Congresso Internacional de Administração
ADM 2020
As Novas Fronteiras da Administração
19 a 21 de outubro Ponta Grossa - PR - Brasil

ESTUDO DA CORRELAÇÃO DA ANÁLISE DE SENTIMENTOS E OS VALORES DAS AÇÕES DA PETROBRAS

STUDY OF THE CORRELATION OF SENTIMENT ANALYSIS AND THE VALUES OF PETROBRAS STOCKS

FINANÇAS: INVESTIMENTO E APREÇAMENTO DE ATIVOS

Eduardo Henrique Kenji Shibukawa, Universidade Estadual de Maringá, Brasil, eduardoshibuka@gmail.com

Wagner Igarashi, Universidade Estadual de Maringá, Brasil, wigarashi@uem.br

Deisy Cristina Corrêa Igarashi, Universidade Estadual de Maringá, Brasil, dccigarashi@uem.br

Resumo

O mercado de ações é conhecido por ser um investimento de alto risco com grande volatilidade nos valores de seus títulos. Com o propósito de prever os valores dessas ações existem duas escolas de pensamento que estudam o mercado financeiro, a escola de análise técnica e a escola de análise fundamentalista. Simplificadamente a primeira escola tem como abordagem a análise de gráficos, busca encontrar padrões para a compra e venda de ações e é mais utilizada para investimentos de curto prazo, enquanto a segunda escola tem como abordagem determinar o valor das ações a partir de fatores externos que afetam os negócios de empresas e suas perspectivas futuras, assim como as demonstrações financeiras da empresa para saber sua situação atual no mercado. Dentro desse contexto este trabalho tem como foco um dos itens da análise fundamentalista, com o objetivo de descobrir se existe correlação entre polaridade de notícias e seu impacto nos valores das ações, assim como a possível classificação de sentimentos de notícias utilizando técnicas da inteligência artificial. Para atingir tal objetivo foram implementados protótipos para cada etapa necessária, sendo eles: a importação de valores de ações, a extração de notícias, a rotulação das notícias e a análise da correlação dos sentimentos com os valores das ações. Utilizou-se dados referentes a Petrobras S.A., no período de maio a outubro de 2017. A análise desses dados mostrou uma correlação baixa, porém consistente entre as polaridades das notícias com os valores das ações da empresa em período referentes. Com os resultados, conclui-se que mais parâmetros devem ser levados em consideração na correlação, tais como peso maior em algumas notícias, um conjunto maior de dados abrangendo mais empresas, com um período maior de extração de dados.

Palavras-chave: Análise fundamentalista, Análise de sentimento, Correlação sentimento, Valor da ação.

Abstract

The stock market is known to be a high risk investment with great volatility in you stock values, with the purpose of predicting the values of the stocks, there are two schools of thought that study the financial market, the technical analysis school and fundamental analysis school, simplifying the first school has as an approach that study charts, seeking to find patterns for buying and selling stocks, commonly used to short term investments, while the second school has as its approach to determine the value of the stocks from external factories that that affect the company business and its future prospects, as well as financial demonstrations of the company to know its current situation in the market. Within the context the current work focuses on one of the items of fundamental analysis, where our goal is to find out whether there is a possible correlation between news feelings and their impact on stock values, as well as classifying sentiments news using artificial intelligence techniques. To achieve such an objective there were implemented prototypes for each stage of the project: an importer of stock values, a news extractor, a news labeler, and an analysis of the correlation of feelings with stock values. Where utilized data from the Petrobras company, from May to October 2017. An analysis of those data showed a low but consistent correlation between

the news sentiments and company stock values in this period. The results lead us to conclude that more parameters need to be taken into account on the correlation, such as greater weight in some news, a larger set of data covering more companies, a longer period of extracted data.

Keywords: *Fundamental analysis, Sentiment classification, Correlation sentiment, Stock market.*

1. INTRODUÇÃO

Em meio à crise que a economia mundial vem passando, muitas pessoas têm procurado meios de conseguir renda extra. Uma dessas formas são os investimentos, segundo Fonseca (2012), “um projeto de investimento pode ser definido como um conjunto de informações que, quando reunidas, possibilitam uma tomada de decisão. Essa tomada de decisão está relacionada à alocação de recursos”. As ações são um tipo de investimento no qual se necessita de conhecimento financeiro e, além disso, do conhecimento sobre a influência do mercado nas empresas em que se quer investir.

Na questão da influência do mercado nas empresas, esta pesquisa enfoca o impacto das notícias na variação de preço de ações. Nesse aspecto utilizou-se subáreas de inteligência artificial tais como: processamento de linguagem natural e análise de sentimento, também chamada de mineração de opinião, área de estudo que tem como objetivo classificar opiniões, sentimentos, avaliações, atitudes ou emoções em relação a produtos, serviços, organizações, problemas, dentro outros (Liu, 2012).

Outra técnica que pode auxiliar nesse processo é a aprendizagem de máquina, que segundo Monard & Baranauskas (2003). Tem como objetivo a construção de sistemas capazes de adquirir conhecimento de forma automática. Para a execução da classificação de sentimento precisa-se antes fazer a extração dos dados, que no contexto do trabalho referem-se a notícias relacionadas à empresa selecionada. Normalmente é utilizado um Web Crawler, um robô que faz a extração automática de dados da web. Com os resultados da classificação do sentimento de notícias de empresas de um período de tempo pré-definido, bem como obtenção da variação dos preços de ações das respectivas empresas, é possível delinear o nível de correlação entre as variáveis envolvidas.

