INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Redes sociais como método de previsão para as decisões dos investidores - caso *Tesla*

Ana Marta Ferreira Alcobia

Mestrado em Business Analytics

Orientadora:
Professora Doutora Márcia Rafaela Cadete dos Santos
Professora Auxiliar, Universidade Lusófona, e investigadora Integrada da BRU - IUL - Business Research Unit (ISCTE- Instituto Universitário de Lisboa)

Outubro, 2023

BUSINESS
SCHOOL

Departamento de Métodos Quantitativos para Gestão e Economia

Redes sociais como método de previsão para as decisões dos investidores - caso *Tesla*

Ana Marta Ferreira Alcobia

Mestrado em Business Analytics

Orientadora:
Professora Doutora Márcia Rafaela Cadete dos Santos
Professora Auxiliar, Universidade Lusófona, e investigadora Integrada da BRU - IUL - Business Research Unit (ISCTE- Instituto Universitário de Lisboa)

Outubro, 2023

Aos meus pais, ao meu irmão e ao meu namorado.

Agradecimento

Esta dissertação não seria possível sem o apoio da minha orientadora a quem deixo uma palavra de agradecimento, Professora Márcia Santos. Também não poderia deixar de agradecer a todos os professores que contribuíram para o aumento dos meus conhecimentos, mas especialmente ao Professor Raul Laureano, a quem deixo um especial agradecimento por todo o enriquecimento a nível pessoal e profissional que proporcionou.

Em especial deixo também um agradecimento aos meus pais e ao meu irmão por toda a compreensão, por me motivarem e, serem pilares fundamentais para a minha vida.

Por fim, quero também agradecer ao meu namorado e aos meus amigos que me apoiaram e deram-me o positivismo necessário para alcançar os meus objetivos.

Resumo

As redes sociais desempenham hoje em dia um papel significativo na disseminação de opiniões e informações sobre empresas, marcas, serviços e produtos. As grandes empresas, em particular, por terem mais presença no mercado são mais interessantes e atraem um interesse massivo do público. O presente estudo procura compreender em que medida o sentimento expresso nas publicações e comentários nas redes sociais pode servir de barómetro para a previsão das decisões dos investidores, focando-se no caso da *Tesla*.

Assim, foram analisadas as informações de mercado do valor da empresa e o sentimento dos comentários e publicações na rede social *Twitter*. Essa análise passou por vários procedimentos. Inicialmente uma revisão sistemática da literatura, com o auxílio de duas bases de dados científicas, a *web of science* e a *scopus*, com a utilização de uma *query*, e analisados 28 artigos científicos sobre o tema.

A metodologia utilizada foi a CRISP-DM. Foi estudada a compreensão do negócio e dos dados, definindo o problema, os objetivos de estudo e a recolha e compreensão dos dados. Os dados reportavam-se ao ano 2018. Seguidamente os dados foram preparados, limpos e transformados para as fases seguintes. Nas fases de modelação e avaliação, criou-se os modelos e avaliou-se, retirando-se conclusões. Por fim, a implementação passou pela escrita desta dissertação.

Pelos resultados conclui-se que existe uma relação entre a variação do valor das ações e o sentimento dos *tweets*, isto é, quando existe um grande número de *tweets* positivos irá aumentar o valor da ação da *Tesla*, e o contrário, nos *tweets* negativos.

Palavras chave: valor da ação; análise de sentimentos; *Twitter*

(JEL Classification System: G12, E30)

Abstract

Nowadays, social networks play a significant role in disseminating opinions and information regarding companies, brands, products and services. Especially large companies, since their market presence is bigger, therefore luring a larger number of stakeholders, attract massive public interest. This study seeks to understand to what extent the sentiment expressed in social media posts and comments can be used for predicting investor decisions, focusing on the Tesla case.

To this end, market information on the company's value and the sentiment derived from the social network Twitter engagements were analysed. In order to develop this study, several procedures were carried out. First, a systematic literature review was conducted using two scientific databases, Web of Science and Scopus, using the same query for both, where 28 scientific articles related to the topic under study were revised.

The methodology employed was CRISP-DM. A business understanding was executed as well as the assessment of the data, where were defined the problem, the objectives of the research and the collection and understanding of the data, including information from January to December 2018. The data was then prepared, cleaned and transformed for the following phases. In the modelling and evaluation phases, the models were created and evaluated, drawing conclusions. Finally, the implementation involved writing this dissertation.

The results show that there is a relationship between change in share value and the sentiment of the tweets, i.e. the larger the number of positive tweets the greater increase in Tesla’s value, and the same linear relation applies for the negative ones.

Key words: share value; sentiment analysis; twitter
(JEL Classification System: G12, E30)

Índice

Agradecimento v

Resumo v

Abstract vii

Índice de figuras xii

Índice de tabelas xiv

Lista de Siglas e Abreviaturas xvii

Capítulo 1: Introdução 19

1.1. Contexto 19

1.2. Twitter 20

1.3. Tesla 21

1.4. Contributos 22

1.5. Estrutura da dissertação 22

Capítulo 2: Revisão da Literatura 23

2.1. Protocolo para a revisão sistemática da literatura 23

2.2. Análise de sentimentos 28

2.3. Cotação da empresa 30

2.4. Contexto de aplicação e Avaliação dos Resultados 30

2.4.1. Contextualização dos artigos 31

2.4.2. Contextualização dos artigos relativos à parte financeira 32

2.4.3. Contextualização dos artigos para a análise de sentimentos 36

2.4.4. Avaliação dos Resultados dos artigos 37

2.5. Avaliação dos artigos científicos relevantes 39

2.6. Contributos da RSL 40

Capítulo 3: Metodologia 41

3.1. Metodologia Utilizada 42

3.1.1. Compreensão do Negócio 42
3.1.2. Compreensão dos Dados 42
3.1.3. Preparação dos Dados 43
3.1.4. Modelação 45
3.1.5. Avaliação 47
3.1.6. Implementação 47
Capítulo 4: Resultados e Discussão 49
Capítulo 5: Conclusões e Recomendações 55
5.1. Conclusão 55
5.2. Limitações 57
5.3. Trabalho futuro 57
Capítulo 6: Referências Bibliográficas 59
Capítulo 7: Anexos 63
7.1. Anexo A 63
7.2. Anexo B 65
7.3. Anexo C 77

