1. Introdução

Um novo processo análise de investimento, que ganhou força nos últimos anos, é o de análise quantitativas, somando-se às antigas técnicas fundamentalistas. Esta era uma técnica que se baseava na análise profunda das condições financeiras da companhia, seus prospectos de crescimento e como o setor ao qual está inserida se comporta. Já um viés mais quantitativo se baseava em análises de séries temporais e outros métodos sistemáticos para que o investimento fosse realizado. Tal processo ganhou relevância por meio dos resultados de especialistas na área, apresentando retornos surpreendentes com uma estratégia mais sistemática e imparcial. Neste âmbito, vale ressaltar os trabalhos e resultados de \cite{Thorp1967}, sendo assumido como um dos primeiros especialistas na área quantitativa \cite{Patterson2011}.

Ao analisar movimentos nos preços de mercado, é de suma importância o entendimento da hipótese da eficiência de mercado (HEM) \cite{Fama1965}. Tal estudo estabelece duas principais hipóteses: o preço praticado nos ativos reflete toda a informação disponível para os agentes de mercado e é impossível alcançar retornos superiores que a média de retornos no mercado. Todavia, métodos como “Value Investing”, defendido por Benjamim Graham - investidor com retornos elevados e consistentes ao longo de vários anos - acaba por conseguir aparentemente enfraquecer tal hipótese. Ainda, outros trabalhos, como \cite{Werner1985} e \cite{Banz1981} ajudam nesse processo. Soma-se a isso a ascensão de estratégias quantitativas - eficientes em trazer retornos elevados acima da média de mercado.

Tais consistentes resultados foram suficientes para Eugene Fama, criador da HEM e laureado com o Nobel de Ciências Econômicas em 2013, duvidar e rever seus conceitos \cite{Fama1996}. Dessa forma, Fama e Kenneth French passaram a considerar outros fatores, além de somente o movimento geral do mercado, para justificar o desempenho de ativos de ações. Por meio de uma análise de regressão, o modelo de Fama-French foi criado \cite{Fama1993}. Este trabalho alega que o movimento do mercado pudesse ser justificado entre 3 fatores: movimento do mercado, tamanho do valor do ativo (*size*) e qualidade financeira da companhia (*value*). Por fim, Fama recentemente publicou outro trabalho \cite{Fama2015}, incorporando outros 2 fatores (capacidade de geração de lucro - *profitability* - e quantidade de capital em investimento - *investment*) na contínua tentativa de justificar o movimento do mercado.

Ainda, com o passar dos anos, foi possível perceber um aproveitamento das técnicas quantitativas a partir do crescente poder computacional disponível. Conjuntamente, estudos mais robustos de ciência de dados foram publicados no meio acadêmico. Com isso, surge a possibilidade de usar informações textuais não estruturadas, como uma nova alternativa de dados para aprimoramento de modelos e de seus resultados.

Assim, partindo dessa visão holística dos dados disponíveis, permite-se a inserção de fatores relacionados com economia comportamental, que traz maior realismo para a modelagem. Tais fatores refletem a reação dos participantes do mercado diante de diversas situações. Em especial, ressalta-se a fuga da racionalidade de investidores individuais em situações de alta insegurança, como é relatado em muitos estudos de Daniel Kanehman, Amos Tversky e Richard Thaler (\cite{Kahneman2013}, \cite{Kahneman2021} e \cite{Thaler2016}). Dessa forma, urge-se a importância da inserção desses vieses humanos nas técnicas quantitativas para melhor compreensão do mercado, tirando maior proveito do registro dos seus movimentos.

Diante o que foi exposto - falta de compreensão plena do mercado e constante desenvolvimento de técnicas de análise de dados - nota-se um ambiente com um grande espaço para aproveitamento dos movimentos do mercado. Com isso, vale focar em trabalhos que buscam avaliar a correlação entre movimentos de mercado e informações qualitativas relacionadas. Tais dados podem ser provindos de sites de notícia, blogs de finanças, redes sociais, divulgação financeira de companhias listadas, etc. Para conseguir a quantificação do efeito de tais dados na tomada de decisão de gestão de ativos, em uma possível abordagem, é conduzida uma análise de sentimento. Tal análise foca em avaliar qual o sentimento que autor do texto avaliado quis expressar (positivo/negativo). Com isso, usa-se a técnica de processamento de linguagem natural, com o objetivo de, a partir do sentimento, posicionar a informação obtida dentro de um espectro quantitativo de compra/venda. Dessa forma, usa-se tal posicionamento como fonte de apoio à tomada de decisão em relação a determinado ativo.

