionais e falha ao lidar com dados discretos e altamente não-estruturados. Em contrapartida, o Named Entity Recognition lida bem com dados discretos e altamente estruturados, pois é mais flexível e possui uma boa acurácia que se aproxima dos métodos convencionais.

No artigo de [14] efetuou-se combinações de indicadores de análise técnica com a análise de sentimento das notícias relacionadas às ações da Apple Inc. (AAPL), Amazon.com Inc.(AMZN), e Netflix Inc.(NFLX), com o intuito de realizar uma predição de tendência do mercado financeiro. Para esse objetivo, extraiu os dados sobre as ações por meio do Yahoo Finance e aplicou os algoritmos de *Machine Learning* Regressão Linear, *Random Forest* e *Gradient Boosting Machine*. A técnica que atingiu melhor desempenho foi o *Random Forest* obtendo uma acurácia de 0,6358.

A implementação do algoritmo de aprendizado profundo (*deep learning*) *Convolutional Neural Networks* (CNN) foi proposta por [15] para a realização de classificação de notícias indonésias, divididas em quatro categorias: "hiburan" (entretenimento), "olahraga" (esporte), "tajuk utama" (manchete), e "teknologi" (tecnologia). Para esta finalidade, o modelo CNN alcançou acurácia média de 90,74%.

Devido à falta de base de dados para realização de pesquisas relacionadas ao processamento na língua bengali, [16] desenvolveu um conjunto de dados baseados em comentários de notícias bengalis do portal *Prothom-Alo*. Como avaliação, para atestar a qualidade desses dados, implementaram os modelos SVM (*Support Vector Machine*, CNN e LSTM (Long Short-term Memory), os quais obtiveram valores de F1-score iguais a 63.97%, 63.68% e 79.291%, respectivamente.

Tendo em vista o impacto que às vezes as notícias tem sobre o mercado financeiro, [17] realizou um estudo com ênfase na análise de sentimentos de notícias relativas ao mercado. Para isto, utilizou uma base com 5.000 sentenças classificadas em positivo, neutro e negativo, a qual foi submetida a um pré-processamento. Com relação à análise de sentimentos, comparou-se a acurácia de algoritmos de *machine learning* e *deep learning* como SVM, KNN (*K-Nearest Neighbors*), árvore de decisão, Random forest, XGBoost (*eXtreme Gradient Boosting*), CNN e LSTM, cujas acurácias atingidas foram 0.73, 0.71, 0.65, 0.71, 0.72, 0.75, 0.72, nesta ordem. Sendo assim, o modelo CNN obteve melhor performance frente aos algoritmos aplicados.

Uma abordagem de implementação de algoritmos de *deep learning* para análise de sentimento de textos de língua amárica é descrita em [18]. Através de um software de exportação de comentários, constituiu-se um conjunto de dados com 10.000 itens adquiridos por meio da página do Facebook da Amhara Media Corporation, uma organização de mídia do governo etíope. As informações foram armazenadas em formato excel. Os modelos implementados atingiram acurácias iguais a 93.49%, 99.10%, 98.98% e 99.09% correspondentes de forma respectiva aos seguintes algoritmos: CNN, LSTM, CNN-LSTM e CNN-GRU.

Com o propósito de construir um preditor de movimento do índice Standard & Poor's 500, [19] propôs a implementação de um modelo RCNN (*Recurrent Convolutional Neural Networks*) baseado em arquiteturas como o CNN e RNN (*Recurrent Neural Networks*). Utilizaram como entradas um conjunto de indicadores de análise técnica e títulos de notícias do mercado financeiro. No que se refere aos resultados, enquanto o RNN se mostrou melhor em entender o contexto nos textos e em modelar características temporais complexas para a previsão do mercado, o CNN exibiu maior eficiência na interpretação da semântica. Além disso, o modelo RCNN proposto utilizou como entrada apenas informações do dia anterior ao dia a ser previsto e se destacou perante outros modelos que usam tanto o dia, como semana e mês precedentes. O resultado atingido fortaleceu a hipótese de que as notícias oferecem impacto por um curto período de tempo sobre o mercado.

Em [20] são exploradas as relações entre os anúncios de *lockdown* realizados pelas autoridades do Chile durante a pandemia do COVID-19 e variáveis como a resposta no Twitter a esses anúncios em um nível municipal, a intensidade de volume de buscas (obtidas por meio do Google Trends) e as reações do mercado financeiro. Com isso, o sentimento social relacionado às postagens no Twitter demonstraram uma relação negativa com o aumento de pessoas confinadas, ao passo que o indicador do Google Trends a demonstrou de forma positiva. Além disso, foi observado que a heterogeneidade dos sentimentos espelha a heterogeneidade das reações do mercado aos anúncios.

A relação entre o volume de buscas no Google Trends (GT) e os mercados financeiros durante o surto de COVID-19 foi estudada por [21]. Os autores adquiriram os índices do Google Trends por país acerca do tópico coronavírus de janeiro a abril de 2020. Os seis países considerados para a análise foram Estados Unidos, Espanha, Itália, Reino Unido, Alemanha e França, devido ao maior impacto referente ao número de casos. Os resultados revelaram que a maior parte dos índices possuem um poder explicativo sobre os mercados financeiros. No entanto, o índice do GT da Itália sobre o COVID-29 se mostrou o mais impulsionador perante os outros mercados levados em conta.

III. Método proposto

O experimento foi realizado em quatro fases, conforme apresentado na Figura 1. A primeira remete à coleta de dados, a segunda exibe a preparação do classificador, a terceira demonstra o processo de rotulação das notícias coletadas e a quarta aborda a análise estatística. As fases e etapas do experimento serão referenciadas por números etre parênteses "( )" desta seção em diante, conforme a Figura 1.

A. Aquisição de dados

As linguagens de programação Python e R são comumente empregadas na construção de ferramentas de *web scraping*, dado que possuem bibliotecas específicas para essa tarefa, sendo estas as linguagens mais populares em ciência de dados, como citado em [12]. Em todas as fases deste trabalho, a

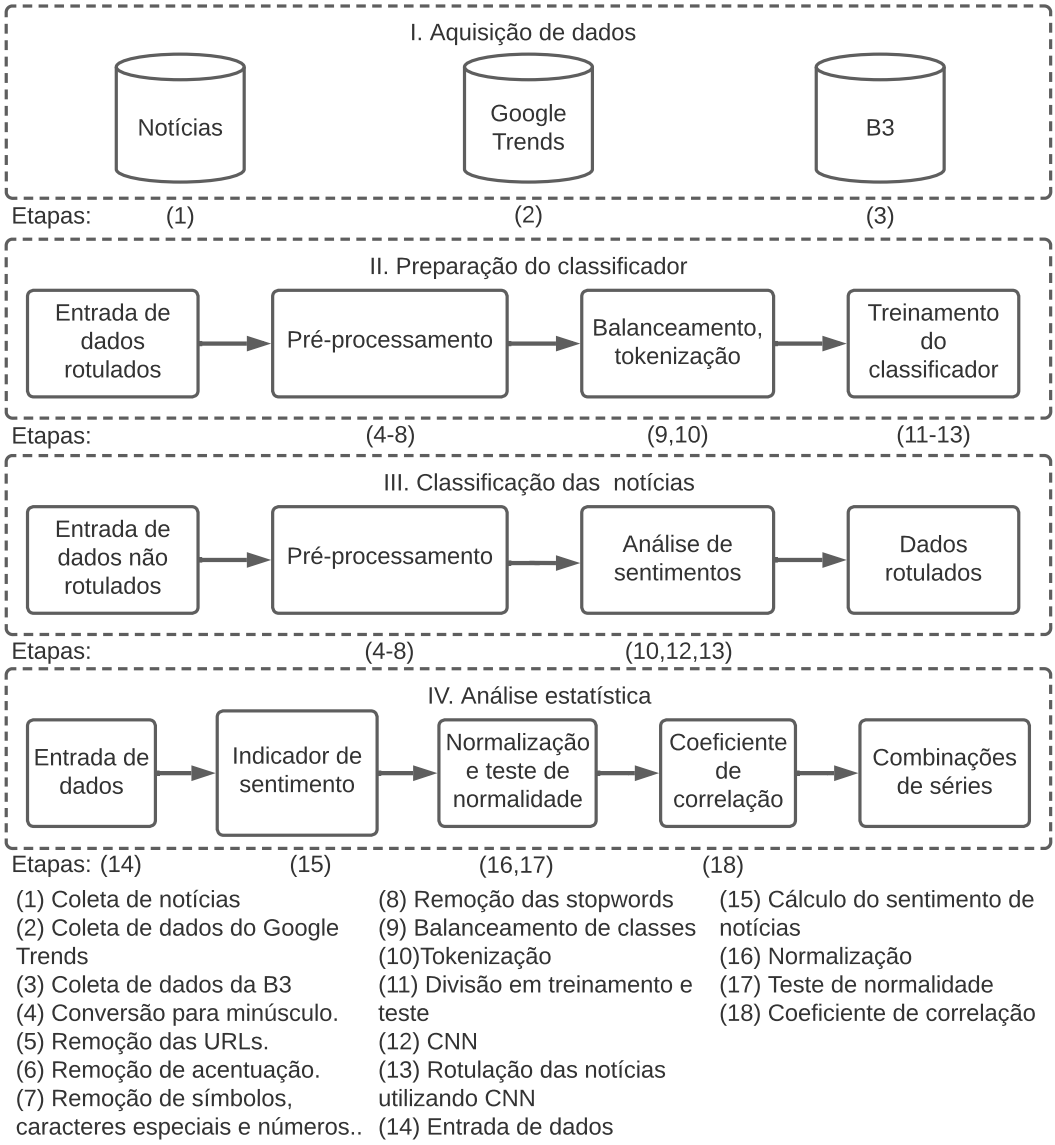

Figura 1: Fases e etapas da metodologia proposta.

linguagem Python foi utilizada em razão de possuir uma comunidade grande, de acordo com [11], o que facilita a resolução de possíveis dificuldades encontradas durante a execução. As bibliotecas Beautiful Soup, Selenium e Requests foram utilizadas na extração de dados, sendo a BeautifulSoup usada para análise de HTML, a Selenium para navegação entre as páginas e a Requests para efetuação de requisições nas páginas web.

