ANÁLISE DE NOTÍCIAS E SEU IMPACTO NO MERCADO FINANCEIRO ATRAVÉS DE INTELIGÊNCIA ARTIFICIAL

NEWS ANALYSIS AND ITS IMPACT ON THE FINANCIAL MARKET THROUGH ARTIFICIAL INTELLIGENCE

MONTICO, Guilherme B.; COSTA, Ivan R. N. ;
Graduandos do Curso de Engenharia da Computação – Universidade São Francisco
guilhermebmontico@gmail.com, **nascimento.rapha@hotmail.com**

RESUMO. O presente trabalho propõe um estudo voltado à Bolsa de Valores com foco em análise do impacto que as notícias no âmbito jornalístico geram nas ações das empresas. Após a atuação no investimento de compra de títulos na Bolsa, foi identificado que logo em seguida que uma notícia é publicada, os acionistas instantaneamente refletem sua intenção com a informação gerada com a venda ou a compra de seus títulos, gerando assim a oscilação das ações. Com isso, é elaborado o desenvolvimento de uma técnica de predição do valor de ações de empresas com capital aberto na Bolsa de Valores através da análise do conteúdo de notícias provenientes de fontes de informações de alto impacto publicadas em mídias sociais, fundamentado em pesquisas bibliográficas com base em acervos eletrônicos (e.g. livros, artigos científicos e monografias) visando a automação de uma técnica computacional que determina uma solução otimizada para servir de apoio ao *trader*, decorrente de um sistema estabelecido onde há uma imensa quantidade de negociação de títulos por segundo. Inicialmente será analisado dados estatísticos registrados no passado dentro do mercado financeiro das empresas Petrobrás e Vale S.A. para iniciar a amostragem de valores para a regressão linear. Logo após é desenvolvido nas notícias o apuramento de análise de sentimento baseado nas informações publicadas através da matéria-prima negociada pelas empresas, resultando em um grau de influência que as informações das mesmas geraram na Bolsa de Valores. Com isto este processo será feito diversas vezes para então ser aplicado a regressão linear obtendo assim a correlação entre preço da ação e notícia. As fontes dos dados serão obtidas através da interface web da B3 - Brasil, Bolsa, Balcão, onde publica diariamente o registro de negociações e o andamento de seus pregões. Diante do apresentado, conclui-se que após os levantamentos quantitativos e proeminentes de uma demanda considerável de recursos, o investidor não dispõe de um suporte adequado com uma ferramenta útil que se adapte à exigência de uma atual tecnologia, a qual alcançou um patamar além do qual o humano pode acompanhar, sendo assim há a exigência de se criar também um método eficaz de apoio ao negociante em tal sistema.

Palavras-chave: Inteligência Artificial, Machine Learning, Regressão Linear, Análise de Sentimento, Mercado Financeiro.

ABSTRACT. The present work proposes a study on the Stock Exchange focused on analysing the impact that news in the journalistic sphere generate on campanie's stock values. After the investment in the Stock Market, it was identified that shortly after a news release is published, shareholders instantly reflect their intention with the information generated by the sale or purchase of their securities, thus generating the oscillation of securities. Thereby it is elaborated a development of a prediction technique of stocks values from companies with public offerings on the Stock Exchange between analysis from news contents coming from high impact informations sources published on social medias, based on bibliographic studies

reasoned on electronic collections (e.g. books, scientific articles and monographs) desiring a computational technique automation that determines an optimized solution to provide support for the trader, in view of a established system where has a large stock market negotiation rounds per second. Initially will be analysed statistics data registered on the past inside the stock market from Petrobrás and Vale S.A. to begin the values sampling to the linear regression. Right after is developed on the news the determination of sentiment analysis based on his published informations, resulting a influence rate that the informations generate on the stock market, and therefore the process will be made several times for then be applied the linear regression, leading thus the correlation between stock price and news. The data source will be obtained from the web interface B3 - Brasil, Bolsa, Balcão, where is daily published the negotiation registers and the progress of the trading floor. In front of the introduced, conclude that after the quantitative collection and prominent from a resource considerable demand, the investor doesn't dispose from a good support with a useful tool that adapts the exigence of the present technology, those who achieve a level beyond that human can do, so there have the exigence to create an effective method to support the investor in those system.

Keywords: Artificial Intelligence, Machine Learning, Linear Regression, Sentiment Analysis, Stock Market.

INTRODUÇÃO

A era da tecnologia e dos dados permitiu ao mercado financeiro e aos que dele usufruem uma quantidade nunca obtida antes de informações. Ao pensar em bolsa de valores, nos remete uma imagem de dezenas de indivíduos em seus telefones falando desordenadamente a respeito de compra e venda de ações uns com os outros, porém este estereótipo foi evoluindo e atualmente há uma corrida sem limites de negociações feitas através de enormes centros de processamento de dados que ocorrem 24h por dia.

Os analistas agora dispõem informações em tempo real e fontes diversas para suas análises e buscas por padrões. Tal desenvolvimento tecnológico é implementado na bolsa de valores e então o operador humano é substituído por *software* e há então uma quantidade massiva de dados sendo gerados na unidade de milissegundos, elaborando assim o *High Frequency Trading* (PARANÁ, 2016). Somente no dia 07 de março de 2019, foram realizadas no período diurno 1.419.141 negociações (BOVESPA, 2019) apenas através da BM&FBOVESPA, Bolsa de Valores oficial do Brasil. Diante dessa afirmação continuar fazendo-as manualmente pode ser perda de oportunidades quando é possível automatizar e potencializar as operações nas bolsas de valores.

Com títulos sendo negociados incessantemente (ARAÚJO; MONTINI 2014), há um fator determinístico de grande influência, as notícias. Fontes de notícias surgem em larga escala, e com novas origens de informações sendo geradas com dados de forma exponencial e incontável, sendo humanamente impossível acompanhar o crescimento da demanda de ler e filtrar as informações relevantes das notícias para modelar um plano de investimento que retorna um lucro dentro da margem aceitável, a tecnologia entra como uma grande auxiliadora nas análises de quantidades massivas de dados.

