Relacionamento estatístico entre indicadores de dados de internet em língua portuguesa e bolsa de valores

Kéthlyn C. Silva[1], Lucas José de Faria[2], Deborah S. A. Fernandes[1], Márcio Giovane C. Fernandes [3], Fabrízzio Soares[1]

[1]EMC e INF, Universidade Federal de Goiás, GO.

[2]Instituto Federal Goiano, GO.

[3]CCET, Universidade Estadual de Goiás, GO.

kethlyncampos@discente.ufg.br, lucas.faria@ifgoiano.edu.br, {deborah.fernandes, fabrizzio}@ufg.br, marcio.giovane@ueg.br

Abstract. *This work presents a statistical analysis between indicators obtained from internet data in Portuguese - news sentiment and Google Trends - and data on the Brazilian stock market through Spearman's rank correlation coefficient. The methodology used to collect, pre-process and obtain each indicator is detailed. Data from the years 2019 to 2021 were obtained. For the sentiment analysis of the news, a CNN model (Convolutional Neural Network) was adopted, which obtained an F1-score of 96%. As a result, some interesting correlations were obtained, among which, an inverse correlation characterized as "moderate" (according to the Cohen scale) between news sentiment and adjusted closing price in 2019; between search volume and closing price, a negative and "very large" and a positive and "large" correlation between trade volume and search volume. In both 2020 and 2021, negative coefficients defined as "large" were found, taking into account the closing price and trading volume.*

Resumo. *Este trabalho apresenta uma análise estatística entre indicadores obtidos de dados de internet em língua portuguesa - sentimento de notícias e Google Trends - e dados sobre o mercado brasileiro de bolsa de valores através do coeficiente de correlação de postos de Spearman. A metodologia utilizada para coleta, pré-processamento e obtençâo de cada indicador é detalhada. Foram obtidos dados dos anos de 2019 a 2021. Para a análise de sentimento das notícias foi adotado um modelo CNN (Convolutional Neural Network) o qual obteve um F1-score de 96%. Como resultados foram obtidas algumas correlações interessantes dentre as quais, uma correlação inversa caracterizada como moderada (segundo a escala de Cohen) entre o sentimento das notícias e preço de fechamento ajustado em 2019; entre volume de buscas e preço de fechamento, uma correlação negativa e "muito grande" e positiva e "grande" entre o volume de negociações e o volume de buscas. Tanto em 2020 como em 2021, constatou-se coeficientes negativos definidos como "grandes", levando em conta o preço de fechamento e volume de negociações.*

1. 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 [Peres et al. 2019] 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 [Fama 1970]. Esta divergência de preços abre margem para oportunidades de lucro, tornando-se atrativo para investidores do mundo todo, como descrito em [Silva and Machado 2020]. 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 forma que mais de 5 milhões são investidores pessoas físicas na custódia da B3 (bolsa brasileira), de acordo com [B3 2022]. 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 com vários temas de pesquisa em aberto.

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.

2. Revisão Bibliográfica

A opinião e o sentimento público têm ganhado cada vez mais relevância tanto no âmbito privado quanto público, uma vez que esses dados podem ser utilizados, por exemplo, para controle de difamação, controle de doenças por parte do governo, para geração de insights sobre o mercado financeiro, controle de qualidade de produtos, etc. A coleta desses dados no contexto do mercado financeiro é feita essencialmente de três formas: notícias, dados de redes sociais e dados de busca na web.

Em [Alves 2015] foi utilizado o Twitter como fonte de dados para simulação de compra e venda de ações no mercado de bolsa de valores brasileiro. Utilizaram dados com e sem pré-processamento, no primeiro caso realizaram filtragens de links, pontuação e de tweets contendo palavras e expressões selecionadas e aplicaram algoritmo classificador de sentimentos par aa obtenção de indicador de sentimentos. No segundo caso, os dados

[1] https://www.internetlivestats.com/internet-users/. Data de acesso: 28 de julho de 2022

sem limpeza foram usados para análise do volume de tweets sobre cada empresa e ação e também para contagem ingênua de mensagens contendo expressões relacioandas à alta e baixa do mercado e compra ou venda de ações. Em [FARIA et al. 2022], 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.