1) Coleta de textos de notícias na web: Para realização da coleta de artigos de notícias da internet, várias estratégias podem ser elegidas. Normalmente, este tipo de site em específico possui uma estrutura muito padronizada, com classes CSS (*Cascading Style Sheets*) únicas para os atributos alvo mais comuns como título, data de publicação, autor e artigo da notícia, o que facilita a construção de um *web crawler* e *web scraper* para navegar entre os links desejados e extrair as informações almejadas, respectivamente. Além das classes CSS, outros atributos da estrutura DOM (*Document Object Model*) da página podem ser utilizados para localização da tag HTML em que se encontra a informação pretendida, tais como id, *name* e nome da tag HTML, ademais a linguagem XPath é uma ferramenta interessante para o acesso aos nós da árvore DOM que contém a informação.

Alguns sites de notícias fornecem um recurso chamado RSS (*Rich Site Summary* ou *Really Simple Syndication*), que é um padrão desenvolvido na linguagem XML para dispor o conteúdo dos sites de maneira resumida. O consumo desse recurso é bem simples de ser realizado e pode ser útil para várias aplicações. Uma das aplicações, que fazem uso do RSS, mais conhecidas são os sistemas de notícias para TV's localizadas em pontos estratégicos, como recepções e elevadores. Não obstante, não são todos os sites de notícias

que oferecem este artifício e ele é, de certa forma, limitado, posto que não possibilita a personalização da busca por data, retornando apenas as notícias mais recentes.

Os métodos descritos acima realizam de forma muito eficaz a tarefa de obtenção dos dados, todavia o uso do RSS é restrito ao fornecimento do recurso pelos sites e é limitado pela data. Já a abordagem que tem como base a estrutura da página, é sensível a alterações na mesma, de maneira que a ferramenta coletora desenvolvida pode tornar-se ineficaz. Desse modo, como descrito em [9], é de grande importância a manutenção da ferramenta, tornando-se esta inviável quando a quantidade de sites diferentes é muito grande em virtude da alta demanda de esforço humano e do custo temporal, uma vez que cada site possui um layout distinto.

Por conseguinte, para projetos com maior escala, nos quais é necessária a extração de dados de diversos sites, a implementação de Inteligência Artificial e Machine Learning contorna o problema de manutenção, uma vez que não é necessário o conhecimento da estrutura DOM da página, consoante com o trabalho realizado em [13].

Para a realização da etapa (1) da fase de aquisição dos dados (I), os sites Yahoo Finanças, Investing.com e Money Times foram escolhidos como fonte devido ao volume de acessos por parte dos usuários. A abordagem de utilização do DOM, neste caso, é interessante, uma vez que o número de sites selecionados é pequeno, o que leva a uma facilidade de manutenção e menor custo temporal para desenvolvimento de código.

A seção relacionada a bolsa de valores do site Yahoo Finanças faz uso do recurso *scroll infinito*, em que o conteúdo é carregado conforme há a rolagem da página. Por meio de métodos da biblioteca Selenium, a solução apresentada por [12] para este cenário foi implementada, cujo algoritmo pode ser visualizado em Algoritmo 1.

Algoritmo 1: Função de rolagem de página

```
def scrollar(driver):
    lastHeight = driver.execute_script
    (''return document.body.scrollHeight
        '')
    while True:
        driver.execute_script(
        ''window.scrollTo
        (0, document.body.scrollHeight)
            ;'')
        sleep(5)
        newHeight = driver.execute_script
        (''return document.body.
            scrollHeight'')
        if newHeight == lastHeight:
            break
        lastHeight = newHeight
```

Desta forma, a página é rolada calculando-se a altura de rolagem, a qual é armazenada em uma variável, sendo assim a altura atualizada pode ser comparada com altura inicial e uma nova rolagem é feita para aquela extensão. Este processo é iterado dentro de um loop, o qual é finalizado à medida que não há mais conteúdo a ser carregado. Os links de notícias são coletados da página através da biblioteca Beautiful Soup utilizando uma classe CSS específica para selecioná-los e usa-se a Requests para fazer uma requisição HTTP em cada link de notícia. A Beautiful Soup é, então, novamente usada para obtenção dos dados das notícias contidas em cada link. O processo descrito pode ser observado na Figura 2. Os dados adquiridos de cada notícia do Yahoo Finanças foram: título, artigo, autor, data e hora de publicação, data e hora de coleta e URL. Considerando que o Yahoo Finanças apenas disponibiliza as notícias mais recentes nesta seção, as 501 notícias adquiridas correspondem apenas a datas próximas ao período de coleta, de forma que o período de coleta se deu de 28/11/2021 às 13:16:22 até 17/12/2021 às 18:51:02 e o intervalo de data e hora de publicação das notícias coletadas é de 23/11/2021 às 17:28:25 até 17/12/2021 às 18:11:56.

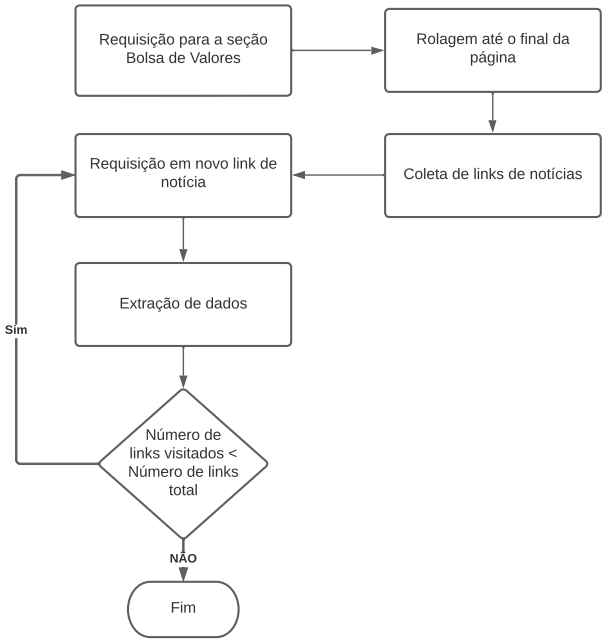

Figura 2: Fluxo de coleta de dados do site Yahoo Finanças

Assim como o Yahoo Finanças, o site Infomoney possui uma estrutura de scroll infinito semelhante, carregando mais conteúdo ao clicar no botão “Ver mais”, até que não haja mais conteúdos a serem carregados. Deste modo, a mesma técnica utilizada no site anterior é aplicada. Ao rolar a página da seção “Mercados”, o botão “Ver mais” é clicado e há uma nova rolagem, esse processo é repetido até que o botão não apareça mais, indicando o fim do conteúdo disponível. Os links das notícias são adquiridos através da biblioteca Beautiful Soup por meio de uma classe CSS, em seguida a biblioteca Requests é usada para realizar requisições HTTP em cada link e, então, os dados são adquiridos por meio da Beautiful Soup. Na Figura 3 é possível visualizar a sequência de passos relatada. Os dados obtidos foram: título, subtítulo, artigo, autor, data e hora de publicação, data e hora de coleta e URL. A data e hora de publicação das 812 notícias adquiridas constitui o intervalo

de 10/11/2021 às 16:33:09 até 17/12/2021 às 19:04:14, tendo o período de coleta se dado de 28/11/2021 às 13:57:19 até 17/12/2021 às 21:34:45.

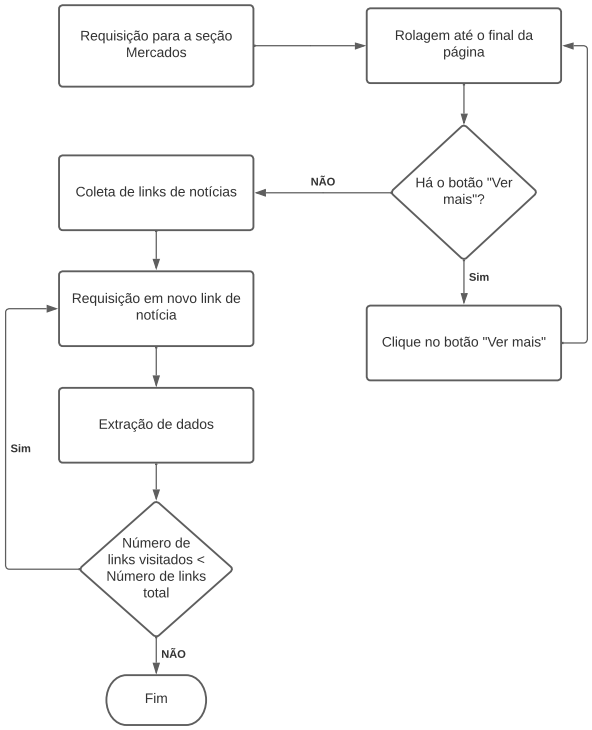

Figura 3: Fluxo de coleta de dados do site Infomoney

O Investing.com faz uso da paginação para dispor o seu conteúdo, portanto é necessária a navegação por meio dessa paginação para visualizar todas as notícias dispostas. Deste modo, na seção “Ações”, há a rolagem até o fim da página, em seguida os links das notícias são coletados por meio da biblioteca Beautiful Soup. Por meio da *Requests* são realizadas requisições HTTP nos links e os dados são obtidos com a Beautiful Soup. Posteriomente, é clicado no botão “Próximo”, o qual carrega uma nova sequência de notícias precedentes. Este processo, cujo fluxo pode ser observado na Figura 4, é repetido até que surjam notícias não pertencentes ao período determinado, ou seja, anteriores a 2019. Foram extraídas 43.143 notícias detalhadas em: título, artigo, data e hora de publicação, data e hora de coleta e URL. O intervalo de data de publicação é de 01/01/2019 às 08:45:00 até 17/12/2021 às 20:45:00, sendo o período de coleta de 03/12/2021 às 20:57:49 até 17/12/2021 às 21:54:29.

