CORRELAÇÃO DA VARIAÇÃO DE PREÇO DE AÇÕES A PARTIR DA ANÁLISE DE SENTIMENTO DE NOTÍCIAS FINANCEIRAS

CORRELATION OF SHARE PRICE VARIATION FROM FINANCIAL NEWS SENTIMENT ANALYSIS

ADMINISTRAÇÃO DA INFORMAÇÃO

William Douglas Costa Silva, Universidade Estadual de Maringá, Brasil, Email

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

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

Resumo

A movimentação do mercado de ações possui diversos atuadores, dentre eles as notícias veiculadas pela mídia, considerando que as altas e baixas dos preços de ativos são responsáveis por substanciais lucros ou prejuízos financeiros, a busca por um meio de encontrar uma relação com estes atuadores é um objeto de constante estudo. Este estudo analisa a situação da possibilidade de encontrar uma relação entre o sentimento contido em notícias publicadas em portais especializados em jornalismo financeiro e o preço de ações de uma determinada empresa. Foram coletadas séries de notícias agregadas a um portal especializado, a partir das quais foi possível explorar uma técnica de análise de sentimento baseado em processamento de linguagem natural, fazendo uso de uma base de termos próprios do mercado financeiro. Foram obtidos valores de avaliação de sentimento para cada notícia, bem como a cotação das ações da Google na bolsa de valores, e deste modo, calculada a correlação por intervalos de tempo considerando o valor da ação no instante da publicação da notícia, 24, 48 e 72 horas depois. Como resultado identificou-se um valor calculado de correlação (próximo a 0,28) porém uma baixa consistência, principalmente na análise gráfica, com poucos pontos visíveis de relação entre as variáveis. Embora os valores de correlação sejam baixos, os resultados demonstram um caminho assertivo dados os resultados de trabalhos anteriores. Considerando que as notícias publicadas possuem um alto teor de abstração linguística e subjetividade textual.

Palavras chave: Preço de ações; Análise de sentimento; Notícias financeiras.

Abstract

The movement of the stock market has several actuators, among them there are the news carried by the media, considering that the highs and lows of asset prices are responsible for substantial financial profits or losses, the search for a way to find a relationship with these actuators is an object of constant study. This study analyzes the situation of the possibility of finding a relationship between the sentiment that the news published in portals specialized in business journalism and the stock price of a certain company. Series of news aggregated from specialized portal were gathered, from which it was possible to explore a sentiment analysis technique based on natural language processing, using a library of terms specific to the financial market. Sentiment valuation values were obtained for each news item, as well as the price of Google shares on the stock exchange, and thus, the correlation by time intervals was calculated considering the share price at the time of publication of the news, 24, 48 and 72 hours later. As a result, we found a calculated correlation value (close to 0.28) but a low consistency, mainly in the graphical analysis, with few visible points of relationship between the variables. Although the correlation values are low, the results demonstrate an assertive path given the results of previous work Considering that the published news has a high tenor of linguistic abstraction and textual subjectivity.

Keywords: *Stock price; Sentiment analysis; Financial news.*

1. INTRODUÇÃO

O mercado de ações flutua constantemente de acordo com um grande conjunto de variáveis, e investidores entusiastas buscam tirar proveito destas oscilações na expectativa de aumentar seu ganho, comprando ativos a um valor mais baixo que serão comercializados em curto ou médio prazo. Para estes atores, a análise do mercado financeiro a partir de notícias de agências especializadas é uma tarefa imprescindível.

Análises fundamentalistas são traçadas a partir de informações publicadas nesses portais, e mesmo que os preços não reajam adequadamente a uma série de dados publicados, um investidor experiente pode tirar proveito disso a partir de uma valorização errônea (Kent, Hirshleifer & Teoh, 2001). Ainda, de acordo com a Hipótese dos Mercados Eficientes (Fama, 1970), os valores de mercado refletem as informações publicadas. Pode-se inferir então que é possível correlacionar a polaridade de notícias publicadas em portais com o movimento (alta ou baixa) dos preços de ações. Com o volume de informações que são publicadas a todo momento na internet, torna-se imprescindível o emprego de técnicas que possam auxiliar nesta análise.

Neste contexto, formula-se então a pergunta de pesquisa “De que modo técnicas computacionais podem ser utilizadas para analisar de modo útil a influência de notícias no mercado acionário? Para responder tal questionamento, este estudo visa explorar a técnica de análise de sentimento baseada em contagem de termos relacionados a um dicionário previamente definido, e traçar sua correlação com a variação de preço de ações.

Considerando trabalhos como o de Valdevieso (2019), este trabalho traz uma contribuição relevante, delineando atividades necessárias a realização da análise de sentimentos de notícias, ao evidenciar que este tipo de análise depende de um vocabulário especializado quando se trata de contextos específicos como o mercado acionário.

2. FUNDAMENTAÇÃO TEÓRICA

Empresas juridicamente constituídas como Sociedades Anônimas têm como principal característica o fato de seu capital social poder ser dividido em ações, que são títulos que representam uma fração da empresa. Estas frações podem ser comercializadas livremente no mercado sem a necessidade de registro público de propriedade.

Estas negociações são intermediadas pelas bolsas de valores que funcionam como um mercado organizado, no qual investidores compram e vendem concorrentemente com as organizações financeiras. (Giampietro, Sarraceni, Montanha, Horita & Indalêncio, 2007)

Uma vez que ações sejam adquiridas por um indivíduo, esta passa a fazer parte do quadro de acionistas desta empresa na condição de investidor em uma pequena parte da mesma. O valor da empresa irá refletir no valor de cada ação, ou seja, caso o valor da empresa suba, os títulos emitidos terão seu valor acrescido na mesma proporção independente de quanto o investidor tenha pago por eles, e o mesmo acontece caso o valor decresça. Assim se configura um ecossistema complexo de negociações, em que o objetivo é comprar ações que possuem potencial de valorização para, futuramente, vender a um valor maior.

