UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE TECNOLOGIA
CURSO DE GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Francielle Vasconcellos Pereira

PROCESSAMENTO DE LINGUAGEM NATURAL PARA DETECÇÃO DE PADRÕES NO MERCADO DE AÇÕES BRASILEIRO

Santa Maria, RS
2021

Francielle Vasconcellos Pereira

PROCESSAMENTO DE LINGUAGEM NATURAL PARA DETECÇÃO DE PADRÕES NO MERCADO DE AÇÕES BRASILEIRO

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do grau de **Bacharel em Ciência da Computação.**

ORIENTADOR: Prof. Joaquim Vinicius Carvalho Assunção

500
Santa Maria, RS
2021

Francielle Vasconcellos Pereira

PROCESSAMENTO DE LINGUAGEM NATURAL PARA DETECÇÃO DE PADRÕES NO MERCADO DE AÇÕES BRASILEIRO

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para obtenção do grau de **Bacharel em Ciência da Computação.**

Joaquim Vinicius Carvalho Assunção, Dr. (UFSM)
(Presidente/Orientador)

Andrea Schwertner Charão, Dra. (UFSM)

Renata Rojas Guerra, Dra. (UFSM)

Santa Maria, RS
2021

ABSTRACT

NATURAL LANGUAGE PROCESSING FOR PATTERN DETECTION IN THE BRAZILIAN STOCK MARKET

AUTHOR: Francielle Vasconcellos Pereira
ADVISOR: Joaquim Vinicius Carvalho Assunção

Through social media, influencers are able to reach thousands of people. In the financial market it is no different, tweets from such people can come to influence the shares's value. In this work, the use of Natural Language Processing (NLP) and the Apriori algorithm is proposed to identify associations between the behavior of stock prices and the sentiment expressed on Twitter. The main approach is to analyze structured and unstructured data, using NLP concepts and association rules to identify the main players capable of influencing the masses and having some impact on stock prices, or even the anomalous social movements capable of causing some impact. Associations were found only in short periods, where in all stocks some periods were identified in which the predominant sentiment is expressed after the change in price, not exceeding a period of 15 days.

Keywords: Natural Language Processing. Stock Market. Association Rules.

LISTA DE ABREVIATURAS E SIGLAS

AMT	Amazon Mechanical Turk
API	Application Programming Interface
ANN	Artificial Neural Networks
AS	Análise de Sentimento
BMF	Bolsa de Mercadorias e Futuros
$BVMF$	Bolsa de Valores, Mercadorias e Futuro
CVM	Comissão de Valores Mobiliários
ELN	Entendimento de Linguagem Natural
EPS	Earnings Per Share
$HTML$	HyperText Markup Language
$LeIA$	Léxico para Inferência Adaptada
MLP	Multilayer Perceptron
NAV	Net Asset Value
$NLTK$	Natural Language Toolkit
NER	Named Entity Extraction
PLN	Processamento de Linguagem Natural
REN	Reconhecimento de Entidades Nomeadas
SAX	Symbolic Aggregate Approximation
URL	Uniform Resource Locator

SUMÁRIO

1 Introdução **7**
1.1 Justificativa 8
1.2 Objetivos 9
1.2.1 Objetivo Geral 9
1.2.2 Objetivos Específicos 9

2 Referencial Teórico **11**
2.1 Mercado Financeiro 11
2.1.1 Bolsa de Valores 12
2.1.2 Rendimentos e Riscos 12
2.2 Análise de Sentimentos 12
2.3 Bibliotecas e Frameworks para Análise de Sentimento 14
2.3.1 TextBlob 14
2.3.2 VADER 15
2.3.3 Spacy 17
2.4 Coleta de dados da Web 18
2.5 SAX 19
2.6 Regras de associação 19

3 Trabalhos Relacionados **22**
3.1 Análise de Sentimento 22
3.2 Predição do preço de ativos 24
3.3 Síntese do Capítulo 25

4 Metodologia e Desenvolvimento da Proposta **27**
4.1 Coleta de dados 27
4.1.1 Coleta das ações 27
4.1.2 Coleta de Tuítes 28
4.2 Classificação de Sentimentos dos Tuítes 30
4.3 Cálculo de Engajamento 31
4.4 Categorização e escolha dos dados 32

5 Estudo Observacional **35**

6 Conclusão **38**

Referências Bibliográficas **40**

APÊNDICES **42**

Apêndice A – Lista das 50 Ações Utilizadas Inicialmente **43**

Apêndice B – Regras Encontradas pelo Apriori **46**

1 INTRODUÇÃO

As mídias sociais tem sido cada vez mais utilizadas como meio de trabalho e disseminação de informação. Através de blogs, sites de notícias, fóruns e redes sociais é possível extrair dados relevantes relacionados a diversos assuntos. A Análise de Sentimento (AS) é uma técnica de PLN que tem como objetivo analisar sentimentos e opiniões de dados desestruturados, como é o caso das mídias sociais.

O desenvolvimento econômico está diretamente relacionado aos investimentos em capital e recursos humanos. Esse desenvolvimento é acelerado quando há investimentos direcionados à alternativas com maiores retornos econômicos e sociais (NOBREGA, 2000). Com base nisso, analisar dados do passado para identificar o comportamento da bolsa de valores, e descobrir informações sobre alteração artificial de preços de ações, é uma estratégia que pode resultar em descobertas significativas.

Há diversos trabalhos desenvolvidos que podem auxiliar investidores a realizarem predições do comportamento do mercado financeiro, seja avaliando dados históricos ou utilizando redes sociais, como é o caso de (NTI; ADEKOYA; WEYORI, 2020), (CAROSIA; COELHO; SILVA, 2019), (SKUZA; ROMANOWSKI, 2015) e (ALKUBAISI; KAMARUDDIN; HUSNI, 2017). Diferentes métricas e características influenciam no comportamento do mercado, como eventos políticos, notícias internas e externas, balanços das empresas etc. Muitas das contribuições supra citadas possuem resultados positivos quando o assunto é a utilização de *Machine Learning* para análise de dados históricos, como indicadores técnicos e preços do mercado.

O desafio que se apresenta é de detectar a movimentação de ações identificando padrões nos preços do mercado e influências via redes sociais, visando a diminuição de riscos em investimentos de curto e/ou longo prazo.

Há uma divisão do estudo do mercado financeiro em três análises: fundamental, que usa dados não estruturados; técnica, baseando-se em dados do passado; e tecnológica, que visa usar um conjunto de dados para aprender padrões que se aproximam da função de geração dos dados (NTI; ADEKOYA; WEYORI, 2019).

Algumas literaturas ainda trazem outra categoria, que tem como foco a previsão de série temporal com modelos tradicionais (CAROSIA; COELHO; SILVA, 2019). Entretanto, a análise de dados temporais pode ser incluída na abordagem de análise tecnológica, que por sua vez pode ser utilizada pelas duas primeiras vertentes.

(ALVES, 2015) encontrou uma relação de mutirão em sua análise, onde o sentimento predominante possuía relação com os lucros/perdas do dia. (SANTOS; LAENDER; PEREIRA, 2015) analisaram apenas o comportamento dos investidores no Twitter ainda em 2015. Afirmando que as pessoas utilizam a rede social como fonte de informação para realizar seus investimentos, mas nesse estudo não houve AS envolvida. (SOUZA;

LUCENA; QUEIROZ, 2019b) afirmaram em seus estudo que o volume diário de tuítes é capaz de impactar o volume de negociações. Além disso, foi verificado que os sentimentos expressos são capazes de influenciar os investimentos a curto prazo.

Este trabalho foca na análise técnica e fundamental com uso de PLN e do Apriori para descrever padrões de possíveis *players* que podem possuir influência suficiente para criar mudanças repentinas, implicando em preços artificiais no mercado financeiro do Brasil. Pode-se ainda, simplesmente identificar fenômenos em redes sociais capazes de alterar o valor de ações sem base em fundamentos técnicos.

Para a realização da análise foram utilizados dados pré-processados de tuítes e os preços históricos de 50 ações categorizados pelo SAX (*Symbolic Aggregate ApproXimation*), listadas na Tabela do Apêndice A.1, no período de Janeiro de 2019 a Agosto de 2021. Foram encontradas associações entre os preços dos ativos e os sentimentos predominantes apenas em alguns períodos analisados. Além disso, as relações são encontradas a curto prazo, podendo influenciar os preços apenas nos dias seguintes, não ultrapassando um período de 15 dias.

1.1 JUSTIFICATIVA

A sociedade encontra-se na era do *Big Data*, onde uma quantidade grande de dados são armazenados diariamente através de sites de notícias, redes sociais, etc. A partir desses dados, é possível compreender melhor produtos, serviços e/ou negócios e refletir sobre a tomada de decisão mais adequada para se obter melhores resultados no contexto analisado.

O mercado financeiro está se adaptando à revolução digital. Toda nossa vida está sendo impactada e transformada por este fenômeno. Ela está revolucionando a forma que nos comunicamos, estudamos e interagimos com o sistema financeiro. Neste processo de adaptação, o mercado financeiro está se reinventando, com um conjunto novo de regras, novos atores e novas ferramentas (SACRAMENTO, 2021).

A criação de preços artificiais pode causar grandes perdas para algumas pessoas, especialmente para investidores menos experientes. Seguindo uma hipótese levantada por (SOUZA; LUCENA; QUEIROZ, 2019b), ao identificar os padrões de atuação de possíveis *players* e como eles podem influenciar o mercado financeiro, espera-se compreender melhor a volatilidade de curto prazo do mercado e a possível relação com a influência das massas via redes sociais.

Os estudos relacionados a análise e predição desses padrões, em sua maioria, são feitos em países desenvolvidos (Estados Unidos ou Reino Unido por exemplo), como pode ser observado na Seção 3. Este trabalho visa analisar o impacto das redes no Brasil, um país emergente com potencial econômico, fazendo parte dos BRICs (agrupamento dos

cinco grandes países emergentes - Brasil, Rússia, Índia, China e África do Sul) e com uma bolsa de valores reconhecida internacionalmente.

Os dados coletados são de um período pós eleitoral (2019) e o período pandêmico (2020 e 2021), onde há uma desvalorização do real e o dólar ultrapassando os R$ 5,00 (cinco reais). A alta do dólar influencia a economia como um todo, prejudicando empresas que importam matérias-primas aumentando o preço de diversos produtos que chegam até o consumidor final em seu cotidiano. Além disso, o dólar é a moeda oficial para operar no mercado financeiro. Algumas aplicações estão atreladas a taxa cambial, fazendo com que os investidores ganhem ou percam investimentos, de acordo com essa variação do dólar.

1.2 OBJETIVOS

1.2.1 Objetivo Geral

O objetivo deste projeto é detectar padrões entre a atuação de possíveis influenciadores e os impactos que eles podem causar no valor das ações mais negociadas da bolsa brasileira. Por meio de PLN e regras de associação, pretende-se analisar os prováveis efeitos que eles podem causar na volatilidade do mercado financeiro através das mídias sociais. Enfatiza-se que a análise visa verificar associações entre os sentimentos e os preços das ações a curto prazo. Não necessariamente os ativos analisados são de curto prazo.

1.2.2 Objetivos Específicos

Mais especificamente, este projeto de trabalho de graduação objetiva:

- Estudar técnicas de PLN usadas para detectar a influência via mídias sociais;
- Realizar o levantamento bibliográfico de pesquisas que se relacionam com a temática do presente trabalho;
- Pesquisar e identificar os *big players* do Twitter relacionados ao mercado financeiro no Brasil;
- Selecionar dados do Twitter e das ações;
- Pré-processar os *datasets* a serem utilizados;

- Analisar possíveis fenômenos sociais que levam a alteração artificial do preço de ações.

Esta monografia está estruturada da seguinte forma. No Capítulo 2, estão os conceitos e tecnologias que serão usadas no trabalho e uma breve explicação sobre elas. No Capítulo 3 são revisados alguns trabalhos que possuem relação com este; no Capítulo 4 a demonstração de como será possível a realização do objetivo deste trabalho, com o uso de técnicas e casos gerais para validar o motivo de cada algoritmo. O Capítulo 5 traz os seus experimentos e resultados encontrados. Com o 6 finaliza-se o trabalho trazendo discussões além de projetos futuros.

2 REFERENCIAL TEÓRICO

Neste capítulo são vistos conceitos que serão pertinentes ao entendimento do desenvolvimento da monografia, seja em questão de algoritmos ou de conhecimento para o assunto tratado.