Em [Thomas and Mathur 2019] 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. Os autores conseguiram realizar a extração de dados referentes a pesquisas frequentes da rede social Reddit e o resultado da análise dos dados foi apresentado na forma de porcentagem, contendo os assuntos mais procurados no site. Em [Thota and Ramez 2021] é proposto um estudo sobre técnicas de *web scraping* com intuito de extrair declarações de líderes de governo sobre o COVID-19 a partir do site de notícias da CNN.

Formas alternativas de realizar o *web scraping* sem a necessidade de conhecimento do DOM (*Document Object Model* - interface de programação para documentos HTML e XML) da página são apresentadas em [Bhardwaj et al. 2021]. Neste, técnicas de Processamento de Linguagem Natural (NLP) e *Machine Learning* (ML) foram aplicadas para contornar este requisito. Utilizaram também técnicas de sumarização de texto e *Named Entity Recognition* (NER) para a construção de um preditor de epidemias. A sumarização do texto não se mostrou tão eficiente quanto os métodos convencionais e falhou ao lidar com dados discretos e altamente não-estruturados. Em contrapartida, o *Named Entity Recognition* suportou os dados discretos e altamente estruturados, sendo mais flexível, possui uma acurácia que se aproxima dos métodos convencionais.

Em [Díaz and Henríquez 2021] foram 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.

3. Etapas do Experimento

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 entre parênteses "( )" desta seção em diante, conforme a Figura 1.

3.1. Aquisição de dados

As linguagens de programação Python e R são comumente empregadas na construção de ferramentas de *web scraping* por possuírem bibliotecas específicas para essa tarefa e comunidade ativa, sendo estas as linguagens mais populares em ciência de dados, como citado em [Thota and Ramez 2021]. Em todas as fases do experimento realizado neste

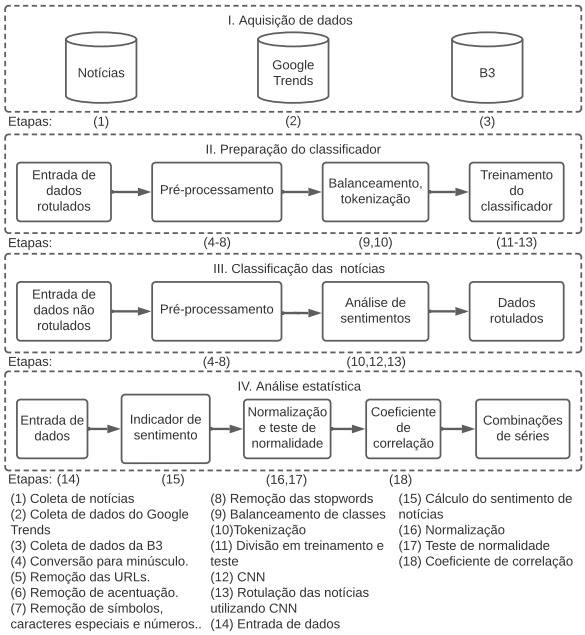

Figura 1. Fases e etapas da metodologia proposta.

trabalho, a linguagem Python[2] 3.10.7 foi utilizada. As bibliotecas Beautiful Soup[3], Selenium[4] e Requests[5] foram adotadas na extração de dados, a primeira utilizada na análise de HTML, a segunda para navegação entre as páginas e a terceira para efetuação de requisições nas páginas web.

3.1.1. Coleta de textos de notícias na web

Para a realização da etapa (1) da fase de aquisição dos dados (I) - Fig. 1, os sites Yahoo Finanças, Investing.com e Money Times foram escolhidos como fonte devido ao volume de acessos por parte dos usuários.

