UNIVERSIDADE FEDERAL DO CEARÁ
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO, ATUÁRIA E CONTABILIDADE
DEPARTAMENTO DE FINANÇAS
CURSO DE FINANÇAS

BRUNO ALVES DE OLIVEIRA

IMPACTO DAS OPINIÕES DOS GRANDES INFLUENCIADORES E O EFEITO TWITTER SOBRE OS INVESTIDORES INDIVIDUAIS

FORTALEZA
2022

BRUNO ALVES DE OLIVEIRA

IMPACTO DAS OPINIÕES DOS GRANDES INFLUENCIADORES E O EFEITO TWITTER SOBRE OS INVESTIDORES INDIVIDUAIS

Monografia apresentada ao Curso de Finanças da Universidade Federal do Ceará, como requisito parcial à obtenção do título de bacharel em Finanças. Área de concentração: Finanças Comportamentais.

Orientador: Prof. Dr. Francisco Gildemir Ferreira da Silva

FORTALEZA

2022

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

O45i Oliveira, Bruno Alves de.
Impacto das opiniões dos grandes influenciadores e o efeito Twitter sobre os investidores individuais / Bruno Alves de Oliveira. – 2023.
42 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Faculdade de Economia, Administração, Atuária e Contabilidade, Curso de Finanças, Fortaleza, 2023.
Orientação: Profa. Dra. Francisco Gildemir Ferreira da Silva.

1. Análise de Sentimento. 2. Twitter. 3. Mercado Financeiro. 4. Modelo de Previsão. I. Título.
CDD 332

IMPACTO DAS OPINIÕES DOS GRANDES INFLUENCIADORES E O EFEITO TWITTER SOBRE OS INVESTIDORES INDIVIDUAIS

Monografia apresentada ao Curso de Finanças da Universidade Federal do Ceará, como requisito parcial à obtenção do título de bacharel em Finanças. Área de concentração: Finanças Comportamentais.

Aprovada em: 21/12/2023.

BANCA EXAMINADORA

_____
Prof. Dr. Francisco Gildemir Ferreira da Silva (Orientador)
Universidade Federal do Ceará (UFC)

_____
Prof. Dr. Vitor Borges Monteiro
Universidade Federal do Ceará (UFC)

_____
Prof. Dr. Paulo André Dias Jacome
Universidade Federal do Ceará (UFC)

AGRADECIMENTOS

Agradeço primordialmente a Deus, por se fazer presente em todo o processo da graduação. A todos os meus familiares próximos, em específico, aos meus pais, Maria Aparecida Alves de Oliveira e José Tarcísio de Oliveira, por todo carinho e estrutura educacional a mim dedicado por todos os anos. Aos meus irmãos Boaz Alves de Oliveira e Bani Alves de Oliveira, por todo o companheirismo e motivação a mim confiado. Amo cada um de vocês.

Ao meu orientador, Dr. Francisco Gildemir Ferreira da Silva, por ter aceitado a proposta visto que não é uma aplicação comum para nossa área. Obrigado pelos atos de incentivo presentes desde o momento em que me tornei seu bolsista e iniciei meus estudos na área da análise de dados e modelagem.

Aos professores Dr. Vitor Borges Monteiro e Dr. Paulo André Dias Jacome, por aceitarem o convite para avaliar o trabalho, e por todas as valiosas críticas e sugestões para a realização dessa pesquisa. Aos meus amigos de dentro e fora da graduação, por serem pessoas iluminadas e nunca faltar palavras de incentivo em situações de grandes dificuldades durante todo o processo da graduação.

RESUMO

A motivação para a presente pesquisa é o fato da possibilidade de compreender interações interpessoais de forma cada vez mais instantânea. Por sua vez, a academia tem contribuído para utilizar o Twitter como uma ferramenta sofisticada para capturar informações em tempo real dos internautas (Fan e Gordo, 2014, p. 76). O objetivo desse estudo foi identificar, por meio das mensagens que são postadas no Twitter, como as informações que são divulgados de forma on-line estão associadas aos movimentos que ocorrem no mercado brasileiro, especificamente no que tange aos retornos e ao volume de negócios. A coleta dos dados relacionados ao Twitter se deu por meio da biblioteca Tweepy. Já os dados financeiros foram obtidos mediante a base de dados *Yahoo Finance*. O índice de sentimento foi empregado após a raspagem dos grandes influenciadores do mercado financeiro na rede social e de suas interações textuais (*tweets*) com seu público. As contas com potencial de influenciar os investidores individuais foram extraídas do perfil do jornalista Sérgio Charlab que possui um algoritmo responsável por realizar este *ranking*. A atribuição do sentimento se deu por meio de *machine learning*, por intermédio da biblioteca *Textblob*. Para atender ao objetivo geral da pesquisa foram estimados dois modelos, o primeiro foi um modelo de regressão Linear Simples e o segundo um uma modelagem de séries temporais VARMAX(8, 0). A análise e intepretação dos dados possibilitou perceber que, em geral, uma modelagem linear simples não iria beneficiar o investidor individual em criar estratégia de investimento. O estudo parte da premissa que o investidor individual é aquele com pouca capacidade em utilizar modelagens robustas para encontrar estratégia de investimento por intermédio da análise de sentimento. Constatou-se ainda que existe a possibilidade de analisar o mercado financeiro através de modelagens não lineares de séries temporais, dado o potencial existente nas variáveis criadas a partir da análise de sentimento, a saber subjetividade e polaridade. Ademais, viu-se que a subjetividade das mensagens possui potencial preditivo ao volume negociado. Assim, os resultados são úteis para mostrar que existe relação entre as informações que são divulgadas na rede social Twitter e os movimentos do mercado acionário brasileiro, trazendo contribuições para a literatura. O estudo também traz contribuições práticas, uma vez que as atividades que ocorrem on-line no Twitter podem ser utilizadas como variáveis em estratégias de investimento, visto que essas estão associadas aos movimentos do mercado quando a modelagem é não-linear.

Palavras-chave: Análise de Sentimento; Twitter; Mercado Financeiro; Modelo de Previsão.

ABSTRACT

The motivation for this research is the possibility of understanding interpersonal interactions in an increasingly instantaneous way. In turn, the academy has contributed to using Twitter as a sophisticated tool to capture real-time information from Internet users (Fan and Gordo, 2014, p. 76). The objective of this study was to identify, through the messages that are posted on Twitter, how the information that is disseminated online is associated with movements that occur in the Brazilian market, specifically with regard to returns and trading volume. The collection of data related to Twitter was done through the Tweepy library. The financial data were obtained through the Yahoo Finance database. The sentiment index was used after scraping the major financial market influencers on the social network and their textual interactions (tweets) with their audience. Accounts with the potential to influence individual investors were extracted from the profile of journalist Sérgio Charlab, who has an algorithm responsible for carrying out this ranking. The attribution of the feeling was done through machine learning, through the Textblob library. To meet the general objective of the research, two models were estimated, the first was a Simple Linear regression model and the second a VARMAX(8, 0) time series model. The analysis and interpretation of the data made it possible to perceive that, in general, a simple linear modeling would not benefit the individual investor in creating an investment strategy. The study starts from the premise that the individual investor is the one with little ability to use robust modeling to find an investment strategy through sentiment analysis. It was also found that there is the possibility of analyzing the financial market through non-linear modeling of time series, given the existing potential in the variables created from sentiment analysis, namely subjectivity and polarity. Furthermore, it was seen that the subjectivity of the messages has a predictive potential for the traded volume. Thus, the results are useful to show that there is a relationship between the information that is disclosed on the social network Twitter and the movements of the Brazilian stock market, bringing contributions to the literature. The study also brings practical contributions, since the activities that occur online on Twitter can be used as variables in investment strategies, since these are associated with market movements when the modeling is non-linear.

Keywords: Sentiment Analysis; Twitter; Stock Market; Forecast Model.

"*What is important in market fluctuations are not the events themselves, but the human reactions to those events.*" (Bernard Baruch)

LISTA DE ABREVIATURAS E SIGLAS

FEAAC	Faculdade de Economia, Administração, Atuária e Contabilidade
UFC	Universidade Federal do Ceará
HME	Hipótese de Mercados Eficientes
NPL	Natural Processing Language

LISTA DE SÍMBOLOS

$	Dólar
%	Porcentagem
~	Aproximadamente
§	Seção
©	Copyright
®	Marca Registrada

SUMÁRIO

1 INTRODUÇÃO

Compreender os fatores endógenos e exógenos que impactam os preços, retornos esperado e volatilidade dos ativos é um estudo intertemporal quando se trata das finanças corporativas e mercado de capitais. A possibilidade de se beneficiar das informações presentes no mercado, surge com a contribuição da teoria da Hipótese do Mercado Eficiente (HME) desenvolvida por Eugene Fama (1970), em outras palavras, seria a capacidade de se beneficiar da presença ou ausência de eficiência de mercado.

Em suma, o mercado é orquestrado por intermédio das expectativas dos agentes econômicos após incorporarem informações presentes nesse mesmo mercado. A HME consiste em estudar a consequência do impacto e a agilidade dos agentes em absorver novas informações relevantes, até o momento em que estas não surtem mais efeitos sobre os preços dos ativos. Posto isto, faz-se necessário encontrar substitutos capazes de fornecer informações referente ao comportamento humano, com o aumento da tecnologia da informação e crescente interação *online* dos investidores, as redes sociais criam alternativa.

Autores como Fan e Gordo (2014) apontam o *Twitter* como uma ferramenta sofisticada para capturar informações em tempo real dos internautas com a finalidade de "*[...] helps identify conversations on social media platforms related to its activities and interests.*" (Fan e Gordo, 2014, p. 76). Esta vantagem ao utilizar as redes sociais, pode ser extrapolada para compreender o impacto entre os agentes influenciadores de opinião e aqueles que recebem essas informações de forma instantânea referente ao mercado acionário através da análise de sentimento, em concordância com os autores já citados acima.

Os estudos comprovam que estresses presentes no mercado podem estar relacionados aos sentimentos dos investidores (DYLIANE, 2020). De acordo com a literatura, autores como Renault (2017), compreendem os impactos das expectativas dos investidores em sentimento é o modelo mais realista da eficiência do mercado em sua forma semiforte - segundo a HME (Hipótese do Mercado Eficiente). As redes sociais, como é o caso do *Twitter* foco deste trabalho, Dyliane (2020, p. 12) afirma que "[...] estudos como o de Mao, Counts e Bollen (2011) comprovam que o sentimento extraído por meio do Twitter, bem como o número de *tweets*, podem ser usados para prever o retorno diário do mercado.".

Portanto, conforme mencionado acima, a análise de sentimento "*[...] is the core technique behind many social media monitoring systems and trend-analysis applications.*" (Fan e Gordo, 2014, p. 76), em outras palavras, a técnica é reconhecida pela academia e pelo mercado como sendo uma forma aceitável de compreender movimentos do mercado acionário brasileiro, surge o seguinte questionamento: **Qual é a capacidade de grandes influenciadores afetarem o sentimento dos investidores individuais via *Twitter*?** Para a presente pesquisa, a análise surgirá da raspagem dos tweets e, posteriormente, será aplicado o método de *deep learning* (*Natural Processing Language* – NPL) e tradução textual para classificar o sentimento do público ao receber as informações dos grandes influenciadores sobre ativos em questão.