Índice de figuras

Figura 2.1: Processo de seleção dos artigos para revisão 25
Figura 3.1: Procedimento de análise de dados no *Orange3* 44
Figura 3.2: Gráfico Regressão Linear 46
Figura 4.1: Nuvem de palavras para sentimento Positivo 51
Figura 4.2: Nuvem de palavras para sentimento Negativo 51
Figura 4.3: Gráfico com a comparação entre *Tweet tone* e Variação do valor da ação 52
Figura 4.4: Gráfico com a comparação entre *Tweet tone* e Variação do valor da ação (3 situações) .. 54

Índice de tabelas

Tabela 2.1: Critérios de inclusão e de exclusão 24
Tabela 2.2: Artigos incluídos na Revisão Sistemática 25
Tabela 2.3: Artigos incluídos na Revisão Sistemática (continuação) 26
Tabela 2.4: Artigos incluídos na Revisão Sistemática (continuação) 27
Tabela 2.5: Critério de qualidade para avaliação dos artigos 28
Tabela 2.6: Análise e metodologia dos estudos 33
Tabela 2.7: Análise e metodologia dos estudos (continuação) 34
Tabela 2.8: Avaliação dos resultados dos artigos na RSL 38
Tabela 2.9: Avaliação da qualidade dos artigos 39
Tabela 3.1: Divisão dos sentimentos, 1º semestre 46
Tabela 3.2: Divisão dos sentimentos, 2º semestre 46
Tabela 4.1: Divisão dos sentimentos sem sentimento neutro, 1º semestre 49
Tabela 4.2: Divisão dos sentimentos sem sentimento neutro, 2º semestre 49
Tabela 4.3: Lista das línguas dos comentários, 1º semestre 50
Tabela 4.4: Lista das línguas dos comentários, 2º semestre 50
Tabela 7.1: Contextualização dos artigos da RSL 63
Tabela 7.2: Contextualização dos artigos da RSL (continuação) 64
Tabela 7.3: Contextualização dos artigos da RSL para a parte financeira 65
Tabela 7.4: Contextualização dos artigos da RSL para a parte financeira (continuação) 66
Tabela 7.5: Contextualização dos artigos da RSL para a parte financeira (continuação) 67
Tabela 7.6: Contextualização dos artigos da RSL para a parte financeira (continuação) 68
Tabela 7.7: Contextualização dos artigos da RSL para a parte financeira (continuação) 69
Tabela 7.8: Contextualização dos artigos da RSL para a parte financeira (continuação) 70
Tabela 7.9: Contextualização dos artigos da RSL para a parte financeira (continuação) 71
Tabela 7.10: Contextualização dos artigos da RSL para a parte financeira (continuação) 72
Tabela 7.11: Contextualização dos artigos da RSL para a parte financeira (continuação) 73
Tabela 7.12: Contextualização dos artigos da RSL para a parte financeira (continuação) 74
Tabela 7.13: Contextualização dos artigos da RSL para a parte financeira (continuação) 75
Tabela 7.14: Contextualização dos artigos da RSL para a parte financeira (continuação) 76
Tabela 7.15: Contextualização dos artigos da RSL para a análise de sentimentos 77
Tabela 7.16: Contextualização dos artigos da RSL para a análise de sentimentos (continuação) 78
Tabela 7.17: Contextualização dos artigos da RSL para a análise de sentimentos (continuação) 79
Tabela 7.18: Contextualização dos artigos da RSL para a análise de sentimentos (continuação) 80

Tabela 7.19: Contextualização dos artigos da RSL para a análise de sentimentos (continuação)..... 81

Lista de Siglas e Abreviaturas

AIC	Critério de informação de *Akaike*
ARDL	Desfasamento distribuído autorregressivo
CRISP-DM	*Cross Industry Standard Process for Data Mining*
DJIA	Índice *Dow Jones Industrial Average*
EDA	Teste de ajuste de dependência bruta
ELM	Máquinas de Aprendizagem Extrema
HML	*High Minus Low*
HQ	Critério de *Hannan Quinn*
LSSVR	Modelos de Regressão Vetorial de Suporte de Mínimos Quadrados
LSTM	Modelos de Aprendizagem de Máquina
MAE	Erro Absoluto Médio
MAPE	Erro Percentual Absoluto Médio
MOM	*Fator Momentum*
MSE	Erro quadrático médio
NLP	*Natural Language Pre-processing*
NMAE	Erro Absoluto Médio Normalizado
Psize	Portfólios formados em tamanho
PSM	Correspondência de pontuação de propensão
RMRF	Excesso de retorno no mercado
RNN	Rede Neuronal Recorrente
R^2	Coeficiente de determinação
RSL	Revisão Sistemática da Literatura
SC	Critério de *Schwarz*
SMB	*Small Minus Big*
SSWN	Modelo *StockSentiWordNet*
SVM	*Support Vector Machine*
VADER	*Valence Aware Dictionary and sentiment Reasoner*
VAR	Vetor Autorregressivo
VIX	*Volatility Index*
WAPE	Erro Percentual Absoluto Médio Ponderado

Capítulo 1: Introdução

1.1. Contexto

Com o crescimento da Internet 2.0, as redes sociais são uma importante fonte de *big data*, uma vez que os utilizadores utilizam grande parte do seu tempo em serviços de redes sociais. Realizar uma análise destes dados poderá permitir uma compreensão mais profunda do comportamento dos utilizadores (Oliveira et al., 2017). As redes sociais são uma forma atual para a obtenção de informações valiosas por meio da análise de informações online ou comentários dos utilizadores (Pai & Liu, 2018).

As redes sociais tornaram-se um importante canal de informação em todo o mundo e definem-se como um conjunto de aplicativos baseados na Internet que se baseiam na *Web 2.0* e que permitem a criação e a troca de conteúdos entre os utilizadores, conteúdos esses que ajudam as empresas a tomar decisões (Yang et al., 2015).

Com o desenvolvimento das novas tecnologias, verificou-se um aumento de volume, velocidade e variedade dos dados disponíveis para os investigadores (Hamraoui & Boubaker, 2022). Muitos destes investigadores demonstram que o humor ou os sentimentos expressos nas redes sociais estão relacionados com os mercados financeiros (Ruan et al., 2018).

Assim sendo, no presente estudo o objetivo principal passou pela previsão do impacto que as publicações em redes sociais podem ter na decisão dos investidores, através da análise do valor das ações da *Tesla* e do sentimento das publicações na rede social *Twitter*. Com este fim, relacionou-se os *tweets*, com as flutuações das ações da *Tesla*.