Por existir uma enorme quantidade de dados que é gerada diariamente. Segundo a IBM (2016), diariamente "são gerados 2.5 Quintilhão de bytes de dados, é uma quantidade tão grande que nos 90% dos dados no mundo foram gerados nos últimos 2 anos", porém para esses dados terem valor enquanto informação eles precisam ser analisados. Uma das formas para a análise de dados é o uso da inteligência artificial. Especificamente no caso desta pesquisa o objeto de estudo é a bolsa de valores – BM&FBovespa – e como ela pode ser diretamente afetada por notícias no dia a dia. Assim justificando-se um estudo, utilizando-se da análise de sentimentos das notícias e sua correlação com os valores das ações na bolsa de valores.

Com base no cenário apresentado, foi estabelecido o objetivo aplicar a técnica de análise de sentimentos de notícias, para verificar a correlação entre notícias sobre uma empresa e as variações do preço de suas ações.

2. FUNDAMENTAÇÃO TEÓRICA

Este estudo foi realizado com base em conceitos da área de inteligência artificial aplicadas no mercado financeiro. Sendo assim, serão destacados na sequência elementos relativos ao contexto do problema desta pesquisa, seguido de conceitos relativos ao ferramental necessário a análise a análise de sentimento.

2.1 Mercado financeiro

Antes de enfocar o mercado financeiro e pesquisa aborda aspectos sobre economia. Segundo Holanda Ferreira (2010) economia é a "Ciência que trata da produção, distribuição e consumo das riquezas de uma nação." Além disso, conforme CVM (2019) a economia está estrutura da seguinte a partir de agentes econômicos, como: famílias, empresas ou o governo que compõe o sistema econômico moderno. As famílias oferecem insumos necessários pela empresa, em troca de rendimentos que são salários, juros, lucros e alugueis. Com isso as famílias conseguem comprar os produtos ou serviços oferecidos pelas empresas, e o governo recolhe impostos das famílias e empresas, e devolve para a sociedade em forma de projetos ou serviços sociais.

Segundo Selan (2015) o sistema financeiro é composto por instituições econômicas que têm como objetivo a intermediação entre poupadores e investidores. Ainda, segundo Mankiw (2013), os poupadores depositam e emprestam dinheiro para instituições econômicas, e estas utilizam este dinheiro para financiar alguns setores da economia que estão precisando de recursos.

O mercado de capitais faz parte do mercado financeiro e, neste, os poupadores destinam seus recursos diretamente ao desenvolvimento econômico de forma direta. Por exemplo, empresas que precisam de recursos conseguem financiamento, por meio da emissão de títulos vendidos diretamente aos poupadores, que agora podem ser chamados de investidores. O mercado de capitais pode ser divido entre cinco tipos, sendo eles: mercado de renda variável: mercado de renda fixa; mercado de câmbio; mercado de derivativos; mercado de fundos de investimento. Em nosso trabalho iremos focar na análise de ações que faz parte do mercado de renda variável (Selan, 2015).

2.2 Mercado de ações

Segundo Abreu (2018), as ações são a menor parcela de capital de uma empresa. Elas são títulos que não garantem um retorno fixo dos recursos alocados pelos investidores, uma vez que sua remuneração é determinada pela capacidade da empresa em gerar lucros.

Ao comprar ações, os investidores se tornam sócios da empresa e, caso o investidor mudar de opinião quanto a capacidade da geração de lucro da empresa, ele pode vende-las. Essa negociação é feita pelo intermédio da bolsa de valores e seu funcionamento é regido pela comissão de valores mobiliários (CVM, 2019).

As ações podem ser divididas em dois tipos, as ordinárias e as preferenciais. Na primeira o investidor tem o direito de voto em assembleias de acionistas da empresa, na segunda ele não possui esse direito (Abreu, 2018).

O preço das ações está relacionado diretamente à sua oferta e procura. Quanto maior sua a procura mais seu preço irá aumentar. Sua procura está relacionada a vários fatores tais como: político, econômico, estratégias das empresas, inovações e aumento de competitividade da empresa no mercado, etc. (Abreu, 2002). Para a compra e venda destas ações é comum se utilizar a análise técnica, ou a fundamentalista.

2.2.1 Análise técnica e fundamentalista

Segundo Murphy (1999) a análise técnica se concentra no estudo do mercado de ações, enquanto na análise fundamentalista o foco são fatores externos, indicadores financeiros da empresa, o mercado econômico, etc, que fazem a demanda pelas ações aumentarem ou diminuírem, assim causando alterações no preço das ações.

Na abordagem fundamentalista, existe uma grande diferença entre o valor das ações da empresa e seu verdadeiro ou potencial valor, também chamado de valor intrínseco. Ele é calculado a partir de vários fatores relevantes do mercado econômico, fatores operacionais da própria empresa, tais como receita, custos, etc, e tentam avaliar como tais fatores pesam no valor real da empresa.

Na abordagem técnica se utiliza várias ferramentas, tais como indicadores e desenhos gráficos na identificação de tendências. As técnicas gráficas têm como objetivo encontrar um padrão de crescimento do histórico de preços, a fim de identificar um bom momento para compra e venda das ações. Nesta pesquisa, tem-se como foco a análise fundamentalista, utilizando as notícias como um fator e verificando se elas possuem alguma correlação com o preço das ações.

2.3 Processamento de linguagem natural

Antes de definir o processamento natural, devemos definir o que é a linguagem, de acordo com Holanda Ferreira (2010) temos como definição de linguagem:

> Linguagem formal, linguagem simbólica que serve de axiomas e leis, bem como de normas especiais, em opos. à linguagem natural; Linguagem natural, o conjunto de sinais que se empregam e interpretam indistintivamente (como a fala, o grito, os olhares, os gestos etc.); Faculdade que têm os homens de comunicar-se uns com os outros, exprimindo seus pensamentos e sentimentos por meio de vocábulos, que se transcrevem quando necessário Voz, grito, canto dos animais: linguagem dos papagaios.