1.1 Objetivo

1.1.1 Objetivo Geral

Compreender o impacto causado pela opinião dos grandes influenciadores brasileiros presentes no Twitter sobre investidores individuais.

1.1.2 Objetivo específico

a) Identificar o impacto no retorno esperado do Índice Ibovespa através dos *tweets* de grandes influenciadores por intermédio da análise de sentimento e seu impacto sobre investidores individuais.

1.2 Justificativa

A tentativa de compreender o mercado a ponto de criar vantagens competitivas surge com Fama (1970) com a Hipótese do Mercado Eficiente, objetivo da pesquisa tinha como determinação encontrar uma explicação se os investidores seriam capazes de se beneficiar das informações ao ponto de alcançar retornos esperados acima da média do mercado. De forma complementar, anos após a publicação deste estudo, autores como Kahneman e Tversky (1976) encontraram problemas na fundação do comportamento humano, dessa forma, buscaram provar que no comportamento humano havia lapsos de irracionalidade e, automaticamente, as decisões estarem sendo influenciadas por vieses cognitivos.

Visto que o comportamento humano é duvidoso e nem todas as informações presentes no mercado poderiam ser incorporadas aos preços das ações (eficiência de mercado), muitos cientistas passaram a questionar qual seria a melhor forma de estudar o sentimento humano e seu impacto no mercado financeiro. Assim como afirmam Fan e Gordo (2014), após a década de 2000, com a grande conectividade das interações interpessoais nas redes sociais, os sentimentos das pessoas passam a ser estudados com a finalidade de compreender cada vez mais o comportamento social e suas implicações no mundo real.

A presente pesquisa tem por motivação utilizar o *Twitte*r para compreender melhor o impacto da opinião dos grandes influenciadores no mercado acionário brasileiro. O *Twitter* - do português, "gorjear" - é uma rede social e um serviço de microblog criado em março de 2006 por Jack Dorsey, Evan Williams, Biz Stone e Noah Glass. O microblog permite aos usuários conexão em tempo real com envio e recebimento de mensagens, atualmente, são mais de 465 milhões de usuários, sendo 19 milhões usuários ativos no Brasil.

A coleta de dados será retirada da API do próprio *Twitter* (Tweepi), sendo assim, a análise de sentimento poderá ser realizada através das *hashtags* criadas pelos próprios usuários sobre um determinado assunto em questão, ou pequenas expressões em textos conhecidos como *tweets* e *retweets*. A partir da construção da base de dados, utilizando a linguagem de programação *Python*, será realizada a análise propriamente dita por intermédio da ferramenta NPL (*Natural Processing Language*) uma técnica de *machine learning* capaz de classificar como negativo e positivo uma expressão textual. Em seguida, será realizada a aplicação desta classificação em modelos econométricos com o intento de encontrar alguma conformidade entre a análise de sentimento e o mercado acionário.

2 REVISÃO DA LITERATURA

2.1 A eficiência do mercado e a racionalidade limitada dos agentes econômicos

A compreensão existente entre a eficiência de mercado e o investidor está presente no trabalho desenvolvido pelo acadêmico Eugene Fama em 1970, em conformidade com o autor, a eficiência deveria ser medida através do retorno esperado de três formas diferentes e fundamentada pelos axiomas da "*expected utility theory*".

A primeira forma possível seria a eficiência de mercado em sua forma 1) fraca, nenhuma nova informação absorvida pelos agentes econômicos seria capaz de contribuir para retornos anormais, já que toda esta informação, em algum momento, fora absorvida pelo mercado e refletida aos preços correntes - esta mesma definição está associada à teoria do passeio aleatório.

Na presença da eficiência 2) semiforte é possível aumentar a rentabilidade ao passo que o investidor obtém informações publicadas, sendo essas informações referente a novas emissões, anúncios de dividendos, anúncios de lucros, anúncios de dividendos, como aponta Camara (2013). A eficiência de mercado em sua forma 3) forte, o investidor sobre as mesmas informações mencionadas anteriormente, pode alcançar retornos anormais, sendo esta informação privilegiada ou não.

No entanto, a HME é uma teoria que tem como base a premissa que os investidores são racionais. Anos depois, Fama (1991) se depara com a academia questionando se os investidores seriam capazes de serem racionais frente às decisões de investimento e seu posicionamento ao demonstra que os investidores estão inseridos em bolhas irracionais ao tentar prever retornos no longo prazo. Em seu trabalho, Fama afirma que "*[...] the predictability of long-horizon returns is the result of irrational bubbles in prices or large rational swings in expected returns.*" (Fama, 1991, p. 1578)

Em conformidade com o que foi exposto anteriormente, Kahneman e Tversky (1979) já lidavam com a fundamentação de que os seres humanos não seriam capazes de sempre tomar decisões sob risco de forma racional por conta de influências do viés. Na obra "*Prospect theory: an analysis of decision under risk*", a título de exemplo, Kahneman e Tversky apontam uma nova forma de abordar as decisões dos seres humanos sob risco, ao invés de interpretar as

decisões tomadas pelas premissas da "*expected utility theory*", os autores passam a propor a "*prospect theory*".

A "*prospect theory*" indica que o "*certainty effect*" se modifica quando os indivíduos são colocados para decidir qual decisão tomar em situações de ganho e perda. Após a aplicação de um questionário, os autores do estudo identificaram que os indivíduos preferem optar por escolher a possibilidade de ganhos certos do que ganhar mais correndo riscos, de outra forma, os indivíduos preferem escolher o benefício de ganhar com certeza do que correr o risco.

Por outro lado, quando estes mesmos indivíduos são colocados em situações de perdas sob as mesmas probabilidades encontradas em situações de ganho, preferem correr o risco de perder mais do que escolher a possibilidade de perdas certas, esta contradição ficou conhecida como o "*Reflection Effect*". Segundo Kahneman e Tversky sintetizam a argumentação da seguinte forma: "*The same psychological principle-the overweighting of certainty-favors risk aversion in the domain of gains and risk seeking in the domain of losses.*" (Kahneman e Tversky, 1979, p. 269). Toda esta consequência de fatos, demonstra a quão contraditória é a racionalidade humana quando posta à prova, permitindo que os retornos não fossem a única forma de capturar o sentimento dos investidores no mercado acionário.

2.2 Análise de sentimento, retornos esperados e volatilidade condicional

A literatura já identificou o papel crucial das redes sociais, já que a "*[...] social media - the mid - 2000 - PR agencies would monitor customers' posts on a business's own website to try to identify and manage unhappy customers.*" (Fan e Gordo, 2014, p. 75). Por exemplo, para modelar a satisfação em tempo real dos clientes e aumentar sua fatia de mercado, as empresas passam a utilizar as redes sociais como vantagem competitiva, isto porque "*Being turned into changing customer tastes and behavior, businesses can anticipate significant changes in demand and adjust accordingly by ramping production up or down.*" (Fan e Gordo, 2014, p. 76).

A aplicação das ferramentas digitais não está restrita ao mercado convencional e suas modelagens, graças a estudos já realizados, o presente trabalho tem como objetivo salientar a ligação existente entre as expectativas dos investidores expressas no mercado acionário em forma dos retornos esperados, volatilidade e previsão. A literatura já permite estabelecer uma

relação plausível entre análise de sentimento e o retorno esperado das ações, autores como Baker e Wurgler, na conclusão trabalho realizado em 2006 afirmam que "*[...] the results suggest that descriptively accurate models of prices and expected returns need to incorporate a prominent role for investor sentiment.*" (Baker e Wurgler, 2006, p. 1677).

O estudo realizado pelos autores já citados acima garante que é possível identificar nos sentimentos classificados como sendo baixo, as ações consideradas pequenas - ou ainda, ações jovens, ações de alta volatilidade, ações não lucrativas, ações que não pagam dividendos, ações de crescimento extremo e ações em dificuldades - obtêm retornos consideravelmente altos do que ações mais antigas - baixa volatilidade de retorno, ações lucrativas, ações que pagam dividendos - e quando o sentimento é alto, os padrões mencionados se invertem.

Ao que se refere à volatilidade e a análise de sentimento, em conformidade com os autores Lee e Indro, *[...] shifts in sentiment are negatively correlated with the market volatility; that is, volatility increases (decreases) when investors become more bearish (bullish).*" (Lee e Indro, 2002, p. 2297). Além disso, os autores descobriram que altos retornos estão associados a uma diminuição na volatilidade condicional resultante de maiores mudanças de alta no sentimento para ações de pequena e grande capitalização e na ocorrência de baixo retornos a constatação é a mesma em sentido contrário.

Por fim, há uma relação positiva existente entre sentimento e o prêmio de risco quando considerado empresas com grande volume de captação e empresas com baixa captação, isto é, *[...] the increase in risk premium associated with the hold-more effect is relatively more important than the negative impact of the price-pressure effect on expected return.*" (Lee e Indro, 2002, p. 2297). Dessa forma, assim como Baker e Wurgler, Lee e Indro acreditam na importância e na robustez em utilizar a análise de sentimento do investidor para explicar a formação da volatilidade condicional e do retorno esperado.

2.3 As redes sociais como uma ferramenta com potencial de análise

Está claro que a análise de sentimento é uma importante ferramenta para modelos de volatilidade e retorno esperado das ações. Para a presente pesquisa, se faz necessário a constatação da utilidade do "*Twitter effect*" no mercado acionário por intermédio da análise de sentimento. Como já citado, os autores Fan e Gordo (2014) relatam a relevância das redes

sociais como uma forma incorporar os interesses dos usuários em tempo real, sendo uma ferramenta capaz de gerar vantagens competitivas às empresas.

Em conformidade, no esforço de compreender com maior clareza o papel do *Twitter* para prever os movimentos do mercado de ações, Nisar e Yeung (2018) salientam a notoriedade do *Twitter* para compreender o sentimento do público referente a duas produções cinematográficas, sendo elas: filmes Bruno e a refilmagem Karate Kid. Por intermédio da análise de sentimento, pôde aferir que o filme Bruno foi recebido pelo público de forma negativa e o filme Karate Kid foi recebido de forma positiva, análise responsável por culminar na refilmagem do filme.

Dado que este fato já está fundamentado pela academia, basta compreender o efeito do *Twitter* no mercado acionário. Para Nisar e Yeung (2018, apud Mao *et al*, 2012, p. 4) faz menção à presença de correlação entre o número diário de *Tweets* e a os indicadores de ações do S&P 500 com 68% de chance de prever movimentos de mercado. Na mesma linha de raciocínio, os autores anteriormente citados (2018, apud Rao *et al*, 2012, p. 5), fazem referência a relação preditiva entre humor do *Twitter* (emoções negativas e positivas) e os movimentos das ações de níveis, porém de forma limitada.