Tendo sido o único site, dentre os escolhidos, que disponibilizava notícias de 2019 a 2021, desta seção em diante as notícias utilizadas nas análises serão apenas provenientes do Investing.com e filtradas por datas comerciais da bolsa.

2) Coleta de dados de buscadores da web - Google Trends: De acordo com o site Internet Live Stats [22], mais de 104 mil pesquisas são processadas pelo Google a cada segundo, ou seja, mais de 8 bilhões de pesquisas são realizadas diariamente no buscador que domina cerca de 65.87% do mercado, segundo o Statcounter GlobalStats [23]. Sendo assim, o Google

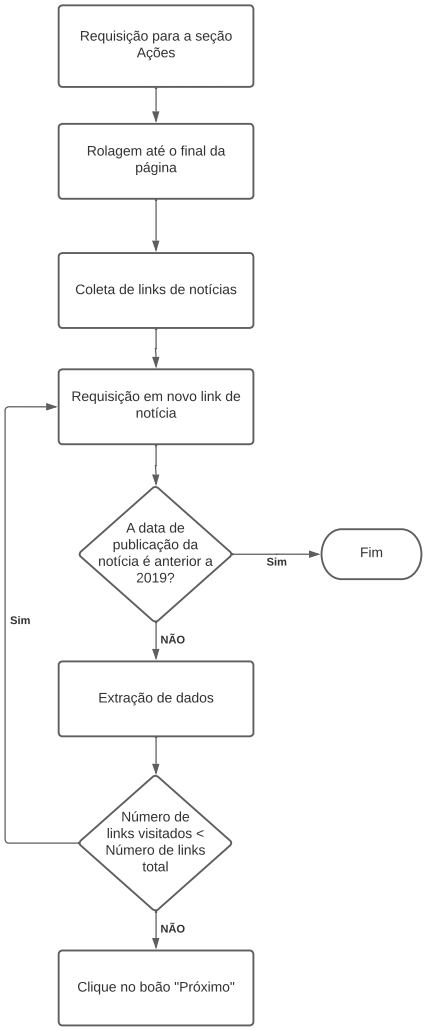

Figura 4: Fluxo de coleta de dados do site Investing.com

Trends é um serviço disponibilizado pelo Google que mostra a intensidade de buscas dos seus usuários por um termo específico em um determinado período e região, a qual é disposta em uma escala de 0 a 100, de modo que 100 representa o pico de interesse e 0 exprime absoluto desinteresse. Ademais, proporciona a exportação dos dados no formato csv, sendo estes arquivos seccionados de acordo a data e região. Esta exportação pode ser feita de forma manual para a pesquisa de poucos termos, contudo no contexto de vários termos é interessante o desenvolvimento de um *web scraper* que realize a extração automática desses dados, reduzindo custo temporal e esforço humano.

Posto isso, foi realizada uma coleta de dados acerca dos termos relacionados ao Ibovespa e às ações que o compõem. Para o desenvolvimento do coletor, a biblioteca Selenium foi utilizada para executar as ações de pesquisar cada termo definido e fazer download das informações disponibilizadas, com relação a período de tempo e à região. Os filtros foram aplicados tendo como região o Brasil e como período de 01/01/2019 à 17/12/2021, que corresponde ao intervalo de data de publicação das notícias obtidas, tendo a coleta se dado no

Tabela I: Dados das notícias adquiridas

Fonte	Quantidade de notícias	Período de publicação	Período de coleta
Yahoo Finanças	501	23/11/2021 17:28:25 - 17/12/2021 18:11:56	28/11/2021 13:16:22 - 17/12/2021 18:51:02
Infomoney	812	10/11/2021 16:33:09 - 17/12/2021 19:04:14	28/11/2021 13:57:19 - 17/12/2021 21:34:45
Investing.com	43.143	01/01/2019 08:45:00 - 17/12/2021 20:45:00	03/12/2021 20:57:49 - 17/12/2021 21:54:29

dia 29/12/2021.

Com o intuito de gerar um indicador a partir dos dados do Google Trends adquiridos, primeiramente foi calculada a média dos índices de cada dia de acordo com os termos pesquisados, conforme descrito em [20], de maneira que quanto maior este resultado maior o interesse dos usuários pelo mercado naquele determinado dia. É possível observar os dados adquiridos com as referidas médias na Figura 5. Em seguida, como o Google Trends disponibilizou os índices com intervalo de 7 dias para o período determinado, foi atribuído a cada dia comercial da B3 o valor da média de volume de buscas do intervalo ao qual o dia em questão estava contido. Desta forma, obteve-se o indicador de intensidade de buscas por dia comercial.

3) Coleta de dados da B3: O processo de tomada de decisão de compra e venda de ações envolve a análise de dados acerca dessas ações. As informações necessárias para o auxílio desse processo são o preço de abertura e de fechamento, preço máximo e mínimo e volume de negociação. Outro dado importante a ser considerado é o preço de fechamento ajustado, que leva em consideração os processos de agrupamento ou desagrupamento das ações e os dividendos providos pela empresa em questão. A aquisição dessas informações pode ser feita de forma manual, entretanto exige demasiado trabalho dependendo da quantidade de ações a serem analisadas e do tamanho do período a ser analisado, pode ser feita de forma automática com a contrução de um *webscraper* ou com o consumo de API's que forneçam estes dados. A bolsa de valores brasileira, B3, onde estão listadas as ações de empresas brasileiras, possui uma API, contudo seu acesso é pago.

Uma alternativa gratuita para adquirir os dados necessários é através do consumo da API Yfinance, que realiza um *scraping* do site Yahoo Finance e disponibiliza os dados históricos e em tempo real de vários mercados financeiros de uma forma simples. A desvantagem na utilização desta API é o fato de não ser uma API oficial, portanto não há garantia de que futuramente possa estar funcionando, contudo é amplamente utilizada.

Devido às vantagens acima mencionadas, a API Yfinance foi utilizada para a obtenção das informações acerca do Ibovespa e das ações que o constituem. O índice Ibovespa (IBOV) envolve as principais ações da B3, por conseguinte é o principal indicador de desempenho das ações negociadas na bolsa. Com relação ao intervalo de 02/01/2019 à 17/12/2021, foram obtidas as informações de volume, preço de abertura, de fechamento, de máximo, de mínimo e de fechamento ajustado, utilizando o intervalo de um dia aderido por muitas análises. Uma amostra destes dados pode ser notada na Tabela II. A Equação (1) foi aplicada para calcular a variação de preços diária da bolsa, utilizada como indicador. É válido ressaltar que neste cálculo empregou-se o preço de fechamento ajustado.

$$\text{Variação } (\%) = \frac{\text{Preço de Fechamento}}{\text{Preço de Abertura}} - 1 \qquad (1)$$

B. Pré-processamento

Com base nas notícias adquiridas do site Investing.com com data de publicação correspondente às datas comerciais da B3, produziu-se a nuvem de palavras exibida na Figura 6, a qual apresenta as palavras com maior frequência no corpus de notícias antes do pré-processamento. Ao observar esta figura, pode-se notar a grande quantidade de pontuação e *stopwords* contidas nos textos. *Stopwords* são palavras vazias dentro de um texto, ou seja, palavras que não agregam significado à frase, como preposições, artigos, etc. Levando em conta o trabalho de [14], na fase (II) de preparação do classificador da Figura 1, os textos das notícias foram convertidos para minúsculo (4) e as URLs (5) bem como a acentuação (6), símbolos, caracteres especiais, números (7) e stopwords (8) foram removidos. A lista de palavras vazias da língua portuguesa da biblioteca NLTK do Python foi aplicada nesta última etapa, no entanto a palavra "não" contida nesta lista foi mantida nos textos das notícias. O resultado do pré-processamento realizado no corpus de notícias pode ser analisado por meio da nuvem de palavras apresentada na Figura 7.

O agrupamento de informações de treinamento empregado constituiu-se da combinação de notícias relacionadas ao contexto do mercado financeiro classificadas em positivo e negativo dispostas por [24], [25] e [26]. Tendo em vista a discrepância na quantidade de notícias positivas e negativas, aplicou-se a técnica *Random Undersampling* (9), que consiste na remoção aleatória de elementos pertencentes à classe majoritária de forma que se equipare a quantidade de elementos da categoria minoritária. A fim de robustecer a base de treinamento, o método aumento de dados foi utilizado, o qual baseia-se na conversão das notícias em vetores, em seguida estes são embaralhados de maneira aleatória gerando assim novas informações. Subsequentemente, usou-se a função Tokenizer da biblioteca Keras para tokenização (10) da base, com o intuito de vetorizar o corpus. Este processo de tokenização envolve a conversão de uma sequência de textos em tokens discretos.