2.1 MERCADO FINANCEIRO

As organizações utilizam recursos financeiros e/ou os serviços de instituições especializadas (agentes econômicos) para operar suas finanças. Segundo (CARRETE; TAVARES, 2019), o mercado financeiro é onde ocorrem operações e serviços oferecidos por instituições financeiras, possibilitando um fluxo de recursos monetários entre agentes econômicos.

(CARRETE; TAVARES, 2019) comenta que para haver equilíbrio entre todos os participantes, esse ambiente deve ser regulado e controlado pela autoridade monetária do país, pois o mercado financeiro é uma peça importante para o desenvolvimento de um país. Ela ainda afirma que o mercado financeiro pode ser dividido em quatro segmentos: mercados monetário, de crédito, cambial e de capitais.

O **Mercado de capitais** envolve as operações de médio e longo prazo com títulos mobiliários emitidos pelas empresas, onde, em geral, instituições financeiras especializadas estruturam as operações e proporcionam o "encontro" entre os agentes superavitários e os demandantes de recursos para investimentos tanto em capital fixo como em capital de giro.

O mercado de capitais visa promover a riqueza nacional e é onde os principais ativos negociados são as ações. As ações são representativos da menor fração do capital das empresas. Ao comprar ações de uma companhia, participa-se ativamente de seus resultados, virando um sócio do negócio e obtendo direitos acionistas. Segundo (NETO, 2021), sociedades anônimas emitentes de ações podem ser de dois tipos: abertas ou fechadas.

Uma empresa aberta é quando tem suas ações partilhadas entre acionistas e que podem ser negociadas na Bolsas de Valores. Além disso, devem ser registradas na CVM (Comissão de Valores Mobiliários) - órgão que regula e fiscaliza o mercado de capitais brasileiro - como capital aberto e fornecem ao mercado, de forma periódica, uma série de informações de caráter econômico, social e financeiro.

As companhias de capital fechado são constituídas de empresas familiares e que possuem as ações partilhadas a um grupo específico de investidores. O presente trabalho utilizou ações de empresas abertas nos dados históricos para realização da análise.

2.1.1 Bolsa de Valores

Conforme o mercado de capitais foi evoluindo, criou-se a Bolsa de Mercadorias e Futuros (BMF), onde negociava-se apenas *commodities* - matérias-primas não industriais distintas que não dependem de origem ou produtor (café, soja, petróleo, etc). Após essa etapa, a BMF tornou-se BMFBovespa em 2008, atuando de forma totalmente eletrônica.

Em 2017 foi criada a Bolsa de Valores Oficial do Brasil (B3), fundindo a Central de Custódia e de Liquidação Financeira de Títulos (CETIP), empresa prestadora de serviços financeiros no mercado de balcão organizado, com a Bolsa de Valores, Mercadorias e Futuros de São Paulo (BVMF), transformando-se de uma associação civil sem fins lucrativos.

A B3 foi formada como uma sociedade de capital aberto emitindo somente ações negociadas no Novo Mercado, onde as ações das companhias têm seu desempenho afetado favoravelmente pelos direitos adicionais concedidos aos acionistas e pela qualidade das informações prestadas ao mercado (NETO, 2021).

2.1.2 Rendimentos e Riscos

Em qualquer investimento, o risco de perdas tem um peso alto e deve ser levado em consideração. O ganho e a perda dependem de diversos fatores, como desempenho da empresa que investiu-se e as condições do mercado e da economia geral.

A identificação do momento certo para investir é muito importante para compensar o alto risco nos investimentos com uma remuneração igualmente alta. (NETO, 2021) afirma que os ganhos de capital oferecidos aos acionistas estão sujeitos ao **comportamento incerto** dos preços de mercado das ações.

(NETO, 2021) ainda divide os riscos em investimento de ações em dois: *risco de empresa*, onde avalia-se o quão atrativo é o negócio e a capacidade financeira em garantir seus deveres com seus investidores; e o *risco de mercado* que está associado à imprevisibilidade em seu comportamento. Essa incerteza quanto ao movimento do mercado se dá por diversos fatores, podendo ser eles políticos, econômicos ou de acordo com o desempenho da empresa.

2.2 ANÁLISE DE SENTIMENTOS

PLN é o processo onde um sistema computacional tenta entender, analisar e abstrair informações da linguagem humana baseando-se em um texto ou em porções desse texto. Há diversas áreas e sub-áreas nesse processamento, como reconhecimento de entidades nomeadas (REN), análise sintática, tradução automatizada etc.

Adjacente ao PLN, há outro campo de estudo na linguística computacional, chamado Entendimento de Linguagem Natural (ELN). É onde são treinados os sistemas computacionais para realizar extração de relação, análise semântica, análise de sentimento (AS) etc. A relação entre os dois campos pode ser visto na Figura 2.1.

Figura 2.1 – Relação Entre PLN e ELN

Fonte: Adaptado e traduzido de (KINOW, 2017)

A análise de sentimento é uma técnica de PLN, mais especificamente, de ELN, que detecta computacionalmente o sentimento atrelado a um trecho de texto, classificando-o como positivo, negativo ou neutro. Esse processo usa técnicas de aprendizado de máquina para avaliar uma parte de um texto e determinar o sentimento por trás dela.

Essa análise pode ajudar a decifrar o humor e as emoções de um público em geral e coletar informações úteis sobre determinado contexto, como o monitoramento dos *feedbacks* de clientes nas redes sociais, de marca ou de campanha. Analisar o sentimento em textos de investidores relacionado a determinada ação pode auxiliar na identificação de padrões no preço das ações.

Alguns exemplos das técnicas mais usadas para AS em trabalhos são: análise léxica, aprendizado de máquina e aprendizado profundo, aplicando-se ANN (*Artificial Neural Networks* - Redes Neurais Artificiais) com diversas camadas (CAROSIA; COELHO; SILVA, 2019).

Há a AS baseada em regras, onde não é utilizado nenhum aprendizado de máquina para descobrir o sentimento do texto. Os sentimentos de uma frase podem ser descobertos contando o número de vezes que o usuário usou uma determinada palavra em seu texto. Isso facilita a implementação de métodos para a AS, entretanto, essa simplicidade também traz desvantagens, já que essa abordagem só se preocupa com palavras individuais,

ignorando o contexto no qual é usado.

Há também a AS automática, onde são utilizadas técnicas de aprendizado de máquina para classificação dos sentimentos. Dado um *dataset* de entrada para servir como treinamento, os demais textos são classificados baseando-se nessas amostras. Para tal, classifica-se características de palavras criando vetores de características que alimentam o modelo de predição.

Embeddings de texto são uma forma de representação de palavras em ELN em que palavras sinonimicamente parecidas são descritas usando vetores semelhantes que, quando representados em um espaço n-dimensional, estarão próximos uns dos outros (ES, 2021). Pacotes e bibliotecas da linguagem de programação Python baseados em incorporação usam esta forma de representação para prever sentimentos. Isso leva a uma melhor representação de texto em ELN e produz um melhor desempenho do modelo (ES, 2021). Por último, a AS híbrida é onde as duas abordagens (baseadas em regras e automática) são mescladas com o objetivo de alcançar melhores resultados em determinados contextos.

2.3 BIBLIOTECAS E FRAMEWORKS PARA ANÁLISE DE SENTIMENTO

Hoje em dia pode-se contar com a simplificação no desenvolvimento com o auxílio de diversas bibliotecas, pacotes e *frameworks* disponíveis gratuitamente. Para a AS não é diferente, existem tecnologias que agilizam todo esse processo, ou ao menos, parte dele.

2.3.1 TextBlob

TextBlob é uma biblioteca da linguagem de programação Python, que suporta análises e operações complexas em dados textuais. Internamente, a biblioteca utiliza *Natural Language Toolkit* (NLTK) para realizar suas tarefas. NLTK é uma biblioteca que dá acesso a mais de 50 recursos léxicos, permitindo que os usuários trabalhem com classificação de textos, WordNet, tokenização, lematização, marcação, análise e raciocínio semântico entre outras tarefas.

Essa ferramenta facilita o processo de desenvolvimento de algoritmos para o entendimento da linguagem humana. Proporciona diversas funcionalidades para estudos e pesquisas na área, auxiliando também no ensino de técnicas de PLN.

O TextBlob oferece acesso à uma *Application Programming Interface* (API) capaz de cumprir diferentes funcionalidades de PLN. Baseada em regras (*rule-based*), ela leva em consideração duas propriedades principais: polaridade e subjetividade.

A extração do sentimento de um determinado texto pode ser feita através de uma

abordagem léxica, capturando um sentimento otimista ou pessimista definido por sua orientação semântica e pela intensidade de cada palavra na frase (SHAH, 2020). Essa análise tem por objetivo associar os sentidos linguísticos das palavras, desconsiderando seu significado, estruturas das frases e contexto.

A polaridade pode variar entre [-1, 1], onde -1 define um sentimento negativo e 1 positivo. A subjetividade quantifica as opiniões pessoais e informações factuais contidas no texto, variando entre [0, 1], sendo 1 significando que o texto contém opinião pessoal em vez de informações factuais (SHAH, 2020).

Uma limitação é que o TextBlob ignora palavras que não estão mapeadas, considerando palavras e frases às quais consegue atribuir polaridade e médias para obter a pontuação final. A biblioteca também não lida com negações, mas quando encontra-se termos de negação a polaridade de cada palavra é multiplicada por -0,5.

Em abordagens baseadas em léxico, um sentimento é definido por sua orientação semântica e pela intensidade de cada palavra na frase. Isso requer um dicionário pré-definido para classificar palavras negativas e positivas. Geralmente, uma mensagem de texto é representada utilizando *bag-of-words*. Após atribuir pontuações individuais a todas as palavras, o sentimento final é calculado por uma operação de agrupamento, como por exemplo, calcular a média de todos os sentimentos (SHAH, 2020).

2.3.2 VADER

Valence Aware Dictionary for sEntiment Reasoning (VADER) é outra biblioteca popular para AS. Assim como o TextBlob, o VADER também é baseado em regras. É utilizada uma lista de características lexicais que são rotuladas como positivas ou negativas de acordo com sua orientação semântica para calcular o sentimento do texto. As abordagens que utilizam dicionários léxicos analisam a pontuação de cada palavra na frase e decidem qual é a categoria de sentimento ou pontuação de toda a frase.

A pontuação do sentimento de um determinado texto é obtida somando-se a intensidade de cada palavra no texto. Essas pontuações são obtidas utilizando o *Amazon Mechanical Turk* (AMT), utilizando a estratégia de multidão para evitar algum resultado tendencioso. Foram utilizados vários vários avaliadores humanos e a partir da média de suas classificações para cada palavra, chegou-se aos *scores*.

O VADER retorna a probabilidade de uma determinada sentença de entrada ser positiva, negativa ou neutra, sendo a soma das três 100% (ES, 2021). Mais especificamente, é retornado um dicionário contendo quatro informações:

- **pos**: percentual positivo do texto
- **neg**: percentual negativo do texto

- **neu**: percentual neutro do texto
- **compound**: sentimento geral normalizado

O *compound* é a métrica mais importante, variando de [-1, 1], funcionando de forma similar a polaridade citada na Seção 2.3.1. Assim como no TextBlob, esse valor é utilizado para determinar o sentimento predominante no texto. A intensidade das emoções contidas no texto, variando de [-4, 4] sendo o ponto médio 0 um sentimento neutro. A normalização desses valores é obtida através da Equação 2.1, onde x é a soma dos scores de cada palavra do trecho do texto e α é um parâmetro de normalização setado com o valor 15.

$$\frac{\sum(scores(x))}{\sqrt{x^2 + \alpha}} \tag{2.1}$$

O VADER também considera palavras que não estão em dicionários, como gírias, emojis ou expressões das redes sociais, como ”kkkk”, ”LOL”, ”nah”, ”:(” etc (GILBERT, 2014). São usadas cinco heurísticas que verificam como os elementos textuais aumentam, diminuem ou negam o sentimento predominante, os valores de cada heurística também foram obtidas com o AMT.

O uso de **pontuações** podem influenciar muito na intensidade de uma mensagem, como uso de reticências, pontos de exclamação e/ou interrogação em excesso ou não. O *score* é então amplificado de acordo com a quantidade de pontuações utilizadas no texto analisado. Primeiro calcula-se o sentimento predominante à frase, caso o *score* seja positivo, o VADER adiciona o valor atrelado a cada ponto de interrogação e/ou ponto de exclamação. Caso contrário, o valor é subtraído. Para casos em que o sentimento é neutro, nada é feito. A Tabela 2.1 exemplifica o uso de pontuações.