Para a realização deste estudo é importante ter em conta que cerca de 70% das emissões de CO_2 do planeta são provenientes do carvão e do petróleo do setor automóvel, por isso, foi necessária a procura de energias renováveis e tecnologias que minimizassem os desafios climáticos. Assim, a alternativa encontrada, no setor automóvel, foram os carros elétricos, a bateria. Quando estes carros elétricos apareceram tinham algumas desvantagens, como: (i) a baixa distância percorrida com uma carga de bateria; e, (ii) a baixa velocidade alcançada. Os carros a combustão continuaram a dominar o mercado. Atualmente, com o desenvolvimento das baterias mais eficientes e duradouras, que utilizam ião lítio e com a emergência da descarbonização para alcançar as metas das mudanças climáticas propostas pelo Comitê de Mudanças Climáticas, os carros elétricos ganharam um lugar no mercado, sendo que muitos países, já tentam acabar com a circulação de carros a combustão (Arangues et al., 2022).

As empresas fabricantes de carros elétricos que mais se destacam são: a *Tesla*, líder de mercado em venda de veículos elétricos, seguida pela *BYD Auto* e pela *SGMW* (estas duas últimas marcas chinesas fabricantes de automóveis) e, por fim, a *Volkswagen* e a *BMW*.

Para a realização deste estudo é importante ter em conta que a *Tesla* é uma empresa de veículos elétricos fundada, não só mas também, por *Elon Musk* em 2003, e que o mesmo é atualmente o diretor executivo da empresa (Nyakurukwa, 2022). Para além da *Tesla*, *Elon Musk* também é dono de algumas mais empresas como a *SpaceX*[1], o *PayPal*[2], a *Neuralink*[3], a *SolarCity*[4] e o *Twitter.*

De forma a estruturar melhor e responder ao objetivo proposto foi definida uma questão de investigação: "Existe uma relação de previsão entre o valor da ação da *Tesla* e o sentimento das publicações na rede social *Twitter*?".

1.2. Twitter

O *Twitter* é uma rede social que serve como ferramenta de comunicação que permite aos seus utilizadores publicar mensagens de 280 caracteres, por cada *tweet* para todos os seus seguidores, permitindo aos seus utilizadores compartilhar e divulgar as suas opiniões (Ruan et al., 2018). O *Twitter* explora a brevidade das mensagens enviadas (Nisar & Yeung, 2018).

O *Twitter* mostrou ser importante para os produtos como filmes, músicas e jogos eletrónicos, pois, estes são os produtos que, normalmente, o sucesso ou o insucesso é instantâneo, nas redes sociais. O *Twitter* transmite, em tempo real, as informações e avaliações sobre a qualidade do produto e, por isso, permite que haja um *feedback* mais rápido. Por ser uma rede social permite que haja um grande número de pessoas a receber essa informação e por isso um *feedback* maior (Nisar & Yeung, 2018).

Com a existência de interfaces de programação de aplicativos (*APIs*) abertas e de fácil utilização, o *Twitter* é muito utilizado como plataforma para analisar e prever as atividades do mercado de ações. Para além dos trabalhos académicos em que existe um grande uso do *Twitter*, esta rede social também é muito utilizada pelas empresas, de modo a interagir com os seus investidores e clientes (Ruan et al., 2018).

Alguns estudos científicos dizem que os utilizadores desta rede costumam publicar noticias e artigos financeiros e compartilhar as suas opiniões no *Twitter*, tornando-se uma rede social amplamente utilizada para analisar e prever os retornos das ações (Ruan et al., 2018).

[1] Empresa focada no desenvolvimento de equipamentos e veículos para transporte espacial
[2] Empresa que faz transações financeiras ao redor do mundo. Mas que em 2022 foi vendida à *Ebay*
[3] Cria implantes para o cérebro humano, de forma a potencializar a atividade do orgão e trata muitas doenças
[4] Empresa de fornecimento de energia solar

Para a utilização do *Twitter* como método de análise do mercado de ações, isto é, os *tweets*, são primeiro analisados por ferramentas de análise de sentimentos para extrair o sentimento inerente e, posteriormente, os sentimentos dos *tweets* são agregados. Estes sentimentos agregados são utilizados para a análise do mercado financeiro(Ruan et al., 2018).

O *Twitter* é uma das principais plataformas de rede sociais e foi considerado como tendo um papel influente para os mercados financeiros através de várias publicações académicas e profissionais nos últimos anos. Em 2012 tinha 517 milhões de utilizadores, que colocavam 340 milhões de *tweets* por dia. Ao longo dos anos, o *Twitter* conseguiu tornar-se um dos 15 principais sites com maior volume. Como prova disso, um estudo demonstrou que 77% das empresas da *Fortune500* tinham contas corporativas no *Twitter* com *tweets* ativos e as empresas do setor da banca tinham 94% de presença ativa no *Twitter*. Assim, verificou-se a existência de uma tendência crescente para as redes sociais terem influência nos mercados (Yang et al., 2015).

1.3. Tesla

A *Tesla* é uma empresa de veículos elétricos fundada por *Elon Musk, JB Straubel, Martin Eberhard, Marc Tarpenning* e *Ian Wright* em 2003. O primeiro é atualmente o diretor executivo da empresa (Nyakurukwa, 2022). Tornou-se a empresa especializada na produção de veículos elétricos utilizando baterias de ião lítio para o armazenamento de energia e auxiliar na instalação de painéis solares. (Liu, 2021).

Antigamente os veículos necessitavam de combustíveis fosseis para o fornecimento da sua energia. Com o aparecimento da *Tesla* começou a ser utilizada a energia renovável para fornecer energia aos carros e isso fez com que a *Tesla* tivesse um rápido desenvolvimento devido ao seu papel influente no mercado de veículos elétricos e inovações tecnológicas avançadas (Liu, 2021).

O rápido crescimento da *Tesla* marcou o início de uma nova era na indústria automóvel, onde muitas empresas foram ofuscadas pelo sucesso da *Tesla* e muitas outras tiveram de gastar biliões de dólares para alcançar a *Tesla*. Apesar de muitas indústrias automóveis terem tentado, a *Tesla* ainda é líder. Em primeiro lugar, por ser a empresa fundadora no algoritmo de direção automática através da aprendizagem de máquina para veículos elétricos, por ser a primeira a obter a produção de baterias potentes a custos mais baixos e, por a sua bateria durar mais tempo. O Modelo S pode viajar até 380 milhas, isto é, mais de 611 quilómetros, o que até agora nenhum outro veículo elétrico conseguiu. Tudo isto leva a que a *Tesla* tenha um domínio indiscutível no mercado de veículos elétricos tendo quase 60% dos veículos elétricos vendidos nos Estados Unidos (Liu, 2021).