C. Análise de sentimentos para notícias

Conforme citado em [27], modelos de *deep learning* são conhecidos por atingirem bons resultados em uma ampla gama de aplicações de inteligência artificial. Embora tenha sido

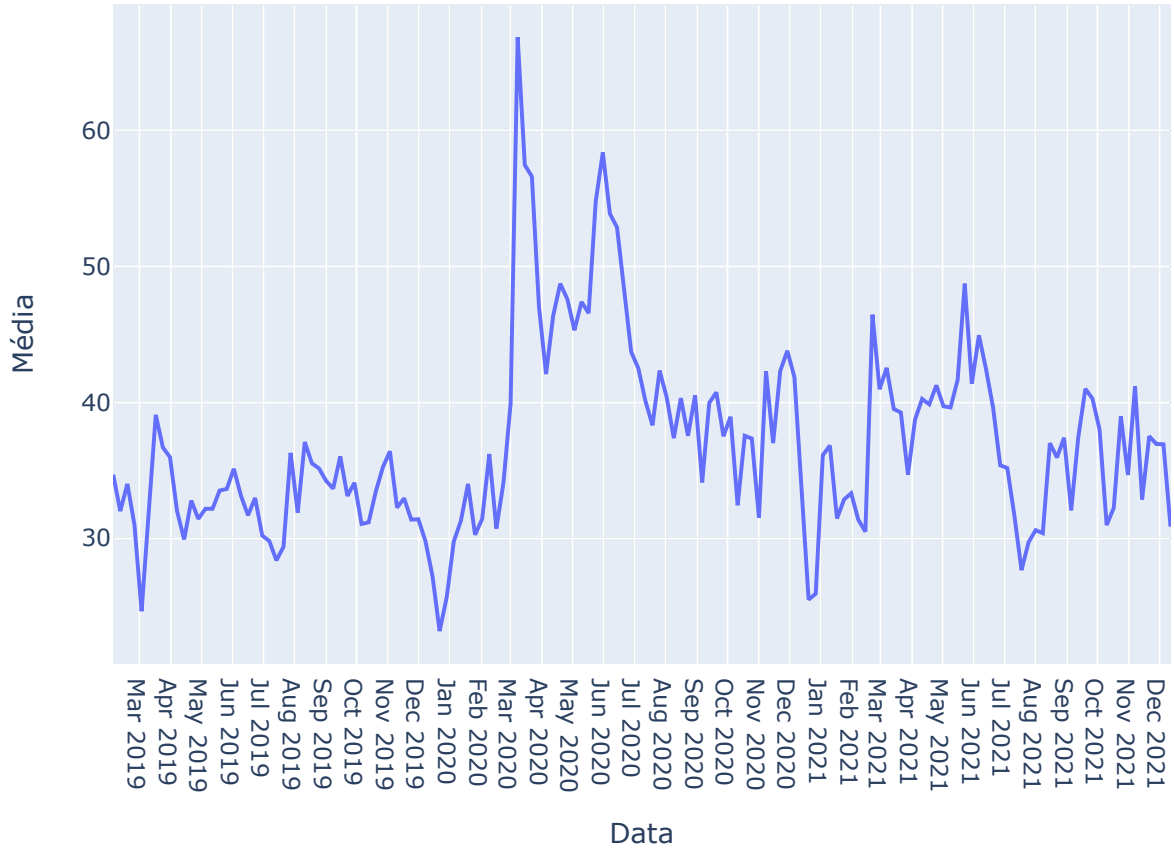

Figura 5: Dados coletados do Google Trends

inicialmente utilizado em aplicações de visão computacional, o algoritmo de *deep learning Convolutional Neural Network* (CNN) tem demonstrado eficiência em tarefas de processamento de linguagem natural (PLN), atingindo resultados satisfatórios em análise de semântica, modelagem de sentença, recuperação de consultas de pesquisas e outras aplicações de PLN, como descrito em [28]. Este algoritmo não necessita de um pré-processamento tão rigoroso, uma vez que dispõe da habilidade de aprender as características do texto, o que pode ser considerado uma vantagem. Ademais, possui padrões semelhantes às conexões dos neurônios humanos em sua arquitetura.

Sendo assim, um determinado modelo pode ser expandido ao fazer uso de várias redes neurais convolucionais paralelas, as quais fazem a leitura do texto empregando diversos tamanhos de kernel. Isto gera uma rede neural convolucional multicanal que utiliza diferentes tamanhos de conjuntos de palavras (*n-gram*) para ler o texto. A quantidade de palavras a ser levada em consideração, conforme a convolução passa pelo texto de entrada, é determinada pelo tamanho do kernel, o qual define um parâmetro de agrupamento.

Um modelo padrão de classificação possui as seguintes camadas: *embedding layer* como entrada, uma CNN unidimensional, uma de *pool* e uma de saída. Dessa forma, uma CNN multicanal para a tarefa de classificação utiliza múltiplas variações do modelo padrão com diferentes tamanhos de kernel, o que leva ao texto ser processado em distintos *n-grams* simultaneamente, ao mesmo tempo em que o modelo aprende a como integrar da melhor forma essas análises.

Para a classificação de notícias, implementou-se o algoritmo[1] CNN multicanal com o auxílio da biblioteca Keras. Isso posto, tendo o conjunto de treinamento sido pré-processado de acordo com as etapas descritas na subseção anterior, foi dividido (11) em treinamento (80%) e teste (20%) utilizando o método *train_test_split* do módulo Scikit-learn, e assim o modelo foi treinado (12).

Posteriormente, a fase (III) apresentada na Figura 1 foi realizada. Conforme descrito anteriormente, as notícias coletadas foram pré-processadas, e rotuladas (13) utilizando o modelo já treinado.

[1]Adaptado de Dhupar, R. (2018). Deep stack models for finance news f1-score .94. URL: https://www.kaggle.com/code/rohandx1996/deep-stack-models-for-finance-news-f1-score-94 .

Tabela II: Amostra dos dados da B3 coletados

Data	Abertura	Máximo	Mínimo	Fechamento	Fechamento Ajustado	Volume	Variação %
02/01/2019	87887.0	91479.0	87536.0	91012.0	91012.0	5615400	0.035557
03/01/2019	91011.0	91596.0	89922.0	91564.0	91564.0	6066200	0.006076
04/01/2019	91577.0	92701.0	90824.0	91841.0	91841.0	5322100	0.002883
07/01/2019	91845.0	92552.0	91288.0	91699.0	91699.0	4565000	-0.00159
08/01/2019	91699.0	92231.0	91064.0	92032.0	92032.0	4496500	0.003631
09/01/2019	92033.0	93626.0	92028.0	93613.0	93613.0	5011400	0.017168
10/01/2019	93599.0	93987.0	93050.0	93806.0	93806.0	4840600	0.002212
11/01/2019	93806.0	93961.0	93359.0	93658.0	93658.0	5330500	-0.001578
14/01/2019	93645.0	94474.0	93335.0	94474.0	94474.0	4486900	0.008853
15/01/2019	94474.0	94695.0	93402.0	94056.0	94056.0	4836400	-0.004424
16/01/2019	94057.0	94393.0	93687.0	94393.0	94393.0	4257400	0.003572
17/01/2019	94393.0	95682.0	93950.0	95351.0	95351.0	4944200	0.010149
18/01/2019	95386.0	96396.0	95386.0	96097.0	96097.0	4884000	0.007454
21/01/2019	96093.0	96093.0	94863.0	96010.0	96010.0	2717600	-0.000864
22/01/2019	96008.0	96070.0	94662.0	95103.0	95103.0	4256000	-0.009426

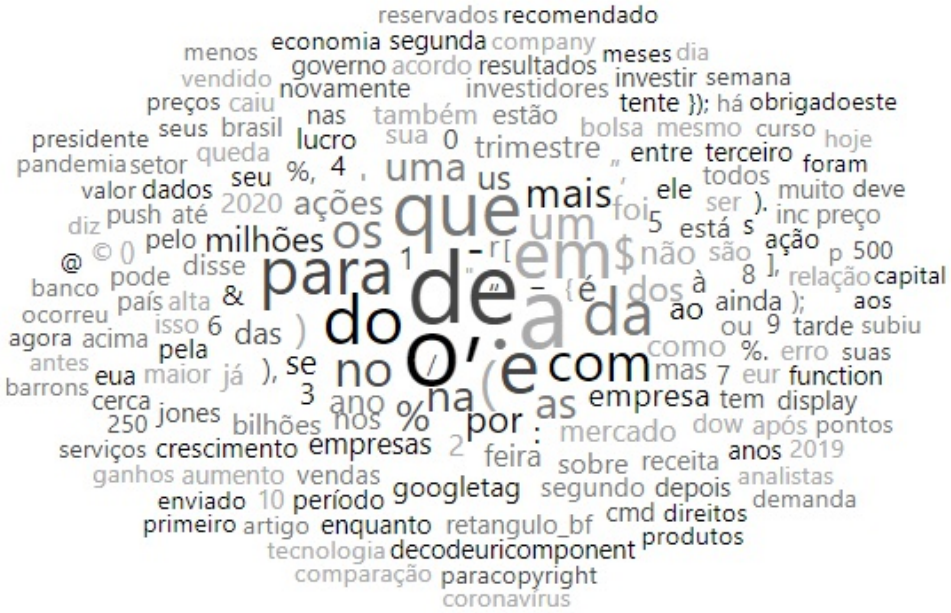

Figura 6: *Corpus* de notícias: nuvem de palavras inicial.