A segunda heurística que o VADER considera é a **capitalização**, heurística na qual diferencia usos de letras maiúsculas e minúsculas para intensificar o sentimento principal. Assim como na heurística de pontuação, aumenta-se ou diminui o *score* final da palavra, mas dependendo se a palavra em questão é positiva ou negativa. O valor na pontuação varia de acordo com a quantidade de pontuações utilizadas, mas na capitalização o valor somado ou diminuído é 0.733 (GILBERT, 2014).

Outra característica utilizada são os **modificadores de grau**. O objetivo dos modificadores é aumentar ou diminuir a intensidade da palavra seguinte ao modificador. O VADER possui um dicionário de reforço, contendo um conjunto de palavras que servem como intensificadores ou amortecedores. A frase “To muito feliz” na Tabela 2.1 exemplifica essa heurística.

A quarta heurística é a de **mudança de polaridade**, onde é levado em consideração contrastes existentes devido o uso de conjunções adversativas, como "mas", "porém", "entretanto", etc. Na prática, todas as palavras que carregam sentimento antes de um “mas” por exemplo, possuem sua valência reduzida para 50% de seus *scores* e as que se

encontram depois do “mas” aumentam em 50% (ALMEIDA, 2018).

A última heurística utilizada pelo VADER é a **negação**. É realizado uma análise do trigrama que antecede um recurso lexical categorizado com um sentimento para capturar a negação de polaridade. É mantida uma lista de palavras negativas, quando a negação é identificada, multiplica-se o *score* de sentimento do léxico carregado de sentimento por um valor determinado de maneira empírica (GILBERT, 2014).

A principal vantagem na utilização dessa biblioteca é que ela foi desenvolvida visando a análise de dados das mídias sociais, como Twitter e Facebook. É capaz de render bons resultados com dados dessas redes, pois é otimizada para trabalhar com tais (ES, 2021), possuindo suporte para tratamento de emojis e pontuações.

Além dessa vantagem, há uma extensão do VADER adaptada para textos em português, chamada LeIA (Léxico para Inferência Adaptada) (ALMEIDA, 2018). A Tabela 2.1 abaixo exemplifica o uso de cada heurística citadas acima, mostrando o dicionário de retorno para cada frase no LeIA.

Frase	**Negativo**	**Neutro**	**Positivo**	**Normalizado**
"To feliz"	0.0	0.196	0.804	0.6249
"To feliz!"	0.0	0.185	0.815	0.6588
"To feliz!!!"	0.0	0.167	0.833	0.7163
"To FELIZ"	0.0	0.167	0.829	0.7034
"To muito feliz"	0.0	0.328	0.672	0.6249
"To feliz mas nem tanto"	0.0	0.611	0.389	0.3716
"To feliz não"	0.301	0.137	0.562	0.4404

Tabela 2.1 – Exemplo do uso de heurísticas do VADER

2.3.3 Spacy

Spacy[1] é uma biblioteca também de código aberto para PLN e escrita em Python. É voltada para uso em sistemas em produção, auxiliando aplicações no processamento de textos. Dá suporte a abordagens que utilizam *Deep Learning* por meio de sua própria biblioteca de aprendizado de máquina Thinc.

Também permite comunicação com modelos estatísticos treinados por bibliotecas como TensorFlow e PyTorch, além de poder ser usada para extração de informações e/ou

[1] https://spacy.io

ELN ou ainda apenas para o pré-processamento de textos. Possui modelos de rede neural convolucional para análise de dependência, REN e categorização de texto em 19 línguas diferentes (até o momento), incluindo o português.

Os componentes utilizados em português incluem token-to-vector, também conhecido como to2vec, morfologizador, *parser*, reconhecedor de frases, REN, lematizador e identificador de atributo de token baseado em regra. O modelo mais amplo do Spacy em português possui cerca de 500 mil vetores de palavras e as precisões de cada componente ficam muito próximas de 1.

2.4 COLETA DE DADOS DA WEB

A coleta de dados automatizada, também chamada de *scraping*, é um processo de recuperação de informações públicas, que podem ser acessadas na web. O ponto a ser ressaltado é que essa coleta não é feita com o auxílio de uma API, mas sim tratando os elementos de um site diretamente através de um navegador web.

Esse processo é feito de forma automática por um programa desenvolvido em alguma linguagem de programação, como Python ou Java. Os dados solicitados encontram-se, geralmente, em HTML ou outros arquivos web. Essa abordagem automatizada é utilizada quando há a necessidade de coletar uma grande quantidade de dados, tornando a coleta manual inviável.

O Yahoo Finanças é um site voltado para notícias, dados, comentários e relatórios financeiros, além de cotações de ações envolvidas na Bolsa de Valores. Os dados históricos dos ativos presentes na B3 também podem ser encontrados no site. Além de disponibilizar essas informações, permitindo realizar *scraping* das iformações no site, é fornecida uma API para a captura desses dados, permitindo ainda realizar consultas e pesquisas em tempo real para mercado de ações, criptomoedas e câmbio.

Além da API do Yahoo Finanças, a B3 também disponibiliza diferentes API's para coleta de informações do mercado financeiro a partir de 2018. Entretanto, o acesso a essas API's são mais restritas, necessitando cadastros e autenticações.

Em Python, há a biblioteca *Yfinance*, que simplifica o uso da API do Yahoo Finanças para acesso à seus dados históricos. Ela permite realizar o *download* dessas informações de maneira prática e rápida, utilizando threads para tal.

Outra ferramenta disponível é **SNScrape**[2], um projeto *open source* para realização de scraping de sites de redes sociais, como o Twitter, Facebook, Instagram, Reddit, Telegram, VKontakte e Weibo. É feita em Python e está constantemente sendo aprimorada. É possível utilizá-la por linha de comando ou através de código em Python.

O SNScrape, como o nome já diz, usa *scraping* para capturar diferentes atributos

[2]https://github.com/JustAnotherArchivist/snscrape

sobre determinado tuíte, como: URL (*Uniform Resource Locator*); data; conteúdo do tuíte; identificador do tuíte; informações do usuário; links externos; quantidade de respostas, retuítes, favoritos e citações; linguagem do tuítes; aparelho de origem (Android, Desktop, Iphone, etc); informações de mídias contidas no tuíte; Id do tuíte original caso o tuíte em questão seja uma citação ou retuíte e; informações de usuários citados no tuíte.

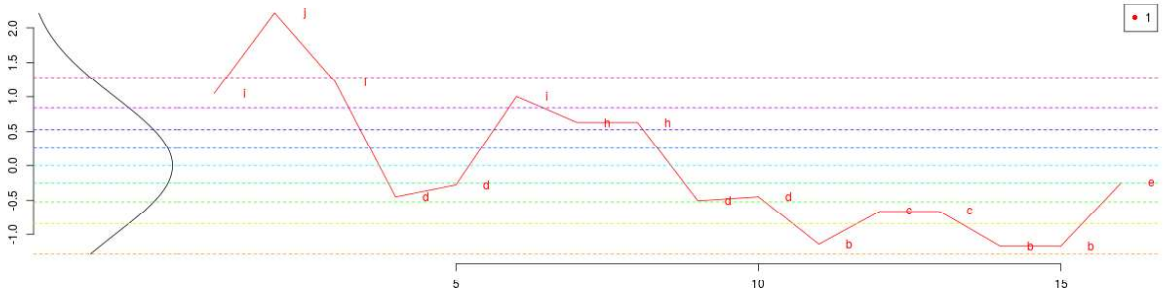
2.5 SAX

Symbolic Aggregate Approximation (SAX) é um algoritmo para representação de série temporais, onde os *inputs* são transformados em *strings*, com o intuito de categorizar uma ou mais série numérica. O algoritmo transforma uma série temporal X de comprimento n em uma *string* de comprimento arbitrário ω, onde geralmente $\omega < n$, usando um alfabeto A de tamanho $A > 2$ (SAX-VSM, 2016).

A representação de séries temporais pode ser dividida em alguns grupos, um deles é a representação adaptativa. Nas representações adaptativas de dados os parâmetros de transformação variam dependendo dos dados disponíveis.

A categorização de dados contínuos em intervalos reduz o ruído e captura a tendência de uma série temporal. O gráfico da Figura 2.2 ilustra a categorização da cotação do ativo Banco Bradesco SA com um alfabeto de tamanho 10.

Figura 2.2 – Categorização do SAX



Fonte: Autora do trabalho

2.6 REGRAS DE ASSOCIAÇÃO

No contexto de mineração de dados, regras de associação são usadas para descobrir elementos que ocorrem em comum dentro de um determinado conjunto de dados.

As regras visam encontrar relacionamentos ou padrões frequentes entre esses conjuntos. Algumas métricas podem ser usadas para avaliação das regras, identificando quais são interessantes para determinado contexto. As mais utilizadas são *thresholds* de suporte e confiança.

O **suporte** de um conjunto é a proporção de transações (itens que foram consultados em uma determinada operação de consulta) que contém esse conjunto, isto é, a frequência em que uma regra é aplicável em um conjunto de dados. A filtragem através do suporte visa descartar regras que podem acontecer por coincidência e que não sejam de interesse da perspectiva do objetivo final da análise.

A confiança mede o quão forte é a inferência feita por uma regra. Dados os conjuntos distintos X e Y, a **confiança** é definida na prática como $conf(X) = suporte(X \cap Y)/suporte(X)$. Essa métrica pode ser entendida como uma estimativa de probabilidade $P(Y|X)$ de encontrar o *Right Hand Side* (RHS) da regra nas transações sobre a condição que essas transações também contenham *Left Hand Side* (LHS).

O componente SE de uma regra de associação é conhecido como antecedente - LHS. O componente ENTÃO é conhecido como o consequente - RHS. O **lift** é definido como a a razão do suporte observado que são esperados se X e Y são independentes. A fórmula $lift(X) = suporte(X)/suporte(Y) \times suporte(X)$ define o lift.

Existem alguns algoritmos responsáveis por realizarem as buscas das regras, um deles é o Apriori. Esse algoritmo funciona identificando os itens individuais frequentes em um conjunto de dados e estende-os a conjuntos de itens cada vez maiores, desde que esses conjuntos de itens apareçam com frequência suficiente na base de dados. Dado um limiar C, o Apriori identifica os conjuntos de itens que são subconjuntos de pelo menos C transações da base de dados.

O algoritmo usa uma abordagem "ascendente", em que subconjuntos frequentes são estendidos um item por vez (uma etapa conhecida como geração de candidatos) e grupos de candidatos são testados em relação aos dados. O algoritmo termina quando nenhuma outra extensão bem-sucedida é encontrada.

O Apriori utiliza internamente busca em largura e uma estrutura de dados em árvore de *hash* para contar os conjuntos de itens candidatos de comprimento k partindo de conjuntos de itens de comprimento $k - 1$. Em seguida, os candidatos infrequentes são podados da árvore. Seguindo o lema do fechamento para baixo, o conjunto candidato contém todos os conjuntos de itens de comprimento k. Depois disso, ele examina o banco de dados de transações para determinar conjuntos de itens frequentes entre os candidatos (BORGELT; KRUSE, 2002).

Algoritmo 1: GERAÇÃO DE REGRAS NO APRIORI

```
1  início
2      para cada conjunto de k itens frequentes f_k, k ≥ 2 faça
3          H1 = {i|i ∈ f_k}
4          chamar geraregras(f_k, H_1)
5      fim
6  fim
```

Os pseudocódigos com os passos do Apriori e da geração de regras pode ser visualizado nos Algoritmos 1 e 2. O segundo algoritmo contém o procedimento auxiliar para geração de regras do Apriori, onde a confiança de cada regra utiliza os contadores de suporte durante a geração dos itens frequentes.

Algoritmo 2: PROCEDIMENTO GERAREGRAS(F_k, H_m)

```
1  início
2      k = |f_k| //tamanho do conjunto de itens frequentes
3      m = |H_m| //tamanho do consequente da regras
4      se k > m + 1 então
5          H_(m + 1) = aprioriger(H_m)
6          para cada h_(m + 1) ∈ H_(m + 1) faça
7              conf = σ(f_k)/σ(f_k − h_(m + 1))
8              se conf ≥ minconf então
9                  gerar regra(f_k − h_(m + 1)) → h_(m + 1)
10             fim
11             senão
12                 apagar h_(m + 1) de H_(m + 1)
13             fim
14         fim
15         chamar geraregras(f_k, H_m + 1)
16     fim
17 fim
```

3 TRABALHOS RELACIONADOS

Este capítulo apresenta trabalhos relacionados ao contexto da análise de sentimento aplicada ao mercado financeiro. Os trabalhos buscam analisar e/ou predizer o comportamento do mercado por meio da AS e/ou dados históricos. Existem muitos *papers* publicados que possuem esse tema, porém, revisões sistemáticas e *surveys* trazem poucos ou nenhum artigo que lide com AS em português ou que envolvem dados do Brasil.