Recentemente, o diretor executivo da *Tesla*, *Elon Musk*, lançou uma sondagem na plataforma *Twitter* propondo vender um décimo da sua participação na *Tesla* se os utilizadores aprovassem a

decisão. Como a decisão foi aprovada por quase 60%, as ações da *Tesla* caíram 5% em poucas horas e representaram mais de 10% de todas as transações das empresas americanas (Nyakurukwa, 2022).

1.4. Contributos

O presente estudo contribui para a compreensão da previsão do impacto que as publicações em redes sociais podem ter na decisão dos investidores, através da análise do valor das ações da *Tesla* e do sentimento das publicações na rede social *Twitter*. Com este fim, relacionou-se os *tweets*, com as flutuações das ações da *Tesla*.

Através do estudo realizado, percebeu-se que esta dissertação veio colmatar um *gap* existente relativo a esta temática, em que apenas um artigo da Revisão Sistemática da Literatura abordava o tema da indústria automóvel, e assim, culminou na produção desta dissertação, escrita em português e baseada em dados retirados da plataforma *Kangle*.

Este estudo foi uma pesquisa importante na medida em que ajudou a perceber se existia uma forma de previsão para a valorização ou desvalorização do valor da ação da *Tesla* através da análise dos conteúdos da rede social *Twitter*.

1.5. Estrutura da dissertação

Posteriormente a esta introdução, no Capítulo 2 é apresentada a Revisão Sistemática da Literatura (RSL), que inclui o protocolo, o processo de pesquisa considerado na RSL e esquematizados todos os conteúdos extraídos dos artigos e, por isso, relevante para ter em atenção no estudo. Por fim, foi feita a avaliação dos artigos presentes na RSL, as suas limitações e contributos para possíveis próximos estudos.

No Capítulo 3, é descrita a metodologia utilizada, a metodologia CRISP-DM (*Cross Industry Standard Process for Data Mining*). Esta é dividida em seis fases e enumeradas as tarefas realizadas em cada uma dessas fases: Compreensão do Negócio, Compreensão dos Dados, Preparação dos Dados, Modelação, Avaliação e Implementação.

No Capítulo 4, são demonstrados todos os resultados obtidos, e feita a avaliação e discussão desses resultados.

No Capítulo 5 são apresentadas as principais conclusões ao estudo, as limitações encontradas e recomendações para futuros estudos.

No final deste estudo, são apresentadas as referências e anexos.

Capítulo 2: Revisão da Literatura

A revisão da literatura aborda a associação que os *tweets* publicados na rede social *Twitter* têm no valor da empresa. Para isso é realizado um protocolo de revisão sistemática da literatura e enumerados todos os passos para obtenção dos artigos analisados. Posteriormente são feitas as análises necessárias de modo a perceber se existem ou não *gaps*.

Para fazer este estudo, é importante ter em conta que, o valor das ações pode variar por diversos fatores. Verificou-se que o valor/cotação da ação pode ser influenciado por catástrofes, como por exemplo, o Crash de 1929 (Mourão, 2018). Para além disso existem diversos fatores que poderão afetar o valor das ações, tais como a inflação, a especulação, onde se utilizam as oscilações dos preços das ações em proveito próprio, e a queda de uma bolsa de valor, isto é, se a bolsa de Lisboa tiver uma queda, a bolsa de *New York* irá ser informada (Jawid, 2022).

2.1. Protocolo para a revisão sistemática da literatura

A revisão sistemática da literatura trata-se de uma investigação científica onde existe uma questão principal e um protocolo predefinido para identificar, criticar e resumir diferentes estudos de pesquisa primária (Baker & Weeks, 2014). O protocolo de revisão sistemática deve apresentar: (i) as bases de dados bibliográficas consultadas; (ii) a *query* utilizada; (iii) o processo de seleção dos artigos, abrangendo os critérios de inclusão e exclusão dos artigos; (iv) a avaliação da qualidade dos mesmos (Abrantes, 2020).

O objetivo desta revisão da literatura é caracterizar os estudos científicos que abordem a relação dos comentários que estão presentes na rede social *Twitter* com o preço das ações das empresas. A questão de investigação decorrente deste objetivo formula-se da seguinte forma: “Como é que os comentários na rede social *Twitter* podem prever o valor da empresa?”. Mas, para isso, devem-se fazer as seguintes questões:

- Qual o objetivo do estudo?
- Como é definido o sentimento da empresa cotada?
- Como é avaliada a cotação da empresa?
- Em que contexto são utilizados os diferentes modelos?
- Como são avaliados os resultados da relação entre o sentimento e a cotação da empresa?

Para a pesquisa de artigos foi utilizada a base de dados científica *scopus*[5], por ser reconhecida como o maior banco de dados de resumos e citações bibliográficas, abrangendo uma ampla gama de assuntos (Md Khudzari et al., 2018). Apesar de ser utilizada a base científica *scopus*, e de modo a obter um maior número de artigos, também foi utilizada a base de dados científica *web of science* [6] , por ser uma base científica muito abrangente, revelante por incluir publicações importantes nas áreas científicas de investigação e por ser muito utilizada nas revisões de literatura (Abrantes, 2020). Assim, para fazer esta pesquisa foi utilizada uma *query* no campo título de modo a ser uma procura mais eficaz.

De modo a ser uma pesquisa mais adequada ao estudo e de modo a procurar o que é mesmo necessário para o estudo, foi necessário pesquisar quais as *stopwords* que dariam os resultados esperados. Foi importante colocar a *stopword Twitter* porque, sendo o trabalho sobre o *Twitter* é importante ter apenas estudos onde é utilizada esta rede social, de forma a entendermos a influência que esta tem no valor da ação da empresa. Assim, são utilizadas as *stopwords: stock market, share value* e *stock price.* Utilizou-se o asterismo de modo aos resultados aparecerem tanto no singular como no plural. São apenas usados estes três sinónimos porque não foi encontrado mais nenhum sinónimo que influenciasse os resultados obtidos.