D. Avaliação do modelo de classificação de notícias

Considerando o trabalho de [29], as seguintes métricas foram utilizadas a fim de avaliar o modelo CNN aplicado: acurácia, precisão, *recall* e F1-*score* (Eq. (5)). Segundo [30], há quatro possibilidades de classes em predições concernentes a problemas de classificação binária. Estas possibilidades são descritas abaixo. Deste modo, nas equações referentes às métricas, VP, VN, FP e FN retratam, nessa ordem, verdadeiro positivo, verdadeiro negativo, falso positivo e falso negativo. Além disso, também foram utilizadas as métricas de perda logarítmica e curva ROC (*Receiver Operating Characteristic*).

- Verdadeiro positivo: quando o modelo atribui um classe positiva e confirma-se que a sentença realmente é desta classe.
- Verdadeiro negativo: quando o modelo rotula positivamente um exemplo, porém ao verifica-se a resposta, a classe correta era negativa.
- Falso positivo: quando a sentença é categorizada pelo modelo como positivo, no entanto a classe é negativa.
- Falso negativo: quando o exemplo é caraterizado como sendo da classe negativa, contudo na realidade a classe era positiva.

A acurácia representa o percentual de classificações corretas obtidas pelo modelo. Sendo assim, é calculada pela soma dos verdadeiros positivos com os verdadeiros negativos dividida

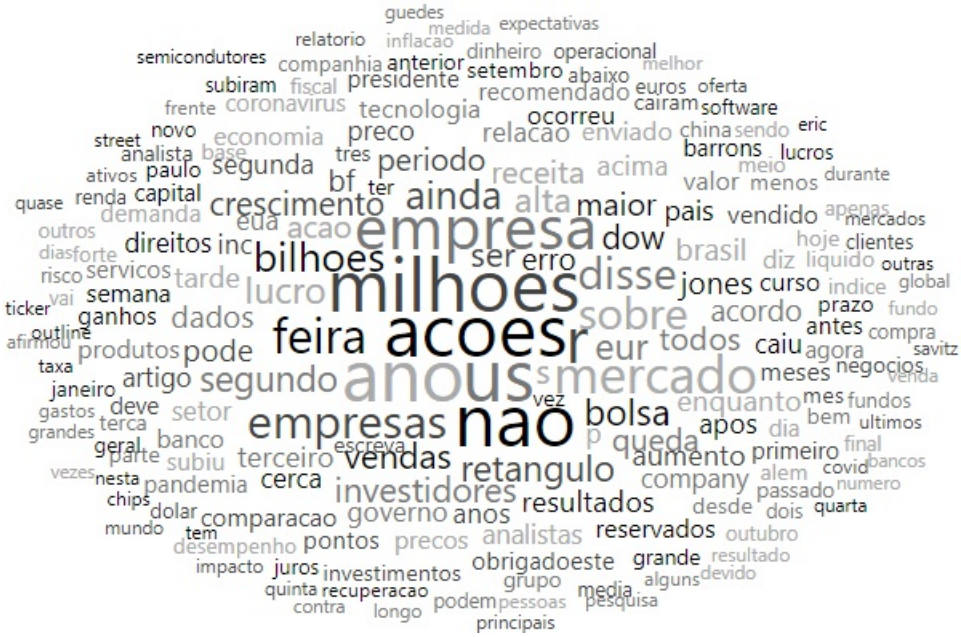

Figura 7: *Corpus* de notícias: nuvem de palavras após etapas do pré-processamento.

pelo total de classificações realizadas (Eq. (2)). Contudo, deve-se levar em consideração que outras métricas devem ser analisadas de forma conjunta, tendo em vista que no caso de haver maior domínio de uma categoria, o modelo pode rotular todas as sentenças com a categoria em questão e obter uma acurácia boa, porém ilusória. Além disso, a acurácia concede o mesmo peso para ambos os erros, desta maneira em um modelo que detecta a existência de tumores, no qual os falsos negativos são mais críticos que os falsos positivos, seria atribuído a mesma relevância para ambos nesta métrica.

A precisão retrata a porcentagem de rotulações positivas corretas obtidas (Eq. (3)). É uma boa métrica para situações onde os falsos positivos são mais danosos que os falsos negativos, sendo o seu cálculo baseado na divisão da quantidade de sentenças classificadas corretamente como positivas e o total de exemplos classificados como positivos. A precisão pode ser enganosa em situações onde muitos exemplos pertencentes à classe positiva não são classificados como positivos, o que ocasionaria uma alta precisão. Por exemplo, no caso de haver 100 exemplos positivos e 100 negativos, um modelo que classifica 10 sentenças como positivas, das quais 8 seriam corretas, obteria uma precisão de 80%, quando de fato definiu corretamente apenas 8% das sentenças.

O *recall* corresponde à razão entre o número de sentenças rotuladas de forma correta como concernentes à classe positiva e o total de exemplos que são realmente positivos. Esta métrica caracteriza a quantidade de instâncias que foram corretamente classificadas como positivas em relação a todas as que realmente pertencem a essa classe. Portanto, é equivalente a medir a proporção de positivos verdadeiros em relação a todos os positivos reais no conjunto de dados (Eq. (4)).

O F1-score é a média harmônica da precisão e do *recall*, de forma que uma característica dessa média é que se a precisão ou o *recall* for muito baixo ou próximo de zero, logo o F1-score será baixo idem. Por outro lado, o F1-score será alto uma vez que a precisão e o *recall* também forem altos, o que implica que o modelo é capaz de realizar predições corretas bem como de recuperar as instâncias da classe de interesse.

A matriz de confusão evidencia a quantidade de classificações obtidas para cada uma das quatro categorias, sendo uma forma simples de visualizar a performance de um determinado modelo.

A perda logarítmica (*Log Loss*) é uma métrica que envolve a ideia de confiança probabilística, portanto seu valor aumenta conforme a probabilidade atribuída pelo modelo a um evento diverge do real rótulo. A Eq. (6) se refere a perda logarítmica para um classificador binário, na qual y representa o valor verdadeiro da variável dependente e p a probabilidade de a sentença pertencer à classe 1.

A curva ROC compreende a taxa de verdadeiros positivos (eixo vertical) em relação à taxa de falsos positivos (eixo horizontal). Portanto, representa a perfomance de um modelo de acordo com a sua taxa de sensibilidade, enquanto a AUC (*Area Under Curve*) é uma forma mais simples de analisar a curva ROC agregando a mesma em um único valor, sendo este a área sobre a curva.

O modelo CNN obteve 0.96 de acurácia e de F1-score, como é possível notar na Tabela III e na matriz de confusão exposta na Figura 8. As curvas de acurácia e perda logarítmica deste modelo a cada época são apresentadas nas Figuras 9 e 10, nesta ordem, ao passo que a curva ROC do mesmo é evidenciada na Figura 11.

$$acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{2}$$

$$precisão = \frac{VP}{VP + FP} \tag{3}$$

$$recall = \frac{VP}{VP + FN} \tag{4}$$

$$F1\text{-}score = 2 \times \frac{precisão \times recall}{precisão + recall} \tag{5}$$

$$Log\ loss = -\frac{1}{N}\sum_{i=1}^{N} y_i \log p_i + (1 - y_i) \log(1 - p_i) \tag{6}$$

Tabela III: Métricas de desempenho do classificador de notícias.

	Precisão	Recall	F1-score	Acurácia
Negativo	0.97	0.95	0.96	
Positivo	0.95	0.97	0.96	
Modelo				0.96

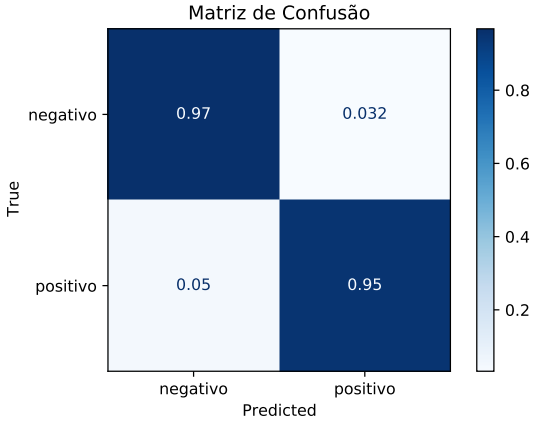

Figura 8: Matrizes de confusão: classificador de notícias.