Os estudiosos ainda concluíram que existe mais potencial preditivo quando analisado o sentimento dos internautas comparado à análise de volume de mensagens do *Twitter*. De forma mais direta, a conclusão dos autores seria que "*3. Contradictory to past research, volume-based predictors (VTOTAL, VPOS and VNEG) have weak ties with VFTSE and CLOSE, highlighting the superiority of sentiment analytics over volume analytics in stock market prediction studies.*" (Nisar e Yeung, 2018, p. 15). Assim sendo, fica provado que a rede social *Twitter* é uma ferramenta usual para além do "*behavioral modelling*" para os negócios, também possui aplicabilidade no mercado acionário.

2.4 As redes sociais como uma ferramenta com potencial de análise

Visto que o Twitter é uma alternativa para capturar os comportamentos dos investidores via análise de sentimento, basta compreender o papel dos grandes influenciadores sociais (investidores institucionais) sobre os investidores individuais. Os autores Fan e Gordo (2014) argumentam que a identificação de comportamentos futuros surge da modelagem das interações

dos influenciadores, assim este seria o primeiro avanço para determinar uma campanha de “viral marketing" em redes sociais como o Twitter.

Desse modo, a presente pesquisa tem como finalidade buscar compreender se esta identificação possui o mesmo efeito no mercado acionário, ou seja, compreender a relação existente entre os operadores de informação (ou grandes influenciadores) e aqueles conhecidos na academia de operadores de ruído que recebem estas informações e a utilizam como tomada de decisão no mercado brasileiro. Segundo a própria literatura “[...] *studies provide strong and consistent support for the hypothesis that stock prices are affected by both individual and institutional investor sentiments.*” (Verma e Verma, 2018, p. 1140).

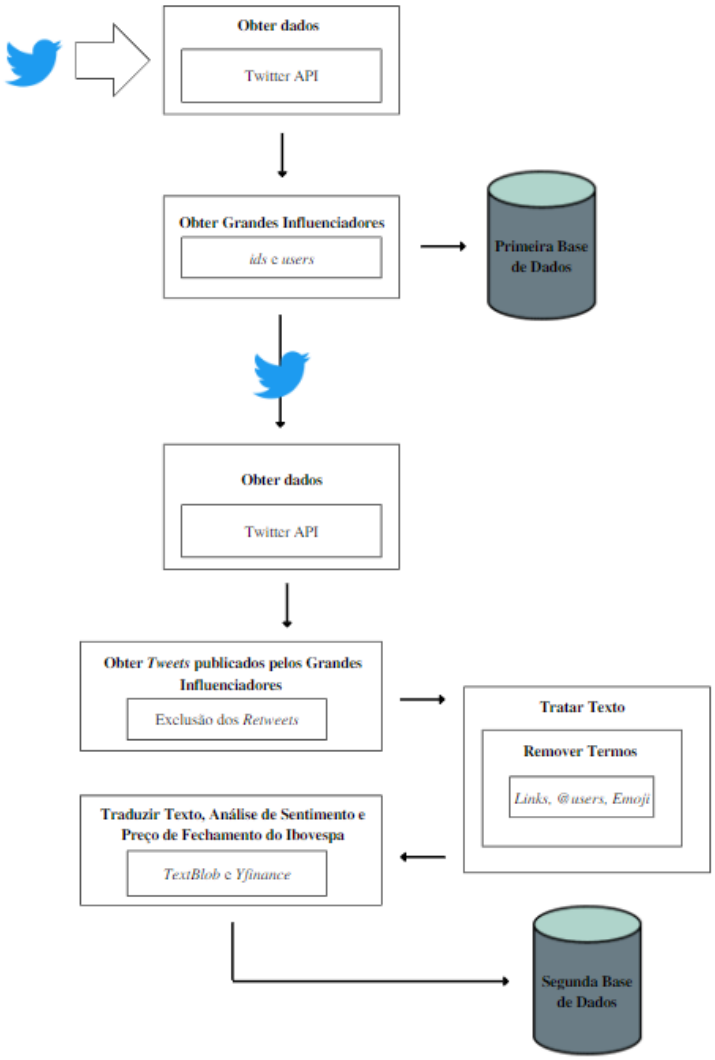
Neste mesmo trabalho, os autores evidenciam que os preços das ações refletem tanto as informações dos operadores de informações, quanto o ruído dos operadores e que os *noise traders* ao agir em conjunto influenciam os preços das ações em equilíbrio, de outro modo, *“DeLong, Shleifer, Summers, and Waldmann (1991) present a model of portfolio allocation by noise traders and show that noise traders as a group can earn expected returns higher than rational investors and can also survive in terms of wealth gain in the long run, due to unpredictability in their sentiments. Similarly, Campbell and Kyle (1993) present a model where stock prices are influenced by competitive interaction between noise traders and rational investors.*” (Verma e Verma, 2018, p. 1141). Posto isso, a produção acadêmica já garante importante relação entre os tipos de operadores ao agir em conjunto e seus impactos através da introdução de risco sistemático precificado nos mercados.

3 METODOLOGIA DA PESQUISA

A presente pesquisa tem por objetivo compreender o impacto dos grandes influenciadores sobre os influenciadores individuais, de outra forma, analisar se há relação entre o sentimento desses mesmos influenciadores e o retorno do Ibovespa. O estudo objetivou analisar o impacto de duas variáveis adquiridas, sendo elas: *polarity* e *subjectivity*. Estas serão variáveis independentes em um modelo matemático de regressão linear múltipla, a finalidade central é compreender quais serão os impactos sobre a tomada de decisão de um investidor individual ao considerar o sentimento geral daqueles que o influencia.

Antes de chegar na aplicação, na Figura 1, há a representação em fluxo de todo o processo de obtenção e tratamento dos dados que será apresentado no decorrer do trabalho.

Figura 1 – Fluxograma



Fonte: Elaboração própria (2022)

3.1 Enquadramento metodológico

A raspagem dos dados foi realizada por intermédio da linguagem de programação *Python* com a utilização da API v2 do *Twitter* com aplicação na modalidade *academic project*, esta responsável por garantir maior flexibilidade na busca dos dados. Na totalidade, foram utilizadas dez bibliotecas para objetivar a raspagem, tratamento e a análise dos dados, sendo elas:

1 – Pandas: manipulação dos dados;
2 – Numpy: manipulação dos dados;
3 – Tweepy: permitiu a conexão com a API v2;
4 – NLTK: análise de sentimento;
5 - TextBlob: análise de sentimento e tradução textual;
6 – Re: tratamento textual;
7 – Yfinance: buscar cotação histórica;
8 – Matplotlib: visualização gráfica;
9 – Pyplot: visualização gráfica.

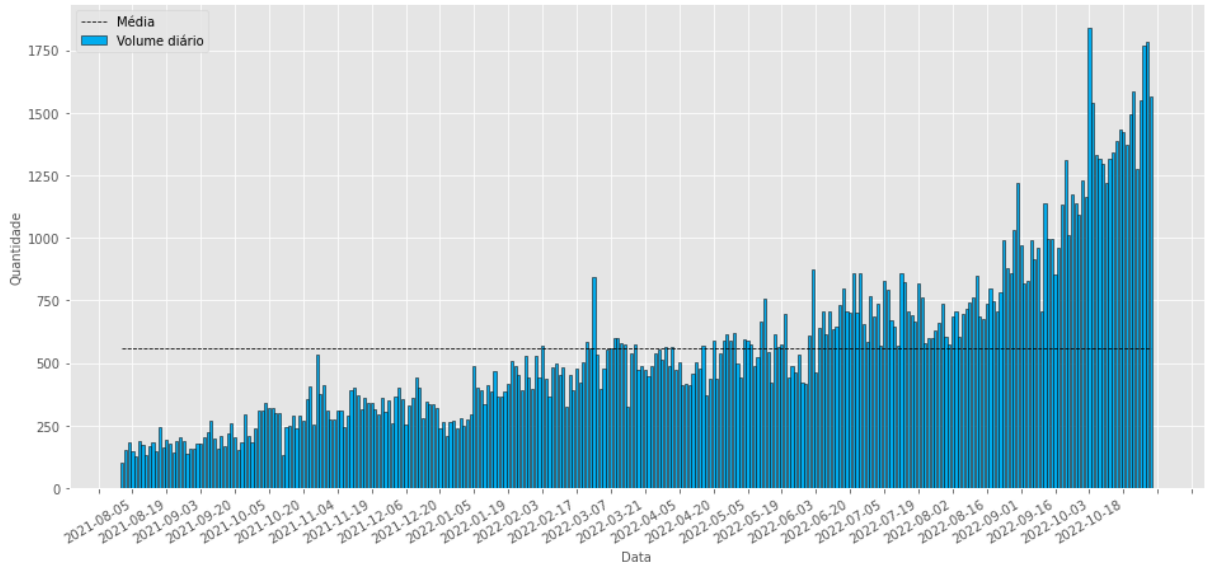
A ideia geral está na busca e tratamento dos *tweets* pelos grandes influenciadores (Primeira Base de Dados), por sua vez, os influenciadores foram definidos através de um *boot* criado pelo jornalista Sérgio Charlab (*id:* @scharlab). De modo geral, mensalmente, o *boot* busca por nomes que tiveram maior interação com o público e cria um *ranking*, este é postado na *timeline* do jornalista. Para a presente pesquisa, a coleta dos influenciadores foi entre a data posição 2021/08/02 até 2022/10/01.

A busca retornou 1.294 influenciadores, porém com resultados em duplicidade, já que no tempo, um único influenciador foi filtrado mais de uma única vez pelo *boot*, assim sendo, a base foi reduzida para 562 influenciadores para o ano de 2022 - representação de ~ 43,43% da base original. Para o ano de 2021, após a remoção de duplicidade a base dos influenciadores para 431. Por fim, para alguns influenciares não foi possível buscar seus respectivos *ids*, condição necessária para obter os *tweets* por meio da API v2 do *Twitter*, a quantidade real dos influenciares totalizou 993.

Em relação ao volume de *tweets* diários (Gráfico 1), houve mais de 680.000 *tweets*

extraídos, porém 169.035 *tweets* foram utilizados para a criação da Segunda Base de Dados após realizar a combinação entre a datas dos tweets e o preço de fechamento do Ibovespa (Figura 1), depois de realizar uma busca com *tweets* acima com 10 caracteres ou mais e considerar como grandes influenciadores aqueles com mais de 15.000 influenciadores.