As decisões de investidores e especuladores são constantemente motivadas por diversas notícias que surgem diariamente, e segundo Chowdhury (2014), os sentimentos desses são impactados de uma forma extremamente importante com as principais notícias,

produzindo de maneira significativa uma mudança no plano de investimento. Zhang e Skiena (2010) trazem também este conceito, onde toda informação conhecida e todos os fatos ocorridos refletem diretamente nos preços atuais das ações. Sendo assim, os investidores e empresários fazem com cautela seus investimentos no mercado financeiro, tendo em vista que os valores das ações são coletados e agregadas por várias informações e alteradas em questão de momentos.

METODOLOGIA

O presente estudo utilizará ferramentas voltadas à ciência dos dados para extração de informações dos dados da Bolsa de Valores. Os registros necessitam de tratamentos e análises para o estudo dos mesmos, pois informam somente as negociações e notícias realizadas. As ferramentas se dispõem na seguinte estrutura:

- Linguagem de programação Python;
- API da Bolsa de Valores cedida pela empresa Alpha Vantage, empresa a qual apoia pesquisas voltadas para o Mercado Financeiro;
- Analisador de sentimentos léxico VADER;
- API do Twitter para acesso aos *tweets* e aquisição dos dados*;*
- Algoritmo de Inteligência Artificial de Regressão Linear do Método dos Mínimos Quadrados;

Este trabalho também fará a realização de experimentos para predição de alterações no preço das ações de companhias adquirindo dados de duas áreas, sendo uma delas fontes de notícias de jornais e a outra preço de ações das companhias na bolsa de valores. A primeira será obtida através de publicações de alta influência através da rede social Twitter por uma Interface de Programação de Aplicações no ambiente de programação do Google Colab, o qual incentiva na pesquisa de *machine learning* e inteligência artificial. O algoritmo adaptado que será utilizado na aquisição de dados é dividido nas partes de: mineração do histórico de preço das empresas; download, onde se baixa do Twitter os dados; limpeza dos dados, para melhores resultados e geração dos resultados da análise.

O serviço de análise de sentimento avalia a entrada de texto e usa um algoritmo de classificação léxico de aprendizado de máquina para gerar uma pontuação de sensibilidade entre zero e um. As pontuações mais próximas de um indicam sentimento positivo, enquanto as classificações mais próximas de zero indicam sentimento negativo. A análise de sentimento produz um resultado de qualidade superior quando é concedida partes menores de texto para trabalhar, por isso através da linguagem Python também são feitos ajustes nos *tweets*, uma vez que é necessário eliminar os termos e dados desnecessários que existem nas redes sociais, mesmo que em canais de mídia oficiais podem aparecer como símbolos, links externos e outros componentes que atrapalham identificar os sentimentos.

Os dados dos valores das ações serão adquiridos através da Bolsa de Valores de São Paulo em um intervalo de dois meses da empresa Petrobrás e Vale, também utilizando a linguagem Python para adquirir o *intraday* e a partir das informações levantadas serão feitos os experimentos e estudos para definir algumas características, as quais:

- Analisar o sentimento gerado na notícia publicada;
- Identificar a variação do valor da ação;
- Estudar o padrão de comportamento e a relação entre o sentimento e a ação;

Após a obtenção dos dados necessários, o padrão de comportamento é dado através da aplicação da técnica de *machine learning* onde um algoritmo de Inteligência Artificial aprenderá a relação entre os dados obtidos e então classificar a polaridade e reação que uma informação possui sobre a outra. Na seção seguinte será explicado como foram feitos os estudos para o aprendizado do sistema de predição do presente estudo.

Dados da bolsa de valores

Devido ao fato da importância sobre a aquisição dos dados históricos de variação de preço das ações das empresas em questão para ser validada a relação da variável preço x variável sentimento, foi solicitada uma chave de API da empresa Alpha Vantage para serem minerados os dados históricos. A Alpha Vantage é uma empresa fornecedora de API's voltadas ao mercado financeiro para pesquisadores, engenheiros e profissionais de negócios. Através de sua interface de requisições de *intraday* foi solicitado os dados da Petrobrás e da Vale com um intervalo de variação de 1 minuto entre as negociações.

Com as requisições da aplicação da Alpha Vantage, foi coletado um conjunto de informações do histórico do preço da Petrobrás e da Vale S.A. Dado o fato da requisição ser feita manualmente e somente serem adquiridos os dados de uma semana anterior até o momento da solicitação, as requisições foram feitas de forma manual semanalmente de cada empresa para ser elaborada a concatenação de data e hora entre os *tweets* e o preço da ação naquela hora. Com a empresa Petrobrás foram levantados valores de ações entre os dias 13/08/2019 e 08/10/2019, gerando uma tabela com mais de 14.000 registros.

Levando em consideração que o minério de ferro é discutido em menor intensidade comparado ao *commodity* de barril de óleo, este negociado pela Petrobrás, as buscas por notícias e informações relevantes ao minério de ferro se tornaram mais dificultosas ao longo do período de análise, ao fim concatenando poucos valores entre Preço da Ação x Notícia. Por isso foi estendido o intervalo de tempo da análise entre 13/08/2019 e 29/10/2019 da empresa Vale S.A. Ao final do levantamento a tabela resultou em mais de 15.000 registros, tendo em vista que o *intraday* informa o registro em seis campos seus valores, sendo eles:

1. Data e hora de abertura da ação;
2. Preço de abertura da ação naquele horário;
3. Maior preço da ação naquele intervalo de tempo;
4. Menor preço da ação naquele intervalo de tempo;
5. Valor de fechamento da ação naquele horário;
6. Volume de negociações feitas no intervalo de tempo.

O valor que será usado para a análise de variação das ações será o preço de abertura da ação, para que, ao executar a associação do *tweet* com o preço da ação, seja visto qual era o valor da ação antes da notícia ser gerada. Obtidas as respostas das requisições do *intraday*, foram minerados os dados da bolsa de valores das empresas Petrobrás e Vale S.A. Na Figura 1 é apresentada uma amostra de informações das empresas em questão, onde na Figura 1 (a) vemos informações sobre a Petrobrás, já na Figura 1 (b) exemplifica sobre a Vale S.A.

