INSTITUTO TECNOLÓGICO DE AERONÁUTICA

**Igor Amâncio Machado Dias**

# PROJETO E AVALIAÇÃO DE MÉTODOS DE ANÁLISE DE SENTIMENTO PARA APOIO À DECISÃO DE COMPRA E VENDA DE ATIVOS

Trabalho de Graduação
2022

## Curso de Engenharia de Computação

CDU ???.??

**Igor Amâncio Machado Dias**

# PROJETO E AVALIAÇÃO DE MÉTODOS DE ANÁLISE DE SENTIMENTO PARA APOIO À DECISÃO DE COMPRA E VENDA DE ATIVOS

Orientador

Prof. Dr. Carlos Henrique Quartucci Forster (ITA)

**ENGENHARIA DE COMPUTAÇÃO**

São José dos Campos
Instituto Tecnológico de Aeronáutica

2022

**Dados Internacionais de Catalogação-na-Publicação (CIP)**
**Divisão de Informação e Documentação**

Amâncio Machado Dias, Igor
Projeto e avaliação de métodos de análise de sentimento para apoio à decisão de compra e venda de ativos / Igor Amâncio Machado Dias.
São José dos Campos, 2022.
32f.

Trabalho de Graduação – Curso de Engenharia de Computação– Instituto Tecnológico de Aeronáutica, 2022. Orientador: Prof. Dr. Carlos Henrique Quartucci Forster.

1. Cupim. 2. Dilema. 3. Construção. I. Instituto Tecnológico de Aeronáutica. II. Título.

## REFERÊNCIA BIBLIOGRÁFICA

AMÂNCIO MACHADO DIAS, Igor. **Projeto e avaliação de métodos de análise de sentimento para apoio à decisão de compra e venda de ativos**. 2022. 32f. Trabalho de Conclusão de Curso (Graduação) – Instituto Tecnológico de Aeronáutica, São José dos Campos.

## CESSÃO DE DIREITOS

NOME DO AUTOR: Igor Amâncio Machado Dias
TITULO DO TRABALHO: Projeto e avaliação de métodos de análise de sentimento para apoio à decisão de compra e venda de ativos.
TIPO DO TRABALHO/ANO: Trabalho de Conclusão de Curso (Graduação) / 2022



---

Igor Amâncio Machado Dias
Rua do H8A, Ap. 131
12.228-460 – São José dos Campos–SP

# PROJETO E AVALIAÇÃO DE MÉTODOS DE ANÁLISE DE SENTIMENTO PARA APOIO À DECISÃO DE COMPRA E VENDA DE ATIVOS

Essa publicação foi aceita como Relatório Final de Trabalho de Graduação

---

Igor Amâncio Machado Dias

Autor

---

Carlos Henrique Quartucci Forster (ITA)

Orientador

---

Prof. Dr. Marcos Ricardo Omena de Albuquerque Maximo
Coordenador do Curso de Engenharia de Computação

São José dos Campos, 23 de junho de 2022.

# Resumo

A partir de técnicas de processamento de linguagem natural, busca-se avaliar o efeito do sentimento de parte dos agentes do mercado na predição de tendência de preços. Neste sentido, primeiramente, é necessário construir e avaliar modelos de análise de sentimento. Para tal, serão considerados diferentes técnicas (tradicionais ou modernas) para a construção do modelo, tais como: dicionários léxicos, encodificação de palavras, SVC, redes neurais profundas e transformers. Este processo usará de um banco de dados contendo publicações da rede social Twitter comentando sobre algumas ações específicas, com parte dos tweets analisados manualmente para o processo de validação. Após a avaliação do processo de análise de sentimento, tais saídas serão usadas em conjunto com dados de série temporal para realizar a previsão de tendência de preço de mercado de forma diária. Com isso, será avaliado se a inserção de dados de sentimento podem contribuir para a previsão de mercado.

# Abstract

Based on natural language processing techniques, we seek to evaluate the effect of sentiment on the part of market agents in the prediction of price trends. Therefore, initially, it's necessary to build and evaluate models of sentiment analysis. To this end, it's going to be considered different techniques (traditional or modern) for the pipeline construction, such as: lexical dictionaries, word embedding, SVC, deep neural networks and transformers. Also, a database will be used, containing posts about specific stocks from the social network Twitter, with part of the tweets analyzed manually for the validation process. After the evaluation of the sentiment analysis, such outputs will be used in conjunction with time series data to perform daily trend forecasting. With that, it will be evaluated if the insertion of data of sentiment can contribute to market forecasting.

# Sumário

# 1 Introdução

## 1.1 Objetivo

Tem-se o objetivo de estudar a correlação da precificação dos agentes de mercado nos ativos negociados na bolsa de valores de Nova York - NYSE com seu grau de sentimento nas redes sociais.

Para tal, será desenvolvido um modelo de análise de sentimento, usando técnicas avançadas de *Word Embeddings*, *Deep Learning* e *Transformers*. Com isso, será avaliada a correlação do resultado de tal modelo com os movimentos de preço de mercado. De forma sumarizada, tem-se a seguir as metas a serem alcançadas:

- Construção e avaliação de modelos de análise de sentimento, usando métodos atualizados de representação textual e modelos classificatórios;
- Analisar o efeito que o resultado de tais modelos podem contribuir em ganho marginal na predição de tendência de preços;

## 1.2 Motivação

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 (THORP; KASSOUF, 1967), sendo assumido como um dos primeiros especialistas na área quantitativa (PATTERSON, 2011).

Ao analisar movimentos nos preços de mercado, é de suma importância o entendimento

da hipótese da eficiência de mercado (HEM) (FAMA, 1965). 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 (BONDT; THALER, 1985) e (BANZ, 1981) 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 (FAMA; FRENCH, 1996). 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 (FAMA; FRENCH, 1993). 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 (FAMA; FRENCH, 2015), 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 inviduais em situações de alta insegurança, como é relatado em muitos estudos de Daniel Kanehman, Amos Tversky e Richard Thaler ((KAHNEMAN, 2013), (KAHNEMAN; SUNSTEIN, 2021) e (THALER, 2016)). 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.

## 1.3 Trabalhos relacionados

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 macroneconômicos (KALAMARA *et al.*, 2022) - PIB, inflação e desemprego. Outros, usam dados de micro-blogs, tais como Twitter, para realizar previsão de preço de Bitcoin (KARALEVICIUS *et al.*, 2018) ou predição de resultados eleitorais (YANG *et al.*, 2018). 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 resultado de tais trabalhos.

### 1.3.1 Técnicas tradicionais de aprendizado de máquina

Dentre os diversos trabalhos na área, tem-se (KHEDR *et al.*, 2017). 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 (SHYNKEVICH *et al.*, 2015) e (BING *et al.*, 2014).

Um modelo um pouco mais complexo é apresentado com o *Media-Aware Quantitative Trader (MAQT)* (LI *et al.*, 2014). Neste caso, é usado como *input* de modelos o sentimento de mini-blogs, 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) (LOUGHRAN; MCDONALD, 2011)), com o intuito de conduzir uma abordagem de sentimento mais assertiva que um dicionário para um ambiente mais geral (e.g. Harvard IV-4 (STONE *et al.*, 1966)). 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 (LEBARON *et al.*, 1999), e apresentando um resultado superior ao da estratégia AZFinText (SCHUMAKER; CHEN, 2009). Estratégia bem semelhante, mas que utiliza um sistema terceirizado - OpinionFinder (MPQA, 2005) - para realizar a classificação positivo/negativo de sentimento.