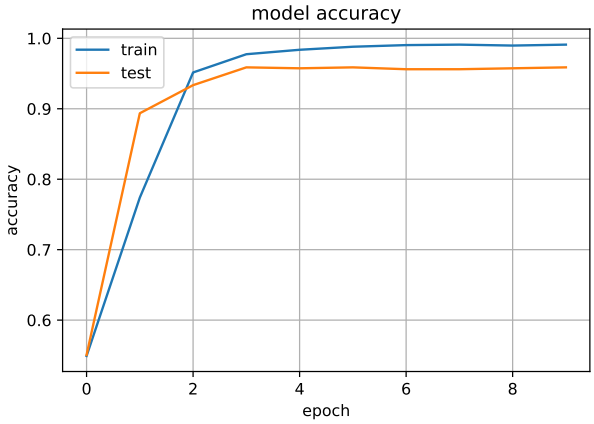

Figura 9: Acurácia a cada época

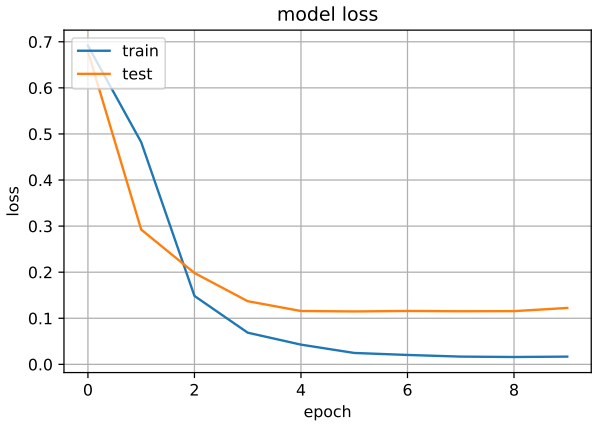

Figura 10: Perda logarítmica a cada época

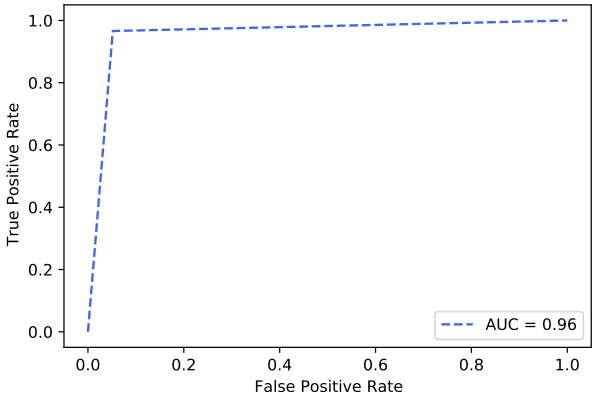

Figura 11: Curva ROC

E. Análise estatística

Após a classificação das notícias, a fase (IV) da Figura 1 iniciou-se. A etapa de entrada de dados (14) foi composta pelas notícias classificadas, o volume de buscas do google trends, e o volume de negociação e variação de preços da bolsa. A fim de medir o sentimento (15) das notícias rotuladas, a equação (7) foi empregue, baseado no trabalho de [31], de forma que $nPos$ e $nNeg$ retratam a quantidade de classificações positivas e negativas, respectivamente.

Na etapa seguinte, para cada dado trabalhado realizou-se a normalização (16) *Min-Max*, cuja fórmula é demonstrada na Eq. (8). A normalização consiste na transformação dos valores das colunas numéricas em uma escala comum (neste caso 0 a 1), de forma que não haja distorção nas divergências dos intervalos. Este procedimento possui o intuito de que se possa realizar comparações entre as colunas removendo o efeito das influências.

Isto posto, aplicou-se o teste de Shapiro-Wilk denotado pela Eq. (9) e apresentado em [32] afim de verificar a hipótese de distribuição normal em todas as variáveis. Nesta equação, a_i reflete as constantes produzidas conforme as médias, covariâncias e variâncias de uma determinada amostra

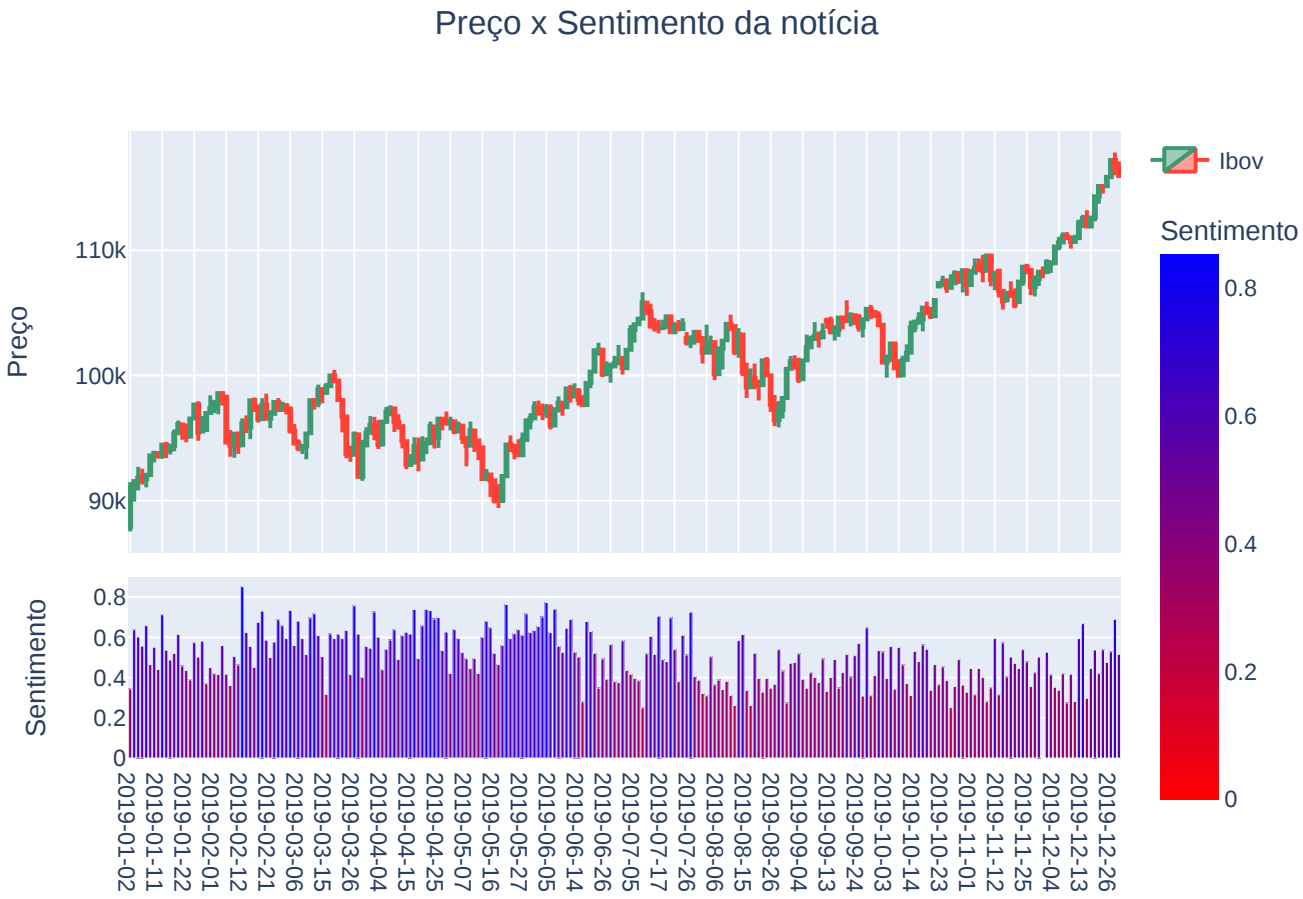

Figura 12: Preço x Sentimento das notícias

de tamanho n de uma distribuição normal e $x_{(i)}$ retrata os valores de amostras ordenadamente.

$$Indicador = \frac{(nPos - nNeg)}{(nPos + nNeg)} \tag{7}$$

$$X^{'} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{8}$$

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{9}$$

$$\rho = 1 - \frac{6\sum_{i=1}^{n} {d_i}^2}{n^3 - n} \tag{10}$$

Uma vez que, de acordo com o teste de Shapiro-Wilk realizado, alguns dos dados em consideração apresentavam distribuições não normais, o coeficiente de correlação de postos de Spearman (ρ) (18) foi usado segundo a fórmula (10), visto que é um método não paramétrico para medir a depêndencia estatística dos postos das variáveis. Desta maneira, n representa a quantidade de pares investigados e d_i a diferença entre os postos dos mesmos.

IV. Resultados e Avaliações

Com o propósito de detalhar de melhor forma, os resultados serão apresentados em três subseções de acordo com os anos analisados, sendo eles 2019, 2020 e 2021. Foram obtidas 64 combinações para cada ano, bem como seus respectivos coeficientes de correlação de Spearman, dentre as quais destacam-se as que compreendem o preço de fechamento ajustado (PF), o volume de negociação (VN), a variação do mercado (VM), o volume de buscas no Google (VB) e o sentimento das notícias (SN). Como uma forma de orientar as discussões, foram elaborados alguns questionamentos, os quais serão evidenciados conforme cada subseção. Pode-se observar um resumo dos resultados na Tabela V.

A. Ano 2019

❏ *Q1: Como se dá a correlação estatística entre o sentimento das notícias e as movimentações do mercado no ano de 2019?*

Os dados referentes ao sentimento das notícias (SN) em relação ao preço de fechamento (PF), variação do mercado (VM) e volume de negociação (VN) no ano de 2019, apresentaram coeficientes de correlação iguais a -0.441031, -0.014518 e -0.172347. De acordo com a Escala de Cohen, definida em [33] e exposta na Tabela IV, estes coeficientes caracterizam