1. Revisão Bibliográfica

Nota-se que o uso de análise de sentimento é abordado de forma ampla, com diferentes métodos e para diferentes ambientes. Com isso, pode-se encontrar trabalhos que usam dados de manchetes de notícias para realizar previsão de dados macroeconômicos \cite{Kalamara2022} - PIB, inflação e desemprego. Outros, usam dados de micro blogs, tais como Twitter, para realizar previsão de preço de Bitcoin \cite{Karalevicius2014} ou predição de resultados eleitorais \cite{Yang2018}. Restringindo-se ao âmbito de finanças, nota-se certo foco em usar técnicas mais basilares para predição de tendência de preços e, mais recentemente, também com técnicas mais atualizadas. Com isso, são apresentados os resultados de tais trabalhos.

Técnicas tradicionais de aprendizado de máquina

Dentre os diversos trabalhos na área, tem-se \cite{Khedr2017}. No estudo, tem-se a união de análise de sentimento de notícias financeiras com os dados de preços históricos de três companhias (Yahoo Inc. - YHOO, Microsoft Corp. - MSFT e Facebook Inc - FB). Usando método de Naive-Bayes e dados de notícias de tais companhias, foi possível alcançar uma acurácia entre 72,73\% - 86,21\% na previsão de sentimento - superando resultados provindos de técnicas como SVM e K-NN. Usando somente os dados de classificação de polaridade para previsão de tendências futura, obteve-se uma acurácia de 59,18\% - 63\%. Todavia, unindo tais dados com o histórico de preços, usando-se do algoritmo de K-NN, alcançou-se uma de acurácia para até 89,90\% para previsão de tendência dos preços. Tal resultado superou o produzido por outros trabalhos, tais como \cite{Shynkevich2015} e \cite{Bing2014}.

Um modelo um pouco mais complexo é apresentado com o *Media-Aware Quantitative Trader* (MAQT) \cite{Li2014}. Neste caso, é usado como *input* de modelos o sentimento de miniblogs, as manchetes de sites de notícias - usando de  *part-of-speech* (POS) para escolha das palavras - e preço histórico. Tais inputs são inseridos no algoritmo SVM para predição de tendências no mercado chinês. Ressalta-se o uso de um dicionário de sentimento específico para finanças (Loughran e MacDonald (LM) \cite{Loughran2011}), com o intuito de conduzir uma abordagem de sentimento mais assertiva que um dicionário para um ambiente mais geral (e.g. Harvard IV-4 \cite{Stone1966}). Ainda, é interessante notar um máximo de previsibilidade se defasar o dia da análise de sentimento e o movimento do mercado em cinco dias, de acordo com os estudos de atraso de captura de informações do mercado de \cite{Lebaron1999}, e apresentando um resultado superior ao da estratégia AZFinText \cite{Schumaker2009}. Estratégia bem semelhante, mas que utiliza um sistema terceirizado - OpinionFinder \cite{OpnionFinder} - para realizar a classificação positivo/negativo de sentimento.

Diversos são outros trabalhos usando técnicas mais tradicionais para análise de sentimento (\cite{Rajeev2021}, \cite{Sprenger2014}, \cite{Grob-Klubmann2015} e \cite{Sanford2019}). Apesar de maior simplicidade nas técnicas usadas, as conclusões dos trabalhos são unânimes em relatar o valor agregado em conduzir análises mais profundas na área.

Técnicas de redes neurais profundas

A ascensão de técnicas como rede neurais recorrentes (RNR) \cite{Rumelhart1986} permitiu no ambiente de processamento de linguagem natural (PLN), por meio das conexões neurais, a melhor avaliação de dados sequencias, trazendo informações passadas para serem considerar nas próximas entradas. Assim, foi possível a aplicação de análises que superassem barreiras de entendimento de contexto, o que antes era bem difícil usando técnicas mais tradicionais de aprendizado de máquina (e.g. SVM e Naive-Bayes). Apesar de promissores, só mais recentemente tais métodos foram explorados.