Diversos são outros trabalhos usando técnicas mais tradicionais para análise de sentimento ((RAJEEV PADMANAYANA, 2021), (SPRENGER *et al.*, 2014), (GROB-KLUBMANN *et al.*, 2015) e (SANFORD, 2019)). 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.

### 1.3.2 Técnicas de redes neurais profundas

A ascensão de técnicas como rede neurais recorrentes (RNR) (RUMELHART *et al.*, 1986) 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 aprenzido de máquina (e.g. SVM e Naive Bayes). Apesar de promissores, só mais recentemente tais métodos foram explorados.

Neste âmbito, vale citar (SOWINSKA; MADHYASTHA, 2020). Usando dados da rede social *Twitter*, consideram-se diferentes modelos de capturas semânticas, tais como *Bag of Words* (BoW), *FastText* (FACEBOOK, 2016) e *BERT* (DEVLIN *et al.*, 2018), seguido de um modelo simples de LSTM. Importante ressaltar que foram considerados diferentes atrasos para avaliar a correlação de impacto nos retorno 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 (LI *et al.*, 2017) e (ZHANG *et al.*, 2018), 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 (KRAUS; FEUERRIEGEL, 2017). 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 (PENNINGTON *et al.*, 2014) 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 (FEUERRIEGEL; PRENDINGER, 2016). 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 (MISHEV *et al.*, 2020). 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 (MIKOLOV *et al.*, 2013), 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) (FRIEDMAN, 2001), usando de *GridSearch* para escolha de hiperparâmetros. Ainda foram estudadas redes neurais profundas (DNN), usando RNN e Redes Neurais Convolucionais (CNN) (KIM, 2014), baseando-se em mecanismos como *Attention* (BAHDANAU *et al.*, 2015), bidireção de redes e taxa de aprendizado adaptável (ATOM) (KINGMA; BA, 2014) para melhora de desempenho. Por fim, vale salientar a consideração também de *transformers* (e.g. BERT, XLM (CONNEAU; LAMPLE, 2019)). Após o estudo desses diversos casos, notou-se um ganho expressivo de acurácia nos algoritmos mais recentes, mas chegando em uma limitação de melhora 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, (DING *et al.*, 2014) 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* (BANKO *et al.*, 2007) - 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 (MILLER, 1995) para *stemming* e VerbNet (SCHULER, 2005) 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 preditação 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.

## 1.4 Organização do trabalho

Além do objetivo, motivação e revisão bibliográfico apresentados no capítulo 1 em questão, tem-se, no capítulo 2, o desenvolvimento do modelo de análise de sentimento, passando pelos pormenores de todas as etapas do *pipeline* escolhido. Já no capítulo 3, serão apresentados e discutidos os resultados encontrados com o modelo construído e aplicado. Por fim, no capítulo 4, tem-se a conclusão, sumarizando o trabalho, sua aplicabilidade na prática, possíveis melhorias e nortes a serem explorados.

# 2 Processamento de Linguagem Natural

## 2.1 Introdução

Processamento de Linguagem Natural ou PLN é um ramo da Inteligência Artificial que usa aprendizado de máquina para realizar processamento de texto e dados. Esta área possui diversas aplicações, tais como: tradução, sumarização, resposta a perguntas, modelagem de tópico, etc. Neste trabalho, conforme já comentado, será focado na finalidade de análise de sentimento de texto.

Em geral, na contrução de um PLN de análise de sentimento, segue-se o *pipeline* apresentado na figura 2.1. Cada etapa será melhor explorada nas seções subsequentes, focando nas técnicas que de fato serão usadas neste trabalho.

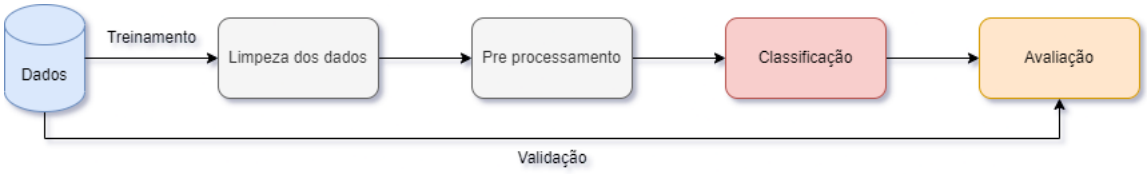


FIGURA 2.1 – *Pipeline* geral para Processamento de Linguagem Natural

## 2.2 Limpeza de dados

Nesta etapa, são recebidos os dados do banco de dados de treinamento. Para que tais dados fiquem permissíveis de se submeterem ao pre-processamento, é necessário realizar alguns ajustes, tornando-os mais genéricos por meio da retirada partes do texto que não grande ganho informacional. Dessa forma, serão realizadas as etapas apresentadas e melhores explanadas a seguir.

### 2.2.1 Tokenização

O processo de tokenização consiste em separar partes de texto em unidades menores conhecidas como *tokens*. Estes podem ser: palavras, caracteres ou subpalavras - conjunto de n caractere (n-gramas).

Para tokenização por palavras, realiza-se a separação de partes do texto baseado em um delimitador (espaço, por exemplo). Esse processo pode se deparar, na etapa de teste, com uma palavra fora do vernáculo conhecido do treinameto. Nesses casos, essas palavras Out-of-Vocabulary (OOV) (ou Fora do Vocabulário) - como definidas - podem ser marcadas como UNK (significando desconhecidas). Apesar de resolver, toda a informação da palavra é perdida e todas as OOV são tratadas de forma igual.

Em outro caso, os *tokens* serão os próprios caracteres. Assim, tem-se um limite do tamanho do vocabulário - número de letras do alfabeta da língua usada. Todavia, torna-se mais complicado encontrar as relações entre os *tokens* que espelham os significados de cada palavra.

Por último, tem-se o caso de tokenização em subpalavras. Neste caso, o vocabulário é construído por *tokens* formados por parte de palavras. Em 2.1, tem-se uma exemplificação de cada caso.

TABELA 2.1 – Exemplificação dos casos de Tokenização

| Tipo de Separação | Entrada | Tokenização |
|---|---|---|
| Palavras | I am the strongest | {I} {am} {the} {strongest} |
| Caracteres | strongest | {s} {t} {r} {o} {n} {g} {e} {s} {t} |
| Subpalavras | strongest | {strong} |

### 2.2.2 Remoção de *Stop Words*

*Stop Words* são, em geral, as palavras mais usadas em uma língua, tais como artigos, preposições, pronomes, etc. Como sua função é de trazer coesão a um texto, em geral, não carregam uma informação importante para análise. Com a sua retirada, o modelo consegue ter mais foco nas palavras que carregam maior valor informacional.

### 2.2.3 *Stemming/Lemmatization*

Em *Stemming*, realiza-se a reduação do *token* em sua forma raiz. Com isso, realiza-se a remoção do sufixo para se gerar somente o *stem* da palavram. Todavia, importante considerar os casos de falso positivo (*overstemming*): classificar “universal”, “universo”e

"universidade" para o mesmo *stem* "univ". Ainda, os casos de falso negativo (*understemming*): classificar diferentemente para "alumnus", "alumnae"e "alumni".