Dentro do contexto de linguagem, Russell & Norvig (2013) definem como uma técnica o processamento de linguagem natural, o qual tem como objetivo adquirir conhecimento ao menos parcial, sobre a linguagem que os humanos usam. O problema é que a linguagem natural é ambígua e complexa e está em constante mutação.

Segundo Jurafsky & Martin (2019) para a resolução deste complexo problema devemos separar/analisar a linguagem natural em seis categorias: fonética e fonologia (estudo da linguagem da fala); morfologia (O estudo dos componentes significativos das palavras); sintaxe (o estudo das estrutura e relações entre as palavras); semântica (o estudo do significado das palavras em seu contexto); pragmática (o estudo de como a linguagem atinge seus objetivos); discurso (o estudo de unidades linguísticas maiores que uma única expressão). Em relação a

estas seis categorias, a análise de sentimentos abrange as categorias morfológica, pragmática e em casos mais avançado a semântica. A análise de sentimentos é então delineada com mais detalhes na seção seguinte.

2.4 Análise de sentimentos

Segundo Liu (2012) análise de sentimento, também chamado de mineração de opinião, é a área que estuda e analisa opiniões, sentimentos, avaliações, atitudes e emoções em relação a entidades como produtos, serviços, organizações, indivíduos, problemas, etc.

Para Tsytsarau & Palpanas (2010) a análise de sentimentos é uma análise de subjetividade mais refinada. As duas possuem a mesma essência: a análise de subjetividade tem como objetivo classificar conteúdos em objetivo ou subjetivo e é composta de três passos: a identificação, classificação e a agregação; já a análise de sentimentos tem como função identificar a opinião expressada sobre um assunto em particular e avaliar a polaridade dessa expressão, no caso distinguir se está é positiva ou negativa.

Segundo Liu (2012) as pesquisas relacionadas a análise de sentimentos são separadas por níveis de acordo com sua granularidade. Seus principais níveis são: (a) nível de documento – tem como objetivo classificar a opinião de um documento inteiro, esse nível de análise leva em consideração que cada documento expressa uma opinião sobre uma única entidade, (b) nível de sentença – tem como objetivo analisar se a sentença foi expressada como positiva, negativa ou neutra, sendo fortemente relacionada com a análise de subjetividade citada anteriormente, (c) nível de entidade ou aspecto – é baseada na ideia de que uma opinião consiste em um sentimento positivo ou negativo e um alvo. Seu objetivo é descobrir o sentimento de entidades e suas diferentes características. É o nível mais complexo de análise.

Para a resolução de nosso problema precisamos definir o que é uma opinião em nosso contexto. A opinião pode ser classificada em dois tipos, a opinião regular e a opinião de comparação. No primeiro tipo de opinião é expressado um sentimento sobre um alvo, já no segundo tipo é expressado um sentimento de um alvo em relação a outro alvo. Na literatura a opinião regular é citada regularmente como opinião, e em nosso trabalho iremos nos focar nesse tipo de opinião.

A opinião consiste em dois componentes principais: um alvo e o sentimento em relação ao alvo ou (g, s), em que g é a entidade meta (*goal*) sobre a qual alguma opinião foi expressada, e s é um sentimento positivo, negativo ou neutro, podendo ser também uma avaliação numérica expressando a intensidade do sentimento.

Hu e Liu (2004; 2010) definem a opinião como uma tupla de cinco elementos $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$, onde e_i é o nome da entidade i, a_{ij} é um aspecto j da entidade i; s_{ijkl} é o sentimento do aspecto a_{ij} da entidade e_i; h_k é o detentor (*holder)* da opinião, e t_l é o momento quando a opinião foi expressada por h_k.

Com essa definição conseguimos definir os objetivos e as principais tarefas da análise de sentimentos. Nosso objetivo pode ser definido como: dado um documento de opinião d queremos descobrir todas opiniões em d, sendo que cada opinião pode ser representada pela tupa $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$.

Para facilitar nossa explicação iremos definir modelos de entidade e e documento de opinião d:

- Modelo Entidade: Uma entidade e_i é representada como um conjunto finito de aspectos $A_i = \{a_{i1}, a_{i2}, \dots , a_{in}\}$, podendo ser representada como qualquer uma das entidades dentro do conjunto final de entidades de expressões $OE_i = \{O_{ei1}, O_{ei2}, \dots , O_{ein},\}$. Cada aspecto $a_{ij} \in A_i$ da entidade pode ser representado como aspecto de expressão contido no conjunto finito de aspectos de expressões $AE_{ij} = \{a_{eij1}, a_{eij2}, \dots a_{eijm}\}$
- Modelo Documento de opinião: Um documento de opinião d contém um conjunto de entidades$\{e_1, e_2, \dots , e_r\}$, e um subconjunto de seus aspectos de um conjunto de detentores de opiniões, em um determinado momento no tempo.

Resumidamente, dado um conjunto de documentos D, a análise de sentimentos consiste em seis tarefas principais. A primeira tarefa é a extração de entidade e categorização, que consiste na extração de todas as entidades de expressão e categorização ou agrupamento de entidades de expressões sinônimas, em suas respectivas categorias. Cada categoria representa uma única entidade e_i.

A segunda tarefa é a extração de aspectos e categorização, que consiste na extração de todos os aspectos de expressão das entidades e a categorização de seus aspectos de expressão em seus clusters. Cada cluster do aspecto de expressão da entidade e_i representa um aspecto a_{ij}. A terceira tarefa é a extração do detentor de opinião e categorização, que consiste na extração dos detentores de opinião do texto ou dos dados estruturados, e sua categorização. A quarta tarefa é a extração de tempo e uniformização, que consiste na extração do tempo na qual opiniões foram obtidas, e na uniformização desses dados. A quinta tarefa é a classificação de sentimento do aspecto, que consiste em determinar se a opinião em um aspecto a_{ij} é positivo, negativo ou neutro, ou atribuir sua avaliação numérica ao sentimento do aspecto. A sexta e última tarefa é a geração da tupla de cinco elementos, que consiste em produzir todas as tuplas de cinco elementos $(e_i, a_{ij}, s_{ijkl}, h_k, t_I)$ contidas no documento d baseado no resultado das tarefas acima.