Neste âmbito, vale citar \cite{Sowinska2020}. Usando dados da rede social Twitter, consideram-se diferentes modelos de capturas semânticas, tais como *Bag of Words* (*BoW*), *FastText* \cite{Facebook2016} e *BERT* \cite{Delvin2018}, seguido de um modelo simples de LSTM. Importante ressaltar que foram considerados diferentes atrasos para avaliar a correlação de impacto nos retornos das ações, considerando 1, 2, 3 e 7 dias à frente. Ainda, para focar nas companhias mais comentadas, delimitou-se o universo de companhias para as 100 empresas de capital aberto com as marcas mais valiosas, considerando o *Financial Times Top 100 Global Brands 2019*. Os resultados encontrados vão ao encontro de \cite{Li2017} e \cite{Zhang2018}, com uma maximização previsibilidade em 3 dias, alcançando uma acurácia de tendência de até 61\%. Ainda, notou-se que usar múltiplas capturas semânticas, tais como *BoW* + *FastText*, podem trazer uma melhora de até 10\% na acurácia.

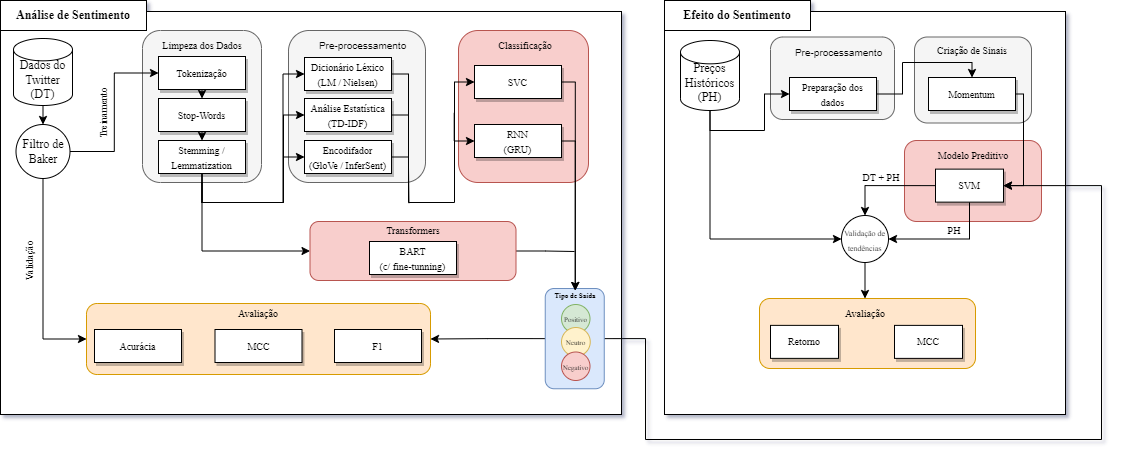
Importante ainda notar trabalhos que focam em analisar o ganho provindo ao usar técnicas de *deep learning*, tais como \cite{Kraus2017}. Neste, realiza-se uma comparação extensa, com o *BoW* sendo usado em conjunto com técnicas como Naive-Bayes, *Ridge Regression*, SVM, *Random Forest* e outras em contraposição às técnicas mais avançadas. Estas seriam: RNN, LSTM com/sem *embeddings* por meio de *GloVe* \cite{Pennington2014} e todas essas três também sendo testadas com aplicação de *transfer learning*. Mesmo após um teste de sensibilidade, nenhuma das técnicas mais tradicionais conseguiram superar o desempenho alcançado pelas de *deep learning* tanto na previsibilidade de tendência, bem como no tamanho dessa tendência. Feuerriegel ainda realizou outra análise, usando em conjunto com os resultados de polaridade, dados de preços históricos e momentuns relacionados \cite{Feurriegel2016}. Apesar de aplicar estratégias mais simples, alcançou uma redução na volatilidade, mas com uma redução no retorno total.

Em outra análise extensa de polaridade, foi possível ampliar ainda mais as técnicas estudadas e suas combinações, focando em predição de sentimento \cite{Mishev2020}. Do lado de representação textual, foram considerados: técnicas de análise léxica (*Count Vectorizer* (CV) ou TD-IDF, usando HIV4 ou LM); *encoders*, para palavras (e.g. *Word2Vec* \cite{Mikolov2013}, *GloVe*) e para sentenças. Já para algoritmo de classificação, para os mais tradicionais de aprendizado de máquina, usou-se SVM e Extreme Gradient Boost (XGB) \cite{Friedman2001}, usando de *GridSearch* para escolha de hiper parâmetros. Ainda foram estudadas redes neurais profundas (DNN), usando RNN e Redes Neurais Convolucionais (CNN) \cite{Kim2014}, baseando-se em mecanismos como *Attention* \cite{Bahdanau2014}, bi direção de redes e taxa de aprendizado adaptável (ATOM) \cite{Kingma2014} para melhoria de desempenho. Por fim, vale salientar a consideração também de *transformers* (e.g. BERT, XLM \cite{Lample2019}). Após o estudo desses diversos casos, notou-se um ganho expressivo de acurácia nos algoritmos mais recentes, mas chegando em uma limitação de melhoria de desempenho. Ainda assim, essa profunda análise reforça que técnicas como DNN são mais eficientes para classificação de sentimentos. Em especial, é importante ressaltar os resultados provindos do método *transformers*, alcançando acurácia superior a 90\% para quase todos os casos.