Por isso, juntamente com o desenvolvimento do próprio mercado financeiro, surge a necessidade de se fazer previsões de preço e movimentações. Embora a Hipótese dos Mercados

Eficientes sugere que os preços das ações sempre refletem as informações disponíveis, de modo que não é possível lucrar de forma sustentável prevendo preços das ações (Fama, 1970), outras hipóteses geram margem ao desenvolvimento de ferramentas de previsão, como por exemplo a hipótese de mercado adaptativo (Lo, 2004). Considerando a hipótese de mercado adaptativo, foram elaboradas diversas técnicas de análise com a finalidade de identificar tendências, estas técnicas estão representadas por dois vieses de pensamento financeiro: análise técnica e análise fundamentalista, (Beyaz, Tekiner, Zeng & Keane, 2018).

Segundo Murphy (1999) a **análise técnica** é o estudo da ação de mercado, faz uso de gráficos de tendência, na intenção de prever o valor futuro deste mercado. Murphy (1999) descreve que é mais provável que uma tendência em movimento se mantenha do que ela tome um sentido inverso. Murphy (1999) também considera que a história se repete. Por isso no estudo da análise técnica são considerados gráficos desenhados por dezenas de anos, dados históricos, buscando padrões que representem o aspecto comportamental do mercado. Isto porque se certos padrões funcionaram no passado, pode-se assumir que continuarão funcionando no futuro, pois representam a psicologia humana que tende a não mudar de modo radical.

Na prática, a análise técnica usa os históricos de preços das ações de uma empresa e as informações do volume de negociação para calcular índices ou indicadores que apoiam a decisão. Essas técnicas conforme Blakey (2001) podem ser divididas em duas categorias: (a) reconhecimento de padrão, busca identificar e extrapolar padrões de comportamento recorrentes, e (b) filtragem, tenta normalizar taxas de ruídos do sinal lido, a fim de facilitar a identificação de padrões e comportamentos subjacentes.

A **análise fundamentalista** busca a relação do preço das ações com base em um estudo detalhado das subjetividades relacionadas ao negócio, por exemplo: lucratividade, eficiência operacional, expertise administrativa relacionada a empresa, além de seus produtos, posicionamento no mercado e a economia como um todo. (Beyaz et al., 2018). Investidores fundamentalistas encontram o valor intrínseco da ação, observando as forças indiretas que podem surtir efeito na variação de preço. A análise fundamentalista como um meio de previsão de mercado de curto prazo, é menos popular que a análise técnica, devido à escassez de ferramentas capazes de gerar com precisão uma previsão totalmente fundamentalista que capture muitas nuances envolvidas, porém a perspectiva futura de uma ação pode ser avaliada por uma análise fundamentalista, e o sucesso de um investimento depende diretamente disso (Islam, Zaman & Ahmed, 2009). Entretanto, este cenário tem mudado e este tipo de análise tem se tornado alvo de diversas pesquisas, principalmente com o aumento do poder de processamento computacional, que permitiu a exploração em larga escala de algoritmos de inteligência artificial e processamento de linguagem natural, capazes de processar grandes quantidades de dados e encontrar relações que humanamente seriam impossíveis de serem detectadas em curto período de tempo.

2.1. Análise de sentimentos

Informações de documentos textuais (relatórios financeiros, notícias, postagens em mídias sociais) são complementares, e, cada vez mais são utilizados como fonte de informação sobre o desenvolvimento do mercado financeiro. Por isso, sistemas automatizados capazes de analisar

textos relacionados as informações de uma empresa ou ramo econômico são importantes. Em geral esses sistemas têm por base análise de sentimentos (Hajek & Barushka, 2018).

A análise de sentimentos pode ser caracterizada computacionalmente como uma atribuição de valor sentimental para uma expressão, oriunda de um dicionário de sentimentos. A construção deste dicionário pode ser: (a) manual - o dicionário é coletado e analisado por especialistas linguísticos, sendo um processo moroso e oneroso porém com um alto nível de acurácia; (b) semiautomática - o dicionário é construído manualmente com algumas palavras chave, e depois é expandido por um conjunto de regras constituindo um novo conjunto de dados ou automática; automático - inicialmente um grande conjunto de dados são coletados da internet, tais documentos são categorizados em positivos ou negativos de acordo com o efeito no mercado. E o dicionário é construído a partir do resultado da categorização (Minh, Sadeghi-Niaraki, Huy, Min & Moon, 2018).

Abstraindo, por hora, a necessidade de pré-processamento, o qual será abordado futuramente, pode-se considerar que a partir deste conjunto léxico pode-se submeter artigos textuais para obter a classificação em termos sentimentais, e essa fase pode ser aplicada em três níveis distintos, a depender da granularidade desejada: (a) nível de documento – todo o texto é considerado uma unidade única de informação. Esta abordagem deve ser considerada se todo o conteúdo é especificamente sobre um único assunto, o objetivo principal é classificar um documento como positivo ou negativo, um exemplo seria a análise da opinião que as pessoas expressam em microblogs; (b) nível de sentença - a polaridade é calculada para cada sentença do texto, e cada uma é considerada uma unidade independente, gerando uma granularidade muito maior de opiniões, pois cada sentença pode ter uma polaridade distinta. Uma das características desta abordagem é que uma sentença pode que ser objetiva (expressa fatos e não possui opinião sobre o objeto), ou subjetiva (apresenta opiniões); (c) nível de aspecto (*feature*) - tem base em uma etapa básica, identificando uma parte do texto como uma característica de algum objeto. Por exemplo: “O preço dos novos processadores está bem acima da média dos concorrentes”, o objeto “processadores”, e o substantivo o qual contém o adjetivo, ou seja, o ponto de opinião "está bem acima da média dos concorrentes” (Kolkur, Dantal & Mahe, 2015). O nível de aspecto exige a preparação dos dados, com rotulamento, extração das características, extração de termos opinativos, etc; para então agrupar características similares, gerando um índice com as análises produzidas. (Kolkur et al., 2015).