O modelo apresentado acima, pode ser considerado um framework para obtermos os dados necessários para a análise de sentimentos dos documentos e dos aspectos alvos. Para a tarefa 5, classificação dos sentimentos do aspecto é necessário o uso de alguma ferramenta para ajudar a resolver o problema de classificação. Em nosso caso iremos nos focar no aprendizado de máquina, que tem esse intuito como um dos seus objetivos.

2.5 Aprendizado de máquina

De acordo com Barber (2016):

> Machine learning is the body of research related to automated large-scale data analysis. Historically, the field was centred around biologically inspired models and the long term goals of much of the community are oriented to producing models and algorithms that can process information as well as biological systems. The field also encompasses many of the traditional areas of statistics with, however, a strong focus on mathematical models and also

> prediction. Machine learning is now central to many areas of interest in computer science and related large-scale information processing domains.

Segundo Barber (2016) o aprendizado de máquina é fundamentalmente sobre extrair informações de grandes conjuntos de dados, muitas vezes tendo como motivação produzir algoritmos capazes de imitar ou melhorar o desempenho humano.

A área de aprendizagem de máquina pode ser dívida em várias subáreas. Os dois principais campos da área podem ser considerados o de aprendizado supervisionado e o aprendizado não supervisionado. A área de aprendizado supervisionado tem o foco de melhorar a acurácia dos resultados da predição, enquanto o outro de encontrar descrições plausíveis sobre os dados.

Nosso foco neste trabalho será no aprendizado supervisionado, que segundo Poole & Mackworth (2010), pode ser abstraído como mostrado a seguir. Assumindo que sejam disponibilizados os seguintes dados ao nosso classificador: um conjunto de atributos de entrada, $(x_1, x_2, \dots, x_n)$; um conjunto de atributos alvos, $(y_1, y_2, \dots, y_n)$; um conjunto de dados de treino, em que os atributos de entrada e atributos alvos são fornecidos para cada exemplo; um conjunto de dados de teste, em que apenas os atributos de entrada são fornecidos; o objetivo é prever os valores dos atributos alvos, para os exemplos de teste e de dados ainda não fornecidos.

Aprendizado é a criação de uma representação que consegue fazer predições baseado nos atributos de entrada de novos exemplos. Outra visão da aprendizagem supervisionada é a de Russell & Norvig (2013), dado um conjunto de treinamento com n pares de entrada x e uma saída y, na qual toda saída foi gerada por uma função desconhecida $y = f(x)$, o objetivo é descobrir uma função hipótese h, que se aproxime da função $f(x)$.

Para a escolha da melhor hipótese h, existem vários modelos de algoritmos, tais como o de regressão linear, as redes neurais, as máquinas de vetor de suporte, o modelo Bayesiano, entre outros. Em nosso trabalho, decidimos escolher o modelo Bayesiano para a resolução do problema de classificação do sentimento das notícias extraídas, pois o mesmo é um modelo simples de ser implementado, e tem resultado satisfatório.

2.6 Aprendizagem Bayesiana

Segundo Poole & Mackworth (2010) a ideia do aprendizado bayesiano é computar posteriormente a distribuição de probabilidades das características de novos exemplos em conjunto com todos os exemplos posteriores. Para Russell e Norvig (2013), a aprendizagem bayesiana simplesmente calcula a probabilidade de cada hipótese, fazendo previsões de acordo com os dados posteriores. Em vez 22de utilizar apenas a melhor hipótese, as previsões são feitas com o uso de todas hipóteses, ponderadas por suas probabilidades.

Supondo que D é a representação dos exemplos posteriores, com valor observado d, a probabilidade de uma hipótese h_i pode ser obtida pela regra de Bayes, $P(h_i|d) = \alpha P(d|h_i)\mathrm{P}(h_i)$, supondo que queremos fazer a previsão de um valor desconhecido X, teremos $P(X|d) = \sum_i P(X|d, h_i)P(h_i|d) = \sum_i P(X|h_i)P(h_i|d)$, então pressupomos que cada hipótese determina uma distribuição de probabilidade sobre X. Com isso temos um equação que mostra que as previsões são médias ponderadas sobre as previsões das hipóteses individuais.

Para facilitar a visualização do funcionamento da aprendizagem bayesiana, vamos considerar um exemplo simples. Temos um pacote de doces, que pode ter até dois sabores de doce dentro dele, cereja ou lima. Os pacotes são sempre produzidos respeitando uma respectiva relação, sendo elas, 100% cereja, 75% cereja e 25% lima, 50% cereja e 50% lima, 25% cereja e 75% lima, 100% lima. Vamos abreviar essas relações respectivamente para $\{h_1, h_2, h_3, h_4, h_5\}$.

Dado um novo pacote de doces, com uma hipótese aleatória H denotando o tipo do pacote de doces, com valores possíveis de h_1 até h_5, e a medida que os doces são retirados do pacote, são revelados os dados $\{D_1, D_2, \dots, D_n\}$, onde cada D_i, é um valor aleatório que contem o sabor do doce, sendo ele cereja ou lima, temos como objetivo prever o sabor do próximo doce retirado do pacote, sabendo que a hipótese a priori sobre $\{h_1, h_2, h_3, h_4, h_5\}$., é $\{10\%, 20\%, 40\%, 20\%, 10\%\}$. (Russell & Norvig, 2013).