(CAROSIA; COELHO; SILVA, 2019) também aponta essa escassez de trabalhos que analisem não só o Brasil, como os demais países considerados em desenvolvimento. (NTI; ADEKOYA; WEYORI, 2020) também traz evidências de um país subdesenvolvido, Gana, mas já com foco na predição do movimento dos preços do mercado de ações.

3.1 ANÁLISE DE SENTIMENTO

Há diferentes formas de analisar o impacto que as mídias sociais podem ter nos preços das ações. (CAROSIA; COELHO; SILVA, 2019) analisaram a influência do Twitter no Brasil através de três vertentes: número absoluto de sentimentos de tuítes; sentimento de tuítes ponderado pela quantidade de favoritos; e sentimento de tuítes ponderado pelo número de retuítes. Já (OLIVEIRA, 2018) utilizou outros atributos para realizar a predição da popularidade de tuítes a curto prazo utilizou outros atributos, como presença de URL's, tamanho do tuítes, o sentimento único atrelado ao texto e banalidade. A escolha dos atributos a serem levados em consideração dependem do contexto e do objetivo final.

(CAROSIA; COELHO; SILVA, 2019) calculam o sentimento predominante em um determinado *time window* de acordo com as vertentes citadas acima. Para cada cálculo, o resultado será um número entre -1 e 1, sendo de -1 a 0 um sentimento predominantemente negativo e de 0 a 1, positivo. Essa abordagem permite descobrir o sentimento atrelado a qualquer tópico em diferentes períodos de tempo.

Para a AS, (CAROSIA; COELHO; SILVA, 2019) realizou um estudo de diferentes algoritmos de aprendizado de máquina utilizando a biblioteca Scikit-Learn[1]. *Multilayer Perceptron* (MLP) foi o que apresentou melhores resultados de acordo com o F1-Score e acurácia, considerando o contexto e a língua portuguesa. A classificação foi feita entre sentimentos positivo ou negativo, utilizando dois *datasets*, Twitter e notícias relacionadas a uma empresa automobilística. Além disso, o intervalo de tempo analisado é bem específico, considerando apenas o período eleitoral no Brasil (2 meses).

(SOUZA; LUCENA; QUEIROZ, 2019b) também fizeram a mesma análise no período eleitoral, considerando três meses, mas apenas com dados do Twitter para a AS. Outra ca-

[1] https://scikit-learn.org

racterística é que o foco foi analisar tuítes relacionados ao IBOVESPA em português e inglês. Não houve estudo aprofundado para realização da AS, apenas utilizou-se métodos da biblioteca Tweetpy para tal. Um ponto importante citado no trabalho de (SOUZA; LUCENA; QUEIROZ, 2019b) é que o sentimento expresso em dias nos quais a bolsa de valores não opera foi acumulado e seu efeito foi analisado no dia útil de negociação seguinte.

Em seu outro trabalho, (SOUZA; LUCENA; QUEIROZ, 2019a) utilizaram a AS com dados do Twitter para investigar se os sentimentos atrelados à essa rede social são capazes de influenciar o mercado de ações e as estratégias de compras e vendas de ações no Brasil. A diferença é que o foco foi o desenvolvimento de um sistema que simule as decisões de compra e venda de ações, além de mapear estratégias para a tomada de decisões baseando-se em tuítes e na análise técnica.

Para os dados históricos, (SOUZA; LUCENA; QUEIROZ, 2019a) utilizou o volume diário de negociações e o preço de fechamento ajustado. Ao todo, ações de 7 empresas brasileiras foram analisadas, capturando dados históricos de 11 ações da B3, embora, após uma filtragem do *dataset* restaram apenas 2 ações de 2 empresas. Para a AS foi utilizada a biblioteca Lingpipe[2]. Uma diferença dos demais trabalhos citados até então é que o período analisado foi de quase 2 anos, sendo de 2013 até 2015, ultrapassando 8 milhões de tuítes coletados envolvendo as 11 ações.

Na tese de (ALVES, 2015), analisou-se o relacionamento entre amostras do Twitter com o mercado de ações brasileiro por meio de um sistema de tomada de decisão de compra ou venda de ativos. O período compreendido foi de 2013 à 2015, quase 2 anos, com dados de ações de 9 empresas brasileiras.

Nos trabalhos citados até então, durante o pré-processamento dos textos são removidas URL's e *hashtags* dos tuítes. Todavia, (ALVES, 2015) ressalta que o conteúdo existente nesses links podem agregar muito valor ao sentimento da mensagem. Nas simulações realizadas pela autora, foram identificadas relações entre o sentimento de mutirão com lucros/perdas dos investimentos.

(SANTOS; LAENDER; PEREIRA, 2015) também analisou as relações entre os sentimentos encontrados no Twitter e o movimento das ações no Brasil. Um dado interessante encontrado nesse estudo foi que a frequência das postagens na rede social inicia juntamente com o pregão da bolsa de valores e se mantém por cerca de três horas após o horário de fechamento.

Também foi concluído que cerca de 45% dos perfis analisados usam o Twitter como fonte de informação para suas tomadas de decisões de investimentos. O período analisado foi de 2013 a 2014, sendo coletados mais de 2 milhões de tuítes durante esse ano para mais de 20 ações do IBOVESPA.

Outra informação relevante encontrada é que 90% dos tuítes coletados eram publi-

[2]http://aliasi.com/lingpipe

cados por apenas 10% dos usuários, ou seja, apenas alguns perfis publicavam em massa sobre os ativos. (SANTOS; LAENDER; PEREIRA, 2015) encontraram um percentual de apenas 3,1% de tuítes que possuem alguma relação com o mercado de ações.

Outra vertente muito seguida é a análise de notícias relacionadas ao mercado de ações. (FU et al., 2008) focam no mercado financeiro de Hong Kong, analisando notícias em chinês. Isso foi feito através da polarização manual de notícias em negativo e positivo e adotaram a AS baseada em regras. Foram escolhidas nove ações para selecionar notícias sobre elas e analisar se seu comportamento no mercado pode ter influência de notícias com um *time window* de 3 dias. A correlação entre as notícias e o comportamento das ações foi positiva para algumas e negativa para outras, não se mostrando muito promissora.

No Brasil, (FALCãO, 2020) analisou a influência de notícias no preços das ações do mercado brasileiro. Para tal, 5 ações (Ambev, Itaú, Magazine Luiza, Petrobrás e Vale) foram utilizadas juntamente com o índice Bovespa. As notícias foram dos sites Jornal Valor Econômico e Folha de São Paulo, compreendendo um período de quase 6 anos de amostras coletadas (2013 à 2019). Utilizou-se 2 dicionários para classificação das notícias, Sentilex e Liwc, ambos apresentando instabilidade nos resultados. O autor afirmou que as notícias também possuem poder de influenciar a movimentação das ações, mas especificamente em períodos de incerteza econômica.

3.2 PREDIÇÃO DO PREÇO DE ATIVOS

O objetivo das análises para encontrar as associações entre sentimentos extraídos e o mercado de ações é, de maneira geral, para verificar se é possível ou não prever os preços com o auxílio desses dados. Há uma revisão sistemática (LI; BASTOS, 2020) onde mais de 30 trabalhos foram estudados. Entre eles, apenas uma publicação envolvia o mercado financeiro brasileiro, ressaltando a ideia da falta de trabalhos dessa temática no país.

(RAY; GANGULI; CHAKRABARTI, 2021) utilizaram o modelo de predição estatística BST (*Bayesian Structural Time*) para identificar se o sentimento transmitido através de notícias no Twitter é capaz de influenciar o mercado financeiro da Índia, país subdesenvolvido.

A abordagem para coleta de dados foi pré-selecionar contas do Twitter com foco em noticiar a economia indiana. Dez contas foram utilizadas, coletando mais de 800.000 tuítes. Os preços de abertura e fechamento, bem como o preço mais alto do dia foram coletados durante o período de 2015 à 2019, totalizando 4 anos de dados históricos. O objetivo principal do trabalho era verificar a acurácia do modelo híbrido proposto para realizar a predição do movimento das ações, tendo um resultado positivo em relação à outros

modelos.

Em (NTI; ADEKOYA; WEYORI, 2020) é feita a predição do movimento das ações de um país também subdesenvolvido, Gana. O estudo utilizou 4 fontes diferentes de dados para a análise, notícias, tuítes, Google Trends e discussões em fóruns. Com esses dados, é feita a predição utilizando ANN com MLP com dados históricos de ativos de 3 empresas em um período de 9 anos (2010 à 2019), além de 5 *time windows* distintos.

Os autores usaram a similaridade de cosseno para detectar duplicatas e o pré-processamento do texto ficou por conta do NLTK. Além da AS polarizada (sentimentos positivo ou neutro) também foi feita uma classificação entre emoções (alegria, tristeza etc). Esse estudo afirmou a possibilidade de prever o mercado de ações através da AS.

(KHAN; ALIN; HUSSAIN, 2011) realizaram a predição do preço de ações do mercado de Bangladeche, um país considerado desenvolvido (MUNDO, 2020). Para tal, foi desenvolvida uma ANN usando *backpropagation* com *feedforward*. Utilizaram 5 parâmetros de entrada da rede: índice geral, relação Preço/Lucro, valor líquido do ativo (NAV - *Net Asset Value*), lucro por ação (EPS - *Earnings Per Share*) e volume. Foi observada uma certa irregularidade quanto a acurácia dependendo da quantidade de entradas, além de ser sensível a predições de curto prazo, pois caso alguma entrada mude repentinamente os preços não serão mais tão próximos quanto o esperado do valor real da ação.

3.3 SÍNTESE DO CAPÍTULO

Como já citado anteriormente, há uma escassez de trabalhos dessa temática que envolvam o Brasil ou demais países emergentes. A maioria dos trabalhos é voltado para países já desenvolvidos. A Tabela 3.1 demonstra essa ausência de trabalhos em três *surveys* diferentes.

Trabalhos	**Total analisados**	**Total sobre o mercado brasileiro**
Carosia, Coelho e Silva, 2019	+100	3
Li e Bastos, 2020	+30	1
Nti, Adekoya e Weyori, 2019	+100	2

Tabela 3.1 – Quantidade de trabalhos no mercado brasileiro estudados em revisões sistemáticas

As pesquisas acima afirmaram haver relação entre o sentimento expresso no Twitter com o movimento do mercado brasileiro. (SOUZA; LUCENA; QUEIROZ, 2019b) notou a relação entre os volumes de tuítes com o volume de negociações feitas em um período de

tempo *t*, afirmando que são capazes de impactar o volume contemporâneo de negociações do IBOVESPA.

No outro trabalho de (SOUZA; LUCENA; QUEIROZ, 2019a) e nos estudos de (ALVES, 2015), (CAROSIA; COELHO; SILVA, 2019), (SANTOS; LAENDER; PEREIRA, 2015) comprovaram a hipótese da relação entre dados do Twitter com os preços das ações no Brasil, afirmando ainda que essas relações podem ser utilizadas para auxiliar investidores em suas tomadas de decisões no momento de comprar e vender ativos. (NTI; ADEKOYA; WEYORI, 2020) e (RAY; GANGULI; CHAKRABARTI, 2021) também tiveram resultados positivos no mercado acionário de Gana e Índia, respectivamente. Todos esses trabalhos focam em países emergentes, totalizando 3 países apenas.

Não é somente nas redes sociais que encontram-se grande volumes de dados com informações financeiras. Assim como (NTI; ADEKOYA; WEYORI, 2020) utilizaram notícias para encontrar associações, (FU et al., 2008) focou sua pesquisa em notícias voltadas ao mercado de ações de Hong Kong. Já no Brasil, (FALCãO, 2020) também utilizou notícias como seu *dataset* principal para verificar sua influência no mercado de ações. (CAROSIA; COELHO; SILVA, 2019) também usaram notícias de uma única empresa para análise. Os estudos encontraram associações com algumas condições ou apenas para ativos específicos.