A linguagem de sentimentos se configura como um meio sistemático para expressão de ideias e sentimentos, e se configura como um grande desafio no campo da análise computacional. A ambiguidade no sentido das palavras, comum na linguagem humana, se torna uma dificuldade quando todo um contexto é necessário para se obter o sentido de uma expressão. Por exemplo a expressão “muito grande” pode representar algo positivo quando se refere ao tamanho de um quarto de hotel, porém negativo se estiver se referindo ao tamanho de um tumor.

Outro desafio é determinar a polaridade em sentenças comparativas. Por exemplo, na frase: “Autonomia do Toyota Prius é melhor do que a autonomia dos Tesla”. O termo “melhor” está presente e a sentença possui um grande potencial opinativo, porém não é fácil para uma máquina identificar qual o interlocutor se refere, nem qual é o objeto em análise.

Termos de negação, também, podem causar resultados completamente errados se não forem manipulados corretamente, por exemplo na frase: “Você não vai precisar recarregar este telefone durante o dia.” O termo “não” é o que torna esta frase positiva.

Existem outros pontos que podem causar problemas por exemplo os advérbios de intensidade, que representam diferentes graus de polaridade e o sarcasmo (Kolkur et al., 2015).

2.1. Coleta de dados e fontes

A Hipótese dos Mercados Eficientes Fama (1970) afirma a impossibilidade de obter retornos positivamente consistentes tomando como base as informações disponíveis no momento em que o investimento é realizado. Porém, Nassirtoussi, Aghabozorgia, Waha e Ngob (2014) explicam que Fama revisa sua obra e inclui três níveis de eficiência em sua hipótese: forte, semi-forte e fraca. Gerando um indicativo de que existem mercados em que a previsibilidade não apenas é plausível como é viável. A eficiência do mercado está correlacionada com a disponibilidade das informações e um mercado e é "altamente eficiente" apenas quando todas as informações estão completamente disponíveis, o que na prática, em geral, não acontece.

Diversos estudos demonstram que a mídia não somente informa os acontecimentos do mercado, mas representa um personagem ativo em sua dinâmica (Robertson, Geva & Wolff, 2006). A interpretação das pessoas sobre as notícias é muito variável, conforme conclui Friesen e Weller (2006), existem vieses cognitivos como excesso de confiança, que são suficientemente grandes para gerar significado econômico.

As informações disponíveis que representam e afetam o mercado financeiro, e que podem ser submetidas a ferramentas de análise de sentimento, são textos opinativos que podem vir de diversas fontes, publicações corporativas, mensagens na internet (redes sociais por exemplo), *blogs* de formadores de opinião entre outros.

As publicações corporativas, que compreendem comunicados oficiais e notas, são uma boa fonte para análises textuais, visto que vem de dentro das empresas, de pessoas que têm um melhor conhecimento da empresa que outras de fora. Estas publicações podem transmitir de maneira confiável informações sobre o desempenho futuro, e tendem a representar as altas ou quedas dos valores de ações. No entanto, a frequência desse tipo de publicação costuma ser baixa demais para constituir uma boa fonte para nossa aplicação.

Por outro lado, notícias veiculadas na mídia costumam compreender uma gama maior de informações, e possuem como característica uma certa credibilidade relacionada a imparcialidade das informações. Estes textos são relevantes para a economia como um todo, pois normalmente vêm acompanhados de opiniões e comentários de especialistas capazes de expressarem pareceres sobre o mercado como um todo ou sobre uma indústria específica.

Isso torna a análise de sentimento baseada em notícias publicadas em portais especializados na internet uma escolha apropriada para o estudo vigente.

Postagens em redes sociais também são potencialmente úteis como fonte para análise de sentimento, pois muitas pessoas despendem tempo diariamente lendo e escrevendo sobre valores de ações, e algumas delas possuem grande visibilidade, tornando-as relevantes na

formação de opinião financeira, podendo até mesmo causar impacto nas reações de outras fontes de notícias, gerando influências no mercado financeiro. (Das & Chen, 2007)

Postagens de menor visibilidade como de pequenos investidores também podem ser uma fonte útil, embora potencialmente possam gerar um nível mais ruidoso de expressão de sentimento, o volume de narrativas é muito maior, e segundo argumenta Black (1986), o ruído sobre um grande número de pequenos eventos, é potencialmente um fator causal mais poderoso que um pequeno número de grandes eventos. (Kearney & Liu, 2014)

2.3. Análise de sentimento em contextos financeiros

Estudos anteriores na área de análise de sentimentos demonstram que a técnica de obtenção do tom de um texto pela contagem de palavras apresenta resultados satisfatórios, porém o cenário se altera quando se trata de notícias e artigos contendo muitos termos de cunho financeiro. Essa técnica se baseia em um dicionário de palavras com dimensões de classificação, normalmente categorizadas como ‘positivas’ ou ‘negativas’.

Loughran e McDonald (2011) descrevem uma fonte comum de classificação de termos o *Harvard Psychosociological Dictionary* (Harvard-IV-4). Uma vantagem de se utilizar dicionários comuns é que os termos não possuem viés interpretativo do pesquisador, por outro lado, uma mesma palavra pode ter significados distintos dependendo da disciplina de análise, e isso gera ruídos consideráveis na obtenção da tonalidade textual para aplicações financeiras. Neste sentido, o estudo de Loughran e McDonald (2011) presta uma importante contribuição evidenciando uma alta frequência de erros de classificação na lista de Harvard, onde 73.8% das palavras classificadas como negativas, não possuem o mesmo tom no contexto financeiro.

2.4. Legibilidade textual

A legibilidade de um texto está relacionada à capacidade do indivíduo de compreender e interpretar uma mensagem, porém quando buscamos extrair alguma métrica computacionalmente útil, temos que recorrer a técnicas mais viáveis do ponto de vista do processamento de linguagem natural.