Com o uso da fórmula de Bayes, supondo que os dados são independentes e identicamente distribuídos, conseguimos calcular a probabilidade de algum dado com a fórmula $P(d|h_i) = \prod_j P(d_j|h_i)$. Para facilitar o entendimento, levando em consideração que retiramos o doce do sabor de lima em 10 iterações, podemos destacar que a partir do oitavo doce de lima retirado do pacote a probabilidade do pacote ser inteiro de lima é de aproximadamente 80% (Russell & Norvig, 2013).

3. DESENVOLVIMENTO

Nesta seção iremos falar sobre as ferramentas utilizadas, sobre o modelo utilizado, e o funcionamento de cada etapa. A Figura 5 ilustra a arquitetura do experimento.

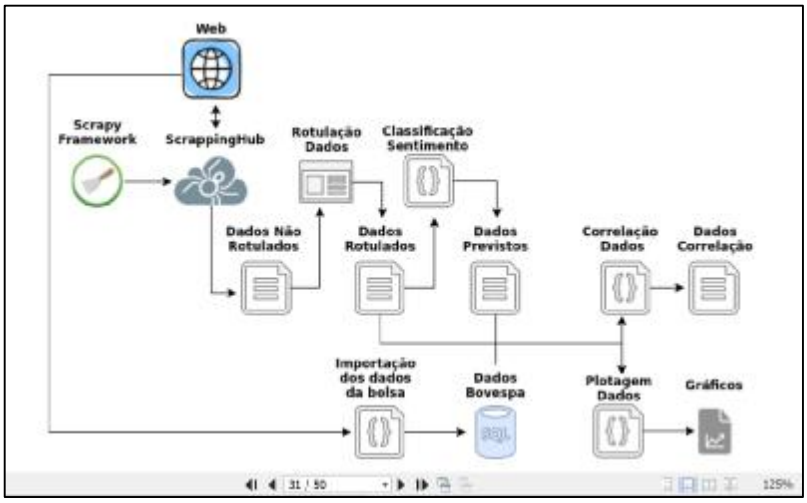

Figura 5 - Arquitetura

De acordo com a Figura 5, primeiramente temos a aplicação de importação de dados da bolsa que insere os dados em um banco de dados. Após isso, temos uma ferramenta criada utilizando o framework Scrapy em conjunto com a plataforma ScrappingHub, que tem a função de extrair as notícias não rotuladas. Em nossa próxima etapa temos um aplicativo que facilita a rotulação de dados que tem como entrada as notícias não rotuladas e como saída as notícias rotuladas. A seguir uma aplicação lê os dados rotulados e os classifica gerando dados com a análise de

sentimentos. Com isso nos resta calcular as correlações tanto dos dados rotulados quanto dos dados de previsões com os valores da bolsa que estão no banco de dados, gerando informações sobre a correlação dos dados, e gerando os gráficos dos dados para facilitar a análise, todas essas etapas estão explicadas com mais detalhes nas seções posteriores.

Para a importação dos dados das ações, utilizamos os arquivos do site da BM&FBovespa, o qual permite obter o histórico de ações em um arquivo, a partir de filtros diversificados, sendo eles por ano, por mês de um ano, ou por um dia específico. Os dados do arquivo podem ser interpretados seguindo um *layout* fornecido no próprio site.

Assim, seguindo esse layout e utilizando a biblioteca *peewee*, foram importados os dados de transações relativa às ações da Petrobras (PETR4) e armazenados em um banco de dados SQLLite nomeado bovespa.db, para ser usado posteriormente.

Par a extração de notícias foi utilizada a linguagem python, em conjunto com o framework Scrapy, sendo criados dois crawlers na plataforma scrapinghub, para extrair os dados do site de notícia G1. Para a obtenção das notícias, utilizamos a pesquisa do google, limitando o período de nossa pesquisa, e utilizando os parâmetros allintitle:string e site:string, conseguimos filtrar respectivamente, os termos pesquisados no título dos sites e buscar as URLs somente de um domínio específico, assim conseguindo uma consulta com os links para a extração.

Para extrair os dados das notícias desta consulta, foi utilizado o framework Scrapy e, seguido da criação de um novo extrator (*spider*) de notícias. A execução do extrator de notícias utilizou como parâmetros “site:G1” e “allintitle: petrobras”, com o período de 01/05/2017 a 31/10/2017, o que gerou um total de 141 notícias. Em relação a cada notícia buscamos basicamente três propriedades para nosso item, que são: o título, a data de atualização, e o conteúdo da notícia.

Para a análise de sentimentos de notícias, utilizando a abordagem de aprendizado de máquina supervisionado, foi necessário rotular todos os dados extraídos o que foi feito de modo manual.

Após a etapa de extração de notícias, foram retirados todos os caracteres que não fossem letras (como números ou outros símbolos e convertendo o texto todo em formato caixa baixa. Na sequência foi realizada a tradução dos textos para inglês utilizando as funcionalidades da biblioteca TextBlob, e realizada a remoção de *stopwords*.

Inicialmente, para a análise de sentimentos foi utilizada a abordagem léxica. Para isso, utilizamos a biblioteca TextBlob, tendo como objetivo buscar a polaridade de cada notícia. Nessa abordagem foi realizada a divisão do documento em uma lista de palavras, calculando a polaridade de cada uma das palavras. Porém não obtivemos um bom resultado. Em um segundo momento, o documento foi divido em sentenças, porém a análise continuou com um resultado insatisfatório. Com base nos resultados, decidiu-se utilizar o método de aprendizado de máquina.

Para utilizar uma abordagem de aprendizagem de máquina foi utilizada a biblioteca scikit-learn, em conjunto com outras bibliotecas, tais como a TextBlob, e a NLTK, a partir das quais foi construído um modelo para classificar os sentimentos entre positivo, neutro ou negativo. Mais especificamente foi selecionado o modelo de Bayes.

O treinamento do modelo de Bayes foi realizado a partir do pré-processamento dos dados rotulados anteriormente, referentes ao período de maio a agosto, a partir dos quais foi criado um pseudomodelo de espaço vetorial no formado *bag of words*.