Gráfico 1 – Quantidade Diária de *Tweets*



Fonte: Elaboração própria (2022)

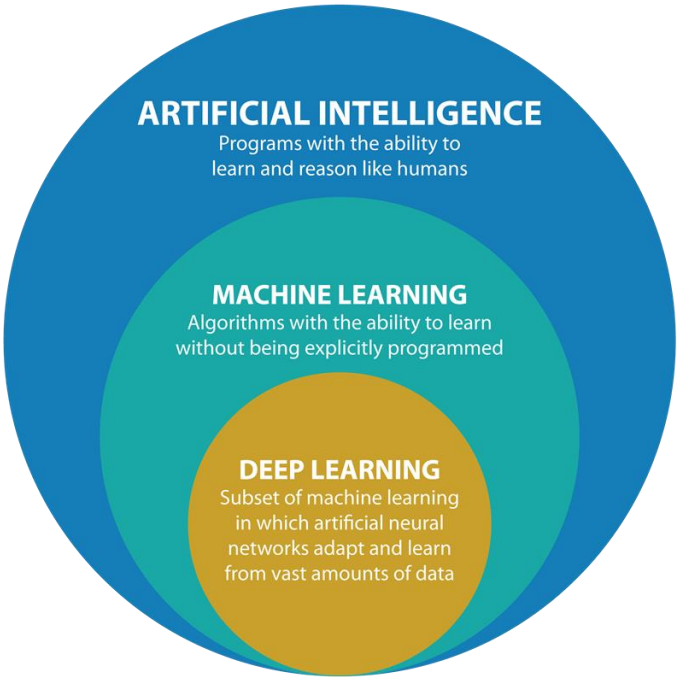
3.2 Modelagem: *Natural Language Processing* (NLP)

O Processamento de Linguagem é uma vertente da Inteligência Artificial contida na subdivisão *deep learning* (Figura 1). Esta é uma área cada vez mais utilizada por grandes conglomerados para desenvolvimento de pesquisa e vantagem competitiva, como por exemplo:

- Diagnóstico médico feito pela classificação de imagens;
- Marketing direcionado pela filtragem de mídia social e análise de dados comportamentais;
- Previsões financeiras feitas pelo processamento de dados históricos de instrumentos financeiros;
- Previsões de demanda de energia e carga elétrica;
- Processo e controle de qualidade;
- Identificação de compostos químicos.

A partir da aplicação dessa técnica a uma grande quantidade de dados e após inúmeras camadas de processamento, os algoritmos conseguem que um computador aprenda por si mesmo e execute tarefas semelhantes às dos seres humanos, tais como a tradução textual, identificação de imagens, o reconhecimento de voz ou a realização de predições, de forma progressiva. Dessa maneira, a NLP é uma técnica que pode limpa, normaliza e converte esse montante de dados textuais em números, para que possam ser processados pelo computador. Para presente pesquisa, foi necessário a aplicação de aprendizado de máquina em dois momentos, 1) tradução textual através da biblioteca *Textblob* e 2) extração da análise de sentimento por intermédio da biblioteca *Natural Language Toolkit* (NLTK).

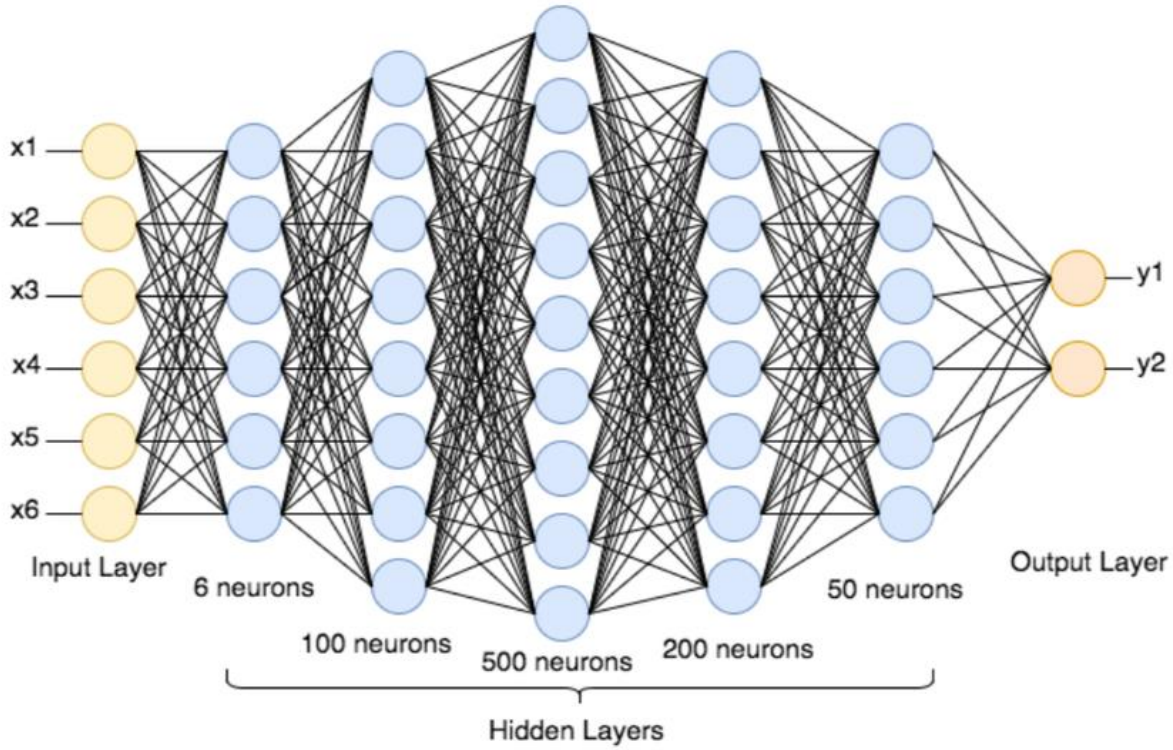
Figura 1 – Subcategorias da Inteligência Artificial



Fonte: Santana, Marlesson (2018)

Um modelo de *deep learning* (Figura 2) capaz de traduzir textos e retornar análise de sentimento de um fragmento textual possui estrutura para processar dados de uma forma inspirada pelo cérebro humano. De outra forma, a rede neural cria um sistema adaptativo que os computadores usam para aprender com os erros e se aprimorar continuamente. As redes neurais artificiais tentam solucionar problemas complicados, como resumir documentos ou reconhecer rostos com grande precisão.

Figura 2 – Arquitetura da Rede Neural Profunda



Fonte: ResearchGate (2018)

De forma simples, a modelagem matemática que fundamenta é conhecida possui várias camadas ocultas, com milhões de neurônios artificiais interligados. Um número, conhecido como peso, representa as conexões entre um nó e outro. A matemática de uma estrutura simples (Figura 3) da Figura 2 é algo como representado na Equação 1.1 e 1.2.

$$y = f(w_1 x_1 + w_2 x_2 + b) \tag{1.1}$$

mais matematicamente, na camada oculta *h* com operação de vetores, este é o produto interno de dois vetores.

$$h = [x_1 x_2 \cdots x_n] \times \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} \tag{1.2}$$

Onde:

x_1 e x_2: valores de entrada;

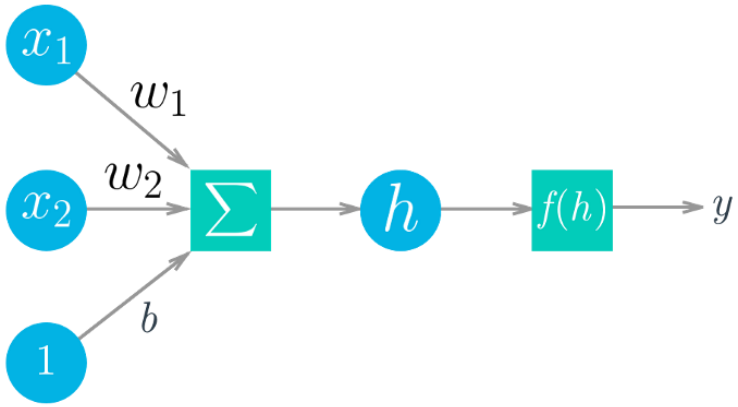
w_1 e w_2: pesos;

b: parâmetro que pode ser aprendido;

h: o *hidden layer* (chamada oculta);

f: função de ativação que receberá a entrada como saída da *hidden layer*.

Figura 3 – Neurônio Simples



Fonte: Sandeep (2021)

Assim sendo, o peso será um número positivo se um nó ativar o outro, ou negativo se um nó desativar o outro. Os nós com valores de peso maiores têm mais influência nos outros nós. Teoricamente, as redes neurais profundas podem direcionar qualquer tipo de entrada para qualquer tipo de saída. Porém, elas precisam de muito mais treinamento do que outros métodos de *machine learning*. Elas precisam de milhões de exemplos de dados de treinamento, enquanto as redes simples talvez precisem de apenas centenas ou milhares.

3.2.1 Análise da polaridade e subjetividade

A *TextBlob* é uma biblioteca Python (2 e 3) para processamento de dados textuais. Ela fornece uma API simples capaz de permitir tarefas comuns de processamento de linguagem natural (NLP), por exemplo, marcação de parte da fala, extração de frase nominal, análise de sentimento, classificação, tradução. Para esta pesquisa será utilizado apenas a função de análise de sentimento e tradução.

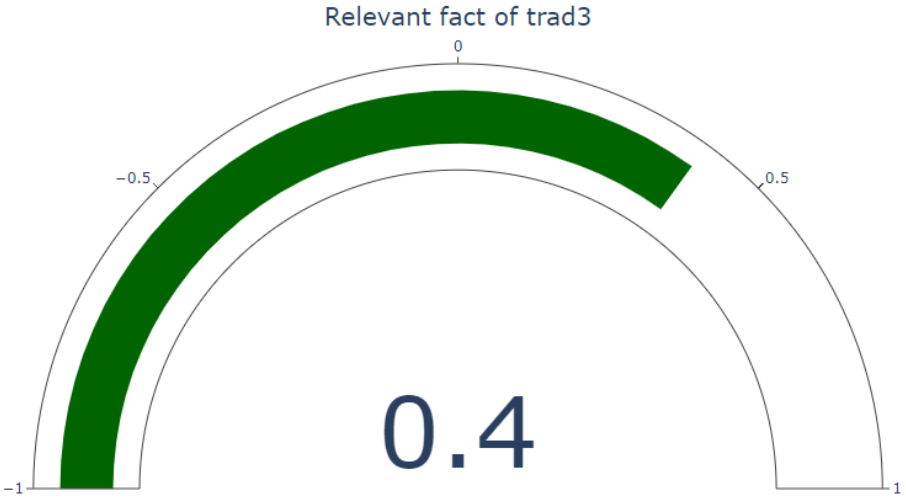
A análise de sentimento é definida por dois parâmetros *polarity* e *subjectivity*. A pontuação de polaridade é uma flutuação dentro do intervalo [-1,0, 1,0]. De outra forma, quanto mais próximo de -1 mais negatividade estará incorporada ao texto e o contrário é verdadeiro, quanto mais perto de 1 mais positividade estará incorporada ao texto e o ponto de neutralidade está presente no marco zero. Nas Figuras 2, 3 e 4 estão representados a classificação dos textos

ao utilizar a biblioteca *TextBlob*.