Para nossa aplicação, onde desejamos obter de forma consistente a polaridade de uma notícia financeira, iremos nos concentrar em uma técnica de extração de significado descrito por Loughran e McDonald (2016) que supõe a independência entre as palavras, ou seja, a ordem e, portanto, o contexto direto de um documento não importa. As técnicas em que a ordem das palavras é desprezada são comumente chamadas de ‘*bag of words*’. Basicamente eles se baseiam na redução do documento em uma matriz de termos, com uma coluna contendo o número de ocorrências de cada palavra, isto permite que um grande documento seja resumido e o processamento destes dados se torna computacionalmente mais acessível. Evidentemente esta perda de contexto gera prejuízos na interpretação textual e portanto, é cabível que sejam empregados meios de compensação de ruído como por exemplo, a exploração das multidimensionalidades dos dicionários.

2.5. Dicionários

Para determinar o sentimento de um artigo financeiro usando a técnica de contagem de palavras, precisamos de um conjunto compilado de termos classificados com sentimentos, o que chamamos de dicionário. Na prática este conjunto compilado consiste em uma planilha onde cada linha representa um termo e cada coluna seus atributos. Vale salientar que estamos considerando o uso de dicionários originalmente desenvolvidos na língua inglesa dada a maior disponibilidade de literaturas e estudos publicados neste idioma.

Dentre os dicionários citados, o proposto por Loughran e McDonald (2011) chama a atenção, pois os autores desenvolveram um dicionário focado em termos da área financeira, baseado em um outro dicionário chamado *2of12inf* (2020). O dicionário de Loughran e McDonald (LM) não contempla abreviações, acrônimos e substantivos próprios, com exceção de '*Scholes*', dada a importância e frequência do termo '*Black-Scholes*', porém inclui as inflexões de termos. Este é um ponto importante a ser considerado, pois a contemplação das variantes de um termo praticamente extingue a necessidade de um pré-processamento textual, fase comum aplicada ao processamento de dados não estruturados, também conhecida como *Data Steaming*.

Para o desenvolvimento deste dicionário, Loughran e McDonald expandiram a lista *2of12inf*, tabulando todos os '*tokens*' de uma amostra de relatórios financeiros (*10-K* e *10-Q form*) coletados no período de 1994 a 2008. Todos os tokens que apareceram pelo menos 100 vezes foram identificados como palavras e foram adicionados ao dicionário.

A lista LM possui 354 palavras positivas e 2355 palavras negativas e possuem a vantagem de serem bastante abrangentes, não faltam termos positivos ou negativos que normalmente aparecem em artigos financeiros.

Além das classificações de palavras positivas e negativas, este dicionário ainda fornece uma série de outras dimensões que podem se tornar relevantes no decorrer do desenvolvimento deste trabalho, como palavras de incerteza, litigiosas e modais (forte, moderado e fraco, exemplo *CLEARLY*, *GENERALLY* e *APPARENTLY*, respectivamente).

Muitas pesquisas já utilizaram a lista de LM para obter a tonalidade em artigos de jornal, como por exemplo Loughran e McDonald (2011), Dougal, Engelberg, Garcia e Parsons (2012) que fez a análise de polaridade de colunistas do *Wall Street Journal* a fim de medir o pessimismo relativo aos mesmos. Vale ressaltar que o dicionário LM não inclui termos de uma única letra, que são muito comuns na língua inglesa como por exemplo 'I' ou 'a', pois normalmente não influenciam na análise além de que são muito usados para enumerar listas.

2.6. Ponderação de termos

Quando tentamos obter alguma característica interpretativa a partir da contagem de palavras, uma série de questões são levantadas a respeito de como as contagens devem ser contabilizadas, por exemplo, se considerarmos a contagem bruta de termos, teremos o tamanho do texto como um critério polarizador (pois evidentemente um grande número de palavras está diretamente relacionado com o tamanho do artigo). Este problema é facilmente resolvido com uma média ou cálculo de proporção, porém em outros casos é plausível considerar pesos distintos para determinados termos, por exemplo, entre as palavras negativas de LM, desfavorável aparece

1.000 vezes mais frequentemente do que ‘expropriar’, ‘desinformar’ ou ‘indiciar’ nos relatórios fiscais divulgados de empresas. Talvez as palavras mais incomuns devam receber um maior peso na tabulação do sentimento negativo (Loughran & Mcdonald, 2016).

2.7. Trabalhos relacionados

Valdevieso (2019) propôs uma análise da correlação do comportamento variante do preço de ativos da bolsa de valores por meio de análise técnica, com um indicador de polaridade das notícias publicadas por portais com foco no mercado financeiro.

O presente estudo difere desta abordagem principalmente pelo emprego de uma abordagem distinta na análise de sentimentos. Valdevieso (2019) explorou uma ferramenta especializada em processamento de linguagem natural a fim de obter escores para cada unidade textual a ser analisada. Por se tratar de uma ferramenta proprietária, com foco de atender grandes volumes de dados corporativos sua customização possui um alto custo, tornando inviável a especialização da mesma para o processamento de termos da área financeira. O autor sugere o emprego de técnicas especializadas na geração da análise de sentimentos.

Minh et al (2018) tentam superar a dificuldade apresentada pelo uso de ferramentas de análise de sentimento não construídas a partir de acervos lexicais comuns ao setor financeiro, propondo um framework de predição de movimento no preço de ações criado a partir de dicionários especializados. Obtiveram resultados expressivos nos experimentos que analisaram índices de preço da S & P 500 e artigos dos portais Reuters e Bloomberg.

Kolkur et al. (2015) discutem sobre uma visão dos diferentes níveis em que a análise de sentimentos pode ser aplicada, com foco em opiniões e comentários sobre produtos de *e-commerce*. São pontuados os principais desafios que tornam a análise de sentimentos uma tarefa pouco trivial. (Kolkur et al., 2015).

Nassirtoussi et al. (2014) busca se aprofundar no estudo teórico da análise textual de mídias sociais e notícias, com foco interdisciplinar que envolve desde os tópicos econômicos e comportamentais até inteligência artificial.