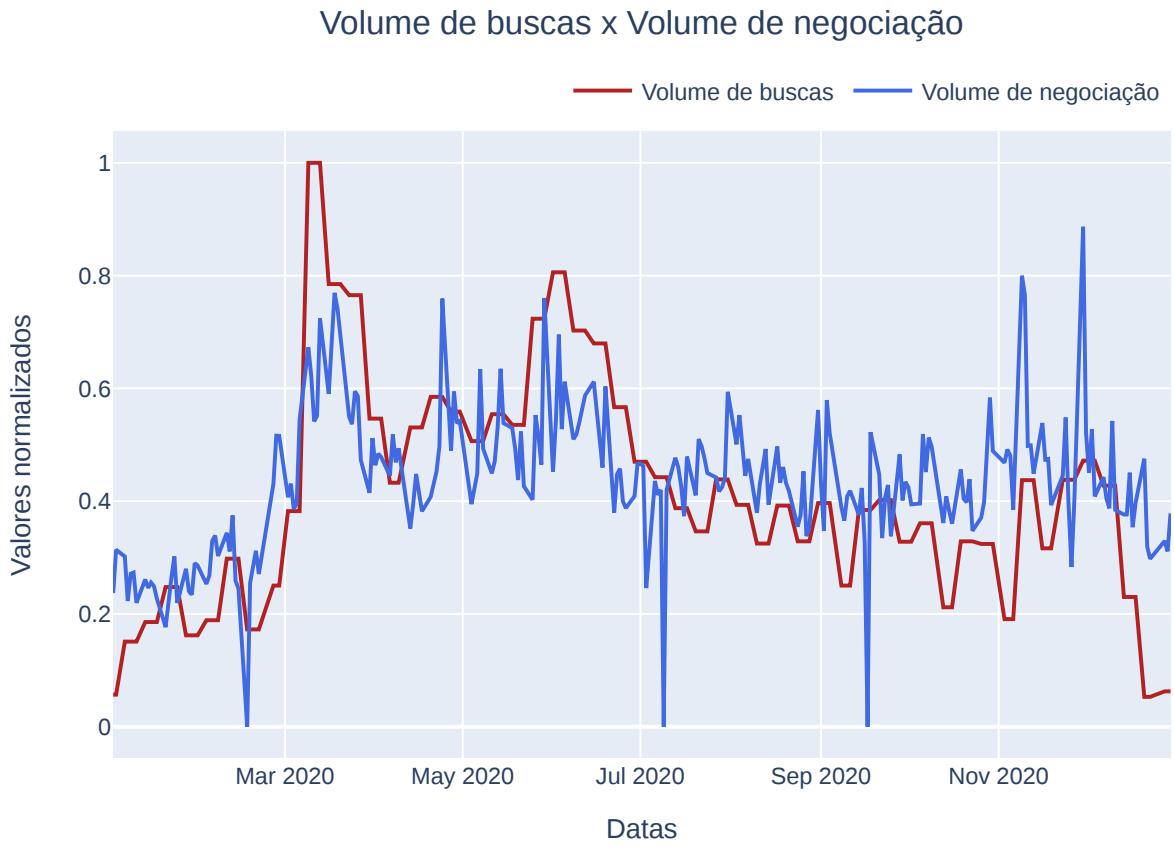

Figura 13: Volume de buscas x Volume de negociação

correlações inversas de caráter moderado, muito pequeno e pequeno, respectivamente. A partir disso, pode-se deduzir que há a possibilidade de que notícias positivas tendam a deixar os investidores mais receosos, o que pode implicar em preços de fechamento menores. Isto pode ser visualizado na Figura 12, que retrata o preço na forma de *candlesticks* e o sentimento das notícias em barras. O gráfico superior é conhecido como gráfico de vela, no qual as velas vermelhas indicam que o preço de fechamento foi menor que o preço de abertura e as velas verdes apontam o contrário. O pavio reflete o mínimo e máximo de preço atingindo no período de tempo determinado (neste caso, diário). No segundo gráfico, pode-se observar a intensidade do sentimento das notícias, de modo que a cor azul representa positividade e a cor vermelha exprime negatividade.

B. Ano 2020

❏ *Q2: Existe correlação estatística entre o volume de buscas no Google e as movimentações do mercado no ano de 2020?*

Os coeficientes de correlação advindos da relação entre o volume de buscas e o preço de fechamento, variação de mercado e volume de negociação obtiveram os valores $\rho = -0.761200$ (muito grande), $\rho = 0.071186$ (muito pequeno) e $\rho = 0.658824$ (grande), nessa ordem. É possível inferir

Tabela IV: Escala de Cohen.

Coeficiente de correlação	Descrição
0,0 a 0,1	muito pequeno
0,1 a 0,3	pequeno
0,3 a 0,5	moderado
0,5 a 0,7	grande
0,7 a 0,9	muito grande
0,9 a 1,0	próximos

que a quantidade de pesquisas feita pelos usuários relaciona-se ao volume de negociação no ano considerado, de maneira que havendo maior interesse pelo mercado consequentemente maior é a atividade no mesmo, como nota-se na Figura 13. Além disso, existe uma correlação negativa forte entre o volume de buscas e o preço de fechamento, o que pode indicar maior atenção por parte dos investidores quando há uma tendência de baixa, conforme observa-se o pico de pesquisas no início da pandemia do COVID-19 na Figura 13. Neste período houve quedas consideráveis no mercado financeiro devido ao pânico das pessoas, que começaram a vender suas ações em massa.

❏ *Q3: De que forma o preço de fechamento se relaciona com*

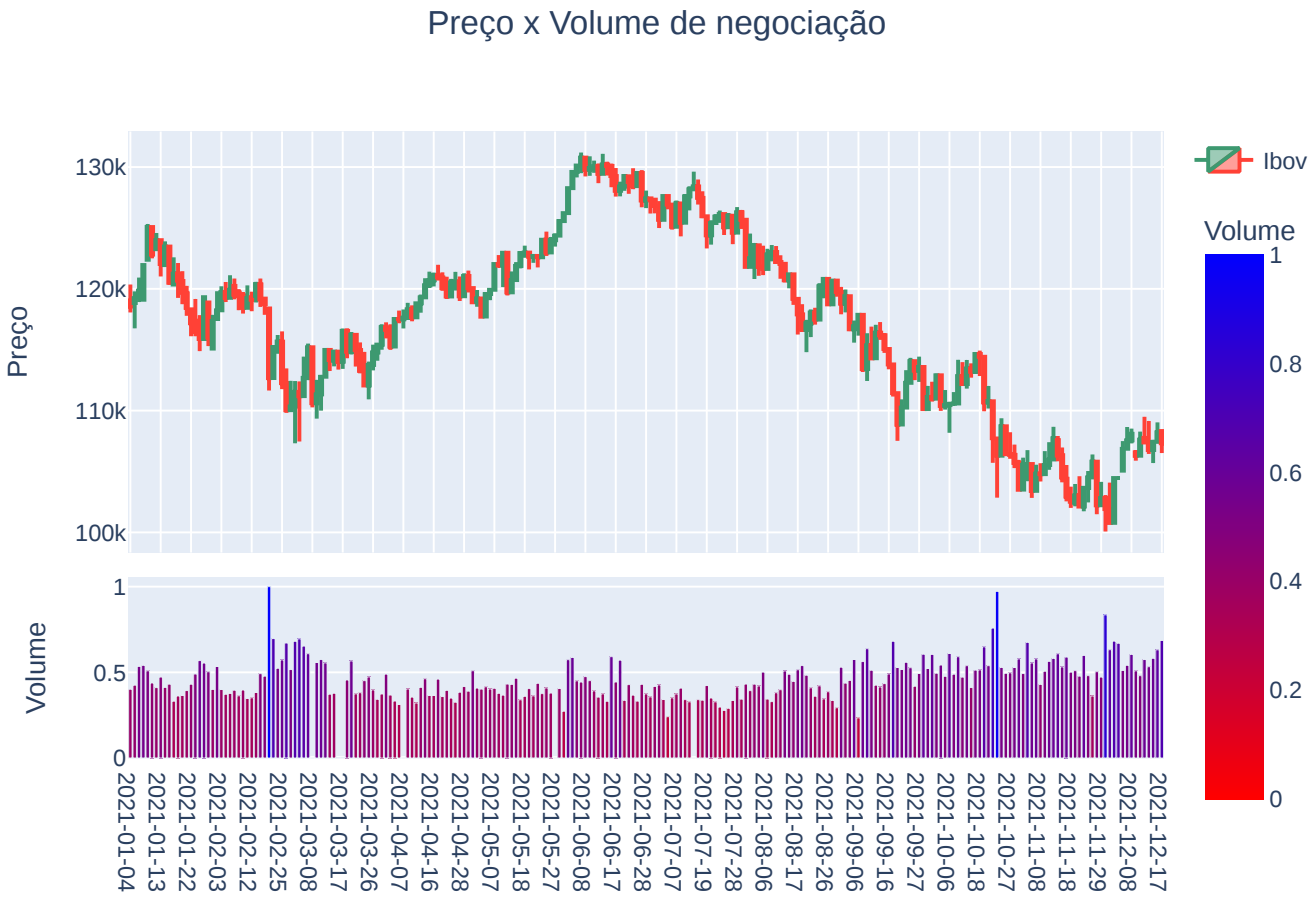

Figura 14: Preço x Volume de negociação

o volume de negociação em 2020?

Referente ao preço de fechamento com o volume de negociação, o resultado foi uma correlação inversa grande de $\rho = -0.588594$. Sendo assim, pode-se afirmar que no ano de 2020, a quantidade de negociações tendem a ser maiores em dias de queda, o que provavelmente se dá devido à uma tendência de baixa no mercado.

C. Ano 2021

❏ *Q4: Há relação entre o sentimento das notícias e o volume de buscas com as movimentações dos indicadores financeiros analisados no ano de 2021?*

Posto que todos os coeficientes de correlação oriundos da relação entre o sentimento das notícias e o volume de buscas com os indicadores financeiros foram menores que 0.2, não há força para afirmar que há correlação estatística entre cada um deles.

❏ *Q5: O preço de fechamento possui correlação estatística com o volume de negociação em 2021?*

O coeficiente de correlação proveniente da relação entre o preço de fechamento e o volume de negociação em 2021 foi -0.600626. Logo, ocorre a mesma situação descrita em 2020, deste modo, provavelmente, esta correlação inversa grande se deve a um tendência de baixa em 2021, como pode-se notar na Figura 14.

Tabela V: Combinações de valores e coeficientes de correlação de Spearman.