Gráfico 2 – Grau de Polaridade Negativo

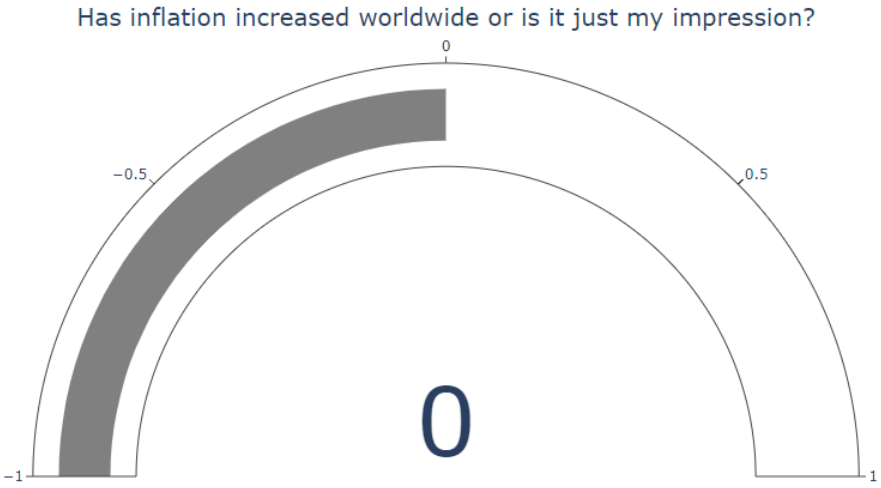
Fonte: Elaboração própria (2022)

Gráfico 3 – Grau de Polaridade Positivo



Fonte: Elaboração própria (2022)

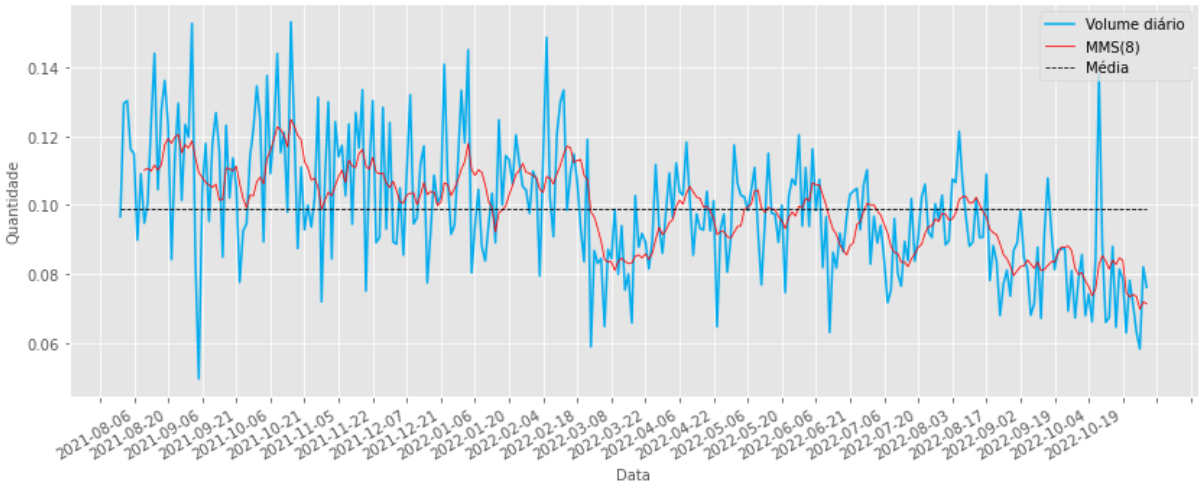
Gráfico 4 – Grau de Polaridade Neutro



Fonte: Elaboração própria (2022)

A polaridade durante a janela de análise se manteve positiva. Através da linha de tendência em vermelho (Gráfico 5) é possível observar que entre 2021-11-01 até 2022-02-18 a polarização se manteve lateralizada acima da média e passa a se manter no movimento sobre a média móvel de oito dias entre a data de 2022-03-01 até 2022-08-17, por fim, a polaridade passa apresentar tendência de queda a partir de 2022-08-17 ao caminhar para uma zona cada vez mais neutra e possível zona de negativa, este fato pode ter relação com o grande volume de postagens causado pelo período das eleições.

Gráfico 5 – Tendência da Polaridade Média Diária



Fonte: Elaboração própria (2022)

Por sua vez, a subjetividade é uma variação dentro do intervalo [0.0, 1.0] onde 0 é muito objetivo e quanto mais próximo de 1 mais subjetivo será a sentença. Veja a seguir, no Gráfico

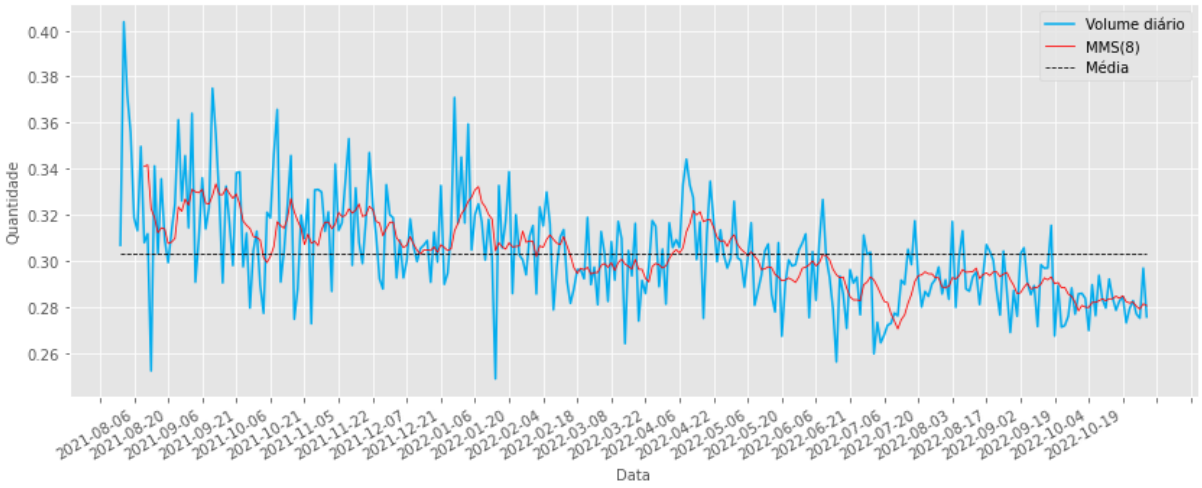
6, a representação da subjetividade da mesma sentença textual utilizada do Gráfico 1.

Gráfico 6 – Grau de Subjetividade

Fonte: Elaboração própria (2022)

Durante a janela de análise, a subjetividade se manteve também positiva, assim como a polaridade (Gráfico 7). É perceptível que a variável supera a média no início do período analisado, mas logo após 2022-08-17, os influenciadores passam a postar conteúdos com menor grau de subjetividade.

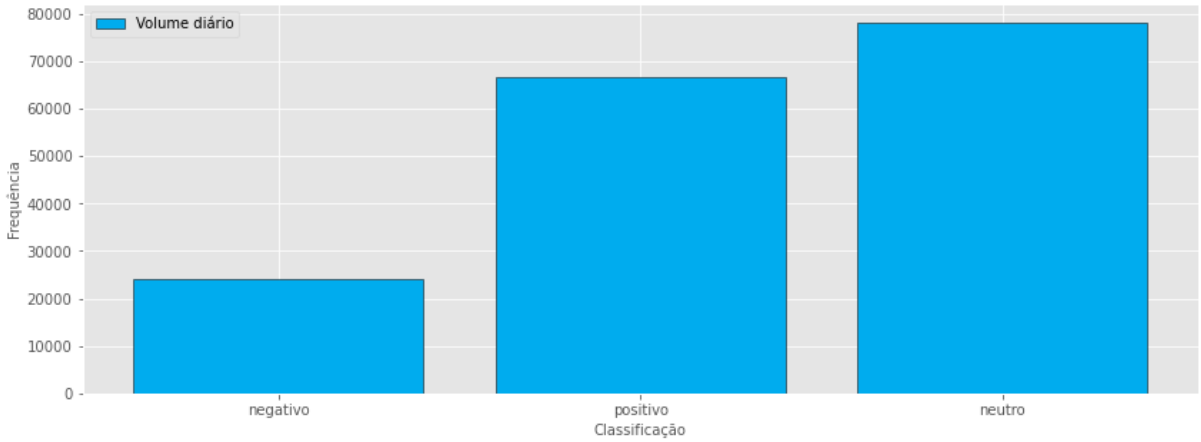
Gráfico 7 – Tendência da Subjetividade Média Diária



Fonte: Elaboração própria (2022)

De modo geral, ao considerar toda a janela de extração dos dados (2021-08-01 até 2022-10-28), a distribuição por grau de polaridade se configurou da seguinte forma: o grau negativo de polarização aparece com 20.420 registros; o grau de polarização positivo registra 55.443 registros; e por fim, a base registra 66.578 para o grau de polaridade neutro.

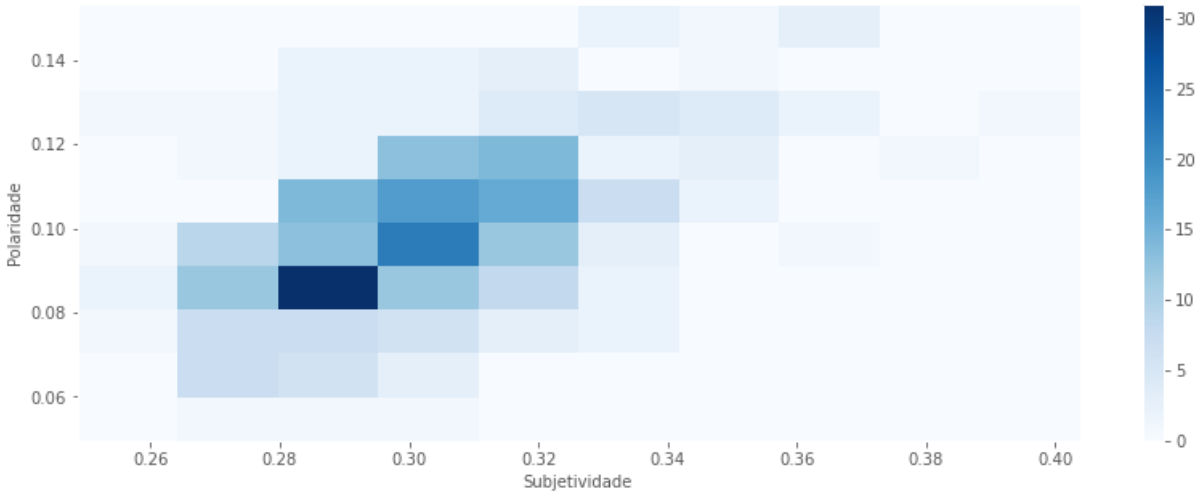
Gráfico 5 – Grau de Polarização



Fonte: Elaboração própria (2022)

Por fim, a subjetividade variou com maior frequência de 0,26 até 0,32 para uma variação de 0,06 até 0,14 na polaridade (Gráfico 8).

Gráfico 8 – Polaridade versus Subjetividade



Fonte: Elaboração própria (2022)

3.3 Modelagem: Regressão Linear Múltipla

O modelo *Naive* utilizado tem como finalidade apenas compreender o comportamento das variáveis independestes em relação ao alvo, não foi encontrado fundamentação teórica que validasse o modelo utilizado estatisticamente, vale ressaltar que a principal motivação é simular a primeira tomada de decisão de um investidor individual em tentar compreender o fenômeno estudado, isto é, se a análise de sentimento é capaz de modelar o retorno do Ibovespa.