Além do emprego e análise de tais ferramentas, \cite{Ding2014} explora a aplicação de eventos estruturados em dados de notícias. Com isso, possui o intuito de ter uma melhor compreensão do contexto ao qual a notícia está inserida. Para tal, primeiramente é necessário retirar os eventos estruturados dos dados, usando técnicas de *Open Information Extraction* \cite{Banko2007} - o que não requer a necessidade de eventos predefinidos. A seguir, aplica-se um processo de generalização, diminuindo o número de eventos. Dessa forma, usa-se WordNet \cite{Miller1995} para *stemming* e *VerbNet* \cite{Schuler2005} para generalizar cada verbo em um nome de uma classe - por exemplo, *add* viraria *multiply\\_class*}. Com isso, realiza uma comparação entre eventos estruturados com *BoW*. Ainda, para etapa de previsão, compara um modelo linear (SVM) contra um não linear. Dos 4 processos possíveis, a acurácia do algoritmo não linear apresentou melhores resultados do que o linear. Analogamente para a análise de eventos contra *BoW*, mesmo ao considerar diferentes atrasos (1 dia, 1 semana e 1 mês). Todos esses resultados reforçam a importância de considerar o contexto para predição de tendência.

Em suma, nota-se que diversos são os trabalhos, produzindo resultados instigantes a uma maior exploração na área. Ainda, técnicas mais avançadas, como *embeddings*, *transformers* e *transfer learning*, trazem um expressivo ganho marginal na acurácia de análise de sentimento. Assim, ampliar a exploração de tais técnicas aparenta ser bem promissor.

1. Fundamentos e Metodologia Proposta



1. Descrição de Fontes de Dados e Justificativa

Conforme, serão utilizados dois bancos de dados. Um deles, será aplicado para o processo de análise de sentimento, enquanto o outro para o processo de predição por série histórica.

Análise de Sentimento

Será usada a base de dados *Stock Market Tweets Data* fornecida pela IEEEDataPort \cite{Taborda2021}. Tal base foi construída por meio do Twitter REST API, usando a biblioteca da linguagem Python “Tweepy (versão 3.8.0)”. Tal API retorna somente os dados dos últimos 7 dias, permitindo filtrar por língua e por *tags* existentes no conteúdo do tweet. Para o banco de dados, foi guardado somente o corpo do tweet e a data de criação.

Dessa forma, realizou-se a coleta de dados entre 9 de Abril de 2020 a até 9 de Julho de 2020, utilizando as seguintes *tags*: #SPX500, #SP500, SPX500, SP500, \$SPX, \#stocks, \$MSFT, \$AAPL, \$AMZN, \$FB, \$BBRK.B, \$GOOG, \$JNJ, \$JPM, \$V, \$PG, \$MA, \$INTC, \$UNH, \$BAC, \$T, \$HD, \$XOM, \$DIS, \$VZ, \$KO, \$MRK, \$CMCSA, \$CVX, \$PEP, \$PFE.

Vale salientar que, seguindo o trabalho de \cite{Baker2006}, entende-se que algumas companhias possuem maior suscetibilidade que seus preços sejam sujeitos a efeitos de sentimento de mercado. Dessa forma, seguindo os 6 indicadores abordados no artigo, será filtrado, tirando os *tickers* relacionado ao index S&P500 - estes serão inseridos, as 2 melhores companhias melhores se encaixem nesse perfil.

Histórico de Preços

Para obter os preços históricos, será usado o \textit{software} Bloomberg \cite{BBG}, retirando dados para o mesmo período - 9 de Abril a até 9 de Julho de 2009. Assim, para as duas companhias selecionados e o ETF SPY (simula a movimentação do S&P500), será obtido suas respectivas cotações diárias.

1. Resultado Preliminar
2. Conclusões