Como citado no início deste capítulo, muitos possuem como objetivo predizer o comportamento do mercado através das associações encontradas. É o caso de (RAY; GANGULI; CHAKRABARTI, 2021), (NTI; ADEKOYA; WEYORI, 2020) e (KHAN; ALIN; HUSSAIN, 2011). Nos dois últimos é usada ANN para as predições enquanto o primeiro estudo se baseia na estatística Bayesiana. Os 3 obtiveram resultados positivos, encontrando ao menos uma configuração capaz de predizer os valores dos ativos efetivamente. A Tabela 3.2 sintetiza as principais publicações relacionadas com o presente trabalho.

Trabalhos	**Métodos**	**Período**	**Conclusão**
Carosia, Coelho e Silva (2019)	MLP - tuítes e notícias	3 meses	Positiva
Souza, Lucena e Queiroz (2019)	Tweetpy	3 meses	Positiva
Alves (2015)	Vocabulário - LingPipe	2 anos	Condicionada
Falcão (2020)	Sentilex e Liwc - notícias	6 anos	Negativa

Tabela 3.2 – Resumo dos principais trabalhos relacionados

4 METODOLOGIA E DESENVOLVIMENTO DA PROPOSTA

O desenvolvimento envolve conceitos de PLN e mineração de dados para encontrar associações entre os sentimentos e opiniões expressas no Twitter com os preços de fechamento das ações. O presente capítulo apresenta os principais aspectos técnicos de captura dos dados, pré-processamento e implementação da AS. A Figura 4.1 resume o processo realizado.

4.1 COLETA DE DADOS

A técnica de *scraping* foi utilizada para capturar o *ticker* das 50 ações mais negociadas, via Yahoo Finanças, e os tuítes, baseados nos *tickers* das ações coletadas. Nas seções abaixo há um detalhamento dos passos realizados para o levantamento dos dados.

4.1.1 Coleta das ações

Os dados históricos são utilizados para verificar a variação do preço das ações ao longo dos dias. Em outros trabalhos, como em (KHAN; ALIN; HUSSAIN, 2011) e (NTI; ADEKOYA; WEYORI, 2020) essas informações são usadas para realizar a predição, baseando-se que os dados do passado são fatores relevantes para determinar seu comportamento futuro.

Foi utilizado o site Yahoo Finanças[1] para capturar os dados das ações. Além de possuir as informações necessárias para o objetivo deste trabalho, em (LI; BASTOS, 2020) é notado que a maioria dos trabalhos revisados utilizam essa plataforma para popular sua base de dados. A escolha dos ativos foi feita através do próprio site, que possui diferentes ranqueamentos para mostrar as ações mais relevantes.

Seguindo o ranking de **ações mais negociadas**, capturou-se o símbolo (identificação da ação na BVMF) das 50 ações mais negociadas naquele momento (agosto de 2021). O site permite escolher alguns critérios de ordenação, o utilizado foi o volume de negociações, que é a ordenação padrão para ranquear as ações. O scraping do nome das ações foi feito com o auxílio das bibliotecas *BeautifulSoup* e *Selenium*. Isso permite que a quantidade de ações pudesse ser diminuída ou aumentada posteriormente de forma automatizada, caso fosse necessário.

Em Python, linguagem majoritariamente utilizada nesse trabalho, há uma biblioteca

[1] https://br.financas.yahoo.com

chamada *yfinance*[2], que utiliza diretamente a API do Yahoo Finanças para acesso a seus dados históricos. Ela permite realizar o download dessas informações de maneira prática e rápida, utilizando threads para tal.

Foi feito o *download* dos dados históricos das 50 ações escolhidas de 2019 até o mês de agosto de 2021. Nessa coleta, seis ações foram descartadas por não terem dados históricos em algum ano ou por ter a maioria deles em branco, tornando a análise viável apenas no ano de 2021.

A biblioteca é capaz de retornar todas as informações históricas de cada ativo, sendo eles preços de abertura, fechamento, fechamento ajustado, mais baixo e mais alto do dia. O preço de fechamento - último preço em que ele foi negociado antes da sessão do dia encerrar - ajustado apenas para desdobramento, foi o parâmetro escolhido para servir de base para a análise da associação entre os preços e sentimentos dos tuítes.

4.1.2 Coleta de Tuítes

Inicialmente, estudou-se o uso da API do Twitter para captura dos tuítes, entretanto, a versão de desenvolvedor não é suficiente para o objetivo desse trabalho. A API permite visualizar os tuítes com um *timeframe* de apenas 7 dias, sendo necessário ser estudante de pós-graduação para ter acesso total ao histórico de tuítes.

Dada essa limitação, buscou-se alternativas para realizar essa coleta por meio de *scraping*. A primeira alternativa encontrada foi a biblioteca Tweetpy, entretanto, ela é apenas um facilitador de uso da API do Twitter, utilizando os mesmos *tokens* de acesso, logo, as limitações supracitadas continuam as mesmas.

O TweetScraper foi outra opção a ser testada, contudo, notou-se muita instabilidade na mesma, pois ao realizar as buscas para cada ação a ferramenta retornava no máximo 5000 resultados, o que não representa o valor real do volume de tuítes.

Como pode ser encontrado em (DESAI, 2021), o GetOldTweets3 foi descontinuado, não possuindo mais suporte aos *endpoints* do Twitter. A última opção verificada, e utilizada sem grandes impedimentos foi a *SNScrape*. Por ser uma ferramenta gratuita e conter tantas informações, o SNScraper foi utilizado para realizar o scraping de tuítes.

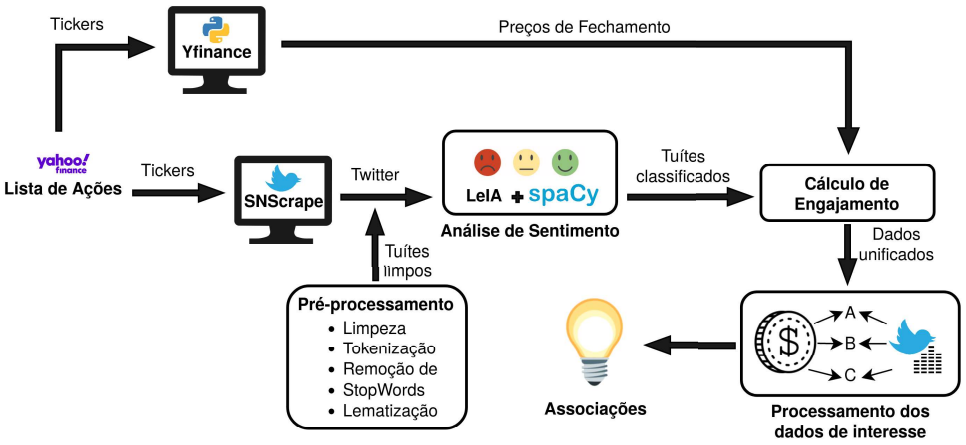
Uma desvantagem é que ele é extremamente lento, pois o *scraping* é feito individualmente para cada tuíte. Além disso, até o momento em que se usou a ferramenta no presente trabalho, o *scraper* se mostrou instável, ao menos utilizando-o diretamente em código Python, levantando exceções durante a execução. As exceções ocorreram várias vezes, sendo necessário retomar a execução posteriormente para o ativo onde houve o ocorrido, pois o erro fazia com que alguns tuítes não fossem capturados.

Foram coletados os tuítes que continham em seu conteúdo alguma citação ao *ticker*

[2]https://pypi.org/project/yfinance

de cada ação escolhida anteriormente, conforme ilustra a Figura 4.1.

Figura 4.1 – Processo de Desenvolvimento da Proposta



Fonte: Autora do trabalho

Em 2019, 44 das 50 ações iniciais foram pesquisadas e coletadas com o SNScraper, o mesmo foi feito para o ano de 2021. O volume de tuítes de alguns ativos foi muito baixo para os 2 anos, não chegando a 500 tuítes em um ano inteiro. Dessa forma, as 20 ações com maior volume de tuítes foram filtradas para captura no ano de 2020, visando acelerar o processo de *scraping* da ferramenta. Além disso, o volume muito baixo de algumas ações poderia dificultar a escolha de períodos para a análise posteriormente, as ações coletadas para os 3 anos podem ser vistas no Apêndice A.1, contendo o volume total. As informações coletadas de cada tuíte foram:

- **Autor**: nome do do autor no Twitter;
- **Seguidores**: quantidade de seguidores que o autor do tuíte possui;
- **Amigos**: quantidade de pessoas que o autor segue;
- **Favoritos**: quantidade de curtidas que o autor realizou em outros tuítes;
- **Status**: quantidade de tuítes realizados pelo autor;
- **Data e hora**: informação importante para unir os tuítes com o preço de fechamento do ativo;
- **URL**: link para o tuíte;
- **Conteúdo**: texto do tuíte com todo o conteúdo que o compõe, como URL's externas, imagens, enquetes, *hashtags*, etc;
- **Respostas**: quantidade de respostas que o tuíte possui;

- **Retuítes**: quantidade de RT's do tuíte;
- **Curtidas**: quantidade de curtidas, ou corações, do tuíte; e
- **Citações**: quantidade de citações do tuíte.

4.2 CLASSIFICAÇÃO DE SENTIMENTOS DOS TUÍTES

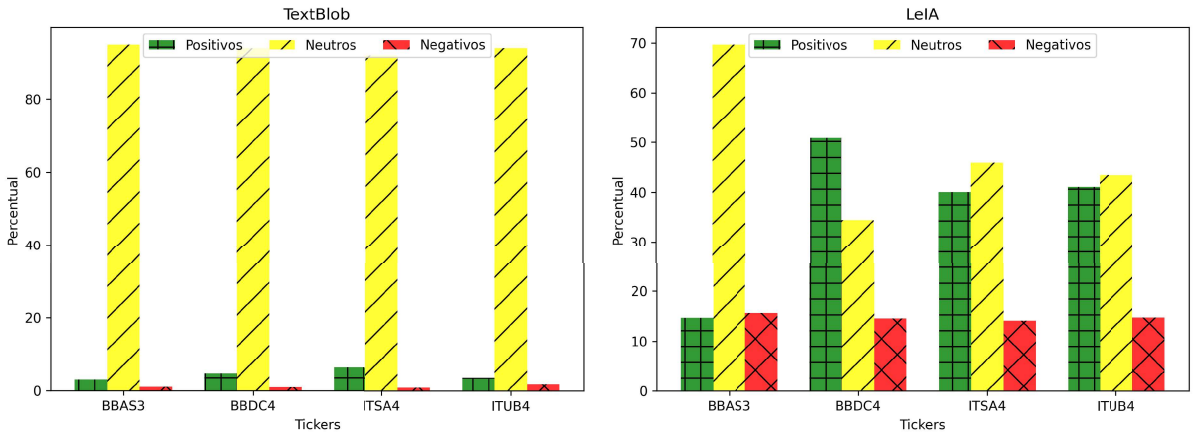
Inicialmente, a biblioteca TextBlob foi utilizada para realizar a classificação dos tuítes, entretanto, a maioria dos tuítes de cada ativo apresentou sentimento neutro, além de haver a necessidade de tradução, pois a biblioteca não tem suporte para o português.

A análise de sentimento foi feita utilizando a biblioteca LeIA, uma adaptação do VADER em português, juntamente com o Spacy 3.2.0 com o modelo largo da língua portuguesa. Primeiramente é feita a limpeza do texto, removendo menções, *hashtags*, RT's e URL's. A tokenização é feita para transformar as frases em uma lista de palavras individuais. Posteriormente, a remoção das *stopwords* - palavras consideradas irrelevantes na compreensão do texto - foi realizada de acordo com o dicionário do Spacy.

Por último, aplicou-se a técnica de lematização para reduzir a variação morfológica. A lematização reduz as formas das palavras a lemas, por exemplo, "perdendo" é substituída pelo seu lema "perder". Após isso, são calculadas as polaridades de cada tuíte. A classificação de sentimentos positivo, neutro ou negativo é obtida pelo valor de *compound*, conforme é explicado na Seção 2.3.2.

O LeIA foi utilizado substituindo o Texblob e houve uma melhora na classificação, já que o uso do Vader é voltado para redes sociais. O gráfico da Figura 4.2 traz um comparativo entre as classificações do TextBlob e do LeIA. Muitos tuítes possuem o teor de apenas noticiar, apenas informando os valores atuais de ativos, por exemplo. Isso faz com que esses tuítes sejam neutros por não possuírem palavras, caracteres ou expressões que indiquem sentimentos ou opiniões.