Figura 1 (a) e (b) - Amostra de informações de preço da Bolsa de Valores

(a)

	date	1. open	2. high	3. low	4. close	5. volume
9382	2019-09-19 10:18:00	14.7300	14.7300	14.7100	14.7100	34795.0
2951	2019-08-23 13:30:00	12.9703	12.9952	12.9702	12.9703	48335.0
11772	2019-10-02 11:19:00	14.0300	14.0350	14.0200	14.0300	50853.0
14646	2019-10-08 14:12:00	13.9000	13.9000	13.9000	13.9000	5407.0
5070	2019-09-04 09:37:00	13.8805	13.8929	13.8579	13.8605	91092.0
3230	2019-08-26 11:47:00	12.8200	12.8300	12.8200	12.8250	11197.0
2101	2019-08-21 12:14:00	13.4600	13.4600	13.4500	13.4600	8833.0
7019	2019-09-11 09:41:00	14.8000	14.8200	14.7900	14.7900	29960.0
11783	2019-10-02 11:30:00	14.0500	14.0700	14.0500	14.0500	12085.0
1036	2019-08-15 13:54:00	13.1600	13.1650	13.1100	13.1100	163026.0

(b)

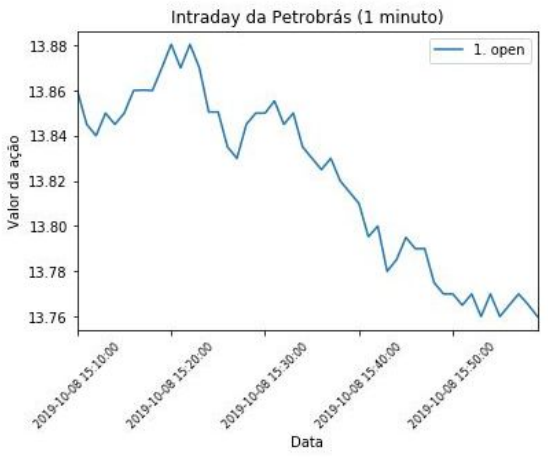
	date	1. open	2. high	3. low	4. close	5. volume
10802	2019-09-30 14:26:00	11.5000	11.505	11.500	11.5050	14161.0
7256	2019-09-12 13:36:00	12.1050	12.115	12.100	12.1100	66188.0
9713	2019-09-20 15:29:00	11.5350	11.535	11.530	11.5300	3610.0
13752	2019-10-28 14:22:00	12.1900	12.195	12.190	12.1900	2978.0
10992	2019-10-01 11:06:00	11.4352	11.449	11.429	11.4302	53053.0
13681	2019-10-28 15:33:00	12.1550	12.155	12.140	12.1450	210017.0
8162	2019-09-16 15:44:00	11.8100	11.810	11.805	11.8100	21045.0
6587	2019-09-09 15:37:00	11.5500	11.565	11.545	11.5650	289480.0
8324	2019-09-17 11:59:00	11.8650	11.870	11.865	11.8650	13721.0
12334	2019-10-04 14:08:00	11.4050	11.410	11.395	11.3950	96382.0

Fonte: Alpha Vantage (2019)

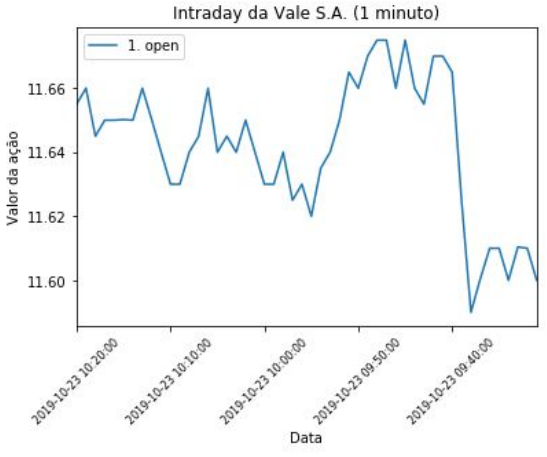
De acordo com a Figura 1 (a) e (b), observa-se que a primeira coluna é o ID do registro, e então os demais valores vem com os detalhamentos do *intraday* já descritos. Podemos também visualizar melhor os dados gerando uma plotagem de uma parcela dos valores para observarmos a variação do preço em um intervalo pequeno de tempo, contemplando que há uma alta taxa de variação. No gráfico da Petrobrás foi gerado no intervalo de 40 minutos do dia 08 de Agosto de 2019, visualizado na Figura 2 (a). Já o gráfico da Vale S.A. observa-se o mesmo tempo porém no dia 23 de Outubro de 2019, o qual encontra-se na Figura 2 (b).

Figura 2 (a) e (b) - Gráfico de *intraday*

(a)



(b)



Fonte: Alpha Vantage (2019)

Ao ser estudada a base de dados de valores de ações, podemos identificar os valores reais das ações das empresas em questão, porém, para se executar o método de Regressão Linear, é necessário que os valores das ações estejam na mesma escala que o valor de sentimento gerado pelos *tweets*, então foi aplicado uma função do TensorFlow que transforma os valores para a escala entre 0 e 1.

Busca de notícias

A API oficial do Twitter tem a limitação incômoda de restrições de tempo, ou seja, não se pode obter *tweets* mais antigos que uma semana. Algumas ferramentas fornecem acesso à *tweets* mais antigos mas, em suma, é necessário gastar algum investimento antes. Ao analisar como o *Twitter Search,* através do navegador, executa normalmente seu fluxo, basicamente funciona da seguinte maneira: ao entrar na página do Twitter, um carregador de rolagem é iniciado; ao rolar para baixo, inicia-se a obtenção de mais e mais *tweets* através de chamadas para um provedor JSON. Aplicado um algoritmo *open source* por Jefferson Henrique[1] que replica esse funcionamento, obtemos a melhor vantagem do *Twitter Search* nos navegadores, podendo-se pesquisar os *tweets* mais antigos e profundos.