Para definição da *query* de pesquisa dos artigos que se enquadram no âmbito do objetivo deste trabalho, a proposta formulada foi discutida com a orientadora e com um investigador externo. A *query* utilizada foi: *TITLE (twitter AND ("stock market*" OR "share* value" OR "stock price*"))* e através da qual foram encontrados 82 documentos na *scopus* e 61 na *web of science*, aos quais foi necessário analisar os critérios de inclusão e exclusão que estão presentes na Tabela 2.1.

Tabela 2.1: Critérios de inclusão e de exclusão

Critérios de inclusão	Artigos que abordam ações e que falam da rede social *Twitter*
Critérios de exclusão	Documentos de conferências, capítulos de livros e *reviews;*
	Artigos que não estejam disponíveis para *download*;
	Artigos que não abordam o assunto em causa;

[5] *https://www.scopus.com*
[6] *https://www.webofknowledge.com*

A aplicação dos critérios de inclusão e exclusão obrigou à leitura do resumo, levando à exclusão de 115 documentos, sendo o processo demonstrado na Figura 2.1.

As tabelas seguintes, Tabela 2.2, Tabela 2.3 e Tabela 2.4, mencionam os 28 artigos utilizados no estudo.

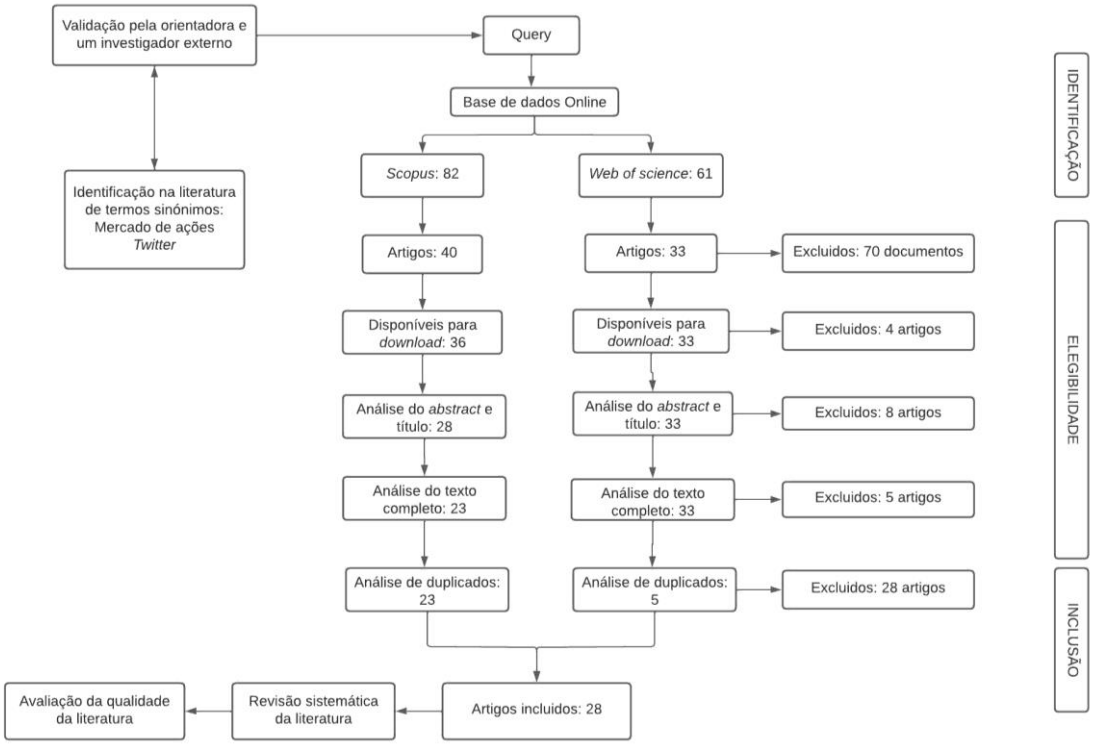

Figura 2.1: Processo de seleção dos artigos para revisão

Tabela 2.2: Artigos incluídos na Revisão Sistemática

ID	Ano	Título	Autores	Journal	Quartil
1	2023	*Are Twitter sentiments during COVID-19 pandemic a critical determinant to predict stock market movements? A machine learning approach*	Jena, P.R., Majhi, R.	*Scientific African*	Q1
2	2023	*Twitter sentiment and stock market: a COVID-19 analysis*	Katsafados, A.G., Nikoloutsopoulos, S., Leledakis, G.N.	*Journal of Economic Studies*	Q1
3	2022	*A Machine Learning Method for Prediction of Stock Market Using Real-Time Twitter Data*	Albahli, S., Irtaza, A., Nazir, T., Mehmood, A., Alkhalifah, A., Albattah, W.	*Electronics (Switzerland)*	Q2
4	2022	*Impact of Twitter sentiment on stock price returns*	Hamraoui, I., Boubaker, A.	*Social Network Analysis and Mining*	Q1

Tabela 2.3: Artigos incluídos na Revisão Sistemática (continuação)

ID	Ano	Título	Autores	Journal	Quartil
5	2022	*User Classification and Stock Market-Based Recommendation Engine Based on Machine Learning and Twitter Analysis*	Achyutha, P.N., Chaudhury, S., Bose, S.C., Kler, R., Surve, J., Kaliyaperumal, K.	*Mathematical Problems in Engineering*	Q2
6	2022	*Does Twitter Affect Stock Market Decisions? Financial Sentiment Analysis During Pandemics: A Comparative Study of the H1N1 and the COVID-19 Periods*	Valle-Cruz, D., Fernandez-Cortez, V., López-Chau, A., Sandoval-Almazán, R.	*Cognitive Computation*	Q1
7	2022	*The wisdom of the Twitter crowd in the stock market: Evidence from a fragile state*	Nyakurukwa, K; Seetharam, Y	*African Review of Economics and Finance-Aref*	NA
8	2022	*Brazilian stock market performance and investor sentiment on Twitter*	Souza, D.M.S., Martins, O.S.	*Revista de Gestao*	Q3
9	2021	*Can Social Media Predict Soccer Clubs' Stock Prices? The Case of Turkish Teams and Twitter*	Tanaltay, A., Langroudi, A.S., Akhavan-Tabatabaei, R., Kasap, N.	*SAGE Open*	Q2
10	2020	*Influence of Twitter activity on the stock price of soccer clubs*	Sóti, A., Ancarani, A., Stéger, J., Vattay, G.	*Social Network Analysis and Mining*	Q1
11	2020	*Does social media elevate supply chain importance? An empirical examination of supply chain glitches, Twitter reactions, and stock market returns*	Schmidt, C.G., Wuttke, D.A., Ball, G.P., Heese, H.S.	*Journal of Operations Management*	Q1
12	2020	*Semantic analysis: Can MD&A word choice increase stock price of american publicly traded banks disseminating using twitter?*	Boylan, D.H., Thomas, M., Matangi, E., Brown, M.	*International Journal of Digital Accounting Research*	Q2
13	2020	*Under his thumb the effect of president Donald Trump's Twitter messages on the US stock market*	Brans, H., Scholtens, B.	*PLoS ONE*	Q1
14	2020	*Emotions in Twitter communication and stock prices of firms: the impact of Covid-19 pandemic*	Dhar, S; Bose, I	*DECISION*	Q1
15	2020	*Does pre-processing affect the correlation indicator between Twitter message volume and stock market trading volume?*	Michalak, J	*Ekonomia I Prawo-Economics and Law*	NA
16	2019	*WHAT TWITTER CAN TELL US ABOUT THE STOCK MARKET*	Hazem, SM; Mohamed, EKA; Ali, H	*ANDULI*	NA
17	2018	*Impact of social media strategies on stock price: the case of Twitter*	Chahine, S., Malhotra, N.K.	*European Journal of Marketing*	Q1