O mesmo processo foi aplicado para as notícias do período de setembro, as quais foram convertidas para o pseudomodelo vetorial. Neste momento, o modelo de Bayes foi utilizado para realizar a análise de sentimento das notícias de outubro.

O coeficiente de correlação de Pearson mede a relação linear entre dois conjuntos de dados. Estritamente falando, a correlação de Pearson exige que cada conjunto de dados seja normalmente distribuído. Como outros coeficientes de correlação, varia entre -1 e +1 sendo 0 uma correlação neutra. As correlações de -1 ou +1 implicam uma relação linear exata. As correlações positivas implicam que, à medida que x aumenta, o mesmo acontece com y. As correlações negativas implicam que, à medida que x aumenta, o valor y diminui. (Scipy, 2017)

Para a correlação dos dados foi utilizada a biblioteca scipy, analisando a correlação entre a análise de sentimentos fornecida pelo modelo de Bayes e os valores das ações. A correlação entre o sentimento de um dia X e o valor da ação no mesmo dia. Também foram realizadas correlações com defasagem de 1 a 7 dias.

4. ANÁLISE DE RESULTADOS

Os registros dos dados da análise de sentimentos obtidos pela abordagem léxica não foram relevantes e tendiam a zero, assim sua análise foi desconsiderada. Então seguiu-se para a utlização do modelo de aprendizagem de máquina, o qual foi treinado e ser testado foi análise em relação as algumas métricas as quais podem ser observadas por meio da matriz de confusão, Tabela 1.

Fonte dos dados	Kappa	Precisão	Recall	FScore
Teste	45.24%	65.66%	66.93%	65.97%
Outubro	39.69%	63.33%	60.00%	60.16%

Tabela 1. Métricas da análise do modelo

De acordo com os dados da Tabela 1, percebe-se que a precisão obtida ao se utilizar o modelo para a classificação de dados de teste e de dados referentes ao mês de outubro ficaram em torno de 63,33% a 65,66%. A métrica Kappa nos indica uma concordância entre moderada e razoável, e a precisão também pode ser vista como razoável em ambos os casos. Os valores das métricas obtidas não foram muito elevadas, entretanto deve-se considerar que o problema de análise de sentimentos tem grau elevado de subjetividade.
Os registros dos dados da análise de sentimentos para o mês de outubro com a abordagem do aprendizado de máquina podem ser observados na Tabela 2.

data	titulo	sentimento
02/10/2017	Petrobras anuncia US$ 63 bi em operações de pagamento renegociação e contratação de dívida Empresa pagou antec ...	1
02/10/2017	Ministro de Minas e Energia prevê que Petrobras será privatizada: É um caminho Fernando Coelho Filho afirmou n ...	1
03/10/2017	Não estamos tratando disso diz ministro sobre privatização da Petrobras Fernando Coelho Filho (Minas e Energia ...	1
04/10/2017	Comitê do setor elétrico pedirá que Petrobras forneça combustível para térmicas paradas Comitê destacou que a ...	1
04/10/2017	Petrobras está perto de atingir meta de desalavancagem A companhia que tem reduzido trimestre a trimestre sua ...	1
05/10/2017	Indústria naval deve recorrer de decisão da ANP que libera Petrobras para construir casco de Libra no exterior ...	0
05/10/2017	Petrobras vai analisar pedido do governo por ajuda a térmicas mas sem subsídio O Comitê de Monitoramento do Se ...	1
05/10/2017	Petrobras vê como positiva flexibilização de conteúdo local em reserva do pré-sal Na véspera ANP flexibilizou ...	0
10/10/2017	Petrobras eleva preço do botijão de gás em 129% a partir desta quarta Estatal estima que preço ao consumidor f ...	0
11/10/2017	Cade vê como complexa compra de ativos da Petrobras por mexicana e pede estudos Venda da Petroquímica Suape e ...	1
11/10/2017	TCU bloqueia bens de Dilma por prejuízo à Petrobras com compra de Pasadena Além da ex-presidente decisão ating ...	-1
16/10/2017	Petrobras pede registro de companhia aberta para BR Distribuidora Pedido de IPO inclui também aval para realiz ...	1
17/10/2017	Moody's eleva nota da Petrobras e muda perspectiva para estável Agência cita redução do endividamento da petrol ...	1
18/10/2017	Produção da Petrobras no Brasil sobe 28% em setembro ante agosto Na comparação com setembro de 2016 houve qued ...	-1
19/10/2017	Governo exclui fatia da Petrobras na Braskem de programa de desestatização Medida também exclui a fatia da pet ...	-1
19/10/2017	Carf decide a favor da Petrobras em processo de R$ 78 bilhões diz estatal Estatal divulgou comunicado nesta qu ...	1
24/10/2017	Petrobras quer vender a BR Distribuidora 'o mais rápido possível' Empresa tem meta de vender ativos que somam ...	1
24/10/2017	Petrobras fará parceria para disputar blocos do pré-sal em leilão na sexta-feira Presidente da estatal adianto ...	0
26/10/2017	Petrobras aprova adesão a Nível 2 de governança da bolsa e corte de gerências Patamar intermediário de governa ...	1
26/10/2017	Petrobras pode incluir ativos de logística em vendas de refinarias; detalhes devem sair no 1º tri diz Parente ...	0
26/10/2017	Petrobras será 'seletiva' mas 'muito firme' no leilão do pré-sal diz Pedro Parente ANP pós à venda 8 blocos de ...	0
26/10/2017	Petrobras adere ao programa de regularização de dívida não tributária Segundo a estatal adesão resultará em im ...	1
27/10/2017	Petrobras 'não podia se dar ao luxo de perder essa oportunidade' diz Parente sobre alta oferta nos leilões Est ...	0
27/10/2017	Com lances altos Petrobras leva 3 blocos do pré-sal e oferece até 80% da produção à União No regime de partilh ...	0
31/10/2017	Petrobras busca parcerias para concluir obras do Comperj diz Parente Presidente da estatal revelou a expectati ...	1
31/10/2017	Parceria da Petrobras com empresa britânica BP deve envolver troca de ativos diz Parente Estatal divulgou cart ...	1

Tabela 2. Sentimento utilizando modelo de Bayes

Os dados da correlação de Pearson e os gráficos relevantes ao mesmo podem ser observados abaixo na Tabela 3.