O algoritmo retorna alguns campos que podem ser usados para refinar a pesquisa através de linha de comando. Os campos são os seguintes: *id*(str), *permalink*(str), *username*(str), *text*(str), *date*(date), *retweets*(int), *favorites*(int), *mentions*(str), *hashtags*(str) e *geo*(str). Foram utilizados nesse trabalho os campos *text* para fazer a análise de sentimento através de um léxico, os campos *retweets* e *favorites* para definir a importância e relevância de cada *tweet* e o campo *date* para concatenar juntamente com a mesma data do preço já mencionada.

Realizada a requisição dos *tweets* através da API que a própria empresa fornece, retornou uma base de dados para os termos de pesquisa "*iron ore*" e "*crude oil*", tendo em vista serem os respectivos termos para pesquisa dos *commodities* da Vale S.A. e Petrobrás. A Figura 3 nos mostra os dados dos produtos da empresa que comercializa minério de ferro, já a Figura 4 nos exemplifica as informações do minério de ferro, matéria prima comercializada pela Vale S.A.

Figura 3 - *Tweets* antes de serem tratados da Petrobrás

	username	date	retweets	favorites	text
2	speedlight	2019-09-16 14:42	0	1	Starker Anstieg im Moment der Ölpreise. Ohh je...
44	billlionnaire	2019-09-15 21:34	0	0	Crude Oil Brent futures pic.twitter.com/2Wj8qc...
8	ISABELNET_SA	2019-09-16 10:18	13	37	Top Three Crude Oil Producers The U.S. is the ...
35	farnazfassihi	2019-09-15 22:41	12	13	Brent crude oil price jumps more than ever on ...
22	CNBC	2019-09-16 10:14	26	25	Brent crude oil spikes the most in history aft...
16	PlattsOil	2019-09-16 00:52	25	32	All eyes are on #oilprices after the attacks o...
23	WtiOil	2019-09-16 09:31	10	6	Today's Crude Oil Prices: WTI: $60.37 (+5.55) ...
42	SchwartzNow	2019-09-15 21:38	0	0	Inside job? - Brent crude #oil futures are tra...
1	CNBC	2019-09-16 15:26	18	23	Oil prices rose after a drone attack on an oil...
26	mari_saita	2019-09-16 06:48	1	5	サウジアラビア、5.7M bbl/dayすなわち世界供給の5%を喪失、Brent は一時$...

Fonte: Twitter (2019)

[1] Disponível em: <https://github.com/Jefferson-Henrique/GetOldTweets-python> Acesso em: 05 de Novembro de 2019

Figura 4 - *Tweets* antes de serem tratados da Vale S.A.

	username	date	retweets	favorites	text
138	GwedeMantashe1	03/09/2019 08:54	31	54	Mining is the strongest performer in the 2nd q...
62	minetransphobic	05/10/2019 20:47	0	0	minecraft iron ore transphobic
175	dpradhanbjp	12/08/2019 13:29	92	569	After a hectic day touring various iron ore mi...
206	Jill_hubley	03/08/2019 21:49	21	83	movement of grain, flour, iron ore , corn, and...
110	bhuvikal	16/09/2019 15:06	23	35	Senior advocate Harish Salve has blamed the Su...
118	MichaelPascoe01	13/09/2019 19:38	58	102	Iron ore gets the headlines but tourism makes ...
52	Mardek15	01/10/2019 10:27	0	0	Steel Outlook Dims as Iron Ore Powerhouse Flag...
65	MagnetiteMines	03/10/2019 23:20	1	2	The drumbeat of voices forecasting India could...
71	OstoulSB	30/09/2019 01:10	0	0	Rio Tinto scraps plans for Canadian iron ore u...
17	baijendra	01/10/2019 13:30	24	89	NMDC to regain iron ore mine in Donimalai as G...

Fonte: Twitter (2019)

É válido ressaltar que, tendo como exemplo a Figura 4, observa-se que o registro com o ID 175 teve um alto índice de *retweets* e *favorites*, indicando a influência da informação ao decorrer do tempo que foi gerada. Já o registro com o ID 71 nota-se o grau de influência da publicação, com nenhuma propagação da mesma informação, tornando-se inútil no impacto das ações na Bolsa de Valores. Essa métrica foi definida para ignorar tais informações irrelevantes que não gerariam nenhum tipo de efeito para as empresas.

Após a aquisição das publicações e dados do Twitter, dentro do próprio ambiente de programação, foram tratados os dados e preparados para o ambiente real, onde alguns ajustes foram necessários, como a alteração de horários e datas dos *tweets* para concatenar com a ação, pois os pregões de negociação da Bolsa de Valores abrem de Segunda à Sexta, a partir das 09:31 e duram até 16:30. Com isso, os *tweets* publicados fora deste horário não influenciaram na aplicação da concatenação, então foram tratados da seguinte maneira:

- Para os *tweets* que foram publicados após 16:30 de Segunda até Quinta, adicionou-se 1 dia à data original e foi modificado também o horário para 09:31;
- Para os *tweets* que foram publicados após 16:30 de Sexta até Segunda antes de 09:31, modificou-se para a Segunda subsequente caso o dia não fosse Segunda e o horário para 09:31.

Essa alteração foi necessária para que a informação publicada fora do horário de pregão fosse associada ao primeiro momento de negociação da ação da empresa, onde o impacto seria oficialmente imposto.

Analisador de sentimentos

Após os *tweets* serem tratados corretamente, é gerado a partir dessa base de dados de *tweets* o analisador de sentimento também na linguagem Python. O VADER (Valence Aware Dictionary e sEntiment Reasoner) é uma ferramenta de análise de sentimentos léxico baseada em regras e sintonizada especificamente com os sentimentos expressos nas mídias sociais. O

VADER usa uma combinação de um léxico de sentimento e uma lista de recursos lexicais (por exemplo, palavras) que geralmente são rotulados de acordo com sua orientação semântica como positivos ou negativos.

Essa ferramenta foi bem sucedida na análise de textos de mídia social, editoriais, resenhas de filmes e de produtos. O Analisador léxico VADER não fala apenas sobre a pontuação de Positividade e Negatividade, mas também indica o quão positivo ou negativo é um sentimento. Seu uso e instalação é bem simples e é uma ferramenta *Open Source* como indica o autor Hutto (2014).