Figura 4.2 – Classificação de Tuítes TextBlob x LeIA



Fonte: Autora do trabalho

4.3 CÁLCULO DE ENGAJAMENTO

Seguindo o cálculo de (OLIVEIRA, 2018) e (PILLAT; PILLAT, 2017), foi computada a taxa de engajamento de cada tuíte. Para (PILLAT; PILLAT, 2017), engajamento nada mais é que a interação de um usuário com um conteúdo seguido por ele (um página, outro usuário, uma *tag*, etc). Quanto mais interações constantes houver nas publicações de um usuário/página, melhor é o engajamento.

No caso do Twitter, algumas métricas simples podem ser utilizadas para determinar a popularidade de um tuíte, como quantidade de curtidas e retuítes, ou quantidade de seguidores. Entretanto, essas métricas não indicam necessariamente o alcance ou popularidade do item postado, ou ainda, não determina a influência que um usuário pode ter.

Supondo um usuário que possui 10.000 seguidores, a ideia padrão é acreditar que esse usuário seja uma pessoa com influência na rede social. Ao analisar o perfil desse usuário, verificou-se que suas postagens não possuem muita interação de outras pessoas. Dessa forma, mesmo tendo vários seguidores, o alcance desse usuário é muito baixo, não o determinando como uma pessoa influente na plataforma, ao menos com relação ao assunto das postagens analisadas.

Em outra suposição, alguém que possua poucos seguidores no Twitter, mas com uma quantidade de interações em suas postagens muito maior, pode ser considerado alguém influente sobre o assunto de seus tuítes.

Para ter um indicador que leve esses casos em consideração, medindo o alcance real de um usuário, a **taxa de engajamento** é calculada. (PILLAT; PILLAT, 2017) expõe em seu trabalho a fórmula da Equação 4.1, voltada para o cálculo no Facebook.

$$TaxE = \frac{curtidas + compartilhamentos + comentarios}{seguidores} * 100 \quad (4.1)$$

Para a realidade do Twitter não há muita distinção ou adaptação necessária, compartilhamentos equivale a quantidade de retuítes e citações do tuíte (*quotes*) e comentários são as respostas. A fórmula utilizada neste trabalho para cálculo da taxa de engajamento pode ser vista na Equação 4.2. Com base nessa taxa e nos sentimentos atrelados ao tuítes, é realizada a escolha de períodos e de quais tuítes analisar.

$$TaxE = \frac{curtidas + (RT's + quotes) + respostas}{seguidores} * 100 \quad (4.2)$$

Através desse cálculo é obtido um percentual onde o valor obtido é positivo e quando maior que 100 indica que a quantidade de interações é maior que o número de seguidores que o autor do tuíte possui. Esse valor pode ser referência para definir se um perfil possui ou não engajamento nos seus tuítes.

4.4 CATEGORIZAÇÃO E ESCOLHA DOS DADOS

As etapas citadas nas seções acima resultam em diversas informações novas. Foi realizada uma filtragem dos dados com o intuito de guardar apenas as informações que seriam relevantes para o trabalho. Após a coleta das ações, coleta e classificação dos tuítes, bem como a taxa de engajamento dos mesmos, o *dataset* de cada ativo para cada ano continha as informações mostradas na Tabela 4.1 abaixo.

Informação	**Descrição**
Data	dia, mês e ano em que o tuíte foi publicado
Conteúdo	conteúdo original do tuíte
Autor	nome do autor do tuíte
Preço	preço de fechamento em reais do ativo no dia do tuíte
Engajamento	valor indicando a taxa de engajamento do tuíte, ou seja, o alcance por ele atingido
Sentimento	classificação do sentimento do tuíte, podendo ser “*Positive*”, “*Neutral*” ou “*Negative*”

Tabela 4.1 – Variáveis relevantes mantidos de cada ativo em cada ano

O SNScrape fornece data e hora de publicação do tuíte, contudo, como mostra a tabela acima, a data e hora foi transformada apenas em data após a união dos tuítes com o preço das ações, pois a hora não é levada em consideração neste trabalho. Inicialmente, foram descartadas as ações FNAM11.SA, RAIZ4.SA, NTCO3.SA, SOMA3.SA, LWSA3.SA,

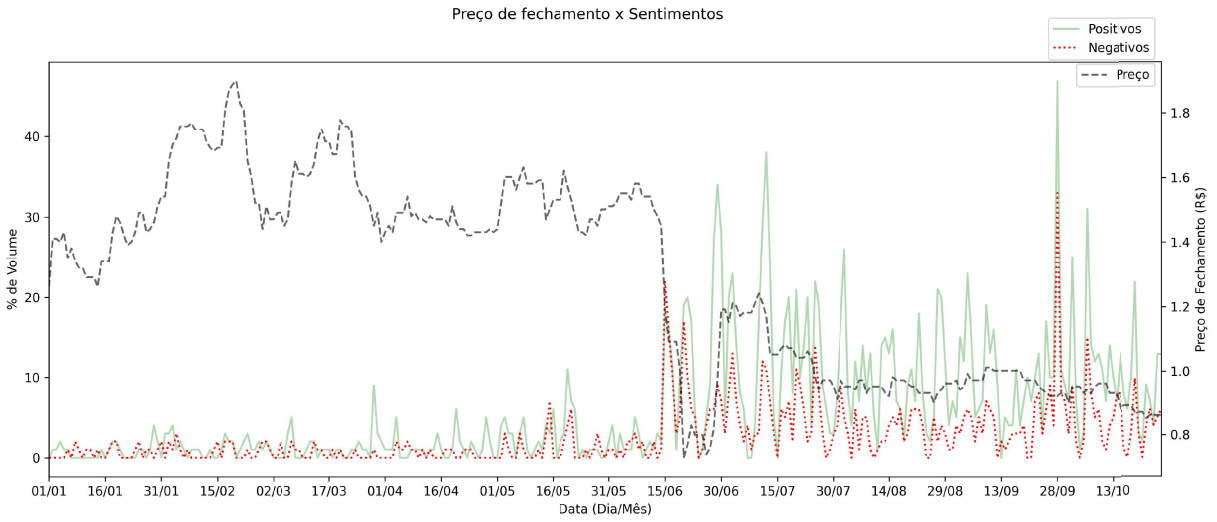
RNEW1.SA, por não possuírem dados históricos dos preços das ações de anos anteriores a 2021 ou por possuírem muitos dias sem informações.

Após o *scraping* dos tuítes de cada ticker, outras 24 ações foram descartadas devido ao baixo volume de tuítes em um ano inteiro. Por exemplo, FNOR11.SA possuía 2 publicações em 2019 e 0 em 2021. LINX3.SA não chegava a 200 tuítes em 2021 e em 2019 possuía apenas 232. Como a escolha das ações foi feita baseada no ranqueamento de mais negociadas em 2021, esse ano foi priorizado para retirar os ativos com volume baixo.

Foram selecionadas as 20 ações com maior volume de tuítes 2021. Sendo PETR4.SA a com maior volume, contendo 16712 e BPAC11 a de menor volume, com 2180 tuítes. A tabela contida no Apêndice A.1 contém as ações que foram mantidas para levantamento de tuítes em 2020, juntamente com o volume total dos 3 anos.

Alguns *datasets* de ativos possuíam mais de 10 mil linhas, sendo necessário capturar apenas períodos e tuítes de interesse. Foram gerados, para cada ação e ano, gráficos com o percentual de volume de tuítes positivos e negativos por dia para verificar períodos que fossem interessantes analisar se há associações. A Figura 4.3 apresenta o gráfico gerado para o ativo OIBR3.SA no ano de 2019.

Figura 4.3 – Volume de Sentimentos Diários x Preço de Fechamento do Ativo OIBR3.SA em 2019



Fonte: Autora do trabalho

No gráfico acima, percebe-se que o volume de tuítes aumentou a partir da metade do ano. Um volume diário abaixo de 10%, a princípio, não seria o suficiente para alterar o preço da ação, a menos que esses tuítes envolvam algum *player* influente sobre aquele ativo, com uma taxa de engajamento alta. Observando o gráfico acima, uma escolha de um período interessante seria de 05 à 20 de Junho, onde em uma queda brusca no preço da OIBR3.SA, a quantidade de tuítes negativos aumentou cerca de 20%, caindo ainda

mais nos dias seguintes.

O período seguinte, de 21 de Junho a cerca de 10 de Julho de 2019 é outro que apresenta certo padrão entre os sentimentos predominantes, onde quando houve um aumento de sentimentos positivos o preço de fechamento também aumentou. Após a escolha dos períodos para cada ativo, filtrou-se, para cada dia, o tuíte que possuía a maior taxa de engajamento. O volume diário de tuítes da maioria dos ativos não é grande, fazendo com que a comparação realizada por (SOUZA; LUCENA; QUEIROZ, 2019b) entre os volumes de tuítes e negociações seja ilógica para este trabalho.

Para ser possível verificar associações entre diferentes dados é necessário categorizá-los de alguma maneira. Da mesma forma que a métrica *compound* foi classificada entre positivo, neutro e negativo, há a necessidade de realizar um processo parecido para o preço das ações e para a taxa de engajamento. Para tal, utilizou-se uma ferramenta[3] de análise de séries temporais para geração do SAX.

Na geração do SAX para o preço das ações e para a taxa de engajamento dos tuítes foi feita a filtragem descrita na Seção 4.4, onde foram gerados gráficos como o apresentado na Figura 4.3 para capturar períodos de interesse que contivessem uma possível associação entre os preços e o sentimento. O SAX classificou o preço das ações e o engajamento em 10 classes, de 'a' até 'j', onde as 5 primeiras letras representam um período de maior baixa do preço da ação e baixo engajamento e as 5 últimas letras representam um período de alta nos preços e no engajamento.

Para cada período escolhido, foram separadas as colunas contendo apenas os preços ao longo dos dias e a taxa de engajamento, pois os *scores* de sentimento foram classificados previamente. As linhas contendo sentimentos neutros foram excluídas e não estão no escopo da análise deste trabalho. O ano de 2019 possui um volume de tuítes com sentimentos atrelados muito baixo para a maioria das ações. Algumas com seu valor máximo não chegam a 10% de tuítes classificados, ainda assim alguns períodos foram escolhidos, pois poderia haver algum usuário com forte influência comentando sobre o ativo.

Todas as ações contendo menos de 30 linhas com dados incluindo todos os períodos de interesse escolhidos foram descartadas para a etapa de geração de regras.

[3] https://joaquim.shinyapps.io/TSmeasTool

5 ESTUDO OBSERVACIONAL

Este capítulo apresenta os experimentos realizados e resultados encontrados através da aplicação do algoritmo Apriori para detecção de regras e verificação de associações entre os preços das ações com os sentimentos expressos no Twitter. Os experimentos visam realizar análises de acordo com as regras encontradas para cada ação em diferentes períodos de tempo.

As regras de associação do algoritmo Apriori definido na Seção 2.6 foi aplicado para geração das regras. O número de transações usados varia conforme a quantidade de tuítes no período analisado. A janela de tempo escolhida foi de 15 dias. Em alguns casos, essa janela foi aumentada, devido a baixa quantidade de transações após a eliminação de linhas conforme a taxa de engajamento, ou por apresentarem períodos maiores com possível relação de acordo com a análise gráfica. Os parâmetros utilizados no Apriori podem ser visualizados na Tabela 5.1 abaixo.

Parâmetros do Apriori	
Suporte mínimo	10%
Confiança mínima	70%
Quantidade mínima de itens	2

Tabela 5.1 – Parâmetros setados na execução do algoritmo Apriori

Para o ano de 2019 foram geradas regras para 8 ativos, escolhidos através da análise gráfica de cada um (conforme apresentado no gráfico da Figura 4.3), onde apresentou-se alguma relação entre a cotação das ações com o sentimento predominante apresentado em dias anteriores. As regras encontradas podem ser visualizadas na tabela do Apêndice B.1.

Para BBS3.SA encontrou-se forte associação entre uma classe que representa um período de alta da ação. Em contrapartida, a classe C com confiança de 71% representa um conjunto de baixa no valor da ação e apresenta associação com um sentimento positivo. O mesmo comportamento é observado nas ações BBDC4.SA e ITSA4.SA, ITUB4.SA, OIBR3.SA e PETR4.SA.