Fonte Dados	Dias transcorrido	Correlação	pvalue
Dados de maio	0	16.96%	53.00%
Dados de junho	0	3.91%	89.45%
Dados de julho	0	-9.19%	76.53%
Dados de agosto	0	4.62%	89.92%
Dados de setembro	0	-16.37%	54.46%
Dados de outubro	0	22.58%	43.77%
Dados da previsão	0	-14.97%	60.95%
Dados de maio	1	45.98%	7.32%
Dados de junho	1	4.50%	87.86%
Dados de julho	1	23.81%	43.34%
Dados de agosto	1	19.04%	59.84%
Dados de setembro	1	-33.18%	20.92%
Dados de outubro	1	43.21%	12.29%
Dados da previsão	1	15.71%	59.16%
Dados de maio	2	43.17%	9.50%
Dados de junho	2	22.21%	44.53%
Dados de julho	2	14.32%	64.08%
Dados de agosto	2	51.83%	12.48%
Dados de setembro	2	-33.38%	20.64%
Dados de outubro	2	32.28%	26.03%
Dados da previsão	2	-0.86%	97.67%

Dados de maio	3	41.98%	10.54%
Dados de junho	3	25.28%	38.33%
Dados de julho	3	10.16%	74.11%
Dados de agosto	3	28.07%	43.21%
Dados de setembro	3	-28.33%	28.77%
Dados de outubro	3	42.80%	12.89%
Dados da previsão	3	9.12%	75.65%
Dados de maio	4	37.04%	15.78%
Dados de junho	4	30.48%	28.93%
Dados de julho	4	6.67%	82.86%
Dados de agosto	4	33.48%	34.43%
Dados de setembro	4	-25.16%	34.72%
Dados de outubro	4	37.30%	18.90%
Dados da previsão	4	13.25%	65.16%
Dados de maio	5	38.57%	14.01%
Dados de junho	5	26.69%	35.63%
Dados de julho	5	4.07%	89.51%
Dados de agosto	5	1.26%	97.25%
Dados de setembro	5	-7.52%	78.21%
Dados de outubro	5	43.15%	12.34%
Dados da previsão	5	11.71%	69.01%
Dados de maio	5	38.57%	14.01%
Dados de junho	5	26.69%	35.63%
Dados de julho	5	4.07%	89.51%
Dados de agosto	5	1.26%	97.25%
Dados de setembro	5	-7.52%	78.21%
Dados de outubro	5	43.15%	12.34%
Dados da previsão	5	11.71%	69.01%
Dados geral	0	18.12%	10.12%
Dados geral	1	24.52%	2.55%
Dados geral	2	24.55%	2.53%
Dados geral	3	24.68%	2.45%
Dados geral	4	24.75%	2.41%
Dados geral	5	25.21%	2.15%
Dados geral	6	22.01%	4.56%
Dados geral	7	17.95%	10.44%

Tabela 3 – Correlação entre a análise de sentimentos e o preço da ação da Petrobras

De acordo com os dados apresentados na Tabela 3, as análises foram realizadas utilizando dados rotulados manualmente, e também utilizando os dados da previsão para o mês de outubro. Analisando os dados percebe-se que os sentimentos de maneira geral possuem uma correlação baixa com os valores da bolsa de ações. Porém alguns meses como maio e outubro possuem uma correlação superior. Os limites das correlações vão de 51,83% a -33,38%.

Ao se agrupar todos os dados temos um correlação baixa, porém relevante com o valor médio de 20%, começando com um valor de 18,12% com 0 dias transcorridos e chegando em seu pico de 25,21% de correlação após 5 dias transcorridos, verificando os valores de p-value, e assumindo um nível de significância de 5% como referência conseguimos observar que com os dados de todos meses agrupados conseguimos um valor bom de 2.15% que é menor do que nosso valor de referência 5%. Verificando os gráficos conseguimos perceber que esses valores fazem algum sentido pois conseguimos perceber alguns padrões onde aparenta-se que os sentimentos impactam os valores das ações após alguns dias. Conseguimos também observar

os dados de nossa previsão e compará-los com os dados do mês de referência que seria o de outubro, verificando assim uma divergência bem alta entre os dois, algo esperado levando a precisão de aproximadamente 60% da classificação dos sentimentos.

Levando em conta todos os fatores analisados supomos que para melhorar o valor da correlação encontrada deveríamos considerar mais parâmetros, e não somente os sentimentos das notícias do dia, além disso precisaríamos de dados de mais fontes e não somente do site G1.

Após a análise de correlação, foram realizadas algumas pesquisas adicionais em sites de busca, e supõe-se que algumas notícias podem ter um impacto maior que outras e isso não pôde ser levado em consideração na análise de correlação. Um exemplo que pode ser citado foi uma notícia de setembro do site da Exame que informa as 10 melhores ações para se investir em setembro segundo 18 corretoras, e dentre estas a ação da Petrobras estava em primeiro e, após essa notícia os valores da Petrobras tiveram um resultado positivo de maneira geral, neste sentido, verifica-se um maior impacto de uma notícia provinda de outra fonte, que não a utilizada no presente trabalho.