Machine Learning baseado nas informações obtidas

Para que os dados coletados se tornem um *insight*, ou seja, uma ideia ser aplicada na tomada de decisão, esses precisam passar por um algoritmo que construirá a Regressão Linear que vai ser utilizada posteriormente para fazer a predição. Nesse ponto, o algoritmo a ser utilizado é Regressão Linear no Método dos Mínimos Quadrados (MMQ) que utiliza o conceito que para cada um dos pontos, deve-se medir a distância vertical entre o ponto e a linha e somá-los, a linha ajustada é aquela que a soma de distâncias é a menor possível.

O algoritmo para o aprendizado será escrito na linguagem de programação Python na ferramenta *cloud* do Google chamada Colaboratory[2] que gera uma máquina virtual com Python instalado e utilizando as principais biblioteca de *machine learning* e *deep learning* chamadas respectivamente de *scikit-learn* e TensorFlow. Com as bibliotecas, é possível passar os dados obtidos no passo anterior e seu processamento, transformando para uma escala única as duas variáveis com o TensorFlow e irá configurar a linha de regressão com o *scikit-learn*. A escolha desta linguagem se devem aos fatos das bibliotecas serem prontamente preparadas para algoritmos de *machine learning* e possuir diversas funções pré definidas que auxiliam no desenvolvimento da técnica. Por fim, os equipamentos físicos necessários foram computadores com conexão com a internet.

Experimento de Predição e exibição

Stevenson (1986) define, "A correlação mede a força, ou grau, de relacionamento entre duas variáveis; a regressão dá uma equação que descreve o relacionamento em termos matemáticos." Também explica que a regressão compreende a análise de dados amostrais para saber se duas ou mais variáveis estão relacionadas numa população e, tem como resultado uma equação matemática que descreve o relacionamento. Portanto a equação pode ser usada para predizer valores futuros de uma variável correlacionada quando se possui os valores da outra variável. Para a aplicação da regressão linear é necessário a descoberta do coeficiente de correlação e é utilizada a fórmula de Pearson (1892) representada na Equação (1) descrita a seguir:

[2] Disponível em <http://colab.research.google.com> Acesso em 12 de Outubro de 2019

$$r = \frac{\Sigma_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\Sigma_i (x_i - \overline{x})^2 \Sigma_i (y_i - \overline{y})^2}}$$

Equação (1)

onde Xi = X1, X2... e Yi = Y1, Y2... são os valores medidos de ambas as variáveis. Além disso, encontramos na Equação 2 as médias aritméticas de ambas as variáveis.

$$\overline{x} = \frac{1}{n} . \sum_{i=1}^{n} x_i \ , \ \overline{y} = \frac{1}{n} . \sum_{i=1}^{n} y_i$$

Equação (2)

A autora ainda explica que a análise correlacional indica a relação entre 2 variáveis lineares e os valores sempre serão entre +1 e -1. O sinal indica a direção, se a correlação é positiva ou negativa, e o tamanho da variável indica a força da correlação. Dancey e Reidy (2006) apontam para uma classificação ligeiramente diferente: r = 0,10 até 0,30 (fraco); r = 0,40 até 0,6 (moderado); r = 0,70 até 1 (forte). Obtendo o valor de R podemos concluir qual a covariância entre as duas variáveis, a direção o quão forte é a correlação entre X e Y.

A análise de regressão após encontrar um coeficiente satisfatório dentre de cada necessidade, a aplicação das Relações funcionais da regressão linear podem ser representadas por $Y = f(X_1, X_2, ..., X_k)$, onde Y representa a variável dependente e os $X_n(X_1, X_2...)$ representam as variáveis independentes como explica Rodrigues (2012).

O modelo estatístico de regressão linear aplicado é mostrado através da Equação 3:

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

Equação (3)

Na Equação 3, y_i é o valor observado para a variável dependente no nível da variável dependente X_i e β_0 é a constante de regressão. β_1 equivale ao coeficiente de regressão e e_i é o erro associado proposto pelo modelo.

Para obter a equação estimada utilizando o método MMQ para minimizar os erros, onde o resultado é a distância entre o ponto e a reta traçada, é elaborada a Equação 4:

$$e_i = Y_i - \beta_0 - \beta_1 X_i$$

Equação (4)

[2] Disponível em <http://colab.research.google.com> Acesso em 12 de Outubro de 2019

Assim elevando a fórmula toda ao quadrado e fazendo a somatória de ambos os lados, é obtida a função na Equação 5:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [Y_i - \beta_0 - \beta_1 X_i]^2$$

Equação (5)

Para se obter o mínimo para a equação deriva-se a mesma em relação a variável de interesse e iguala-se a zero. Derivando a Equação 3 em relação a e , e igualando ambas a zero são obtidas duas equações, β_0 e β_1 , que formam o sistema de equações normais fornecidos na Equação 6:

$$\hat{\beta}_1 = \frac{\Sigma x_i y_i - \frac{\Sigma xi - \Sigma yi}{n}}{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}} = \frac{SPD_{xy}}{SQD_x}, \quad \hat{\beta}_0 = \hat{Y} - B_1 \hat{X}$$

Equação (6)

Uma vez obtida as estimativas, a equação estimada pode ser escrita na forma encontrada na Equação 7:

$$\hat{Y}_i = \beta_0 + \beta_1 X_i$$

Equação (7)

A visualização dos dados obtidos será em um gráfico de dispersão com o traçado da linha da análise de regressão utilizando o método explicado nas Equações 3 e 4. O gráfico é exibido pela ferramenta Google Colab e pode ser exibido com o comando *plot*. Esse método é padrão nessa ferramenta e provém das bibliotecas da linguagem Python traduzida pela sua IDE. A exibição do gráfico nesse ponto é feita para uma avaliação visual de correlação, onde é analisado se a reta corresponde ao método proposto.

RESULTADOS E DISCUSSÃO

Para a análise de sentimento, foi gerado o código do analisador léxico onde analisou todos os registros de *tweets* dentro de um laço até finalizar o *dataset* de informações. A função responsável pelo retorno do valor de polaridade do sentimento se encontra na primeira linha do laço, identificada por *analyser.polarity_scores*, onde podemos visualizar através da Figura 5.