O modelo utilizado terá como variável independente o $Retorno_{simples}$ do Ibovespa. As variáveis independentes são Volume, *subjectivity* e *polarity* (Equação 1.3).

$$Retorno_{simples} = \beta_0 + \beta_1 \times Volume_t + \beta_2 \times subjectivity_t + \beta_3 \times polarity_t + \varepsilon \quad (1.3)$$

Onde:
Volume: volume de negociação no dia *t*;
$subjectivity$: subjetividade dos grandes influenciadores em *t*;
$polarity$: sentimento positivo ou negativo dos grandes influenciadores em *t*;
ε: termo de erro.

O $Retorno_{simples}$ variáveis foi derivado do preço de fechamento (Equação 1.4).

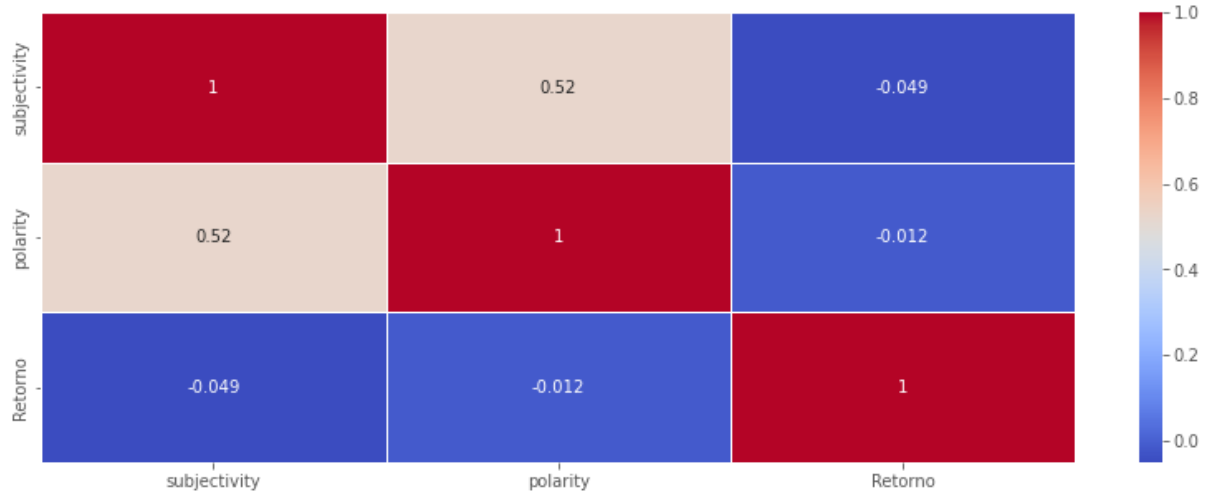
$$Retorno_{simples} = \frac{(Preço\ de\ fechamento_t - Preço\ de\ fechamento_{t-1})}{Preço\ de\ fechamento_{t-1}} \quad (1.4)$$

Onde:
Preço de fechamento$_t$: preço do Ibovespa no dia *t*;
Preço de fechamento$_{t-1}$: preço do Ibovespa em *t-1*.

4 ANÁLISE DOS RESULTADOS

Essa seção traz a análise estatística descritiva das variáveis; uma análise dos gráficos das principais variáveis de interesse da pesquisa; e os resultados sobre a relação entre o sentimento obtido por meio do Twitter sobre o retorno do mercado e da associação entre o volume de *tweets* e o volume de negócios do mercado brasileiro. Há no Gráfico 9, está a representação da correlação entre as variáveis para a melhor compreensão do comportamento entre elas. De modo geral, as variáveis subjetividade e polaridade possuem correlação alta e positiva entre si de 0,52 e correlação baixa e negativa entre as demais.

Gráfico 9 - Matriz de Correlação



Fonte: Elaboração própria (2022)

No Gráfico 10, a partir do dia 2022-09-22 até 2022-10-31, o Índice Ibovespa mantém sua tendência de alta e ao observar este mesmo período para o retorno, no dia 2022-10-04, o retorno do Ibovespa chega a ∼ 8,0% em toda a janela analisada. A volatilidade segue em tendência de alta de 2022-10-28 em diante alcançando resistência em ∼ 2,5% na data 2022-10-04.

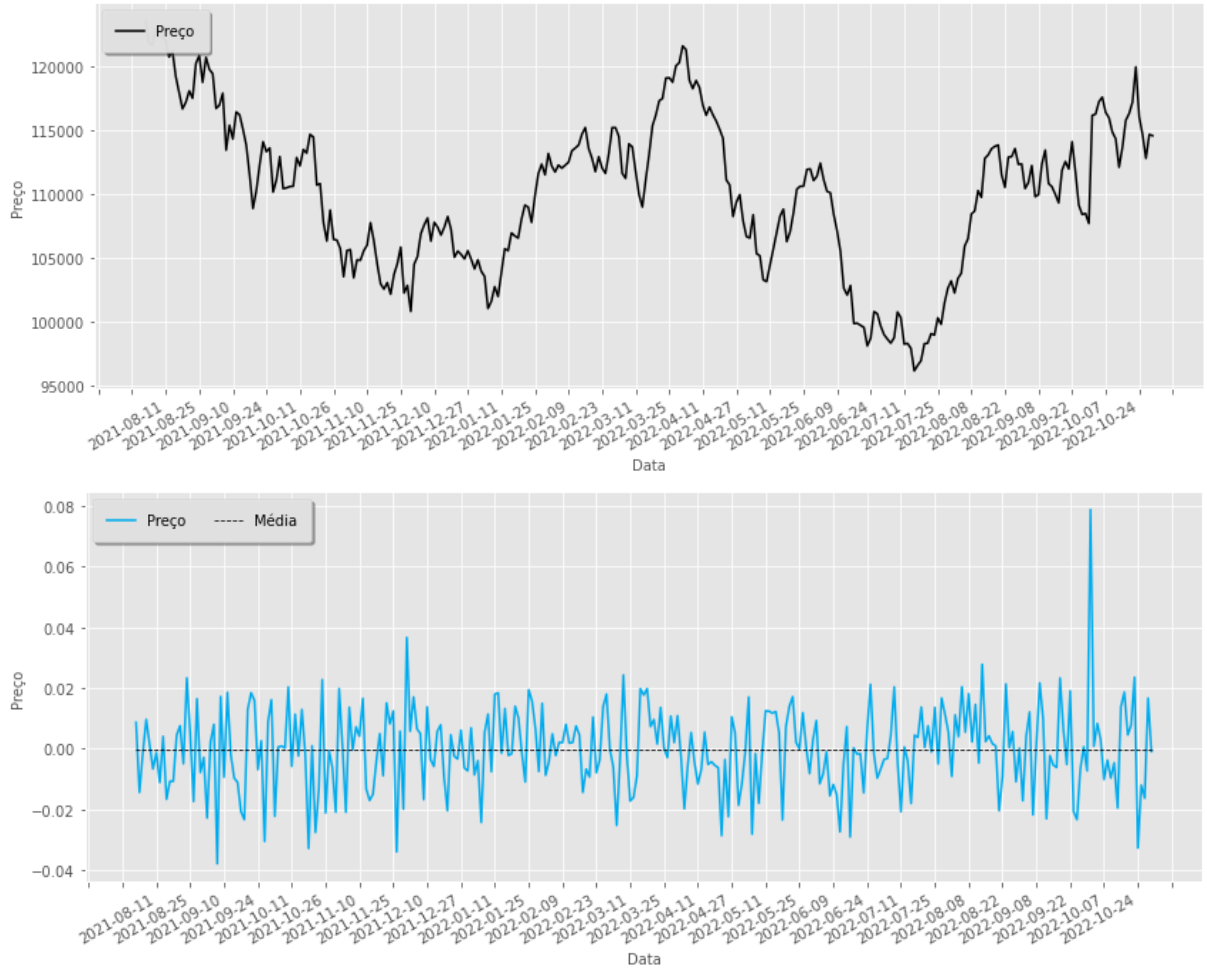
4.1 Treinamento e Teste do Modelo

Para compreender a qualidade de previsão do modelo, a base de dados foi dividida em base de treinamento do dia 2021-08-02 até 2022-07-31 e teste do dia 2022-08-01 até 2022-10-28. As variáveis independentes não possuem relação linear com relação, por conta do baixo grau de ajuste presentado pelo coeficiente de determinação múltipla igual a 0,002 (Tabela 2).

Pelo lado do comportamento das variáveis independentes em relação ao retorno se dá da seguinte forma, as variáveis Volume e *polarity* possuem comportamento positivo em relação ao retorno, na medida em que a *subjectivity* é a única variável do modelo que afeta o retorno de forma negativa (Tabela 2).

Todavia, as variáveis não são estatisticamente significantes, pode-se concluir que as variáveis não possuem capacidade de explicar o fenômeno analisado. De outra forma, as variáveis obtidas da análise de sentimento não são capazes de explicar a variação do Ibovespa. Ao analisar o Gráfico 11 e 12, é possível inferir o mesmo resultado, as variáveis não possuem relação linear com o fenômeno estudado, por esse motivo, o modelo não é capaz de capturar a variação condicional presente no retorno do Ibovespa.

Gráfico 10 – Preço de fechamento e Retorno



Fonte: Elaboração própria (2022)

Tabela 2 – MQO Summary

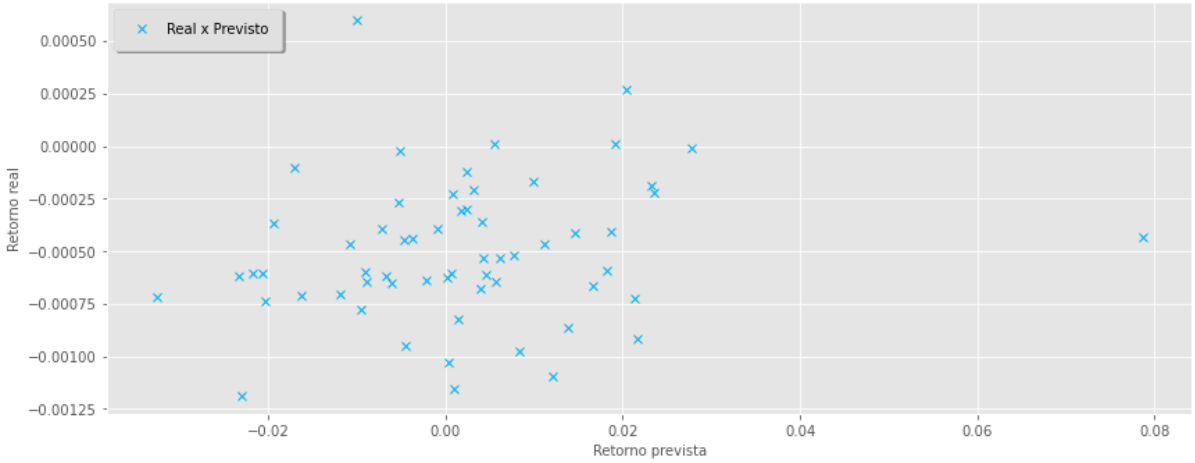
Dep. Variable:	Retorno	**R-squared:**	0.002
Model:	OLS	**Adj. R-squared:**	-0.011
Method:	Least Squares	**F-statistic:**	0.1263
Date:	Sat, 17 Dec 2022	**Prob (F-statistic):**	0.944
Time:	20:39:35	**Log-Likelihood:**	707.37
No. Observations:	241	**AIC:**	-1407.
Df Residuals:	237	**BIC:**	-1393.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0031	0.012	0.258	0.797	-0.021	0.027
Volume	4.986e-11	3.1e-10	0.161	0.873	-5.62e-10	6.61e-10
subjectivity	-0.0218	0.040	-0.541	0.589	-0.101	0.058
polarity	0.0231	0.053	0.438	0.662	-0.081	0.127