Neste sentido, *Lemmatization* vem como um desenvolvimento de *Stemming*. Apesar de serem similares, aquele traz o contexto em consideração, agrupando palavras com significado semelhantes. Assim, por exemplo, "melhor" pode virar "bom".

Em geral, acaba-se por usar os dois processos nessa etapa. Apesar de *Lemmatization* ser preferível antes de um método que foca na simples retirada de um sufixo, ele acaba por demorar bem mais. Isso pode ser importante ao considerar documentos textuais bem longos.

## 2.3 Pre-processamento

Os dados textuais serão transformados em dados numéricos para que possam ser usados nos modelos classificatórios. Para tal, pode-se usar diferentes métodos, tais como dicionários léxicos (DL), análise estatística (AE), encodificação de palavra (EP) e de sentença (ES).

### 2.3.1 Dicionários Léxicos

São dicionários construídos para auxiliar no processo de análise de sentimento ao avaliar manualmente seu vernáculo. Este é constituído de palavras presentes na língua considerada do DL e, em geral, possuem um foco específico para determinada área, como negócios ou checadores de escritas (BRAASCH, 2013). Assim, para cada palavra do vocabulário, é dado uma métrica (em geral +1 para um sentimetno positivo ou -1 caso contrário).

Exemplos de DL são o dicionário específico de finanças LM (LOUGHRAN; MCDONALD, 2011) e o genérico Nielsen (NIELSEN, 2011). Neste tipo de análise, o que vale é a relação semântica entre os termos, baseada na avaliação provinda de cada palavra para cada dicionário. Assim, tem-se os seguintes dicionários de palavras:

- **Loughran and McDonald**: Atualizado anualmente, realiza uma avaliação binária entre as palavras de seu vernáculo, classificando-as como positiva (pontuação +1) ou negativa (pontuação -1). No dicionário, existem outras classificações, mas essas duas são as mais relacionados com sentimento. Sua construção é feita ao escolher palavras conhecidas do dicionário inglês e arquivos do EDGAR. Este, um acrônimo para *Eletronic Data Gathering, Analysis, and Retrieval* (ou Agregador, Analisador e Recuperar de Dados Eletrônicos), é uma ferramenta da *Securities and Exchange*

*Commission* (SEC) (ou Comissão de Ativos e Câmbio) dos Estados Unidos que permite angariar todos os dados e relatórios submetidos por companhias que são o obrigadas por lei a fazerem. Com isso, consegue realizar uma avaliação mais restrita ao ambiente de finanças, tal como dar a palavra "*bankrupt*"uma pontuação -1.

- **Nielsen**: Neste caso, ao invés de usar uma dicotomia entre +1/-1, baseia-se em um espectro de -3 a até +3. Com isso, para palavras com o mesmo sentimento, podem pontuar de forma diferente para demonstrar sua intensidade. Um exemplo seria que *excellent* pontua mais positivamente que *good*. Todavia, urge-se considerar que não é um dicionário específico para finanças. Além disso, em um ambiente de micro-blogs como Twitter, as pessoas podem preferir por usar palavras menores por ser mais rápido e estar dentro do limite de 140 caracteres, mesmo querendo passar uma maior intensidade.

### 2.3.2 Análise Estatísitca

Dentro do espectro de uma abordagem mais estatística, as duas técnicas mais comuns são *Bag-of-Words* e *TF-IDF*. Este é o que será usado no modelo e melhor explicado a seguir.

*Term frequency - inverse document frequency* (TF-IDF) (ou Frequência de termo - frequência inversa no documento) é um algoritmo que busca avaliar a relevância de uma palavra em um documento. Em *Bag-of-Words*, cada palavra do vocubulário se torna um atributo e para cada sequência analisada, as palavras que estão presentes recebem um contador - quantificando todas as vezes que aparece na sequência. Fugindo dessa abordagem de uma simples contagem, TF-IDF considera os pesos de cada atributo a partir da multiplicação de duas métricas: frequência do termo (TF) - o que calcula o número de ocorrência do termo na sequência em questão (equação 2.2) - e frequência inversa no documento (IDF) - penaliza a pontuação do atributo se aparece mais na sequência do que em comparação com todas as sequências do curpos de treinamento (equação 2.3). Nas seguintes equações, $t$ é o termo, $s$ a sequência considerada e $S$ o conjunto de todas as sequências do corpus (e.g. todas as manchetes de notícias, todos os twitters, etc.).

$$tfidf(t, s, S) = tf(t, s) * idf(t, S) \tag{2.1}$$

$$tf(t, s) = log(1 + freq(t, s)) \tag{2.2}$$

$$idf(t, S) = log(\frac{N}{count(s \in S : t \in s)} \tag{2.3}$$

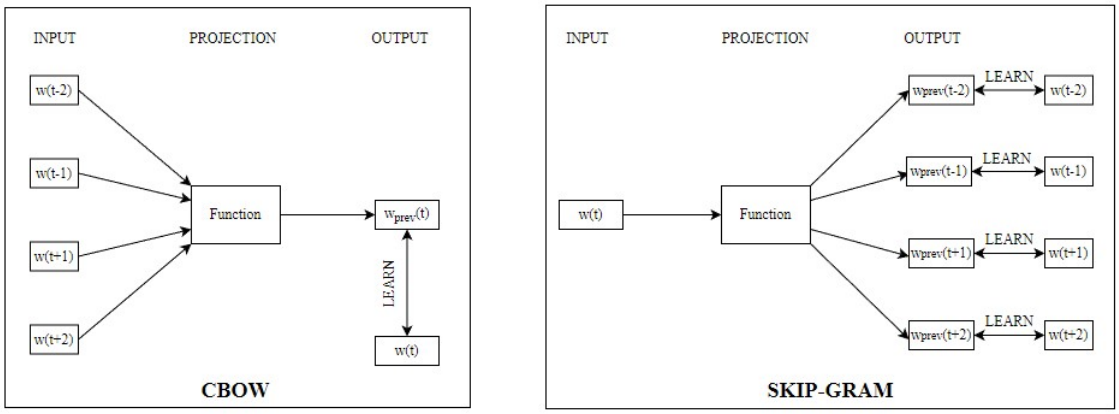


FIGURA 2.2 – As duas possíveis arquiteturas para Word2Vec

### 2.3.3 Codificadores

Com os métodos apresentados até então, a semântica do contexto não é considerada nos modelos. Com isso, palavras que possuem até significados parecidos, podem ser tratadas de forma bem distintas por não ter sido considerado o contexto das sentenças aos quais cada palavra estava inserida.

Para superar esse desafio, foram criados modelos baseados na conversão de palavras em vetores de alta dimensão - tais modelos ficaram conhecidos como *word embeddings* (ou codificadores de palavras) - EP. Neste caso, baseado no principio de hipótese distribuída (HARRIS, 1954), a semântica é considerada na formação de tais vetores. Assim, palavras com ideias semelhantes são assimilados vetores "próximos"no espaço n-dimensional.