Figura 5 - Código do analisador léxico VADER

```
i=0 #counter

compval1 = [ ]  #empty list to hold our computed 'compound' VADER scores

while (i<len(crude_oil)):

    k = analyser.polarity_scores(crude_oil.iloc[i]['text'])
    compval1.append(k['compound'])
    i = i+1
#converting sentiment values to numpy for easier usage

compval1 = np.array(compval1)

len(compval1)
```

```
crude_oil['VADER score'] = compval1
```

Fonte: Autor Próprio.

Na linha 1 do código é criada a variável para ser o contador do laço, e em seguida na linha 3 é criada uma lista denominada *compval1* para armazenar os valores de polaridade de sentimento gerados pelo texto do *tweet*. Na linha 5 temos o laço para ler todas as linhas do *dataframe* sobre o óleo. Iniciando o comando do laço, visualizamos na linha 7 uma variável temporária *k* para armazenar a polaridade gerada pelo texto do *tweet* do registro em questão. Seguindo para a linha 8 é usada a função para adicionar o elemento da variável *k* à lista *compval1*, e então logo após na linha 9 é somado mais um ao contador para seguir para o próximo registro, onde fará o procedimento da linha 7 à 8 até finalizar a análise completa no *dataframe*. Completo o laço, na linha 12 é aplicada uma função para transformar os valores da lista *compval1* para o formato *array* do *numpy*, biblioteca do Python. Logo após observamos na linha 14 a função para contar o tamanho da lista *compval1*, e, finalizando os valores dessa lista, na linha 15 são alimentados para o *dataframe* original *crude_oil* em uma nova coluna *VADER score* respectivamente aos registros.

Após analisados e processados os *tweets* para que fossem eliminados caracteres especiais, *hashtags* e outras formas de dados que não eram texto, e executada a análise de sentimento, a qual retornou uma nova coluna na base dados com os valores de sentimentos de -1.0 a 1.0, visualizamos a adição da coluna em ambos os *dataframes*, na Figura 6

relacionados a pesquisa "*crude oil*" e na Figura 7 para os *tweets* relacionados a pesquisa "*iron ore*".

Figura 6 - *Tweets* após serem tratados da Petrobrás

	date	retweets	favorites	text	VADER score
14	2019-09-11 18:20	15	28	Crude oil prices hover around $60 per barrel, ...	0.0387
17	2019-09-05 13:42	4	46	By tying AB economy to crude oil prices , and ...	-0.9118
8	2019-09-16 09:04	15	73	Drone attack in Saudi has highly impacted crud...	-0.7951
18	2019-09-03 08:15	3	39	We were number in Africa from 2012-2013 becaus...	0.1911
4	2019-09-19 21:11	10	27	Only a limited impact on inflation is expected...	-0.8658
16	2019-09-06 20:16	10	20	Crude Oil prices will jump to $120 - Make Saud...	0.1027
9	2019-09-16 02:57	4	20	Report: The rupee tumbled by 68 paise to Rs 71...	-0.5719
23	2019-08-31 10:09	44	143	I made Indian Economy 3rd largest in the world...	-0.8934
1	2019-09-16 08:37	3	13	Crude oil prices have gone up, Nigeria will ma...	-0.3291
22	2019-09-01 01:30	72	406	Manmohan Singh, 10.08% GDP when, - Coalition G...	-0.9260

Fonte: Autor Próprio.

Figura 7 - *Tweets* após serem tratados da Vale S.A.

	date	retweets	favorites	text	VADER score
1	05/10/2019 15:04	18	110	Just one more. Northants churches are special ...	0.4019
79	14/08/2019 13:38	7	71	Look at Europe, where there's a much more robu...	0.8338
22	24/09/2019 01:01	17	132	I forgot this wild concept art was at the #Dis...	0.4374
80	12/08/2019 15:56	1524	1802	A 2-MW windmill is made of 260tons of steel th...	0.2732
102	05/08/2019 00:18	124	200	Shock of the day: Rand at 14.86 to dollar. Chi...	-0.7351
11	01/10/2019 00:25	4	32	As part of annual Reward & Recognition Program...	0.9081
26	21/09/2019 06:49	38	62	. @JSPLCorporate 's TRB Iron Ore Mines in Tens...	0.9168
18	27/09/2019 04:30	10	21	Great call from @Barchart .!! Zanaga Iron Ore ...	0.6892
36	14/09/2019 18:33	53	105	Enugu- Coal, Iron ore , beautiful topography(t...	0.8481
108	31/07/2019 23:18	22	124	Our 3 largest exports are iron ore , coal & ed...	-0.7351

Fonte: Autor Próprio.

A base de dados recebeu a coluna *VADER score* da figura 6 e 7 usando como indexador a coluna *date*. Com essa operação foram concatenados os dados relevantes das bases de dados de preço da ação e da base de dados onde continham os *tweets* com as análises de sentimento, ambas as colunas extraídas do mesmo período.

Em seguida os dados passaram por outro pré-processamento para a transformação de uma mesma escala dos valores. Então após todos os tratamentos com os dados, os mesmos foram alimentados na função da biblioteca do *scikit-learn* de Regressão Linear, encontrado na Figura 8.

Figura 8 - Código da escala e Regressão Linear

```
from sklearn.preprocessing import StandardScaler
scaler_x = StandardScaler()
variation = scaler_x.fit_transform(variation)

from sklearn.linear_model import LinearRegression
model = LinearRegression()
model.fit(sentiment, variation)
```

Fonte: Autor Próprio.