Tabela 2.4: Artigos incluídos na Revisão Sistemática (continuação)

ID	Ano	Título	Autores	Journal	Quartil
18	2018	*Twitter as a tool for forecasting stock market movements: A short-window event study*	Nisar, T.M., Yeung, M.	*Journal of Finance and Data Science*	Q1
19	2018	*Using Twitter trust network for stock market analysis*	Ruan, Y., Durresi, A., Alfantoukh, L.	*Knowledge-Based Systems*	Q1
20	2018	*Predicting vehicle sales by sentiment analysis of twitter data and stock market values*	Pai, P.-F., Liu, C.-H.	*IEEE Access*	Q1
21	2018	*Conceptual framework for stock market classification model using sentiment analysis on twitter based on Hybrid Naïve Bayes Classifiers*	Alkubaisi, G.A.A.J., Kamaruddin, S.S., Husni, H.	*International Journal of Engineering and Technology(UAE)*	NA
22	2017	*The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices*	Oliveira, N., Cortez, P., Areal, N.	*Expert Systems with Applications*	Q1
23	2016	*The Wisdom of Twitter Crowds: Predicting Stock Market Reactions to FOMC Meetings via Twitter Feeds*	Azar, PD; Lo, AW	*JOURNAL OF PORTFOLIO MANAGEMENT*	Q1
24	2015	*Twitter financial community sentiment and its predictive relationship to stock market movement*	Yang, S.Y., Mo, S.Y.K., Liu, A.	*Quantitative Finance*	Q1
25	2015	*The effects of twitter sentiment on stock price returns*	Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M., Mozetič, I.	*PLoS ONE*	Q1
26	2015	*Using Twitter to Predict the Stock Market: Where is the Mood Effect?*	Nofer, M., Hinz, O.	*Business and Information Systems Engineering*	Q1
27	2014	*Analysis of twitter users' mood for prediction of gold and silver prices in the stock market*	Porshnev, A., Redkin, I.	*Communications in Computer and Information Science*	Q4
28	2011	*Twitter mood predicts the stock market*	Bollen, J., Mao, H., Zeng, X.	*Journal of Computational Science*	Q1

NA: Não aplicável

Após a seleção dos artigos, foi necessário avaliar a sua qualidade, de acordo com alguns critérios pertencentes a cada questão mencionada acima, conforme Tabela 2.5.

Tabela 2.5: Critério de qualidade para avaliação dos artigos

Objetivo do estudo	Q1: Define claramente o objetivo?
	Q2: O objetivo implica comparar com resultados de outros estudos?
Sentimento da empresa cotada	Q3: O sentimento é claramente definido?
	Q4: São comparados vários conceitos de sentimento?
Cotação da empresa	Q5: O modo como é medida a cotação da empresa é justificada e clara?
	Q6: O estudo compara vários modelos para ver a cotação da empresa?
Contexto de aplicação	Q7: O período temporal em estudo é adequado?
	Q8: O estudo feito foi na indústria automóvel?
	Q9: A metodologia está devidamente apresentada?
	Q10: O *software* é adequadamente justificado?
	Q11: Justifica a eficácia do modelo no contexto da aplicação?
	Q12: É definida uma relação entre a análise de sentimentos e a cotação da empresa?
	Q13: O estudo compara vários modelos para ver a relação entre a análise de sentimentos e a cotação da empresa?
Avaliação dos resultados	Q14: Realiza métricas de avaliação?
	Q15: Descreve as limitações do estudo?
	Q16: Descreve os contributos do estudo?

Esta avaliação irá ser feita pelo conjunto de critérios acima mencionados, em forma de pergunta, onde cada critério terá uma pontuação de:

- 0 - não responde à pergunta;
- 0,5 - responde parcialmente à pergunta;
- 1 - responde perfeitamente à pergunta.

2.2. Análise de sentimentos

Os estudos de análise de sentimentos concentram-se no uso de sentimentos nas conversas do *Twitter* de modo a prever um comportamento. Através de um estudo realizado, verificou-se a existência de uma relação preditiva entre o humor do *Twitter* e os movimentos das ações. Por outro lado, a literatura demonstra uma correlação positiva entre o volume de *tweets* positivos e os preços. Mas num intervalo de 48 horas, os *tweets* negativos são os melhores indicadores (Nisar & Yeung, 2018).

A análise de sentimentos ajuda uma empresa a analisar as opiniões e *feedback*, dos seus clientes. A utilização desta análise permite que haja uma melhoria dos serviços e da qualidade dos produtos no

futuro, de forma a ajudar a atrair novos clientes, mas também a fidelizar os atuais clientes (Abdulsattar et al., 2018). Faz também com que as empresas anunciem rapidamente qualquer notícia e estejam ligados tanto aos seus *stakeholders* como aos investidores, em tempo real (Hazem et al., 2019). O grande desafio, na análise de sentimentos, é a escolha do algoritmo ideal para medir esse sentimento com a maior precisão possível e, por isso, a escolha assentar na natureza da informação (Yang et al., 2015).