Apenas MGLU3.SA e USIM5.SA possuem todas as regras conforme o esperado, onde ações com valores altos apresentam um volume de sentimentos positivos predominantes e negativos para valores baixos. BBS3.SA possui um comportamento esperado no preço classificado com a categoria **I** (valor da ação em alta), estando associado a um sentimento positivo com 100% de confiança. Entretanto, apresenta uma regra onde uma classe com período de baixa no valor da ação atrelado a um sentimento também positivo.

Em 2020 foram geradas regras para 10 ativos, também filtradas por meio da análise

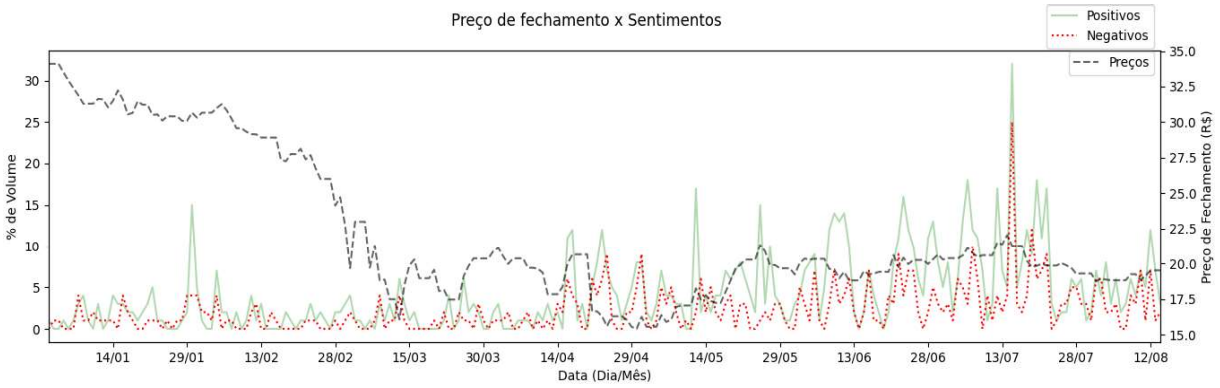
gráfica. As regras encontradas podem ser visualizadas na tabela do Apêndice B.2. No ano de 2020, o percentual de volume de tuítes classificados aumentou, refletindo em um aumento de ações com uma possível relação entre preço x sentimento. Nesse ano, apenas a ação MGLU3.SA encontrou regras com um comportamento esperado. Não foram encontradas regras para o ativo EMBR3.SA, com 32 transações. As demais 8 ações possuem alguma associação oposta ao esperado, como o caso de OIBR3, onde duas classes representando uma baixa da ação está associada com 78% e 88% de confiança a um sentimento positivo, e outras duas classes de alta da ação a um sentimento positivo.

Assim como 2020, 10 ações foram analisadas em 2021 por apresentaram alguma relação entre preço e volume de sentimentos de tuítes na análise gráfica. As regras encontradas podem ser visualizadas na tabela do Apêndice B.3. Não encontrou-se nenhuma regra para o ativo VALE3 contendo 30 transações. Nesse ano, encontrou-se regras coerentes para a ação BBDC4.SA, onde valores altos estão associados a um sentimento positivo, não sendo as regras encontradas totalmente contraditórias com o esperado. O mesmo que ocorreu com OIBR3.SA em 2020, ocorreu com CIEL3.SA em 2021, apresentando um comportamento esperado em 3 das 4 regras encontradas. Nesse ano não houve nenhuma ação contendo todas as regras que possuam 100% do que era esperado da hipótese levantada neste trabalho.

O comportamento no Twitter para algumas ações acompanham o preço, mas não são capazes de influenciar seu valor futuro, mas sim, de refletir o preço passado de cada ativo. Por exemplo, um ativo que se encontra em queda em um período de dias consecutivos terá uma quantidade crescente de tuítes negativos relacionados a ele nos dias seguintes. Também há comportamentos diferentes em anos distintos, por exemplo, para MGLU3.SA, não encontrou-se associações entre preço dos ativos e volume de sentimentos negativos ou positivos no ano de 2021. Em 2019 e 2020, há forte associação entre as 2 variáveis.

A maioria das ações não possuem um padrão que identifique uma possível influência no preço provocada pelo Twitter. Por exemplo, em 2019 e 2020, mesmo com sequências de dias de tuítes predominantemente positivos com relação a BBDC4.SA, o preço seguiu despencando, como pode ser observado no gráfico da Figura 5.1. Relações foram percebidas na análise gráfica no ano de 2021, porém as regras encontradas negam a forte associação esperada no ano.

Figura 5.1 – Volume de Sentimentos Diários x Preço de Fechamento do Ativo BBDC4.SA em 2020



Fonte: Autora do trabalho

Quando há influência no preço através dos sentimentos expressos, ela é refletida no valor da ação logo nos dias subsequentes. Considerar sentimentos para médio e longo prazo para analisar o preço das ações não deve surtir um efeito positivo, visto que as relações encontradas nesse trabalho possuem associações a curto prazo.

6 CONCLUSÃO

A análise descrita neste trabalho é inicial e buscou outras alternativas de realizar os processos que envolvem encontrar associações no mercado financeiro brasileiro. A pesquisa e os procedimentos seguidos podem servir de base para outros estudos além de aprimoramentos na corrente pesquisa. Utilizando bibliotecas do Python descritas na Seção 2.3, houveram diversos facilitadores que não tornam necessário a implementação de um novo algoritmo para análise de sentimentos, classificadores, *scraping*, etc.

Como ideia inicial, considerava-se utilizar tuítes e notícias diversas com relação aos ativos selecionados para realização da análise de sentimentos. Entretanto, os resultados encontrados em (ALVES, 2015) e (FALCãO, 2020) não foram positivos, verificou-se que o uso de notícias não possui forte impacto no mercado.

Em todas as ações analisou-se que o sentimento é expresso após a alteração no preço, por exemplo, uma ação que teve uma queda repentina em seu valor apresentou um volume de tuítes predominantemente negativo nos dias subsequentes, não sendo configurada uma influência devido a rede social, apenas um reflexo da queda anterior. Apenas um ativo possuiu regras com associações esperadas, onde o preço de fechamento termine o dia em alta conforme o volume de sentimentos positivos, ou em baixa se o volume predominante for negativo.

Foram encontradas associações apenas em alguns períodos, isso pode indicar que nesses *frames* específicos há alguma anomalia ou evento que recaiu sobre os preços das ações e nos sentimentos expressos no Twitter. A rede social é muito utilizada para compartilhamento de notícias, muitos tuítes possuem links para sites externos. Alguma notícia impactante pode ter causado as associações encontradas nos períodos analisados, ou ainda, provocar o efeito contrário, ou seja, desassociando as duas variáveis.

(CAROSIA; COELHO; SILVA, 2019) também encontraram relações entre o sentimento predominante e os preços das ações em seu trabalho, entretanto, para considerar o engajamento os sentimentos foram ponderados por quantidade de curtidas e retuítes, não considerando outras opções que o Twitter possui (comentários, citações e quantidade de seguidores do autor do tuíte) e que foram utilizadas neste trabalho.

(SOUZA; LUCENA; QUEIROZ, 2019b) levantou a hipótese de que possíveis *big players* do mercado acionário poderiam ter maior influência na rede social, variando assim o preço das ações conforme suas opiniões. Entretanto, neste trabalho não encontrou-se usuários que realmente se destaquem. O destaque de algum perfil se daria através do volume de tuítes sobre determinado ativo com uma taxa de engajamento alta, o que não ocorreu, ao menos para os tuítes das ações selecionadas.

Apesar de possuir um viés diferente deste trabalho, em (NTI; ADEKOYA; WEYORI, 2020) é levantada uma dificuldade no desenvolvimento da proposta (predição do movi-

mento do mercado utilizando AS) devido ao fato do baixo volume de tuítes existentes. A mesma dificuldade foi encontrada neste trabalho, onde mais de 20 ativos foram descartados devido ao baixo volume de tuítes em um ano inteiro. Isso mostra que os investidores nesses países não compartilham suas opiniões sobre mercado. Dessa forma, utilizar outras fontes para analisar o movimento do mercado pode resultar em associações melhores.

O primeiro processo que pode ser melhorado é a análise de sentimento para obtenção de uma classificação melhor, levando em conta emojis não mapeados, como setas diagonais para indicar que o preço de um ativo aumentou ou diminuiu, pois muitos tuítes possuem apenas teor informativo, sendo classificados como neutros. Uma outra abordagem é focar em tuítes que envolvam alguma notícia relacionada a um ativo.

A detecção desses tuítes é complexa, pois pode envolver links externos à plataforma, mas como observou-se na Tabela 3.2, o uso individual de notícias não possuem impacto relevante nas negociações de ativos. As notícias podem não influenciar diretamente, mas opiniões expressas nas redes sociais referentes às notícias publicadas em diferentes veículos de comunicação podem impactar de alguma forma.

A taxa de engajamento não ajuda a definir um *big player* através de um único tuíte, mas se um conjunto de tuítes de um usuário sobre um ou mais ativos possui uma taxa relativamente alta (acima de 50% por exemplo), ele pode sim ser considerado um *player* com potencial para influenciar pessoas. Identificar esses usuários e focar na coleta de dados e análise desses perfis é outra linha que pode ser seguida.

Outro estudo que pode ser realizado é utilizar *clustering* hierárquico para identificar que ações estão mais próximas e verificar possíveis padrões em comum. Melhorias mais simples e diretas deste trabalho envolvem realizar o tratamento dos dados do preço das ações e taxa de engajamento com diferentes níveis do SAX, além de utilizar diversos parâmetros disponibilizados pelo algoritmo Apriori, por exemplo, verificar se o sentimento predominante consegue influenciar o preço em diferentes janelas de tempo.

Uma hipótese levantada, observando as regras encontradas, é que o sentimento negativo não possui muita influência no preço das ações, mas talvez possua no volume de negociações. (SOUZA; LUCENA; QUEIROZ, 2019b) afirmou haver relação direta entre os volumes de tuítes e de negociações no mercado financeiro, o mesmo pode se aplicar ao volume atrelado aos sentimentos predominantes.

Tanto este trabalho como o de (CAROSIA; COELHO; SILVA, 2019) e (SOUZA; LUCENA; QUEIROZ, 2019b) analisaram períodos de instabilidade e incertezas na economia no mercado brasileiro. Investigar associações e possíveis correlações entre o sentimento predominante e o preço das ações em um período de maior estabilidade também pode ser um caminho para trabalhos futuros, bem como o uso de outras redes sociais, como fóruns e Google Trends.

REFERÊNCIAS BIBLIOGRÁFICAS

ALKUBAISI, G.; KAMARUDDIN, S.; HUSNI, H. A systematic review on the relationship between stock market prediction model using sentiment analysis on twitter based on machine learning method and features selection. **Journal of Theoretical and Applied Information Technology**, v. 95, p. 6924–6933, 12 2017.

ALMEIDA, R. J. A. **LeIA - Léxico para Inferência Adaptada**. [S.l.]: GitHub, 2018. <https://github.com/rafjaa/LeIA>.

ALVES, D. S. **Uso de técnicas de Computação Social para tomada de decisão de compra e venda de ações no mercado brasileiro de bolsa de valores**. 2015. Tese (Tese) — UNIVERSIDADE DE BRASÍLIA, 2015.

BORGELT, C.; KRUSE, R. Induction of association rules: Apriori implementation. In: **COMPSTAT**. [S.l.: s.n.], 2002.

CAROSIA, A. E. O.; COELHO, G. P.; SILVA, A. E. A. Analyzing the brazilian financial market through portuguese sentiment analysis in social media. **Applied Artificial Intelligence**, v. 34, n. 1, p. 1–19, 2019.

CARRETE, L. S.; TAVARES, R. **Mercado Financeiro Brasileiro**. [S.l.]: Atlas, 2019.

DESAI, R. **How to scrape millions of tweets using snscrape**: A comprehensive guide to scraping tweets coherent with twitter. Medium, 2021. Último Último acesso em 09 nov. 2021. Disponível em: <https://medium.com/dataseries/how-to-scrape-millions-of-tweets-using-snscrape-195ee3594721>.

ES, S. **Sentiment Analysis in Python: TextBlob vs Vader Sentiment vs Flair vs Building It From Scratch**. NeptuneBlog, 2021. Último acesso em 01 nov. 2021. Disponível em: <https://neptune.ai/blog/sentiment-analysis-python-textblob-vs-vader-vs-flair>.