Neste âmbito, o encodificador que deu o passo inicial foi o Word2Vec. Um time de pesquisa do Google criou duas possíveis arquiteturas de modelos: *Continous Bag-of-Word* (CBOW) e *Continuous Skip-Gram*. Em CBOW, como mostrado na figura 2.2, o foco está em conduzir um modelo de aprendizagem não supervisionado usando o contexto para prever a palavra, enquanto no Skip-Gram, a representação distribuída da palavra é usada para prever o contexto. Para melhor entendimento de CBOW e Skip-Gram, olhar (MIKOLOV *et al.*, 2013).

Neste trabalho, será dado o foco em GloVe (PENNINGTON *et al.*, 2014), metodologia proposta em 2014 por pesquisadores da Universidade de Stanford baseada nos trabalhos de Word2Vec. Nesta nova metodologia, usa-se dois métodos em conjunto de representação distribuída de palavras. O primeiro deles é a matriz global de fatorização, responsável por capturar as estatísticas gerais e relações entre as palavras. Já o segundo, o Skip-gram, conseguindo capturar melhor o contexto local na sentença à qual a palavra está inserida.

Ainda, será aplicada outra metodologia de encodificação, mas focado em sentenças - ES. InferSent (CONNEAU *et al.*, 2017) foi desenvolvido pelo grupo de pesquisa de inteli-

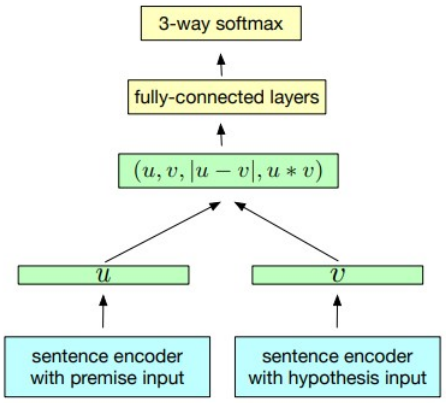


FIGURA 2.3 – Arquitetura genérica de treinamento por ILN

gência artificial do Facebook ao qual é baseado na inferência de linguagem natural (ILN). Em ILN, tem-se a tarefa de determinar se uma hipótese é verdadeira, falsa ou neutra dado uma premissa. Neste processo, apresentado na figura 2.3, tem-se um aprendizado supervisionado o qual tenta entender atributos mais gerais para cada setença ao compará-la com a hipótese. Para tal, usa-se do banco de dados Inferência de Linguagem Natural de Stanford (ILNS), o qual possui 507 mil pares de sentenças inglesas, com suas relações manualmente avaliadas.

Note que a primeira etapa é a codificação, o qual neste trabalho será por meio do GloVe. A seguir, realiza-se mensurações entre os dois vetores $u$ e $v$. Com isso, tem-se métricas como: concatenação ($(u, v)$), produto por elemento ($u*v$) e módulo por elemento ($|u - v|$). Tais métricas serão entradas no encodificador. A arquitetura escolhida para produzir a encodificação da sentença é o BiLSTM com votação máxima, o qual apresenta empiricamente o melhor resultado.

## 2.4 Modelos Classificatórios

São diversos os algoritmos possíveis de serem usados nesta etapa. Em geral, pode-se dividir entre os modelos de aprendizado mais tradicionais e as redes neurais profundas. Além desses, tem-se o caso dos *transformers*.

### 2.4.1 Aprendizado de Máquina

No espectro dos mais tradicionais, será dado o enfoque ao Support Vector Classifier (SVC) - (ou Classificador dos Vetores de Suporte). SVC é uma generalização da técnica da Maximal Margin Classifier (MMC) (ou Classificador por Margem Máxima). Em MMC, busca-se criar um hiperplano que consiga separar as observações. Em SVC, esse grau

de separação ganha maior flexibilidade, permitindo que haja um "invasão"da margem de separação. Tal processo se consegue por meio da inserção de um novo parâmetro ($C$ na equação 2.4) que controlorá o grau de permissividade na margem de separação.

$$\begin{aligned} \underset{\beta_0,\beta_1,...,\beta_p,\epsilon_0,\epsilon_1,...,\epsilon_n,M}{max} & M \\ sujeito & \sum_{j=1}^{p} \beta_j^2 = 1 \\ & y_i(\beta_0 + \beta_1 x_{i1} + ... + \beta_1 x_{ip}) \geq M(1 - \epsilon_i) \\ & \epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C \end{aligned} \quad , \tag{2.4}$$

### 2.4.2 Redes Neurais Profundas

Aqui impõe-se considerar dois tipos de arquiteturas: redes neurais recorrentes (RNR) (RUMELHART *et al.*, 1986) - bom desempenho para modelagem sequencial e captura de dependências distantes - e redes neurais convolucionais (RNC) - mais eficiente na captura de correlação espacial ou temporal. Para maior profundidade em RNC, busque (KIM, 2014).

Será dado o enfoque na RNR, a qual permite que a conexão entre neurônios formem ciclos. Estes se baseiam no que a rede conseguiria passar de informação na passagem de uma entrada $x_i$ para outra $x_{i+1}$. Com isso, considerando uma modelagem não sequencial, tem-se a arquitetura na figura 2.4.

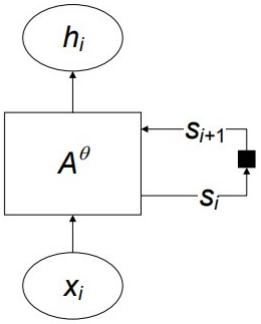


FIGURA 2.4 – Esquema da estrutura de uma RNR

Neste caso, na $i$-ésima interação: tem-se a entrada $x_i$, o estado escondido $s_i$, a saída $h_i$ e a rede *feedforward* $A^\theta$ - parametrizada com $\theta$. Ao passar da interação $i$ para $i+1$, é realizada o cálculo da saída $h_{i+1}$ considerando:

$$h_{i+1} = A^\theta(s_i, x_{i+1}) \tag{2.5}$$

Numa abordagem sequencial - mais perto do que é usado na prática - é apresentado na figura 2.5, com a passagem do estado escondido $s_i$ para a próximo $A^\theta$ da sequência.

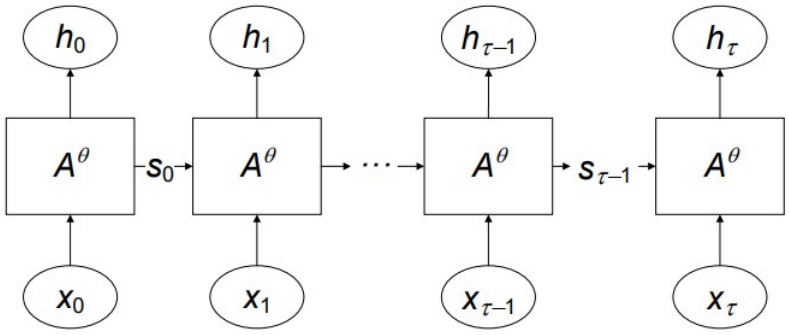


FIGURA 2.5 – Esquema estrutural de RNR sequencial com as entradas $x_0, x_1, ..., x_\tau$, estados $s_0, s_1, ..., s_{\tau-1}$, saídas $h_0, h_1, ..., h_\tau$ e redes neural com *feedforward* $A^\theta$

Todavia, na prática, dois problemas são os responsáveis por atrapalhar uma melhor desempenho do RNR (GOODFELLOW I.; BENGIO, 2017). Um deles seria o que ocorre com os cálculo de gradientes, apresentando grande instabilidades ao convergir para valores muito grandes ou muito pequenos. Ainda, a informação só persiste por algumas interações na memória - não é duradoura (BENGIO *et al.*, 1994).