Os autores contribuem na formação de um quadro mais claro de discussão a partir de pesquisas publicadas na área, tomando como base três aspectos comuns: pré-processamento, aprendizado de máquina e mecanismo de avaliação, todas com várias sub discussões. Segundo concluem, os avanços no campo da mineração de texto com finalidade preditiva do mercado de ações podem ter (entre outras) implicações muito significativas, como o aumento da visão dos mercados financeiros, pois a falta dela pode afetar negativamente os meios de subsistência de milhões de pessoas em todo o mundo. (Nassirtoussi et al., 2014)

Por fim, Schumaker e Chen (2009) propuseram uma abordagem preditiva baseada em aprendizado de máquina para a análise de notícias financeiras usando representações textuais distintas. Usando esta abordagem fizeram a análise de 9211 notícias e milhões de cotações de ações da S & P 500, gerando uma estimativa de preço das ações vinte minutos após o lançamento de uma notícia. Com isso obtiveram uma precisão direcional com 57,1% de precisão em um simulador de negociações.

3. MATERIAIS E MÉTODOS

O estudo foi realizado com base em levantamento teórico sobre a técnica de análise de sentimentos e estudos correlatos para o entendimento do problema de pesquisa. A partir do levantamento, foram buscadas bases de dicionários do domínio financeiro que pudessem contribuir na análise de sentimento.

Para tal, foi selecionado e avaliado o dicionário de Loughran e McDonald (LM). Após esta etapa, foram verificados portais de notícias e realizada a seleção do yahoo!finance, que fornece um serviço agregador de conteúdo, reunindo notícias de acordo com um tema escolhido, de modo a extrair notícias de uma determinada ação. Foram coletadas notícias entre os dias 28 de setembro e 04 de novembro de 2020 sobre a Alphabet Inc, uma holding responsável por gerir o portfólio de empresas da Google.

A partir das notícias realizou-se a análise de sentimentos e a coleta dos respectivos preços no mesmo momento, e temporalmente defasados em 24, 48 e 72 horas após. Com base nos dados obtidos foi realizada a análise de correlação entre a polaridade das notícias e o preço da ação.

Para a realização da pesquisa foi necessária a utilização de recursos computacionais como: linguagem Python - Linguagem multiparadigma de propósito geral e de alto nível; Anaconda - Framework para computação científica; Spyder: IDE para programação em Python; computador Intel Core i5 de 8ª geração com 16Gb de memória ram e armazenamento SSD.

4. EXPERIMENTOS

A fim de avaliar o dicionário de Loughran e McDonald (LM), utilizou-se um algoritmo para analisar uma série de termos de controle com a intenção de averiguar a abrangência e a coerência deste dicionário. O algoritmo realiza a ponderação por contagem simples de termos, resultando em uma análise com diversos parâmetros, sendo os mais relevantes para o momento: *Number of words* (contagem de termos relacionados)*; % Positive; % Negative; % Uncertainty*.

O experimento processou a lista de controle contendo 6.577 termos e expressões típicos da área financeira, do total de 16.663 palavras. O objetivo da análise foi validar a abrangência do acervo léxico da biblioteca LM. Essa lista foi coletada do site *Investopedia*, um portal de informações financeiras, que oferece diversas ferramentas para investidores e entusiastas, dentre elas um dicionário com significados de expressões financeiras. Como resultado, obtivemos 14.986 termos relacionados ao dicionário, ou seja 89.9% de correspondência.

4.1. Coleta de dados de notícias

A fonte escolhida para este estudo são notícias de portais especializados. Notícias sobre o mercado financeiro estão disponíveis em diversos sites na Internet, para esta fase foi utilizado o yahoo!finance, que fornece um serviço agregador de conteúdo, reunindo notícias de acordo com um tema escolhido.

Foram escolhidos alguns parâmetros para a coleta e análise dos dados, os quais são descritos na sequência: Período - foram coletadas notícias entre os dias 28 de setembro e 04 de novembro de 2020; empresa - observando empiricamente as movimentações de ações da NASDAQ (*National Association of Securities Dealers Automated Quotations*) uma das mais relevantes

bolsas dos Estados Unidos, escolhemos a Alphabet Inc, uma *holding* responsável por gerir o portfólio de empresas do Google, sendo seu maior subsidiário o próprio Google Inc, as ações são negociadas na bolsa de valores sob a sigla 'GOOG'; Portais -O agregador escolhido foi o yahoo!finance, o critério utilizado nesta escolha foi a relevância da página em sites de busca.

Durante o período de buscas via coleta manual de notícias, foi gerado um conjunto de arquivos textuais contendo cada um deles: o título e subtítulo, o conteúdo da notícia, a data e hora de publicação. No total foram gerados 193 arquivos textuais.

4.2. Coleta de dados do preço da ação alphabet inc

A obtenção da cotação de ações segue um padrão parecido com o tópico anterior, diversos sites se propõem a monitorar as operações de compra e venda de papéis na bolsa, obtendo as cotações e disponibilizando para o público. Existem ainda os que disponibilizam os dados em interfaces de programação de aplicações (*Application Programming Interface* - API), facilitando o desenvolvimento de sistemas que fazem proveito dessas informações.

Dentre estes, destacamos a *Alpha Vantage*, uma comunidade focada em fornecer esse tipo de serviço. Tendo como um dos objetivos a democratização da informação, é considerada hoje uma das mais relevantes fontes de cotação via API, moeda estrangeira (*Forex*) e criptomoeda. Dado o baixo volume de consultas necessárias para a realização dos experimentos deste trabalho, a utilização deste serviço é gratuita. Sua relevância, disponibilidade e facilidade de uso justificam seu emprego.

O acesso aos dados foi feito utilizando uma biblioteca própria da *Alpha Vantage*, que permite o acesso aos dados de maneira estruturada, com granularidade parametrizável. O intervalo escolhido inicialmente foi de 1 minuto, usando como parâmetro a sigla desejada. Obtivemos como retorno um *dataset* que pôde ser escrito em arquivo, utilizando formato tabular.