5. Desafios e Dificuldades

No decorrer do estudo foram encontradas várias dificuldades. A primeira foi a dificuldade em encontrar dados referentes a sentimentos de notícias rotulados. Foi realizada a busca em vários sites, porém o único lugar que foi encontrado algo do gênero foi no site da kaggle. Um dataset que possuía as 25 notícias mais populares do dia no site reddit, a data e o sentimento. Porém, não possível utilizar os dados, pois as notícias não possuíam nenhuma relação com a Petrobras.

A falta de uma base de notícias rotuladas, direcionou o estudo para a atividade de extração de notícias, na qual foram encontradas outras dificuldades. A principal delas foi filtrar os dados por determinado período e de uma fonte específica de forma prática. Isto foi possível por meio do mecanismo de pesquisa do Google, utilizando a plataforma ScrapingHub.

No decorrer do estudo foi verificado que a seleção da fonte de dados pode ter um impacto significativo na qualidade da classificação das notícias em relação aos sentimentos.

6. Considerações Finais

O mercado de ações é complexo e volátil impactado por diversos fatores. Um desses fatores pode ser encontrado na análise fundamentalista, como notícias, que podem ser analisadas para tentar prever essa volatilidade. É nesse ponto que o objetivo geral deste trabalho foca, criando um classificador de sentimento de notícias a fim de se obter a correlação destas com os valores das ações.

Para atingir o objetivo deste trabalho, foi realizada a extração de notícias do site G1 em relação às notícias da Petrobras. Para tal, foi realizada uma pesquisa com as técnicas para a extração de dados de sites da internet, assim como suas ferramentas. Foram escolhidas a linguagem Python em conjunto com o framework Scrapy, utilizando a plataforma ScrapyCloud da empresa ScrapingHub.

Para a "Análise de sentimentos das notícias extraídas", foi feito um estudo sobre as técnicas de classificação de sentimento e, primeiramente, decidimos utilizar a técnica léxica em conjunto

com uma ferramenta Python chamada TextBlob. Porém, após resultados não satisfatórios e mais estudos, decidimos usar o aprendizado de máquina, utilizando o modelo de Bayes para a solução do problema. Utilizando a ferramenta scikit-learn chegamos a resultados razoáveis com uma precisão de 63,33% e um Kappa de 39,69%.

A "Análise da correlação dos dados da análise de sentimentos com os dados de valorização ou desvalorização das mesmas" foi feita utilizando a biblioteca SciPy. Para essa análise foi preciso importar os dados das ações da BM&FBOVESPA. Utilizamos a biblioteca Peewee para armazenar os dados em um banco de dados SQLite. Com isso fizemos a análise dos valores das ações da Petrobras com os sentimentos rotulados para o treino da análise de sentimento, e com os valores da previsão do mês de outubro. Os resultados mostraram uma correlação consistente, porém baixa com os dados de todos os meses agrupados com uma média 20%.

De modo geral, se pode concluir que os sentimentos de notícias são subjetivos, e a análise de sentimentos pode auxiliar a indicar o impacto destas na variação de preços de uma ação. Entretanto, ela não deve ser utilizada de forma isolada, pois há outras variáveis que podem impactar na variação de preço de uma ação.

Como trabalhos futuros, pode-se sugerir a ampliação deste trabalho para um número maior de empresas de tempo de análise. Extração de dados de mais fontes de informação como a Exame e a InfoMoney. Rotulação das notícias por profissionais da área financeira. Agregação da análise de sentimentos aliada a outros ferramentas como indicadores técnicos.

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*. Tese de Doutorado, PGEA, Universidade de Brasília, Brasília, DF, Brasil.

Holanda Ferreira, A. B. (2010). *Dicionário Aurélio da língua portuguesa*. Editora Positivo.

Barber, D. (2016). *Bayesian reasoning and machine learning*. Cambridge University Press.

Mankiw, N. G. (2019). *Introdução à Economia*. Editora Cengage Learning.

Abreu, J. C. (2018). *Administração financeira II*: finanças para empreendedores experientes e executivos financeiros. Editora FGV.

CVM – Comissão de Valores Mobiliários. (2019). *Mercado de Valores Mobiliários Brasileiro*. 4ª edição.

Fonseca, J. W. F. da. (2012). Análise e Decisão de Investimentos. Curitiba, PR? IESDE Brasil.

Hu, M.; Liu, B. Mining and summarizing customer reviews. In: *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2014. Disponível em: /doi.acm.org/10.1145/1014052.1014073>.

IBM. Entenda porque o Big Data é o petróleo do século 21. 2016. Disponível em: /www.ibm.com/blogs/robertoa/2016/03/entenda-porque-o-big-data-e-o-petroleo-do-seculo-21/>.

Jurafsky, D.; Martin, J. H. (2019). *Speech and Language Processing*: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. MIT Press.

Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers.

Monard, M. C.; Baranauskas, J. A. (2003). *Sistemas inteligentes*: fundamentos e aplicações. Manole Ltda.

Murphy, J. J. (1999). *Technical analysis of the financial markets*: A Comprehensive Guide to Trading Methods and Applications. New York Institute of Finance.

Poole, D. L.; Mackworth, A. K. (2010). *Artificial Intelligence*: foundations of computational agents. Cambridge University Press.

Russell, S.; Norvig, P. (2013). *Inteligência Artificial*. 3ª edição. Elsevier.

SCIPY. Scipy | scipy.stats.pearsonr. 2017. Disponível em: /docs.scipy.org/doc/scipy-0. 19.1/reference/generated/scipy.stats.pearsonr.html>.SCRAPY. Scrapy | A Fast and Powerful Scraping and Web Crawling Framework. 2017. Disponível em: /scrapy.org/>.

Selan, B. (2015). *Mercado Financeiro*. Universidade Estácio de Sá, SESES.

Tsytsarau, M.; Palpanas, T. (2010). Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, 24, 478–514.