Omnibus:	3.877	**Durbin-Watson:**	2.007
Prob(Omnibus):	0.144	**Jarque-Bera (JB):**	3.892
Skew:	-0.308	**Prob(JB):**	0.143
Kurtosis:	2.909	**Cond. No.**	8.04e+08

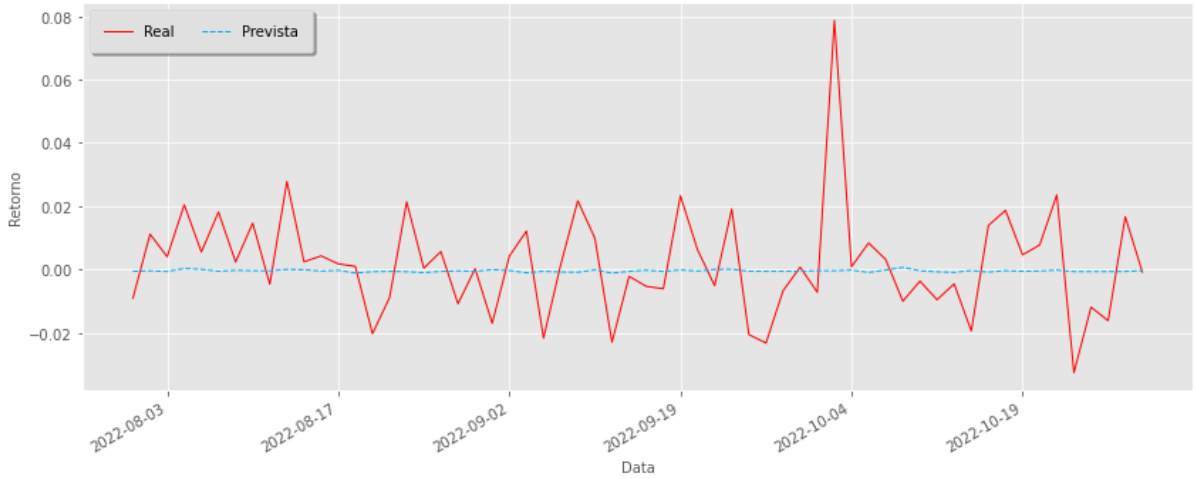
Fonte: Elaboração própria (2022)

Gráfico 11 – Retorno Real versus Retorno Previsto



Fonte: Elaboração própria (2022)

Gráfico 12 – Previsão do Modelo na Base de Teste



Fonte: Elaboração própria (2022)

Para além do coeficiente de determinação, há duas outras medidas para analisar o grau de ajuste de um modelo, o MAE (*Mean Absolute Error*) e o RMSE (*Root Mean Squared Error*), respectivamente, Equação 1.8 e Equação 1.9. O MAE é igual a 0,001 para base de treinamento e se manteve 0.01 e para base de teste igual a 0,012, isto significa que os valores previstos não estão próximos ao valor real, quanto menor este número melhor. Já o RMSE é igual a 0,013 para a base de treinamento e 0,017 para a base de teste, valores distantes, consequentemente, pode-se afirmar que o modelo não está bem ajustado.

$$MAE = \frac{\sum_{i=1}^{n} |\varepsilon_i|}{n} \tag{1.5}$$

Onde:

n: número de observações;

ε_i: erro em i.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y})\text{^}2}{n}} \tag{1.6}$$

Onde:

n: número de observações;

y_i: valor predito em i;

$\hat{y}$: valor real em i.

4.1.1 Análise dos Resíduos

Os resíduos do modelo não estão normalmente distribuídos, dessa forma, não seguem a hipótese de homoscedasticidade e não autocorrelação (Gráfico 13). Outra forma possível, para além da visualização gráfica, seria o teste de Shapiro para provar se a distribuição dos resíduos é normal, para o presente estudo o *p-value* é igual a ~ 0,045, rejeita-se a hipótese nula de que os resíduos são normalmente distribuídos. O teste de Breusch-Pagan com *p-value* igual a ~ 0,011, significa dizer que não há variação constante na variância dos resíduos, assim se rejeita a hipótese nula.

Hipóteses para o teste de Shapiro:

H_0: normalidade na distribuição dos dados → *p-value* ≥ 0,05
H_1: não normalidade na distribuição dos dados → *p-value* ≤ 0,05

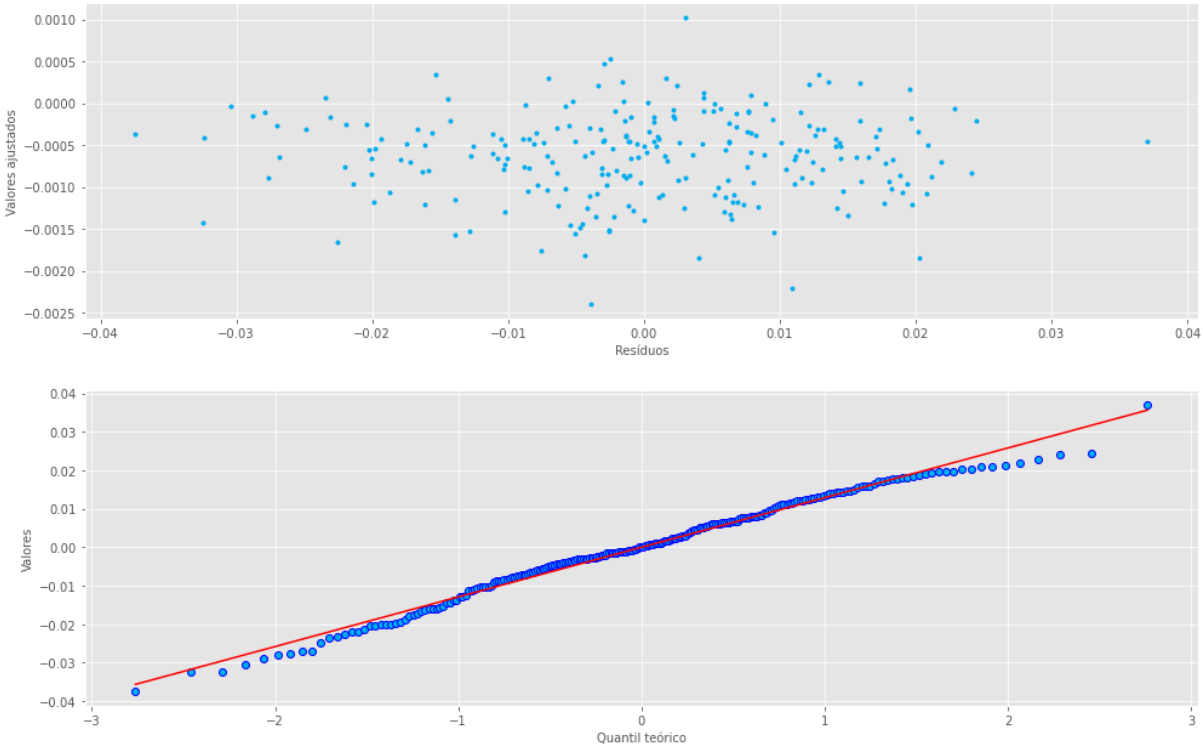
Onde:
H_0: hipótese nula;
H_1: hipótese alternativa.

Hipóteses para o teste de Breusch-Pagan:

H_0: presença de homoscedasticidade → *p-value* > 0,05
H_1: não presença de homoscedasticidade → *p-value* ≤ 0,05

Onde:
H_0: hipótese nula;
H_1: hipótese alternativa.

Gráfico 13 – Resíduos versus Valores Ajustados, Normal Q-Q



Fonte: Elaboração própria (2022)

5 LIMITAÇÕES

A tradução dos *tweets* do português para o inglês é condição necessária para realizar a análise de sentimento ao utilizar as bibliotecas NLTK e TextBlob. Porém, com a base de 2022, houve meses em que o *script* responsável pela tradução passou mais 5h30min em execução, na medida que havia falha de conexão todo o processo era perdido e reiniciado do ponto de partida. Por conta de toda a dificuldade no processamento textual, a criação da base de dados para modelagem foi prejudicada. Embora a quantidade de *tweets* foi satisfatória para a presente pesquisa, no momento de cruzar os dados da análise de sentimento com os dados do preço de fechamento do Ibovespa muitas informações foram perdidas já que os dias corridos foram substituídos por dias úteis.

Outro ponto de crucial de importância foi encontrar a modelagem mais adequada para estudar o comportamento da premissa dado as limitações técnicas dos investidores individuais. A premissa possui ligação direta com o investidor individual recebendo informação de grandes influenciadores para criar racional técnico na tentativa de diminuir vieses cognitivos nas tomadas de decisão, desse modo, a pesquisa objetivou buscar compreender de forma simples o fenômeno estudado por intermédio de uma modelagem clássica estudada na academia, a regressão linear múltipla. Todavia, há grandes limitações em modelar o mercado financeiro com as premissas deste modelo, como já foi exposto no decorrer deste trabalho. Em resumo, as premissas do modelo são:

1 - A forma funcional da relação entre Y e X é:

$$Y = X\beta + \varepsilon \tag{1.7}$$

2 - O valor esperado dos termos de erros aleatórios é zero.

$$E(\varepsilon) = \begin{bmatrix} E(\varepsilon_1) \\ E(\varepsilon_2) \\ \vdots \\ E(\varepsilon_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{1.8}$$

3 - Os erros aleatórios são homocedásticos e são não-autocorrelacionados, e têm a seguinte matriz de variância-covariância, $Var(u)$.

$$Var(\varepsilon) = \sigma^2 \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \quad (1.9)$$

$$Var(\varepsilon) = \sigma^2 \, I \quad (1.10)$$

sendo σ^2 uma constante positiva; I_n uma matriz identidade de ordem n; Var abreviatura de variância; Cov abreviatura de covariância; i e j duas observações diferentes do termo de erro; e n o tamanho da amostra.

4 - A matriz X, de ordem nxK, é não-aleatória; e apresenta colunas (e linhas) linearmente independentes, isto é, essa matriz possui posto igual ao seu número de colunas K. Isto significa que não há multicolinearidade perfeita no modelo. O tamanho da amostra é superior ao número de parâmetros a serem estimados no modelo.