Para cada registro de cotação, o dataset ocupa uma linha, e contém data e hora, valor de abertura, máximo no período, mínimo do período e fechamento. Foi feita a coleta cobrindo o mesmo período das notícias, bem como valores referentes a mesma empresa ('GOOG').

4.3. Processamento e construção dos dados de análise

A última fase do experimento, é o processamento das bases de dados com o intuito de obter os dados de relevância ao estudo. Sendo assim, dividimos esta etapa em duas partes: Processamento da base de notícias e estruturação dos resultados para cálculos de correlação.

4.3.1 Processamento da base de notícias

A fase de coleta de dados gerou um conjunto de arquivos de texto aptos a serem processados pelo algoritmo de ponderação de termos, a partir do qual extraímos o valor relativo a análise de sentimento. Este algoritmo foi construído com a intenção de processar um conjunto de arquivos de textos de um determinado diretório.

Todos os 193 arquivos foram processados e salvos em um arquivo CSV.. A escrita do arquivo em formato CSV possibilitou a tabulação dos dados em software de planilha eletrônica de maneira direta.

4.3.2. Modelagem e correlação

Considerando que as saídas que obtivemos foram em formatos tabulares, optamos inicialmente por modelar os dados em software de planilha eletrônica. Pela disponibilidade e simplicidade de uso, escolhemos a suíte Google, que dispõe gratuitamente de diversas ferramentas de trabalho, incluindo planilhas.

Partimos da tabela com as análises de sentimento das notícias, inicialmente excluímos os parâmetros desnecessários para esta análise, mantendo somente o nome do arquivo (que contém o *timestamp* da notícia), os percentuais positivo, negativo e incerteza. O *timestamp* indica em que tempo algo ocorreu (dia, mês, ano, hora, minuto, segundo, milissegundo).

Importamos a tabela com os dados das ações minuto a minuto extraídos da API citada anteriormente, como esta possui um número grande de registros, adicionamos em outra página da planilha. Outro passo importante foi a conversão dos dados, sendo estes de texto para valores numéricos, e a extração das informações de *timestamp* para um formato de data e hora, isto foi importante para a automação de algumas tarefas descritas a seguir.

Uma vez que temos a data e hora da publicação da notícia, podemos buscar na tabela de preços o valor da ação naquele momento. Isso foi realizado de maneira automática utilizando uma *query* de funcionamento bastante simples, retornando o valor na coluna de preço, no registro onde a data e hora (de publicação da notícia) correspondem a data e hora da cotação da ação:

```
=QUERY(stocks;"SELECT A, B, C WHERE A= timestamp '"&TEXTO(A5;"yyyy-mm-dd HH:mm:ss")&"'";0)
```

Embora os registros de preços obtidos possuírem resolução de minuto a minuto, existem diversos intervalos em que não houve negociação alguma, fazendo com que os preços não se alterassem, quando isto acontece esses registros são abstraídos, o que é uma forma eficiente de economizar volume de dados desnecessários, porém implica na adaptação da nossa fórmula a fim de obter a cotação no instante mais próximo do momento da publicação da notícia.

Utilizando a mesma abordagem, adicionamos a busca considerando o valor da ação 24, 48 e 72 horas após a publicação da notícia. Adicionamos uma coluna contendo o valor da diferença entre o teor positivo e negativo na intenção de obter um índice numérico de sentimento da notícia. Por último, extraímos a variação padronizada do preço em cada um dos intervalos de tempo, e também da coluna de diferença.

Por padronização da variação entende-se a fórmula simples de obtenção da variação proporcional onde o valores deslocam o mínimo possível de zero:

$$f(x) = \frac{x - min(x)}{max(x) - min(x)}$$

5. RESULTADOS

O principal objetivo deste estudo, é estabelecer a correlação entre o teor de uma determinada notícia, com a variação do valor da ação no qual a mesma se refere. O cálculo de correlação entre duas variáveis pode ser obtido pela fórmula de coeficiente de correlação de Pearson:

$$r = \frac{\sum_{i=1}^{n} (x-\underline{x})(y-\underline{y})}{\sqrt{\sum_{i=1}^{n} (x-\underline{x})^2(y-\underline{y})^2}}$$

Onde x e y são as médias do intervalo a ser calculado.

A ferramenta escolhida para análise implementa esta função com a seguinte assinatura:

PEARSON(ax:ay; bx:by)

Onde a e b são as variáveis, x e y são os intervalos a serem calculados. Esta função torna a análise simples dado o conjunto de dados disponível.

5.1. Correlação geral

De maneira preliminar, é executado o cálculo de correlação entre a diferença de proporção negativa e positiva, e o preço da ação em todo o conjunto de dados, ou seja, considerando todos os dias como intervalo (Tabela 1).

Tempo decorrido (horas)	média geral	desvio padrão	correlação
0	1623,00	55,26	0,2537
24	1631,80	60,91	0,0792
48	1647,95	52,68	0,0816
72	1665,27	57,46	0,1974

Tabela 1. Cálculos considerando todo o intervalo

Esta análise inicial demonstra um coeficiente baixo de correlação, porém considerando os resultados de análises desta área, houve motivação para uma análise mais detalhada.

5.2. Correlação por intervalo

Com o objetivo de obter dados consistentes, fizemos a análise de correlação por dia, obtendo os seguintes valores (Tabela 2):

	0 hrs	24 hrs	48 hrs	72 hrs
27/10/2020	-0,2654	-0,2357	-0,0917	-0,2002
28/10/2020	0,0837	0,0481	0,3117	-0,0063
29/10/2020	0,2833	-0,1777	-0,2493	0,0909
30/10/2020	0,1123	-0,0333	-0,1590	-0,0367
2/11/2020	0,1805	-0,2546	-0,0736	0,0721
3/11/2020	-0,0537	-0,0581	-0,0496	0,0088

Tabela 2. Correlação por dia

Ao se analisar a Tabela 2, verifica-se que os valores positivos indicam uma correlação entre as notícias e o preço das ações, enquanto os valores negativos indicam uma correlação negativa, ou seja, quando a notícia é positiva o preço cai e quando a notícia é negativa o preço sobe. Tais valores negativos indicam inconsistência em relação às notícias e seus impactos no preço da ação. Com base nos dados apresentados verifica-se que há uma maior correlação entre notícias e o preço das ações da Google se analisarmos as informações no mesmo horário, pois cerca de 66,67% dos dias apresentaram correlação positiva com média de 0,16 para as correlações positivas. Para um período de defasagem de 24 a 48 horas as correlações positivas foram encontradas em 16,67% dos dias. Um fato diferenciado ocorreu na defasagem em 72 horas, quando a correlação positiva ocorreu em 50% dos dias analisados.