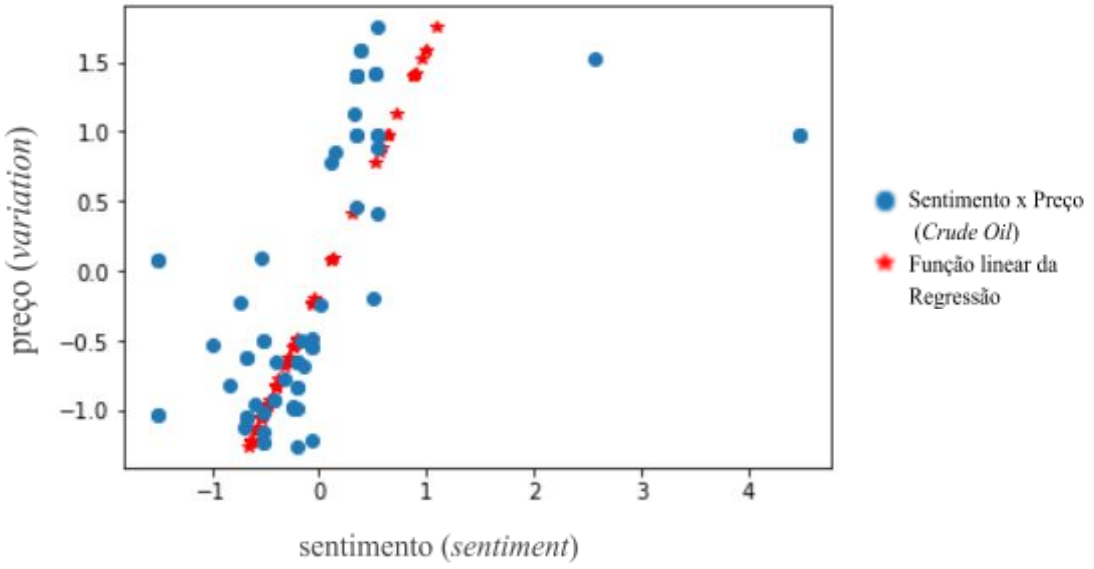
Observamos na linha 16 a importação da função de Escala Padrão da biblioteca de pré-processamento do *scikit-learn*, e na linha 17 a alocação da função à uma variável temporária *scaler_x*. É criada através da linha 18 a variável *variation* para transformar os valores contidos nesse *dataframe* onde estão localizados os preços das ações para alterar para a mesma escala do sentimento. Identificamos na linha 19 a importação da função de Regressão Linear também da biblioteca *scikit-learn*, e através da linha 20 é alocado para a variável *model* a função da regressão linear, e logo em seguida é aplicado com a função *model.fit* a regressão linear dos valores, recebendo como parâmetro nossos dados de treinamento *sentiment* e nossos valores alvo da variável *variation*.

O algoritmo utilizou os métodos citados anteriormente de escalonamento e tratamento da relevância dos dados e seu impacto considerando o número de *favorites* e *retweets.* Após a normalização, foi treinado e alimentado com os *datasets* obtidos através da utilização de funções de *machine learning* da biblioteca *scikit-learn*. Foram descritos no parágrafo seguinte, os valores dos resultados de r encontrados na Equação 1, ou seja, a nota que a biblioteca retorna de acordo com a quantidade de acertos e a proximidade com os resultados reais.

No resultado do teste feito com o modelo treinado a predição da regressão linear obteve uma assertividade de 55,02% nos testes realizados com ações da Petrobrás e 0,18% em relação a Vale SA. Essa diferença na assertividade se dá devido à alguns fatores, como por exemplo o número de *tweets* relacionados a *Iron ore* se comparado com os *tweets* sobre *Crude* oil. Outro ponto relevante é que durante o período avaliado, o *commoditie* do petróleo sofreu grande volatilidade de preço e de notícias, o que facilita a identificação de sentimento através do Algoritmo.

A medição foi realizada com parte extraída da própria base de dados, com 20% dos dados, foram colocados na variável dependente X e observados os resultados gerados em Y conforme a Figura 9. Pode-se observar os valores correspondente a relação sentimento (Y) e variação de preço (X) em azul, e o valor de predição para cada ponto de sentimento ele retorna um ponto de variação em vermelho.

Figura 9 - Relação entre treino e predição de Sentimento (*Crude Oil*) X Preço (Petrobrás).

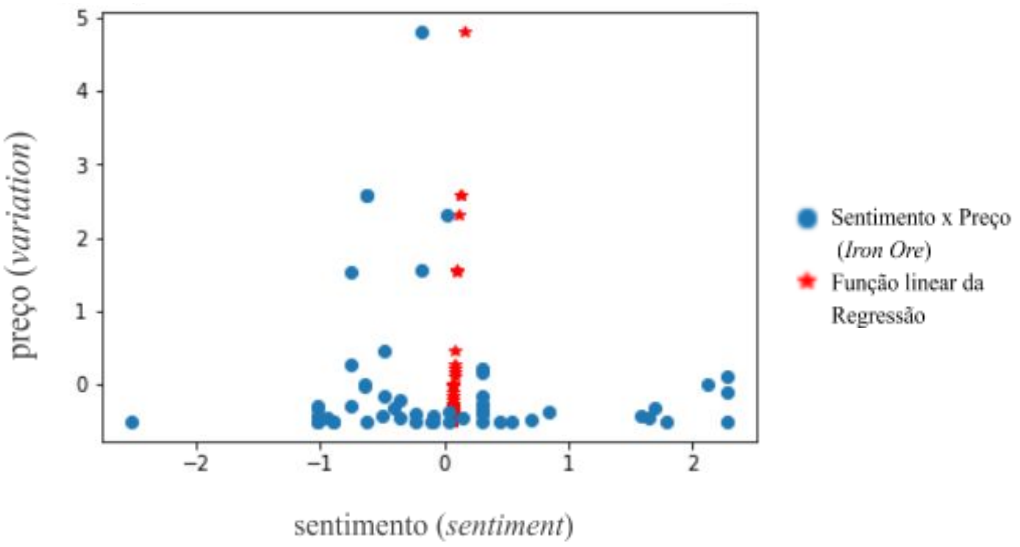


Fonte: Autor Próprio.

No eixo vertical observamos a variação de preço em valores transformados para uma escala a qual o algoritmo de aprendizado pudesse aplicar o Método dos Mínimos Quadrados, pois realizada a aplicação do processo de identificação das variáveis base e preditora, visualizada no eixo horizontal de sentimento de *tweet*, os valores não se tornaram condizentes utilizando o preço da ação original. Devido a isso, os valores foram convertidos para a escala adequada, tendo em vista que, após a transformação, os valores dos dois eixos ficaram na mesma proporção.