A seção relacionada a bolsa de valores do site Yahoo Finanças faz uso do recurso *scroll infinito*. Neste, o conteúdo é carregado conforme há a rolagem da página. Por meio de métodos da biblioteca Selenium, a solução apresentada por [Thota and Ramez 2021] para este cenário foi implementada para o experimento. 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 novamente adotada para a obtenção dos dados das notícias contidas

[2]https://www.python.org/
[3]https://www.crummy.com/software/BeautifulSoup/bs4/doc/
[4]https://selenium-python.readthedocs.io/
[5]https://requests.readthedocs.io/en/latest/

em cada link. 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 disponibiliza apenas as notícias mais recentes nesta seção, as 501 notícias adquiridas correspondem apenas a datas próximas ao período de coleta. O período de coleta foi iniciado em 28/11/2021 às 13:16:22 e encerrado em 17/12/2021 às 18:51:02. A data e hora das publicações coletadas inicia em 23/11/2021 às 17:28:25 e finaliza em 17/12/2021 às 18:11:56.

Os site Infomoney, também utilizado para a coleta, assim como o Yahoo Finanças, possui uma estrutura de scroll infinito. Ao clicar no botão "ver mais", o conteúdo é carregando enquanto pressionado, até que não haja mais para ser apresentado. Deste modo, a mesma técnica utilizada no site anterior foi aplicada, um iterador é foi implementado para para pressionar o botão, a cada rolagem, repetidamente, até que o final do conteúdo esteja disponpível. 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. Em seguida, os dados são adquiridos por meio da Beautiful Soup. Os dados obtidos para este site foram: título, subtítulo, artigo, autor, data e hora de publicação, data e hora de coleta e URL. O intervalo de data e hora das 812 notícias coletadas das 812 foi de 10/11/2021 às 16:33:09 até 17/12/2021 às 19:04:14. A coleta foi iniciada em 2021-11-28 às 13:57:19 e finalizada em 2021-12-17 às 21:34:45.

O terceiro site utilizado foi o Investing.com que faz o uso de paginação para dispor seu conteúdo. Neste caso, faz-se necessária a navegação por meio dessa paginação com a finalidade de visualizar todas as notícias dispostas. Este site foi o único, dentre os três escolhidos para a realização deste experimento, que permitia a obtenção de notícias publicadas desde 2019 como foi definido para este trabalho. O passo a passo para a coleta é feito da mesma forma descrita anteriormente para o site Infomoney. 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 e hora de publicação foi de 2019-01-01 às 08:45:00 até 2021-12-17 às 20:45:00, sendo o período de coleta iniciado em 2021-12-03 às 20:57:49 e finalizado em 2021-12-17 às 21:54:29. Sendo este 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 de valores B3.

3.1.2. Coleta de dados de buscadores da web - Google Trends

De acordo com o site Internet Live Stats[6], mais de 8 bilhões de pesquisas são realizadas diariamente no buscador que domina cerca de 65.87% do mercado, segundo o Statcounter GlobalStats [GlobalStats 2022]. Sendo assim, o Google 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. Essa intensidade é disposta em uma escala de 0 a 100, de modo que 100 representa o pico de interesse e 0 exprime absoluto desinteresse. Ademais, o serviço proporciona a exportação dos dados no formato csv, sendo estes arquivos seccionados de acordo a data e região. Posto isso, foi realizada uma coleta de dados acerca dos termos relacionados ao Ibovespa e às ações que o compõem.

[6]https://www.internetlivestats.com/one-second/#google-band. Data de acesso: 19 de julho de 2022

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 ao período de tempo e à região. Os filtros foram aplicados tendo como região o Brasil e período de 01/01/2019 à 17/12/2021, o mesmo das notícias coletadas.

Com o intuito de gerar um indicador a partir dos dados do Google Trends, primeiramente foi calculada a média dos índices de cada dia de acordo com os termos pesquisados, conforme descrito em [Díaz and Henríquez 2021]. Neste, quanto maior o valor da média, maior o interesse dos usuários pelo tema buscado no determinado dia. 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.

3.1.3. 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 ou automática. A primeira opção exige demasiado trabalho dependendo da quantidade de ações e do tamanho do período a ser analisado. Na segunda opção pode ser realizada com o desenvolvimento de um *webscraper* ou com o consumo de API's que forneçam estes dados. A bolsa de valores brasileira B3 possui uma API para acesso automatizado, contudo é pago.