Dessa forma, a arquitetura Gate Recurrent Unit (GRU) (ou Unidade do Portão de Recorrência) (CHUNG JUNYOUNG; GULCEHRE, 2014) surge para superar tais problemas. A partir de mecanismos internos - portões - consegue regular o fluxo de informações. Essa regulação é melhorada a partir do momento que os portões aprendem quais informações são importantes para se manter ou jogar fora. Assim, consegue-se manter e transmitir informações relevantes na rede.

No GRU, existe duas portas que são basicamente vetores contendo valores entre 0 - aquela posição não tem importância - e 1 - máximo de importância. Tais vetores mulitplicam os dados de entrada e/ou estado oculto e conseguem escolher quais são elementos importantes ou não, gerando uma saída e um novo estado oculto. Sua arquitetura unitária poderá ser vista na figura 2.6.

Aprofundando nas portas, o *update gate* (ou porta de atualização), tem-se o foco de determinar a quantidade de informação anterior que será passada para o próximo estado. Para tal, realiza-se o seguinte cálulo de $z_t$ na equação 2.6, produzindo um valor entre 0 e 1 provindo de uma função sigmoid $\sigma$ ($b_z$ representa o viés da conexão).

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z) \tag{2.6}$$

Já no *reset gate* (ou porta de esquecimento), busca-se saber quanto do passado que

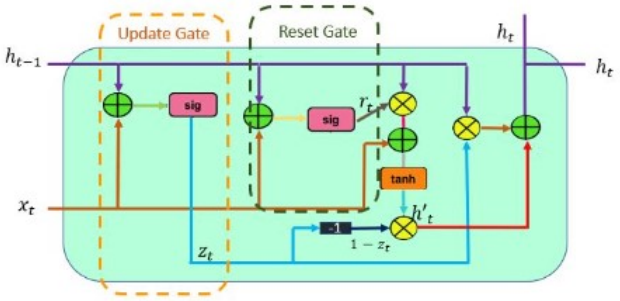


FIGURA 2.6 – Arquitetura unitária do GRU. O estado escondido $h_{t-1}$ e a entrada $x_t$ passam pelos cálculos dos portões, produzindo o próximo estado $h_t$. O símbolo $\oplus$ determina uma concatenação entre vetores e $\otimes$ uma multiplicação de vetores a nível de elemento. Ainda, tem-se o uso das funções tangente hiperbólica (*tanh*) e sigmoid (*sig*).

pode ser esquecido. Para tal, apresenta um método bem parecido que o portão anterior como mostrado na equação 2.7, como mostrado a seguir, mudando o peso para $W_r$ e o viés para $b_r$.

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z) \tag{2.7}$$

Ainda, como forma de melhorar o desempenho do GRU, tem-se o uso da técnica bidirecional. Dessa forma, são coletados dados dos atributos nas duas direções. Enquanto em um sentido coletado dados do começo ($x_1$) ao fim ($x_n$) - $\overrightarrow{h}$, tem-se também a coleta do sentido inverso $\overleftarrow{h}$. Com isso, numa camada interna, recebe-se a concatenação dos dois sentidos: $h_i = \overrightarrow{h}_i \oplus \overleftarrow{h}_i$.

Outro mecanismo é o Attention (BAHDANAU *et al.*, 2015). Por meio dele permite que a GRU foque em determinadas partes da sequência de entradas, podendo ressaltar aquelas que trazem maior importância para avaliação do contexto da sequência de palavras. Considerando o GRU, com os dois mecanismo citados - bidiração e attention, tem-se sua arquitetura na figura 2.7.

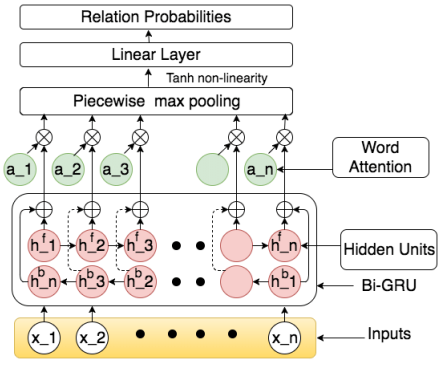


FIGURA 2.7 – Arquitetura para GRU biderecional com o attention

### 2.4.3 Transformers

Por fim, vale levar em consideração os *trasformers*. Este método usa dois modelos para realizar tratar uma sequência de entrada para uma sequência de saída: um encodificador e um decodificador. Introduzido por meio do artigo "Attention is all you need"(VASWANI *et al.*, 2017), tem-se a construção de uma arquitetura focada no mecanismo Attention. Com isso, sua estrutura é apresentada na figura 2.8.

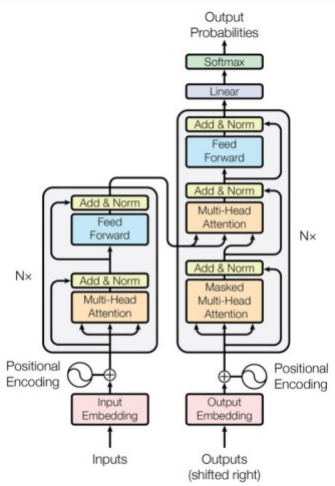


FIGURA 2.8 – Arquitetura Geral de um Transformers. Na parte esquerda tem-se o encodificador, podendo ser composto por uma pilha de N arquiteturas apresentadas em Nx. Já do lado direito, tem-se o decodificador, sendo o processo de milhas ser igualmente aplicado em seu respectivo Nx.

Nesta arquitetura demonstrada, é importante considerar que a posição da palavra de entrada em relação a sua sequência faz parte do processo de representação textual - demonstrado como *Positional Encoding* (ou Encodificação Posicional). Isso ocorre pelo fato de se não usar RNR e, com isso, não ter a comodidade de saber como a sequência está alimentado o modelo.

Outro importante ponto é entender melhor como funciona o processo de *Multi-Head Attention* (ou Attention por Múltiplas Camadas). Para tal, considere a figura 2.9, com o processo da parte da esqueda sendo simplificado usando a equação 2.8.

$$Attention(Q, K, V) = a \cdot V \tag{2.8}$$

$$a = softmax(\frac{Q \cdot K^T}{\sqrt{d_k}}) \tag{2.9}$$

Em que $Q$ é uma matrix que contem a representação vetorial de uma palavra da sequência, $K$ todos os vetores das palavras existentes na sequência e $d_k$ é a dimensão dos vetores. Na equação 2.9, tem-se o cálculo do peso $a$, o qual representa como a influência de todas as outras palavras da sequência influenciam determinada palavra. Ao aplicar

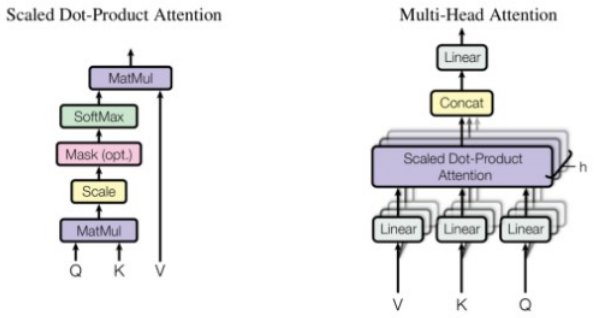


FIGURA 2.9 – À esquerda, uma camada do Attention por Múltiplas Camadas e na direita, as várias camadas em paralelo

a função SoftMax, tem-se um resultado entre 0 e 1, o qual será aplicado sobre $V$. Este possui definição similar ao $K$, mas se diferencia quando recebe sequências provindas do decodificador e encodificador. Nesse caso, a sua sequência é o universo provindo das sequência dos dois lados.