Na aplicação da Regressão Linear para o *dataset* da Petrobrás podemos visualizar a reta do modelo de forma mais clara, sendo o resultado esperado através do processo de regressão, encontrado através dos pontilhados em vermelho. Apesar da pesquisa se basear em *commodities* da empresa, os dados quantitativos da requisição foram suficientes para a aplicação do modelo, retornando, assim, a reta de predição e a correlação entre as variáveis. Contemplamos também que existem dois pontos discrepantes de Sentimento x Preço na Figura 9 encontrados pelos pontos em azul mais à direita do gráfico, e esses são pertencentes ao processo de tratamento da base de dados e que, ao fim da regressão, ainda restaram poucos resquícios, mas, apesar disso, o resultado da reta foi satisfatório nos entregando uma correlação de 55,02%.

Figura 10 - Relação entre treino e predição de Sentimento (*Iron Ore*) X Preço (Vale S.A).



Fonte: Autor Próprio.

Como já mencionado, foi aplicado também a transformação do valor da ação original da empresa Vale S.A, encontrado no eixo vertical, para uma escala dentro da proporção do valor de sentimento, encontrado no eixo horizontal. Como explanado na Figura 9, temos também poucos valores que se tornaram discrepantes após a aplicação da Regressão, estes encontrados com os pontos em azul com valores maiores em relação ao eixo vertical preço. É considerável também ressaltar que, visualizando de uma maneira mais compreensível através da Figura 10, a reta que se esperava alcançar não foi possível devido aos valores obtidos das variáveis que foram divergentes e pouco satisfatórios após a aplicação da regressão. Também teve um grande impacto a requisição dos *tweets, que* não retornaram eficazmente o quantitativo para o modelo, pois o *commoditie* da empresa em questão não é discutido com frequência na rede social.

Os resultados obtidos no presente estudo sugerem que a correlação entre as variáveis de valor de ação de uma empresa na bolsa de valores e o sentimento gerado por uma notícia com grande impacto obtém uma devida coerência entre o grau de aceitação da informação com a taxa de variação do valor da ação, exemplificado de maneira mais eficaz no tratamento dos valores da empresa Petrobrás e do *commoditie* de óleo bruto, pois, para cada variação que seja superior, houve um sentimento positivo gerado por uma notícia publicada no âmbito jornalístico, sendo de uma considerada proporção entre o sentimento e o preço, como também a variação inferior do preço teve sua relação para com o sentimento negativo provocado pela notícia publicada.

CONCLUSÃO

O uso de tecnologia em outros setores é o futuro, inclusive um futuro que já está presente. Tanto na medicina quanto na agronomia já se utiliza tecnologia de ponta diariamente e no setor da economia não é diferente, porém ainda existem mais possibilidades de aplicação e vertentes ainda pouco exploradas como o caso desse trabalho em que se aplica Machine Learning como tentativa de ser mais rápido e preciso na análise das notícias que influenciam o mercado de ações do que a análise de especialistas e economistas que precisam estudar de fato cada notícia e seu impacto, desperdiçando tempo em micro análises enquanto poderiam focar em suas macro análises.

Os investidores podem utilizar para tomar decisões antes que a euforia no preço se forme, tendo assim uma segurança maior e um risco menor, ao observar as notícias quais podem afetar seus investimentos e tendo a noção de valorização ou desvalorização dos mesmos. Como a maioria da tecnologia no mercado financeiro de alta frequência observa os gráficos e não preveem o impacto de notícias e eventos no ativo em que atuam perdem essa capacidade de diminuir seus riscos e aumentar seus ganhos. Por consequência, quem utiliza esse tipo tecnologia estará um passo à frente de seus concorrentes ou colegas de profissão.

Após apresentados os fatos e dentro das limitações referentes aos dados disponíveis, conclui-se que a previsão do preço de ações através de algoritmos de *machine learning* não é indicado em todos os commodities. Apesar de satisfazer os objetivos do trabalho, o resultado não garante ao investidor confiança necessária para o investimento apenas baseado nesses resultados, pelo motivo de não serem somentes especialistas comentando sobre o assunto. Para trabalhos futuros, podem ser utilizados fontes de dados mais confiáveis, com um período de tempo analisado maior e ferramentas exclusivas para predição de ações.

REFERÊNCIAS

ARAÚJO, A. C. D.; MONTINI, A. D. A. **High Frequency Trading: Preço, Volume e Volatilidade em uma Nova Microestrutura**, São Paulo, Outubro 2014. ISSN 2177-3866.

B3. BM&FBOVESPA, 07 Março 2019. Disponivel em: <http://www.bmfbovespa.com.br/pt_br/servicos/market-data/consultas/boletim-diario/boletim-diario-do-mercado/>. Acesso em: 21 Março 2019.

CHOWDHURY, S. R. . S. C. News Analytics and Sentiment Analysis to Predict. **International Journal of Computer Science and Information Technologies**, Kolkata, 2014. 10. Disponivel em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.433.1426&rep=rep1&type=pdf>. Acesso em: 15 Março 2019.

DANCEY, Christine & REIDY, John. (2006), Estatística Sem Matemática para Psicologia: Usando SPSS para Windows. Porto Alegre, Artmed.

Hutto, C.J. & Gilbert, E.E. (2014). **VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text.** Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

PARANÁ, E. **A finança digitalizada: informatização a serviço da mundialização financeira**, Brasília, 12 Setembro 2016. Disponivel em: <https://revistas.face.ufmg.br/index.php/novaeconomia/article/view/3362/2421>. Acesso em: 14 Março 2019.

PEARSON, Karl.; **The grammar of science**. London, J. M. Dent and Company.(1892)

RODRIGUES, S. C. A. **Modelo de Regressão Linear e suas Aplicações**. Covilhã: [s.n.], 2012. Disponivel em: <https://ubibliorum.ubi.pt/bitstream/10400.6/1869/1/Tese%20Sandra%20Rodrigues.pdf>.

STEVENSON, W. J. (1986). **Estatística aplicada à administração**. São Paulo: Harbra

ZHANG, W.; SKIENA S. Trading Strategies to Exploit Blog and News Sentiment. **Association For The Advancement Of Artificial Inteligence**, Nova York, 2010. Disponivel em: <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1529/1904>. Acesso em: Março 15 2019.