Através das finanças comportamentais demonstra-se que as decisões financeiras, muitas vezes, são influenciadas por emoções ou humor. Um grande número de análises de sentimento do *Twitter* envolve a limpeza e filtragem intensiva dos dados. Assim, é necessário haver um pré-processamento dos dados, onde em primeiro lugar, o algoritmo verifica a existência de erros ortográficos e gramaticais, posteriormente existe a divisão das palavras e a remoção ou substituição por *tags* especiais de *links*, *hashtags*, *retweets* e nomes de utilizadores (Yang et al., 2015). As ferramentas de análise de sentimentos mais utilizadas são os resultados binários, isto é, se é um sentimento positivo ou um sentimento negativo. Embora algumas análises de sentimento, possam gerar sentimentos neutros (Ruan et al., 2018).

As emoções ou os sentimentos dos investidores podem influenciar o mercado de ações, isto é, sentimentos negativos nas redes sociais podem fazer com que haja uma queda no preço das ações. Sentimentos positivos evidenciarão o contrário, isto é, aumento do preço das ações. Serão necessários métodos de processamento de linguagem natural para analisar as emoções dos investidores.

Existem muitas ferramentas de análise de sentimentos, que podem ser divididas em duas categorias: (i) estratégia de análise de contagem de palavras e (ii) estratégia de aprendizagem de máquina.

- A estratégia de análise de contagem utiliza dicionários para determinar o sentimento de cada palavra e agrega o sentimento de palavras. Os dicionários mais utilizados são o dicionário *Hardvard-IV* e o dicionário financeiro de *Loughran* e *McDonald's*.

- As estratégias de aprendizagem de máquina na sua maioria são os classificadores, como o classificador *Naive Bayes* e classificador *SVM (Support Vector Machine)*. Um dos grandes problemas desta estratégia é que ela requer um conjunto de dados de treino classificados, o que requer uma enorme carga de trabalho manual (Ruan et al., 2018). Para além disso, estes últimos métodos têm um problema na classe neutra, uma vez que existe uma relação conflituosa entre as classes neutras e positivas, dando sentimentos positivos a partes que poderão ser neutras (Abdulsattar et al., 2018).

As *reviews* dos clientes online são algo muito importante para uma empresa, uma vez que se consegue perceber melhor as opiniões dos clientes relativamente a produtos ou serviços (Pai & Liu,

2018) e poderão estar presentes tanto nas plataformas de redes sociais da empresa como nas redes sociais próprias do cliente (Nyakurukwa, 2022).

2.3. Cotação da empresa

O mercado de ações é um tema que, cada vez mais, está nos assuntos do momento do *Twitter* e, por isso, existem diversos grupos sobre o mercado de ações, tais como por exemplo, o *StockTwits*, *FinancialTimes* e o *Market-Watch* (Ruan et al., 2018). Vários estudos analisaram as flutuações das empresas que estão no mercado de ações, mas até ao momento não foi encontrado nenhum método confiável e que considere a evolução do mercado de ações (Sóti et al., 2020).

A evolução do valor da ação depende da interação entre as empresas e os clientes, o que leva, muitas vezes, a que o preço das ações, apresente comportamentos imprevisíveis. Assim, uma vez que as previsões costumam ser baseadas na interpretação das condições do mercado, nas especulações e na informação disponível, a questão é relevante para a ciência e para as empresas (Sóti et al., 2020).

O preço do mercado de ações é definido como o preço atual no mercado da bolsa de valores pelo qual é negociado. Quando novas informações são disponibilizadas no mercado sobre um respetivo ativo, o preço dessas ações terá de ser ajustado, aumentado ou diminuído, conforme as novas noticias (Hazem et al., 2019). Os retornos das ações das empresas podem depender do desempenho operacional da mesma, como por exemplo, dos lucros ou da sua posição financeira (Tanaltay et al., 2021). É comum os investidores comparem o preço de mercado das ações com o valor estimado das mesmas, para decidir a compra de ações da empresa (Hazem et al., 2019).

Tanaltay et al., (2021), demonstrou que uma equipa ao vencer uma competição poderá fazer com que haja um aumento do preço das ações do clube, tornando-se esse clube, num ativo valioso e, contrariamente, quando perde, poderá causar uma depreciação das ações, levando a que hajam perdas de milhões de dólares. Também Schmidt et al., (2020) demonstrou que por exemplo, a reação no *Twitter* a um acidente aéreo poderá influenciar o preço da ação da empresa.

2.4. Contexto de aplicação e Avaliação dos Resultados

De modo a analisar os contextos de aplicação, foram realizadas as tabelas presentes nos anexos A, B e C, e a Tabela 2.8 onde estão mencionadas as limitações presentes no estudo.

2.4.1. Contextualização dos artigos

Os artigos incluídos na RSL, são na sua grande maioria, artigos relativos à área das Finanças, à exceção de 7 artigos que falam sobre: (i) área desportiva (Tanaltay et al., 2021) e (Sóti et al., 2020); (ii) momentos eleitorais (Brans & Scholtens, 2020) e (Nisar & Yeung, 2018); (iii) Covid-19 (Dhar & Bose, 2020; Jena & Majhi, 2023; Katsafados et al., 2023); (iv) venda de veículos automóveis (Pai & Liu, 2018); e, (v) área de Finanças e está também relacionado com a área das tecnologias (Abdulsattar et al., 2018).

Relativamente às empresas em estudo, apenas 15 artigos abordam essa informação, mas somente 4 mencionam o número de utilizadores que estão presentes no estudo (Brans & Scholtens, 2020; Ruan et al., 2018; Yang et al., 2015; Bollen et al., 2011).

Em termos de número de *hashtags* utilizados, apenas 3 artigos o mencionam. O artigo de Hamraoui & Boubaker, (2022) utiliza 22 *hashtags,* cada *hashtag* corresponde ao nome de uma empresa. Os artigos de Nisar & Yeung, (2018) e Michalak, (2020) utilizam 3 *hastags* cada.

Analisando o número de *tweets* em estudo, conclui-se: a grande maioria dos artigos analisa um grande número de *tweets*, sendo essa uma realidade favorável, uma vez que um estudo com uma grande amostra evidencia um estudo bem feito e fundamentado e que as informações presentes são, em princípio, corretas. No entanto, existem 3 estudos que abordam um pequeno número de *tweets* que são: (i) o Hamraoui & Boubaker, (2022) que analisa 856 *tweets*; (ii) o Brans & Scholtens, (2020) no inicio começa por analisar mais de 5600 *tweets* mas que elimina os *retweets* e os *tweets* de outras empresas, o que leva a que, no final, acabe com apenas 100 *tweets;* e, (iii) o Hazem et al., (2019) começa com 75.154 *tweets,* mas como apenas queriam *tweets* financeiros, um *tweet* de cada dia e como existiam *tweets* com falta de disponibilidade dos dados do preço das ações, ficaram 833 *tweets*.