Com isso, diferentes $Q$, $K$ e $V$ são analisados na etapa de paralelismo, ponderados por um peso $W_i^x$ (equação 2.11). O tratamento nessa etapa de paralelismo é discrimida na equação 2.10, com os pesos $W_i^x$ precisando ser aprendidos. Após, seus resultados funcionam como entradas para uma camada de *feed-forward*.

$$MultiHead(Q, K, V) = [head_1, head_2, ..., head_h] \cdot W^0 \tag{2.10}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{2.11}$$

Dos vários modelos que foram criados a partir dessa ideia, neste trabalho será dado o enfoque no BART (LEWIS *et al.*, 2020). Proposto por um grupo de pesquisa do Facebook, este modelo leva bem em consideração relações bidirecionais - importante para avaliação de contexto de sequências. Além disso, por meio do seu modelo de attention cruzado, consegue criar saídas que sejam bem relacionadas com sua entrada.

# 3 Metodologia

## 3.1 Pipeline

Com isso, será seguido o seguinte pipeline:

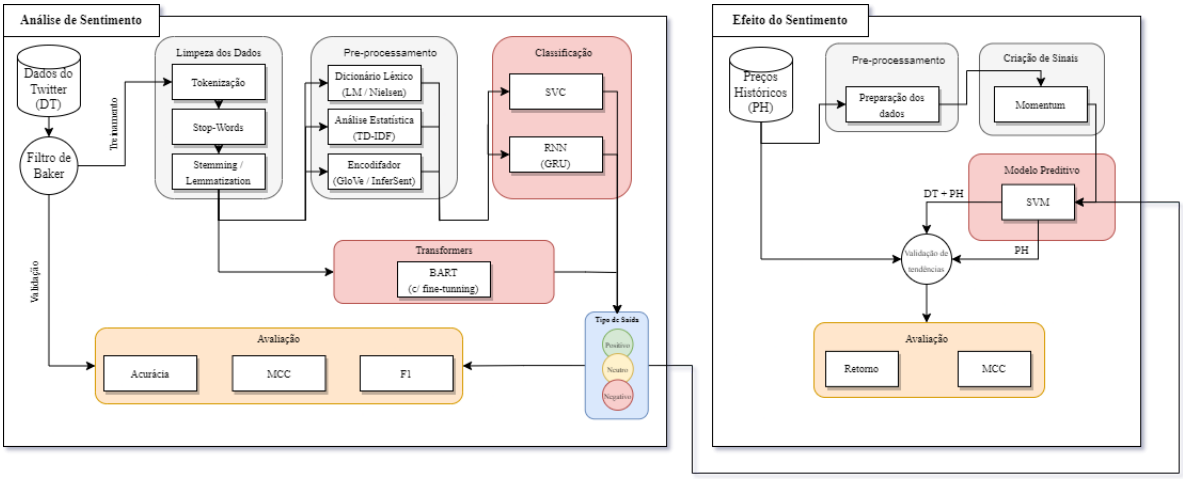


FIGURA 3.1 – Arquitetura completa para análise de sentimento e seu efeito em predição de tendências

## 3.2 Base de Dados

Conforme apresentado em 3.1, 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.

### 3.2.1 Análise de Sentimento

Será usada a base de dados *Stock Market Tweets Data* fornecida pela IEEEDataPort (TABORDA *et al.*, 2021). 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 contéudo 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 (BAKER; WURGLER, 2006), 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 SP500 - estes serão inseridos, as 2 melhores companhias melhores se encaixem nesse perfil.

### 3.2.2 Histórico de Preços

Para obter os preços históricos, será usado o *software* Bloomberg (BLOOMBERG, 2022), 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 SP500), será obtido suas respectivas cotações diárias.

## 3.3 Sinais pela Série Histórica

Além dos sinais positivos, negativos e neutros da análise de sentimento também serão usados sinais provindos da série temporal. Dessa forma, em conjunto, funcionarão como entrada no modelo de SVM para realizar previsão de tendência.

Com isso, será usada uma análise de momentum. Esta estratégia se baseia na ideia de que existe uma certa inércia nos movimentos de mercado. Com isso, ações que estão se valorizando, tendem a continuar seguindo essa tendência - analogamente para o sentido oposto. Para tal, considere $p_{i,t}$ - preço de certa ação $i$ no tempo $t$. Com isso, tem-se a construção da métrica a Razão de Mudança (RM), apresentado na equação 3.1. Tal métrica quantifica o movimento do mercado para ação analisada.

$$RM_t = \frac{p_{i,t} - p_{i,t-1}}{p_{i,t-1}} \tag{3.1}$$

## 3.4 Avaliação

Seguindo as avaliações usadas nos trabalhos citados em 1.3, serão considerados os seguintes avaliadores de resultados no processo de análise de sentimento:

- Acurácia: Nesta métrica, será avaliado o nível de certo do modelo, considerando todos as classificações corretas e dividindo por todas as classificações. Com isso, um acurácia igual a zero significa que errou todas as predições, enquanto uma igual a um quer dizer que acertou todas.

- Coeficiente de Correlação de Matthews (MCC): Métrica muito importante para casos de classificação binária, é calculada por meio da equação 3.2. Ela retorna um valor entre (-1) e (+1), em que um coeficiente de (+1) representa uma predicao perfeita, (0) representa uma predicao aleatoria media, e (-1) uma predicao inversa. Para tal, considere: VP e VN como verdadeiro positivo e negativo. Ainda, FP e FN como falso positivo e negativo.

$$MCC = \frac{VP \cdot VN - FP \cdot FN}{\sqrt{(VP + FP) \cdot (FN + PN) \cdot (FP + VN) \cdot (VP + FN)}} \tag{3.2}$$

- F1-Score: métrica que computa a média harmônica entre precisão e recall, como mostrado em 3.3. Dessa forma, essas duas métricas são computadas em conjunto. Na precisão tem-se de tudo que foi previsto como positivo, o quanto de fato foi certo. Já em recall, de tudo que era pra ser positivo, quanto de fato foi previsto corretamente.

$$F1 - Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{3.3}$$

Já, na avaliação de efeito na predição com o preço histórico, será realizado duas predições serão efetuados as seguintes análises:

- Retorno: mensuração do retorno - variação % da alocação inicial, após o término do período analisado. Para calcular, basta considear a equação 3.4.

$$R = \prod_{t=1}^{T} (1 + RM_t \cdot a_t) \tag{3.4}$$

$$com\ a = \begin{cases} 1, \text{ caso tenha acertado a previsão de tendência} \\ 0, \text{ caso contrário} \end{cases}$$

- MCC: Mesmo coeficiente apresentado anteiormente, mas agora avaliando a tendência.