De modo a ilustrar a variação das notícias e o preço das ações da Google, é apresentado o Gráfico 1 referente ao dia 29/10/2020. Para a análise gráfica optou-se por utilizar os valores padronizados, movendo a escala da variação para próximo de zero, de modo que os valores são apenas a taxa proporcional da variação, e não valores absolutos. Cada ponto do eixo horizontal representa o instante de publicação de uma notícia e, no eixo vertical está representado o valor obtido pela análise de sentimento. Logo temos a linha vermelha representando a variação da polaridade das notícias publicadas ao longo do dia. Decidimos pelo uso desse tipo de gráfico com linhas suavizadas para que fosse mais fácil de se observar a tendência que os valores demonstram, cuja comparação é importante para nossa análise.

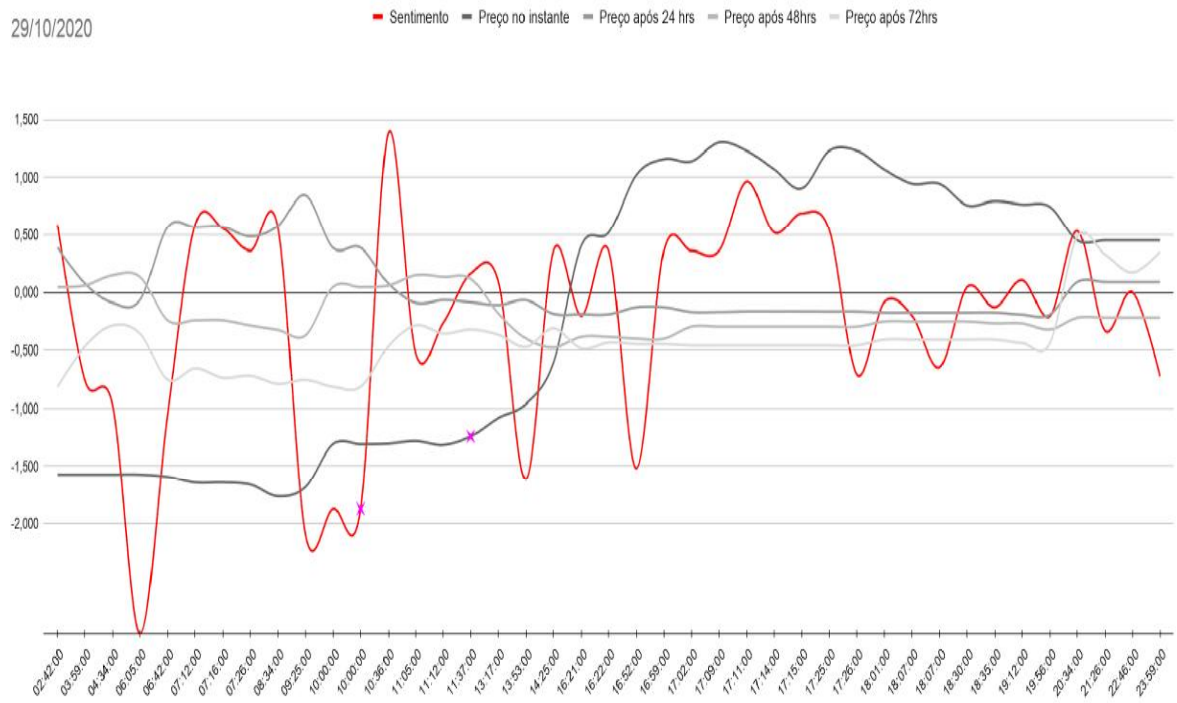

Gráfico 1. Variação do preço e sentimento 29/10/2020

No gráfico 1, de acordo com os cálculos realizados, obtivemos uma taxa de correlação alta neste dia (29/10/2020) para as cotações das ações no instante da publicação da notícia, sendo esta de

0,2833. Ainda em relação ao gráfico 1, destacam-se no gráfico dois pontos na cor magenta, onde possivelmente exista uma relação entre a publicação de uma notícia com alto teor de polaridade positiva, e a ascensão do preço da ação, visível entre o final da manhã (11:37) e o final da tarde (17:11).

CONCLUSÃO

Com o crescimento de participantes ativos e diretos no mercado de ações, a utilização de ferramentas de análise de sentimentos tem sido um objeto de pesquisa em diversos estudos. Esse tipo de análise tem como subsídio informações textuais em linguagem natural, então a junção do processamento e estruturação deste tipo de dado com o desempenho dos preços é imprescindível.

De modo, para atingir o objetivo deste estudo, a coleta de notícias foi executada utilizando um agregador capaz de centralizar notícias de diferentes portais. Para tais notícias foi então realizada a análise de sentimentos para obtenção das respectivas polaridades; bem como foram obtidas as cotações de determinada ação com baixa granularidade, permitindo a correspondência com os horários de publicação das notícias. Com os dados gerados foi realizada a correlação entre os valores da cotação da ação em diferentes momentos: no momento da publicação da notícia, 24, 48 e 72 horas após, a fim de verificar a histerese do impacto da notícia com o valor do ativo.