Uma alternativa gratuita para adquirir os dados necessários, é através do consumo da API Yfinance. Esta realiza um *scraping* do site Yahoo Finance e disponibiliza os dados históricos e em tempo real de vários mercados financeiros. A desvantagem está no fato de esta não ser uma API oficial, portanto não há garantia de que futuramente possa estar funcionando, no entanto é amplamente utilizada. Para este experimento a API Yfinance foi utilizada.

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. A equação $\text{Variação}(\%) = \frac{\text{Preço de Fechamento ajustado}}{\text{Preço de Abertura}} - 1$ foi aplicada para calcular a variação de preços diária da bolsa e utilizada como indicador.

3.2. 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 2. Nela são apresentadas as palavras com maior frequência no corpus de notícias antes do pré-processamento. Ao observá-la, 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 [Kabbani and Usta 2022], 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[7] 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 3.

Figura 2. *Corpus* de notícias: nuvem de palavras inicial.

Figura 3. *Corpus* de notícias: nuvem de palavras após etapas do pré-processamento.

O conjunto de dados de treinamento utilizado constituiu-se da combinação de notícias relacionadas ao contexto do mercado financeiro classificadas em positivo e negativo[8]. 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[9] para tokenização (10) da base, com o intuito de vetorizar o corpus.

3.3. Análise de sentimentos para notícias

Para a classificação de notícias, implementou-se o algoritmo[10] de *deep learning Convolutional Neural Network* (CNN). O CNN 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. 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%), e assim o modelo foi treinado (12).

Posteriormente, a fase (III) apresentada na Figura 1 foi realizada. Sendo assim, as notícias coletadas foram pré-processadas, conforme descrito anteriormente, e rotuladas

[7]https://www.nltk.org/

[8]Dispostas em https://bit.ly/3xY4oXN, https://bit.ly/3OGdi26 e https://bit.ly/3vLnyxz

[9]https://keras.io/

[10]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 .

(13) utilizando o modelo já treinado. O modelo obteve 0.96 de acurácia e de F1-score, como é possível notar na Tabela 1.

Tabela 1. 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

3.4. Análise estatística

A fase (IV) da Figura 1 expõe a análise estatística efetuada. 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 $Indicador = \frac{(nPos-nNeg)}{(nPos+nNeg)}$ foi empregada. Esta, baseada no trabalho de [Carosia et al. 2019], 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* $X' = \frac{X-X_{min}}{X_{max}-X_{min}}$.Isto posto, aplicou-se o teste de Shapiro-Wilk denotado por $W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$, [Shapiro and Wilk 1965]. Nesta equação, a_i reflete as constantes produzidas conforme as médias, covariâncias e variâncias de uma determinada amostra de tamanho n de uma distribuição normal e $x_{(i)}$ retrata os valores de amostras ordenadamente.

Uma vez que alguns dos dados em consideração apresentavam distribuições não normais, o coeficiente de correlação de postos de Spearman $\rho = 1 - \frac{6\sum_{i=1}^{n} {d_i}^2}{n^3-n}$ foi usado , visto que é um método não paramétrico. Desta maneira, n representa a quantidade de pares investigados e d_i a diferença entre os postos dos mesmos.

4. Resultados e Discussão

Os resultados serão apresentados de acordo com os anos dos dados coletados e 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 3. Para o ano de 2019 foi levantada a seguinte questão:

❑ *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 [Cohen et al. 2003] e exposta na Tabela 2, estes coeficientes caracterizam correlações inversas de caráter moderado, muito pequeno e pequeno, respectivamente. A partir disso, pode-se inferir 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 4, 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.

Tabela 2. 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

Para o ano de 2020 foram levantadas as seguintes questões:

❑ *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 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 5. 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 5. 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 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 foram maiores em dias de queda, o que provavelmente se dá devido à uma tendência de baixa no mercado.

Para o ano de 2021 foram levantadas as seguintes questões:

❑ *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 6.