# 4 Próximos passos

Para os próximos passos a serem seguidos, tem-se:

- Operacional
    1. Construção, aplicação e avaliação dos modelos, seguindo as múltiplas alternativas para as diferentes partes do pipeline de análise de sentimento;
    2. Inserção de dados históricos, em conjuntos com os resultados de polaridade, para predição de tendência de preços;
    3. Avaliação dos efeitos da inserção de dados qualitativos na previsão de tendência de preços;
- Relatório
    1. Revisão dos capítulos de Introdução, Ambientação e Metodologia;
    2. Escrita dos capítulos de Resultados, Discussões e Conclusão;

# Referências


BAHDANAU, D.; CHO, K.; BENGIO, Y. Neural machine translation by jointly learning to align and translate. In: BENGIO, Y.; LECUN, Y. (Ed.). **3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings**. [s.n.], 2015. Disponível em: <http://arxiv.org/abs/1409.0473>.

BAKER, M.; WURGLER, J. Investor sentiment and the cross section of stock returns. **Journal of Finance**, v. 61, n. 4, 2006. ISSN 1645-1680. Disponível em: <https://www.hbs.edu/ris/Publication\%20Files/sentiment_5e81d758-b344-4c0b-bf6a-3eccbf4cf1b6.pdf>.

BANKO, M.; CAFARELLA, M.; SODERLAND, S.; BROADHEAD, M.; ETZIONI, O. Open information extraction from the web. In: . [S.l.: s.n.], 2007. p. 2670–2676.

BANZ, R. W. The relationship between return and market value of common stocks. **Journal of Financial Economics**, v. 9, n. 1, p. 3–18, 1981. ISSN 0304-405X. Disponível em: <https://www.sciencedirect.com/science/article/pii/0304405X81900180>.

BENGIO, Y.; SIMARD, P.; FRASCONI, P. Learning long-term dependencies with gradient descent is difficult. **IEEE Transactions on Neural Networks**, v. 5, n. 2, p. 157–166, 1994.

BING, L.; CHAN, K. C.; OU, C. Public sentiment analysis in twitter data for prediction of a company's stock price movements. In: **2014 IEEE 11th International Conference on e-Business Engineering**. [S.l.: s.n.], 2014. p. 232–239.

BLOOMBERG. **Gráfico de histórico de preços para ações - GP**. 2022. Disponível em: <https://www.bloomberg.com/professional/solution/bloomberg-terminal/>.

BONDT, W. F. M. D.; THALER, R. Does the stock market overreact? **The Journal of Finance**, v. 40, n. 3, p. 793–805, 1985. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1985.tb05004.x>.

BRAASCH, A. 82. generic dictionaries for multiple booktitle = Supplementary Volume Dictionaries. An International Encyclopedia of Lexicography: Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography, publisher = De Gruyter Mouton, pages = 1186–1194. In: _____. [s.n.], 2013. Disponível em: <https://doi.org/10.1515/9783110238136.1186>.

CHUNG JUNYOUNG; GULCEHRE, C. C. K. e. B. Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. In: **NIPS 2014 Workshop on Deep Learning, December 2014**. [S.l.: s.n.], 2014.

CONNEAU, A.; KIELA, D.; SCHWENK, H.; BARRAULT, L.; BORDES, A. Supervised learning of universal sentence representations from natural language inference data. In: . [S.l.: s.n.], 2017. p. 670–680.

CONNEAU, A.; LAMPLE, G. Cross-lingual language model pretraining. In: _____. **Proceedings of the 33rd International Conference on Neural Information Processing Systems**. Red Hook, NY, USA: Curran Associates Inc., 2019.

DEVLIN, J.; CHANG, M.; LEE, K.; TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. **CoRR**, abs/1810.04805, 2018. Disponível em: <http://arxiv.org/abs/1810.04805>.

DING, X.; ZHANG, Y.; LIU, T.; DUAN, J. Using structured events to predict stock price movement: An empirical investigation. In: . [S.l.: s.n.], 2014. p. 1415–1425.

FACEBOOK, I. **fastText: Library for fast text representation and classification**. [S.l.], 2016. Disponível em: <https://github.com/facebookresearch/fastText>.

FAMA, E. F. The behavior of stock-market prices. **The Journal of Business**, University of Chicago Press, v. 38, n. 1, p. 34–105, 1965. ISSN 00219398, 15375374. Disponível em: <http://www.jstor.org/stable/2350752>.

FAMA, E. F.; FRENCH, K. R. Common risk factors in the returns on stocks and bonds. **Journal of Financial Economics**, v. 33, n. 1, p. 3–56, 1993. ISSN 0304-405X. Disponível em: <https://www.sciencedirect.com/science/article/pii/0304405X93900235>.

FAMA, E. F.; FRENCH, K. R. The capm is wanted, dead or alive. **The Journal of Finance**, [American Finance Association, Wiley], v. 51, n. 5, p. 1947–1958, 1996. ISSN 00221082, 15406261. Disponível em: <http://www.jstor.org/stable/2329545>.

FAMA, E. F.; FRENCH, K. R. A five-factor asset pricing model. **Journal of Financial Economics**, v. 116, n. 1, p. 1–22, 2015. ISSN 0304-405X. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0304405X14002323>.

FEUERRIEGEL, S.; PRENDINGER, H. News-based trading strategies. **Decision Support Systems**, v. 90, p. 65–74, 2016. ISSN 0167-9236. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167923616301191>.

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 29, n. 5, p. 1189 – 1232, 2001. Disponível em: <https://doi.org/10.1214/aos/1013203451>.

GOODFELLOW I.; BENGIO, Y. e. C. A. **Deep Learning**. [S.l.]: MIT Press, 2017. ISBN 9780262035613.

GROB-KLUBMANN, A.; EBNER, M.; KöNIG, S. Structure in the tweet haystack: Uncovering the link between text-based sentiment signals and financial markets. 10 2015.

HARRIS, Z. S. Distributional structure. In: _____. Word, 1954. v. 10, p. 146–162. Disponível em: <https://doi.org/10.1080/00437956.1954.11659520>.

KAHNEMAN, D. **Thinking, Fast and Slow**. 1st. ed. [S.l.]: Farrar Straus Giroux, 2013. ISBN 0374533555.

KAHNEMAN, D.; SUNSTEIN, C. **Noise: A Flaw in Human Judgment**. 1st. ed. [S.l.]: Little, Brown Spark, 2021. ISBN 0316451401.

KALAMARA, E.; TURRELL, A.; REDL, C.; KAPETANIOS, G.; KAPADIA, S. Making text count: economic forecasting using newspaper text. **Journal of Applied Econometrics**, 2022. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.2907>.

KARALEVICIUS, V.; DEGRANDE, N.; WEERDT, J. D. Using sentiment analysis to predict interday bitcoin price movements. **The Journal of Risk Finance**, v. 19, n. 1, p. 93–105, 2018. ISSN 1526-5943. Disponível em: <https://doi.org/10.1108/JRF-06-2017-0092>.