5 - O vetor de termos de erros aleatórios u, de ordem $nx1$, tem distribuição normal multivariada dada por $u \sim N(0, \sigma^2 I)$.

5.1 Recomendações

Ainda que a modelagem utilizada possua limitações na sua estrutura para compreender o retorno do Ibovespa com a análise de sentimento dos *tweets* dos grandes influenciadores, isso não significa que as variáveis são desapropriadas para estudar o fenômeno. Ao realizar uma rápida análise com um modelo multivariado mais robusto de séries temporais, o VARMAX(8, 0) é possível afirmar que há modelos capazes de suprir a complexidade do fenômeno estudado nesta pesquisa.

Após realizar o teste de ADF e fazer a primeira diferença das variáveis para as tornar estacionária é provável determinar causalidade entre as variáveis que compõe o modelo (Equação 1.3) por intermédio do teste de Causalidade de Granger. Este tipo de análise não foi possível de realizar com o modelo de regressão linear múltipla já que os parâmetros do modelo não se mostraram estatisticamente significantes, em outras palavras, variáveis incapazes de explicar a variável dependente ($Retorno_{simples}$).

Não é possível dizer que a variável retorno causa *subjectivity* e vice-versa, a variável *polarity* pode causar retorno na sexta defasagem em diante com 5% de significância (Figura 4). A relação de causalidade entre *subjectivity* e volume possui significância estatística na primeira, segunda e terceira defasagem em 5% (Figura 5).

Figura 4 – *Polarity* causa Retorno

```
Granger Causality
number of lags (no zero) 6
ssr based F test:         F=3.4041  , p=0.0029  , df_denom=282, df_num=6
ssr based chi2 test:   chi2=21.3662 , p=0.0016  , df=6
likelihood ratio test: chi2=20.6279 , p=0.0021  , df=6
parameter F test:         F=3.4041  , p=0.0029  , df_denom=282, df_num=6
```

Fonte: Elaboração própria (2022)

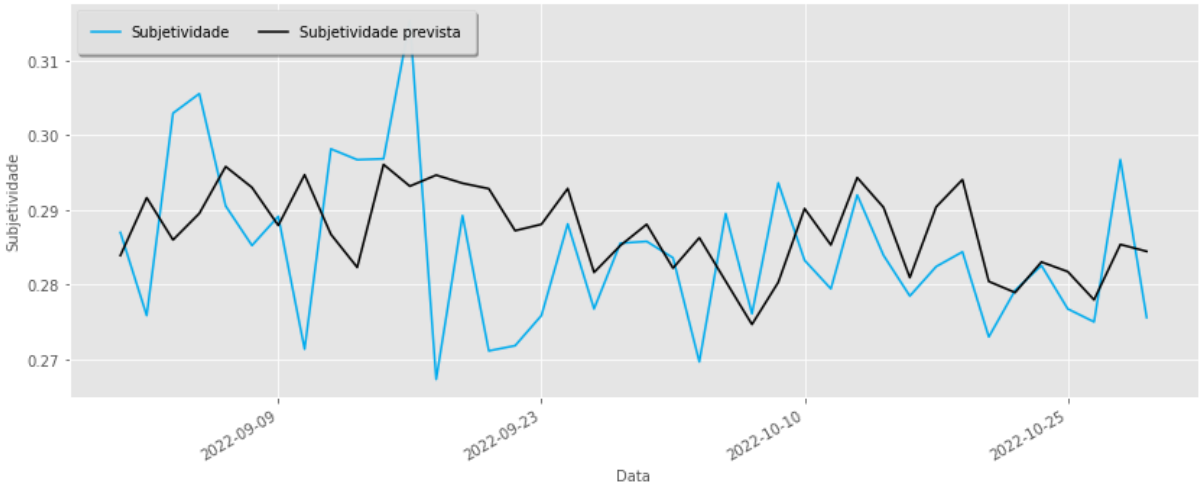
Figura 5 – *Subjectivity* causa Volume

```
Granger Causality
number of lags (no zero) 1
ssr based F test:         F=21.2610 , p=0.0000  , df_denom=297, df_num=1
ssr based chi2 test:   chi2=21.4757 , p=0.0000  , df=1
likelihood ratio test: chi2=20.7419 , p=0.0000  , df=1
parameter F test:         F=21.2610 , p=0.0000  , df_denom=297, df_num=1

Granger Causality
number of lags (no zero) 2
ssr based F test:         F=8.4334  , p=0.0003  , df_denom=294, df_num=2
ssr based chi2 test:   chi2=17.1536 , p=0.0002  , df=2
likelihood ratio test: chi2=16.6796 , p=0.0002  , df=2
parameter F test:         F=8.4334  , p=0.0003  , df_denom=294, df_num=2

Granger Causality
number of lags (no zero) 3
ssr based F test:         F=5.3000  , p=0.0014  , df_denom=291, df_num=3
ssr based chi2 test:   chi2=16.2823 , p=0.0010  , df=3
likelihood ratio test: chi2=15.8531 , p=0.0012  , df=3
parameter F test:         F=5.3000  , p=0.0014  , df_denom=291, df_num=3
```

Fonte: Elaboração própria (2022)

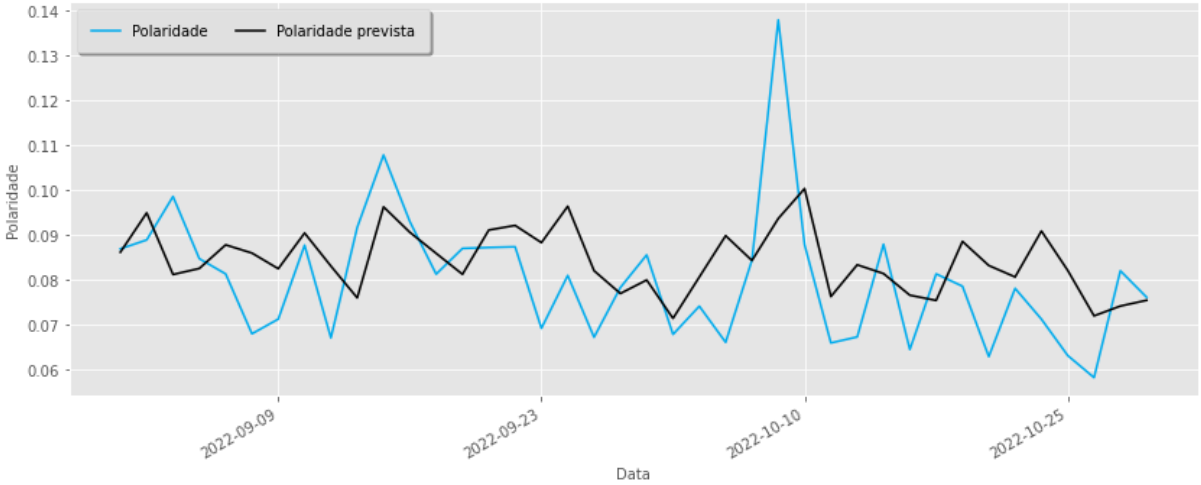
Ademais, ao utilizar a mesma modelagem é possível identificar graficamente (Gráfico 14 e 15) que as variáveis criadas a partir da análise de sentimento (subjetividade e polaridade) possuem características preditivas e, por conta disso, há possibilidade de avançar o estudo nesta área.

Gráfico 14 – Subjetividade Real versus Subjetividade Prevista



Fonte: Elaboração própria (2022)

Gráfico 15 – Polaridade Real versus Polaridade Prevista



Fonte: Elaboração própria (2022)

6 CONSIDERAÇÕES FINAIS

A análise de sentimento sobre a verificação do comportamento dos *tweets* dos grandes influenciadores sobre os investidores individuais como tomada de decisão no mercado acionário brasileiro cumpriu com a proposta da presente pesquisa. As variáveis independentes criadas da análise de sentimento com a aplicação do processo de aprendizado de máquina NLP (*Natural Language Processing*) foram possibilitaram a segunda modelagem do presente estudo (Equação 1.3).

Frente as análises, para minimizar o viés cognitivos ao se basear nos palpites dos grandes influenciadores e criar estratégia de investimento, o estudo comprovou que o investidor individual deverá utilizar modelos com maior rigor matemático. A regressão linear múltipla foi incapaz de capturar qualquer comportamento capaz de gerar eficiência semiforte ou forte ao investidor individual. Ainda assim, as variáveis *subjectivity* e *polarity* podem ser responsáveis por parâmetros significantes para modelar a hipótese quando aplicado o modelo mais adequado.

A sugestão seria utilizar modelos de séries temporais capazes de gerar melhor compreensão do fenômeno estudado. O modelo VARMAX() capturou efeito de causalidade entre as variáveis de sentimento (*subjectivity* e *polarity*) sobre volume e retorno, desse modo, mostrou-se mais eficiente do que o de regressão linear múltipla. O modelo de heteroscedasticidade condicional autorregressiva generalizada (GARCH) ou o DCC-GRACH multivariado poderiam ser outra oportunidade de compreender melhor o fenômeno investigado nesta pesquisa.

REFERÊNCIAS

FAMA, E. F. **Efficient capital markets II. Journal of Finance**, v. 46, n. 5, p. 1575-1617, 1991

FAMA, E. (1970). **"Efficient capital markets: a review of theory and empirical work". The Journal of Finance**. Cambridge, v. 25, p.383-417, 1970

KAHNEMAN, D.; TVERSKY, A. **Prospect Theory an Analysis of Decision under Risk.** Econometrica, v. 47, p. 263-291, 1979

KAHNEMAN, D.; TVERSKY, A. **Investor sentiment and the cross-section of stock returns. Journal of Finance**, v. 61, p. 1645–1680, 2006

LEE, W.; JIANG, C..; INDRO, D. **Stock market volatility, excess returns, and the role of investor sentiment. Journal of banking & Finance**, v. 26, n. 12, p. 2277-2299, 2002

FAN, W.; GORDON, M. **The power of social media analytics. Communications of the ACM**, v. 57, n. 6, p. 74–81, 2014

VERMA, R.; VERMA, P. **Are survey forecasts of individual and institutional investor sentiments rational?**. International Review of Financial Analysis, v. 17, n. 5, p. 1139-1155, 2008

MAO, Y.; WEI, W.; WANG, B., LIU; B. **Correlating S&P 500 stocks with Twitter data**, p. 69-72, ACM, 2012

ROCHA, Francisco José Sales. **Introdução ao modelo de regressão linear clássico. Fortaleza: Edições Universidade Federal do Ceará**, 2016. 124 p. ISBN 978-85-7282-692-1.

PINHEIRO, Nina. **Introdução ao processo de Linguagem Natural – Natural Language Processing (NLP)**, Medium, 2019

SANTANA, Marlesson. **Deep Learning: do Conceito às Aplicações**, Medium, 2018

INTEL. **Deep Learning for Natural Language Processing (NLP)**, Intel, 2018

KIRWAI, Sandeep. **Understand Deep Learning with a sample exercise – PyTorch**, Medium, 2021