Tabela 3. 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

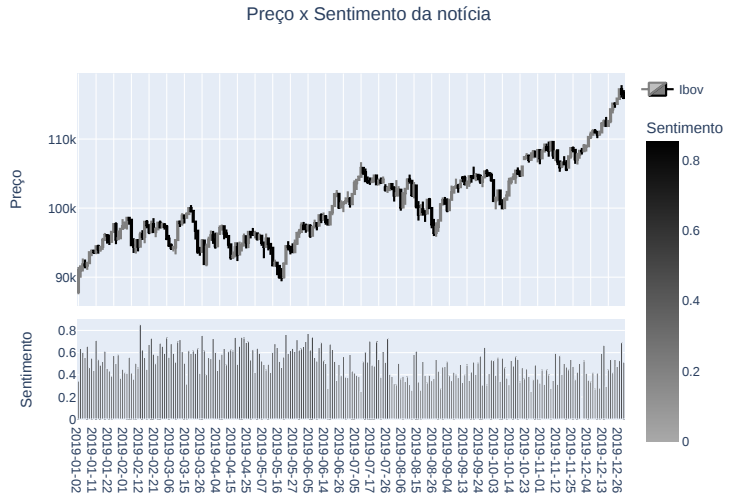

Figura 4. Preço x Sentimento das notícias

5. 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

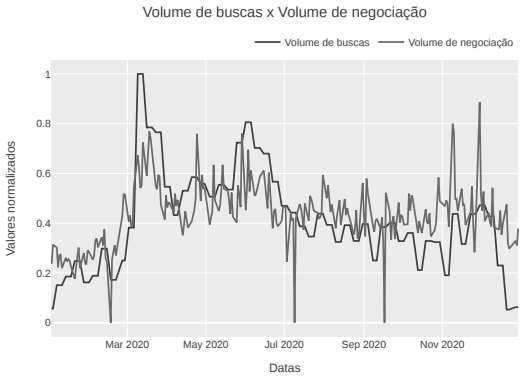

Figura 5. Volume de buscas x Volume de negociação

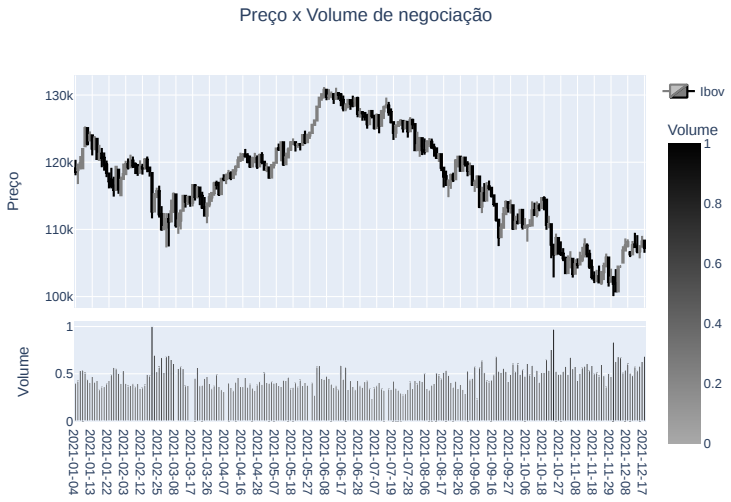

Figura 6. Preço x Volume de negociação

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

Alves, D. S. (2015). *Uso de técnicas de computação social para tomada de decisão de compra e venda de ações no mercado brasileiro de bolsa de valores*. PhD thesis, Universidade de Brasília, Brasília.

B3 (2022). Data de acesso: 28 de julho de 2022.

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

Carosia, A. E. O., Coelho, G. P., and Silva, A. E. A. (2019). Analyzing the brazilian financial market through portuguese sentiment analysis in social media. *Applied Artificial Intelligence*, 34(1):1–19.

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

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

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *Journal of Finance*, 25(2):383–417.

FARIA, L. J., SILVA, K. C., Fernandes, M. G. C., FERNANDES, D. S. A., and SOARES, F. (2022). 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.

GlobalStats, S. (2022). Data de acesso: 19 de julho de 2022.

Kabbani, T. and Usta, F. (2022). Predicting the stock trend using news sentiment analysis and technical indicators in spark.

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

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611.

Silva, C. and Machado, M. (2020). The effect of foreign investment flow on commonality in liquidity on the brazilian stock market. *Revista Contabilidade & Finanças*, 31.

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

Thota, P. and Ramez, E. (2021). 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*, PETRA 2021, page 306–314, New York, NY, USA. Association for Computing Machinery.