KHEDR, A. E.; SALAMA, S. E.; YASEEN, N. Predicting stock market behavior using data mining technique and news sentiment analysis. **I.J. Intelligent Systems and Applications**, v. 9, n. 7, p. 22–30, 2017. ISSN 0304-405X. Disponível em: <https://www.mecs-press.org/ijisa/ijisa-v9-n7/IJISA-V9-N7-3.pdf>.

KIM, Y. Convolutional neural networks for sentence classification. In: **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Doha, Qatar: Association for Computational Linguistics, 2014. p. 1746–1751. Disponível em: <https://aclanthology.org/D14-1181>.

KINGMA, D.; BA, J. Adam: A method for stochastic optimization. **International Conference on Learning Representations**, 12 2014.

KRAUS, M.; FEUERRIEGEL, S. Decision support from financial disclosures with deep neural networks and transfer learning. **Decis. Support Syst.**, 2017.

LEBARON, B.; ARTHUR, W.; PALMER, R. Time series properties of an artificial stock market. **Journal of Economic Dynamics and Control**, v. 23, n. 9, p. 1487–1516, 1999. ISSN 0165-1889. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0165188998000815>.

LEWIS, M.; LIU, Y.; GOYAL, N.; GHAZVININEJAD, M.; MOHAMED, A.; LEVY, O.; STOYANOV, V.; ZETTLEMOYER, L. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**. Online: Association for Computational Linguistics, 2020. p. 7871–7880. Disponível em: <https://aclanthology.org/2020.acl-main.703>.

LI, B.; CHAN, K. C.; OU, C.; RUIFENG, S. Discovering public sentiment in social media for predicting stock movement of publicly listed companies. **Information Systems**, v. 69, p. 81–92, 2017. ISSN 0306-4379. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0306437916304860>.

LI, Q.; WANG, T.; GONG, Q.; CHEN, Y.; LIN, Z.; SONG, S. kwang. Media-aware quantitative trading based on public web information. **Decision Support Systems**, v. 61, p. 93–105, 2014. ISSN 0167-9236. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167923614000232>.

LOUGHRAN, T.; MCDONALD, B. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. **The Journal of Finance**, v. 66, n. 1, p. 35–65, 2011. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2010.01625.x>.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient estimation of word representations in vector space. **Proceedings of Workshop at ICLR**, v. 2013, 01 2013.

MILLER, G. A. Wordnet: A lexical database for english. **Commun. ACM**, Association for Computing Machinery, New York, NY, USA, v. 38, n. 11, p. 39–41, nov 1995. ISSN 0001-0782. Disponível em: <https://doi.org/10.1145/219717.219748>.

MISHEV, K.; GJORGJEVIKJ, A.; VODENSKA, I.; CHITKUSHEV, L. T.; TRAJANOV, D. Evaluation of sentiment analysis in finance: From lexicons to transformers. **IEEE Access**, v. 8, p. 131662–131682, 2020.

MPQA. **OpinionFinder (version 1.0)**. 2005. Disponível em: <http://mpqa.cs.pitt.edu/opinionfinder/opinionfinder_1/>.

NIELSEN, F. A new anew: Evaluation of a word list for sentiment analysis in microblogs. **CoRR**, 03 2011.

PATTERSON, S. **The Quants: How a New Breed of Math Whizzes Conquered Wall Street and Nearly Destroyed It**. 1st. ed. [S.l.]: Crown Business, 2011. ISBN 0307453383.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: **Empirical Methods in Natural Language Processing (EMNLP)**. [s.n.], 2014. p. 1532–1543. Disponível em: <http://www.aclweb.org/anthology/D14-1162>.

RAJEEV PADMANAYANA, H. D. A. -stockguru:-smart-way-to-predict-stock-price-using-machine-learning-. **International Journal of Innovative Research in Computer Science  Technology (IJIRCST)**, v. 9, n. 4, p. 48–52, 2021. ISSN 2347 - 5552. Disponível em: <https://www.ijircst.org-/view_abstract.php?year=vol=9primary=QVJULTYwMw==>.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning Representations by Back-propagating Errors. **Nature**, v. 323, n. 6088, p. 533–536, 1986. Disponível em: <http://www.nature.com/articles/323533a0>.

SANFORD, A. Does perception matter in asset pricing? modeling volatility jumps and returns using twitter-based sentiment indices. **Journal of Behavioral Finance**, 4 2019. Disponível em: <http://dx.doi.org/10.2139/ssrn.3180950>.

SCHULER, K. **VerbNet: A broad-coverage, comprehensive verb lexicon**. Tese (Doutorado) — University of Pennsylvania, 01 2005.

SCHUMAKER, R. P.; CHEN, H. Textual analysis of stock market prediction using breaking financial news: The azfin text system. **ACM Trans. Inf. Syst.**, Association for Computing Machinery, New York, NY, USA, v. 27, n. 2, mar 2009. ISSN 1046-8188. Disponível em: <https://doi.org/10.1145/1462198.1462204>.

SHYNKEVICH, Y.; MCGINNITY, T.; COLEMAN, S.; BELATRECHE, A. Stock price prediction based on stock-specific and sub-industry-specific news articles. In: **2015 International Joint Conference on Neural Networks (IJCNN)**. [S.l.: s.n.], 2015. p. 1–8.

SOWINSKA, K.; MADHYASTHA, P. **A Tweet-based Dataset for Company-Level Stock Return Prediction**. 2020.

SPRENGER, T. O.; TUMASJAN, A.; SANDNER, P. G.; WELPE, I. M. Tweets and trades: the information content of stock microblogs. **European Financial Management**, v. 20, n. 5, p. 926–957, 2014. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-036X.2013.12007.x>.

STONE, P.; DUNPHY, D.; SMITH, M.; OGILVIE, D. **The General Inquirer: A Computer Approach to Content Analysis**. [S.l.: s.n.], 1966.

TABORDA, B.; ALMEIDA, A. de; DIAS, J. C.; BATISTA, F.; RIBEIRO, R. **Stock Market Tweets Data**. IEEE Dataport, 2021. Disponível em: <https://dx.doi.org/10.21227/g8vy-5w61>.

THALER, R. H. **Misbehaving: The Making of Behavioral Economics**. 1st. ed. [S.l.]: W. W. Norton  Company, 2016. ISBN 039335279X.

THORP, E.; KASSOUF, S. **Beat the Market: A Scientific Stock Market System**. 1st. ed. [S.l.]: Random House, 1967. ISBN 0394424395.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L. u.; POLOSUKHIN, I. Attention is all you need. In: GUYON, I.; LUXBURG, U. V.; BENGIO, S.; WALLACH, H.; FERGUS, R.; VISHWANATHAN, S.; GARNETT, R. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2017. v. 30. Disponível em: <https://proceedings.neurips.cc/paper-/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>.

YANG, X.; MACDONALD, C.; OUNIS, I. Using word embeddings in twitter election classification. **Information Retrieval Journal**, v. 21, 2018. ISSN 1573-7659. Disponível em: <https://doi.org/10.1007/s10791-017-9319-5>.

ZHANG, X.; ZHANG, Y.; WANG, S.; YAO, Y.; FANG, B.; YU, P. S. Improving stock market prediction via heterogeneous information fusion. **Knowledge-Based Systems**, Elsevier BV, v. 143, p. 236–247, mar 2018. Disponível em: <https://doi.org/10.1016\%2Fj.knosys.2017.12.025>.