FALCãO, M. C. dos S. **ANÁLISE DE SENTIMENTO DE NOTÍCIAS DO MERCADO FINANCEIRO**. 2020. Dissertação (Dissertação) — UNIVERSIDADE FEDERAL DO RIO DE JANEIRO, 2020.

FU, T.-C. et al. Discovering the correlation between stock time series and financial news. In: . [S.l.]: IEEE, 2008. p. 880–883.

GILBERT, C. H. E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In: **Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf**. [S.l.: s.n.], 2014.

KHAN, Z. H.; ALIN, T. S.; HUSSAIN, M. A. Price prediction of share market using artificial neural network (ann). In: **International Journal of Computer Applications**. [S.l.]: IEEE, 2011. v. 22, n. 2, p. 0975–8887.

KINOW. **Natural Language Processing and Natural Language Understanding**. Kinow, 2017. Último Último acesso em 02 nov. 2021. Disponível em: <https://kinoshita.eti.br/2017/06/03/natural-language-processing-and-natural-language-understanding.html>.

LI, A. W.; BASTOS, G. S. Stock market forecasting using deep learning and technical analysis: A systematic review. **IEEE Access**, v. 8, p. 185232–185242, 2020.

MUNDO, B. N. **O país que já foi o 'menos desenvolvido do mundo' e hoje supera a China em crescimento.** UOL, 2020. Último Último acesso em 18 nov. 2021. Disponível em: <https://economia.uol.com.br/noticias/bbc/2020/03/05/o-pais-que-ja-foi-o-menos-desenvolvido-do-mundo-hoje-supera-a-china-em-crescimento.htm>.

NETO, A. A. **Mercado Financeiro**. [S.l.]: Atlas, 2021.

NOBREGA, M. F. da. **O mercado de capitais : sua importância para o desenvolvimento e os entraves com que se defronta no Brasil**. [S.l.]: BOVESPA, 2000. 46 p.

NTI, I. K.; ADEKOYA, A. F.; WEYORI, B. A. A systematic review of fundamental and technical analysis of stock market predictions. **Artificial Intelligence Review**, Springer Nature, v. 53, p. 3007–3057, 2019.

____. Predicting stock market price movement using sentiment analysis: Evidence from ghana. **Applied Computer Systems**, Sciendo, v. 25, n. 1, p. 33–42, 2020.

OLIVEIRA, L. L. de. **PREDIÇÃO DA POPULARIDADE DE TUÍTES UTILIZANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA.** 2018**.** Dissertação (Monografia) **—** UNIVERSIDADE FEDERAL DE SANTA MARIA, 2018.

PILLAT, V. G.; PILLAT, V. G. Comparação entre duas fórmulas utilizadas para o cálculo da taxa de engajamento utilizando como base a porcentagem de visualizações e o total de fãs. **Revista Brasileira de Pesquisas de Marketing, Opinião e Mídia**, v. 10, p. 298–302, 2017.

RAY, P.; GANGULI, B.; CHAKRABARTI, A. A hybrid approach of bayesian structural time series with lstm to identify the influence of news sentiment on short-term forecasting of stock price. **IEEE Transactions on Computational Social Systems**, PP, p. 1–10, 06 2021.

SACRAMENTO, G. **Entenda como funciona a transformação digital no mercado financeiro**. RockContent, 2021. Último acesso em 20 dez. 2021. Disponível em: <https://rockcontent.com/br/blog/transformacao-digital-mercado-financeiro/#:~:text=Enquanto%20outros%20nichos%20j%C3%A1%20experimentaram,e%20de%20prefer%C3%AAncia%20dos%20usu%C3%A1rios>.

SANTOS, H.; LAENDER, A.; PEREIRA, A. A twitter view of the brazilian stock exchange market. In: **International Conference on Electronic Commerce and Web Technologies.** [S.l.: s.n.], 2015. p. 112–123. ISBN 978-3-319-27728-8.

SAX-VSM. **Symbolic Aggregate approXimation**. 2016. Último Último acesso em 31 dez. 2022. Disponível em: <https://jmotif.github.io/sax-vsm_site/morea/algorithm/SAX.html>.

SHAH, P. **Sentiment Analysis using TextBlob**: Sentiment analysis using textblob and its working! Towards Data Science, 2020. Último acesso em 04 out. 2021. Disponível em: <https://towardsdatascience.com/my-absolute-go-to-for-sentiment-analysis-textblob-3ac3a11d524>.

SKUZA, M.; ROMANOWSKI, A. Sentiment analysis of twitter data within big data distributed environment for stock prediction. In: **Federated Conference on Computer Science and Information Systems (FedCSIS)**. [S.l.]: IEEE, 2015. p. 1349–1354.

SOUZA, D. M. S. de; LUCENA, W. G. L.; QUEIROZ, D. B. de. Decision-making simulator for buying and selling stock market shares based on twitter indicators and technical analysis. In: **International Conference on Systems, Man and Cybernetics (SMC)**. [S.l.]: IEEE, 2019.

____. O efeito do sentimento do investidor expresso via twitter sobre o comportamento do mercado acionário brasileiro durante o período eleitoral. In: **USP INTERNATIONAL CONFERENCE IN ACCOUNTING**. [S.l.]: IEEE, 2019.

APÊNDICE A – LISTA DAS 50 AÇÕES UTILIZADAS INICIALMENTE

Tabela A.1 – Tabela de ações

Ticker	Nome	Volume total
FNAM11.SA	Fundo de Investimento da Amazônia Closed Fund	–
FNOR11.SA	Fundo de Investimentos do Nordeste	–
PETR4.SA	Petróleo Brasileiro SA Petrobras Preference Shares	51.204
PETR3.SA	Petróleo Brasileiro SA Petrobras	11.110
B3SA3.SA	Brasil Bolsa Balcão	–
COGN3.SA	Cogna Educação	10.511
VVAR3.SA	Via Varejo On	19.590
OIBR3.SA	Oi	23.955
MGLU3.SA	Magazine Luiza	10.895
CIEL3.SA	Cielo	10.860
BBDC4.SA	Banco Bradesco SA Preference Shares	9.159
LAME4.SA	Lojas Americanas SA Preference Shares	–
HAPV3.SA	Hapvida Participações e Investimentos SA	–
ITUB4.SA	Itaú Unibanco	9.383
ITSA4.SA	Itausa SA Preference Shares	8.089
BBAS3.SA	Banco do Brasil SA	13.670
RAIZ4.SA	Raizen SA	–
EMBR3.SA	Embraer SA	8.305
BPAC11.SA	Banco TG Pactual SA Brazilian Units	3.987
JBSS3.SA	JBS	6.683
ABEV3.SA	Ambev	–
BRML3.SA	brMalls	–
BEEF3.SA	Minerva Foods	6.995
RAIL3.SA	Rumo SA	–
IRBR3.SA	IRB Brasil RE	13.609
CPLE6.SA	Companhia Paranaense de Energia	–
VIVR3.SA	Viver Incorporadora e Construtora SA	–
UGPA3.SA	Grupo Ultra	–
VALE3.SA	Vale SA	21.711
LREN3.SA	Lojas Renner	–
AMER3.SA	B2W Digital	–
BTOW3.SA	B2W Digital On	–
USIM5.SA	Usinas Siderúrgicas de Minas Gerais A Pref. Shs.	6.119

Continuação da Tabela A.1		
Ticker	**Nome**	**Volume de tuítes**
CCRO3.SA	Grupo CCR	–
BRDT3.SA	Vibra Energia	–
NTCO3.SA	Natura & Co Holding SA	–
PRIO3.SA	PetroRio	5.944
SOMA3.SA	Grupo de Moda Soma SA	–
CSNA3.SA	Companhia Siderúrgica Nacional	5.424
LINX3.SA	Linx On NM	–
CYRE3.SA	Cyrela	–
CSAN3.SA	Cosan SA	–
MRVE3.SA	MRV	–
LWSA3.SA	Locaweb	–
LCAM3.SA	Locamerica	–
RNEW1.SA	Renova On	–
WIZS3.SA	Wiz Soluções e Corretagem de Seguros SA	–
POSI3.SA	Positivo Tecnologia	–
BRF3.SA	BRF	–
SUZB3.SA	Suzano SA	–

APÊNDICE B – REGRAS ENCONTRADAS PELO APRIORI

Tabela B.1 – Regras para 8 ações em diferentes períodos de 2019

BBAS3.SA - 33 transações				
Antecedente	**Consequente**	**Suporte**	**Confiança**	**Lift**
I	Positivo	0.151	1	1.434
C	Positivo	0.151	0.714	1.02
BBDC4.SA - 48 transações				
D	Positivo	0.166	1	1.230
B	Positivo	0.145	0.875	1.076
C	Positivo	0.166	0.888	1.094
E	Positivo	0.166	0.727	0.895
ITSA4.SA - 48 transações				
D	Positivo	0.166	1	1.23
B	Positivo	0.145	0.875	1.07
C	Positivo	0.166	0.888	1.094
E	Positivo	0.166	0.727	0.895
ITUB4.SA - 33 transações				
I	Positivo	0.121	1	1.375
D	Positivo	0.151	0.714	0.982
MGLU3.SA - 39 transações				
I	Positivo	0.179	1	1.39
A	Negativo	0.102	0.800	2.83
H	Positivo	0.102	0.800	1.114
OIBR3.SA - 48 transações				
I	Positivo	0.145	0.875	1.20
A	Positivo	0.145	0.777	1.06
H	Positivo	0.187	0.750	1.02
PETR4.SA - 51 transações				
I	Positivo	0.136	1	1.466
F	Negativo	0.181	1	3.142
G	Positivo	0.227	1	1.466
D	Positivo	0.103	0.750	0.144
B	Positivo	0.172	0.714	1.090
USIM5.SA - 31 transações				
H	Positivo	0.161	0.833	1.435

Tabela B.2 – Regras para 9 ações em diferentes períodos de 2020

BBAS3.SA - 30 transações				
Antecedente	**Consequente**	**Suporte**	**Confiança**	**Lift**
B	Positivo	0.166	1	1.875
BBDC4.SA - 51 transações				
C	Positivo	0.313	0.800	1.275
J	Negativo	0.156	0.727	1.952
CIEL3.SA - 38 transações				
B	Positivo	0.157	0.857	1.357
COGN3.SA - 56 transações				
D	Positivo	0.142	0.800	1.200
F	Positivo	0.125	0.777	1.088
A	Positivo	0.125	0.777	1.088
CSNA3.SA - 44 transações				
B	Positivo	0.113	0.714	1.257
F	Positivo	0.181	0.800	1.408
ITUB4.SA - 47 transações				
I	Positivo	0.212	1	1.382
B	Positivo	0.255	0.705	0.975
MGLU3.SA - 42 transações				
J	Positivo	0.142	0.857	1.285
I	Positivo	0.166	0.777	1.166
OIBR3.SA - 43 transações				
E	Positivo	0.186	0.888	1.158
B	Positivo	0.255	0.785	1.023
F	Positivo	0.116	0.714	0.930
J	Positivo	0.116	0.714	0.930
VALE3.SA - 31 transações				
E	Positivo	0.129	1	1.291
I	Positivo	0.225	0.875	1.130
C	Positivo	0.161	0.714	0.922

Tabela B.3 – Regras para 9 ações em diferentes períodos de 2021

BBAS3.SA - 38 transações				
Antecedente	**Consequente**	**Suporte**	**Confiança**	**Lift**
D	Positivo	0.184	0.7	1.156
BBDC4.SA - 56 transações				
J	Positivo	0.160	0.9	1.172
C	Positivo	0.178	0.833	1.082
B	Positivo	0.125	0.7	0.911
BEEF3.SA - 52 transações				
C	Positivo	0.115	1	1.268
F	Positivo	0.134	1	1.268
CIEL3.SA - 38 transações				
J	Positivo	0.184	1	1.187
C	Positivo	0.157	0.857	1.017
I	Positivo	0.105	0.800	0.950
B	Positivo	0.210	0.727	0.863
COGN3.SA - 35 transações				
H	Positivo	0.114	1	1.458
C	Positivo	0.228	1	1.458
ITUB4.SA - 51 transações				
D	Positivo	0.176	0.9	1.639
MGLU3.SA - 29 transações				
C	Positivo	0.103	1	1.260
H	Positivo	0.137	1	1.260
I	Positivo	0.241	0.875	1.103
OIBR3.SA - 31 transações				
D	Positivo	0.129	0.800	1.127
PETR3.SA - 34 transações				
F	Negativo	0.147	0.714	1.278
G	Negativo	0.147	0.714	1.278