Ao se analisar os resultados, verificou-se a existência de níveis de correlação entre os sentimentos das notícias e a variação de preço do ativo analisado no estudo. Por outro lado, também foram encontradas discrepâncias nas correlações calculadas nos intervalos definidos. O que revela uma possível carência de um refinamento na análise de intervalos entre a publicação da notícia e a cotação da ação. Outro ponto de atenção, é o fato que os métodos utilizados não levam em consideração nenhuma subjetividade textual, o que pode gerar avaliações pouco assertivas de notícias, ou ainda, considerando que a notícia se refira ao objeto de busca, uma avaliação de um texto generalista seja adicionado à análise. Uma limitação deste trabalho se deve ao escopo de tempo e número de ativos utilizados na análise.

Considerando as dificuldades encontradas em relação a subjetividade textual, podem ser considerados como possibilidades de estudos futuros: desenvolvimento de uma forma de análise capaz de considerar a subjetividade linguística; obter os valores de cotação de maneira linear, a fim de obter a correlação de intervalos menores; aumentar o escopo de tempo e de ativos para se estudar a correlação de sentimentos de notícias e seu possível impacto no preço.

REFERÊNCIAS

Alpha Vantage. (2020, novembro 12). *Alpha Vantage offers free stock APIs in JSON and CSV formats for realtime and historical equity, forex, cryptocurrency data and over 50 technical indicators*. Disponível em: https://www.alphavantage.co. Acesso em: 20/09/2020.

Beyaz, E., Tekiner, F., Zeng, X., & Keane, J. (2018). Comparing Technical and Fundamental Indicators in Stock Price Forecasting, *IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems* (HPCC/SmartCity/DSS), Exeter, United Kingdom, p. 1607-1613.

Black, F. (1986). *Noise. Journal of Finance*, 41, 529–543.

Blakey, P. (2001). Wireless investor [fundamental and technical analysis]. IEEE Microwave Magazine, v. 2, n. 4, p. 18-24.

Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management science*, v. 53, n. 9, p. 1375-1388.

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

Finantial Terms DICTIONARY – Investopedia (2020). [S. l.], Disponível em: https://www.investopedia.com/financial-term-dictionary-4769738. Acesso em: 14 set. 2020.

Giampietro, A. C. T., Sarraceni, J. M., Montanha, R. D. C. L, Horita, R. Y., & Indalêncio, T. C. (2007). As Bolsas de Valores Estão Cada Vez Mais Fazendo Parte da Vida dos Brasileiros?, Lins Sp. *I Encontro Científico e I Simpósio de Educação – Unisalesiano.* Disponível em: http://www.unisalesiano.edu.br/encontro2007/trabalhosaceitos.html. Acesso em: 15 jul. 2020.

Hajek, P., & Barushka, A. (2018). Integrating sentiment analysis and topic detection in financial news for stock movement prediction. *Proceedings of the 2nd International Conference on Business and Information Management*, p. 158-162.

Islam, A., Zaman, H., & Ahmed, R. (2009). Automated fundamental analysis for stock ranking and growth prediction. *12th International Conference on Computers and Information Technology*, Dhaka, p. 145-150.

Kearney, C., & Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, v. 33, p. 171-185.

Kent, D., Hirshleifer, D., & Teoh, S. H. (2002). Investor Psychology in Capital Markets: Evidence and Policy Implications. *Journal of Monetary Economics*, v. 49, n. 1, p. 139-209.

Kolkur, S., Dantal, G., & Mahe, R. (2015). Study of different levels for sentiment analysis. *International Journal of Current Engineering and Technology*, v. 5, n. 2, p. 768-770.

Lo, A. W. (2004). The adaptive markets hypothesis: market efficiency from an evolutionary perspective. *Journal of Portfolio Management*,30:15–29.

Loughran, T., & Mcdonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, v. 66, n. 1, p. 35-65.

Loughran, T., & Mcdonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, v. 54, n. 4, p. 1187-1230.

Minh, D. L., Sadeghi-Niaraki, A., Huy, H. D., Min, K., & Moon, H. (2018). Deep Learning Approach for Short-Term Stock Trends Prediction Based on Two-Stream Gated Recurrent Unit Network. *IEEE Access*, v. 6, p. 55392-55404.

Murphy, J. J. (1999). *Technical analysis of the financial markets*. EUA: New York Institute of Finance.

Nasdaq Composite. (2020). *Daily Stock Market Overview, Data Updates, Reports & News.* Nasdaq, Disponível em: https://www.nasdaq.com. Acesos em 14/09/2020.

Nassirtoussi, A. K., Aghabozorgia, S., Waha, T. Y., & Ngob, D. C. L. (2014).Text mining for market prediction: A systematic review. *Expert Systems with Applications*, v. 41, n. 16, p. 7653-7670.

Reuters S. Facebook‘s EU-US data transfer mechanism ‘cannot be used‘, Irish regulator says. Reuters.com, Disponível em: https://www.reuters.com/article/us-facebook-privacy/facebooks-eu-us-data-transfer-mechanism-cannot-be-used-irish-regulator-says-idUSKBN2602X4. Acesso em: 14/09/2020.

Robertson, C., Geva, S. & Wolff, R. (2006). What Types of Events Provide the Strongest Evidence that the Stock Market is Affected by Company Specific News. In Li, J, Simoff, S J, Kennedy, P J, Christen, P, & Williams, G J (Eds.) *Proceedings of the Fifth Australasian Data Mining Conference 2006*. Australian Computer Society Inc., Australia, p. 145-154.

Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems (TOIS)*, v. 27, n. 2, p. 1-19.

The 2of12inf dictionary is documented online. (2020). Disponível em: http://wordlist.aspell.net/12dicts-readme/. Acesso em: 14/09/ 2020.

Valdevieso, G. S. (2019). Análise de Sentimentos e indicadores técnicos: uma análise da correlação dos preços de ativos com a polaridade de notícias do mercado de ações. 2019. 54 f. *TCC (Graduação) - Curso de Informática, Departamento de Informática*, Universidade Estadual de Maringá.

yahoo!finance. (2020). *Search for news, symbols or companies. finance.yahoo.com*, Disponível em: https://finance.yahoo.com. Acesso em: 28/11/2020.