Ano	Combinações	ρ	Classificação de Cohen
2019	SN vs PF	-0.441031	moderado
2019	SN vs VM	-0.172347	pequeno
2019	SN vs VN	-0.014518	muito pequeno
2020	VB vs PF	-0.761200	muito grande
2020	VB vs VM	0.071186	muito pequeno
2020	VB vs VN	0.658824	grande
2020	PF vs VN	-0.588594	grande
2021	PF vs VN	-0.600626	grande

V. Conclusão

Neste artigo foram apresentados os procedimentos de coleta de notícias em língua portuguesa sobre o mercado financeiro, de informações de motores de busca do Google Trends e de dados da bolsa brasileira B3 provindos da internet. Para os dados coletados foram realizados processamentos para a obtenção de indicadores. Além disso, realizou-se uma análise estatística do relacionamento dos indicadores levantados à

partir dos dados obtidos, considerando os coeficientes de correlação.

Todos os dados coletados foram armazenados em formato CSV, de maneira que pudessem ser manipulados com facilidade para outros trabalhos. Ademais, foi possível notar que o sentimento das notícias esteve correlacionado de forma moderada, mesmo que inversamente, com o preço de fechamento em 2019, contudo para os outros indicadores essas relações se apresentaram pequenas ou muito pequenas. Mostraram-se interessantes os coeficiente obtidos a partir do volume de buscas com o preço de fechamento (inversa e muito grande) e com volume de negociação (grande) em 2020, visto que demonstraram a maior atenção dos investidores quando há uma tendência de baixa e a atividade do mercado conforme o interesse dos mesmos. Tanto em 2020 quanto em 2021, pôde-se verificar a existência de correlação estatística grande entre o preço de fechamento e o volume de negociação, levando à conclusão de que nesses anos as negociações foram superiores em períodos de queda. Parte do trabalho realizado neste experimento foi utilizado para outro experimento no qual foram avaliados os relacionamentos estatísticos dos indicadores obtidos para as notícias e dados da B3 com outros de sentimentos de dados coletados do Twitter. Como trabalho futuro, espera-se agregar também o indicador obtido do Google Trends ao experimento deste último trabalho citado.

REFERÊNCIAS

[1] I. L. Stats, data de acesso: 28 de julho de 2022. [Online]. Available: https://www.internetlivestats.com/internet-users/

[2] V. Peres, R. Vieira, and R. Bordini, "Análises de Sentimentos: Abordagem lexical de classificação de opinião no contexto mercado financeiro brasileiro," *Workshop of Artificial Intelligence Applied to Finance*, 2019.

[3] E. F. Fama, "Efficient Capital Markets: A Review of Theory and Empirical Work," *Journal of Finance*, vol. 25, no. 2, pp. 383–417, May 1970.

[4] C. Silva and M. Machado, "The effect of foreign investment flow on commonality in liquidity on the brazilian stock market," *Revista Contabilidade & Finanças*, vol. 31, 05 2020.

[5] B3, data de acesso: 28 de julho de 2022. [Online]. Available: https://www.b3.com.br/pt_br/market-data-e-indices/servicos-de-dados/market-data/consultas/mercado-a-vista/perfil-pessoas-fisicas/faixa-etaria/

[6] D. S. Alves, "Uso de técnicas de computação social para tomada de decisão de compra e venda de ações no mercado brasileiro de bolsa de valores," Ph.D. dissertation, Universidade de Brasília, Brasília, 2015.

[7] L. J. FARIA, K. C. SILVA, M. G. C. Fernandes, D. S. A. FERNANDES, and F. SOARES, "Tweet and news sentiment indicators and the behavior of the brazilian stock market," in *Proceedings of the 21st ACM IEEE International Conference on Industrial Informatics*. Perth, Australia: IEEE, 2022.

[8] N. Kamanwar and S. Kale, "Web data extraction techniques: A review," 2016. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84994134167&doi=10.1109\%2fSTARTUP.2016.7583910&partnerID=40&md5=98de0ed226640d956675d44e3a25d695

[9] M. Parvez, K. Tasneem, S. Rajendra, and K. Bodke, "Analysis of different web data extraction techniques," 2018. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85059372411&doi=10.1109\%2fICSCET.2018.8537333&partnerID=40&md5=cc823d1fb406ee50ce18f1c1234d169d

[10] A. Maududie, W. E. Y. Retnani, and M. A. Rohim, "An approach of web scraping on news website based on regular expression," in *2018 2nd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, 2018, pp. 203–207.

[11] D. M. Thomas and S. Mathur, "Data analysis by web scraping using python," in *2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2019, pp. 450–454.

[12] P. Thota and E. Ramez, "Web scraping of covid-19 news stories to create datasets for sentiment and emotion analysis," in *The 14th PErvasive Technologies Related to Assistive Environments Conference*, ser. PETRA 2021. New York, NY, USA: Association for Computing Machinery, 2021, p. 306–314. [Online]. Available: https://doi-org.ez49.periodicos.capes.gov.br/10.1145/3453892.3461333

[13] B. Bhardwaj, S. I. Ahmed, J. Jaiharie, R. Sorabh Dadhich, and M. Ganesan, "Web scraping using summarization and named entity recognition (ner)," in *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, vol. 1, 2021, pp. 261–265.

[14] T. Kabbani and F. Usta, "Predicting the stock trend using news sentiment analysis and technical indicators in spark," 01 2022.

[15] M. Ramdhani, D. Maylawati, and T. Mantoro, "Indonesian news classification using convolutional neural network," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 19, 08 2020.

[16] M. A.-U.-Z. Ashik, S. Shovon, and S. Haque, "Data set for sentiment analysis on bengali news comments and its baseline evaluation," in *2019 International Conference on Bangla Speech and Language Processing (ICBSLP)*, 2019, pp. 1–5.

[17] M. Omarkhan, G. Kissymova, and I. Akhmetov, "Handling data imbalance using cnn and lstm in financial news sentiment analysis," in *2021 16th International Conference on Electronics Computer and Computation (ICECCO)*, 2021, pp. 1–8.

[18] B. T. Abeje, A. O. Salau, H. A. Ebabu, and A. M. Ayalew, "Comparative analysis of deep learning models for aspect level amharic news sentiment analysis," in *2022 International Conference on Decision Aid Sciences and Applications (DASA)*, 2022, pp. 1628–1633.

[19] M. R. Vargas, B. S. L. P. de Lima, and A. G. Evsukoff, "Deep learning for stock market prediction from financial news articles," in *2017 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, 2017, pp. 60–65.

[20] F. Díaz and P. A. Henríquez, "Social sentiment segregation: Evidence from twitter and google trends in chile during the covid-19 dynamic quarantine strategy," *PLoS ONE*, vol. 16, 2021.

[21] M. Costola, M. Iacopini, and C. R. M. A. Santagiustina, "Google search volumes and the financial markets during the covid-19 outbreak," *Finance Research Letters*, vol. 42, p. 101884, 12 2020.

[22] I. L. Stats, data de acesso: 19 de julho de 2022. [Online]. Available: https://www.internetlivestats.com/one-second/#google-band

[23] S. GlobalStats, data de acesso: 19 de julho de 2022. [Online]. Available: https://gs.statcounter.com/

[24] C. A. d. Medeiros, G. Rodrigues, K. Felix, M. Barbosa, P. V. Souza, and R. Magner, Sep 2020. [Online]. Available: https://bit.ly/3xY4oXN

[25] M. Picanço, "Stock market news data in portuguese: Sentiment analysis dataset for financial news in brazilian portuguese[dataset]," 2021. [Online]. Available: https://bit.ly/3OGdi26

[26] A. Sinha, 2020. [Online]. Available: https://bit.ly/3vLnyxz

[27] Y. Qi, D. Sachan, M. Felix, S. Padmanabhan, and G. Neubig, "When and why are pre-trained word embeddings useful for neural machine translation?" 01 2018, pp. 529–535.

[28] Y. Kim, "Convolutional neural networks for sentence classification," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 08 2014.

[29] F. Benevenuto, F. Ribeiro, and M. Araújo, "Métodos para análise de sentimentos em mídias sociais," *Sociedade Brasileira de Computação*, 2015.

[30] L. De Castro and D. Ferrari, *Introdução à Mineração de Dados: Conceitos Básicos, Algoritmos e Aplicações*, 05 2016.

[31] A. E. O. Carosia, G. P. Coelho, and A. E. A. Silva, "Analyzing the brazilian financial market through portuguese sentiment analysis in social media," *Applied Artificial Intelligence*, vol. 34, no. 1, pp. 1–19, Oct. 2019.

[32] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965.

[33] J. Cohen, P. Cohen, S. G. West, and L. S. Aiken, *Applied multiple regression/correlation analysis for the behavioral sciences*, 3rd ed. Mahwah, N.J.: L. Erlbaum Associates, 2003.

Kéthlyn C. Silva é uma estudante de Engenharia de Computação pela Universidade Federal de Goiás. Atualmente atua na área de engenharia e mineração de dados, e possui grande interesse em Ciência de Dados e Processamento de Linguagem Natural. Possui ainda experiência em desenvolvimento *web Front-end*.

Deborah S. A. Fernandes trabalha no Instituto de Informática da Universidade Federal de Goiás – Campus Samambaia, Goiânia, Goiás. Além disso, é Cientista da Computação (Pontifícia Universidade Católica de Goiás), mestre em Engenharia Elétrica com ênfase em Visão Computacional (Escola de Engenharia – Universidade de Brasília/DF) e doutora em Engenharia de Sistemas Eletrônicos e Automação (Escola de Engenharia -Universidade de Brasília/DF) com ênfase em análise de dados de redes sociais e apoio à tomada de decisão.

Márcio Giovane C. Fernandes possui graduação em Ciência da Computação pela PUC-Goiás. Cursou o mestrado em Engenharia Elétrica e de Computação na UFG-Goiás. Atua na área da docência e seus principais interesses na computação e informática são: detecção de eventos em grandes volumes de dados, alfabetização de alunos em lógica de programação, desenvolvimento de software, computação social - análise de tendência em textos de microblogs.