Em relação ao período, todos os estudos apresentam uma periocidade diferente, variando entre meses a anos. À exceção do artigo escrito por Nisar & Yeung, (2018) que tem uma periocidade de 6 dias, pois esta análise foi realizada para ver as emoções dos utilizadores antes, durante e após as eleições eleitorais, e também do artigo escrito por Albahli et al., (2022) que tem uma periodicidade de 21 dias, de 1 de março a 21 de março de 2021. Por fim e relativamente ao país da amostra, esta, na sua maioria, foi nos Estados Unidos da América mas também na Tunísia (Hamraoui & Boubaker, 2022), na China (Valle-Cruz et al., 2022), no Zimbábue (Nyakurukwa, 2022), no Brasil (Souza & Martins, 2022), na Turquia (Tanaltay et al., 2021), na Itália, Inglaterra e Holanda (Sóti et al., 2020), Reino Unido (Nisar & Yeung, 2018; Yang et al., 2015), Coreia (Pai & Liu, 2018), Alemanha (Nofer & Hinz, 2015), Índia (Jena & Majhi, 2023) e Bélgica, Alemanha, França, Itália, Espanha e Reino Unido (Katsafados et al., 2023).

2.4.2. Contextualização dos artigos relativos à parte financeira

Verifica-se que alguns artigos utilizam o modelo de Regressão Linear no seu estudo, como por exemplo (Azar & Lo, 2016; Bollen et al., 2011; Dhar & Bose, 2020; Nisar & Yeung, 2018; Oliveira et al., 2017; Ruan et al., 2018; Sóti et al., 2020; Yang et al., 2015). Também é utilizado o modelo Vetor Autorregressivo (VAR) (Hamraoui & Boubaker, 2022; Katsafados et al., 2023; Nofer & Hinz, 2015; Souza & Martins, 2022) e o Algoritmo Naive Bayes (Abdulsattar et al., 2018; Achyutha et al., 2022; Michalak, 2020; Pai & Liu, 2018; Ranco et al., 2015; Tanaltay et al., 2021).

Para além dos modelos anteriores, são utilizados modelos SVM (Nyakurukwa, 2022; Oliveira et al., 2017; Porshnev & Redkin, 2014; Ranco et al., 2015; Tanaltay et al., 2021), além disso Valle-Cruz et al., (2022) tenta detetar as correlações através dos máximos valores de correlação, já Schmidt et al., (2020) utiliza diversos modelos tais como: Modelo de 4 fatores, *OLS*, Modelo ajustado ao mercado e ajustado à média para estimar os retornos esperados do mercado de ações e o Modelo de Regressão de efeitos aleatórios, Brans & Scholtens, (2020) e Hazem et al., (2019) utilizam modelos de retornos ajustados ao mercado e ao risco.

Para além dos modelos acima mencionados, (Jena & Majhi, 2023) utiliza modelos de aprendizagem de máquina (LSTM) e modelos de séries temporais, (Albahli et al., 2022) utiliza o modelo StockSentiWordNet (SSWN) para treinar máquinas de aprendizagem extrema (ELM) e rede neuronal recorrente (RNN) e também modelo de espaço vetorial, e por fim, Pai & Liu, (2018) utiliza modelos de regressão vetorial de suporte de mínimos quadrados (LSSVR) e modelos de séries temporais.

Verifica-se a existência de diversas métricas de avaliação utilizadas para cada artigo, sendo as mais utilizadas: (i) a Correlação de *Pearson* (Hamraoui & Boubaker, 2022; Jena & Majhi, 2023; Michalak, 2020; Nisar & Yeung, 2018; Ranco et al., 2015; Ruan et al., 2018); (ii) o R^2 (Dhar & Bose, 2020; Jena & Majhi, 2023; Nisar & Yeung, 2018; Ruan et al., 2018; Schmidt et al., 2020; Tanaltay et al., 2021); (iii) o *p-value* (Brans & Scholtens, 2020; Dhar & Bose, 2020; Katsafados et al., 2023; Sóti et al., 2020; Valle-Cruz et al., 2022; Yang et al., 2015); (iv) o teste *Dickey-Fuller* (Azar & Lo, 2016; Hamraoui & Boubaker, 2022; Ranco et al., 2015; Souza & Martins, 2022); (v) o teste de Causalidade de *Granger* (Bollen et al., 2011; Hamraoui & Boubaker, 2022; Katsafados et al., 2023; Nofer & Hinz, 2015; Ranco et al., 2015; Souza & Martins, 2022).

Relativamente ao tipo de análise feita e à metodologia de estudo:

Tabela 2.6: Análise e metodologia dos estudos

Autor	Análise feita	Metodologia de estudo
Valle-Cruz et al., (2022)	Análise da correlação entre os índices financeiros e a polaridade dos *tweets*	Metodologia baseada na análise do sentimento financeiro e nas correlações dos diferentes índices financeiros
Nyakurukwa, (2022)	Relação entre o sentimento médio dos *tweets* e os retornos das ações da *Old Mutual Zimbabue Limited*	metodologia passa por uma análise dos dados e a utilização da análise *Wavelet*[7]
Schmidt et al., (2020)	Relação linear entre o volume de *tweets* e o sentimento desses *tweets*	Metodologia de estudo de eventos
Boylan et al., (2020)	Entender se o sentimento expresso nos *tweets* irá ter influência no preço das ações da empresa	Metodologia da teoria financeira[8] e a metodologia de estudo de eventos[9]
Brans & Scholtens, (2020)	Análise textual	Metodologia de estudo de eventos
Hazem et al., (2019)	Avalia o efeito que os anúncios de eventos financeiros têm no rendimento das ações do *Financial Times Stock Exchange 100 Index*	Metodologia de estudo de eventos
Nisar & Yeung, (2018)	Análises de correlação e regressão para comparar a análise de sentimentos com as variações no preço do *FTSE 100*	Seleção da amostra, do método de cálculo do sentimento e métodos de análise estatística
Abdulsattar et al., (2018)	Análise temporal e espacial	Recolha dos dados, pré-processamento, caracterizar os sentimentos e avalia-los
Oliveira et al., (2017)	Avaliação do valor dos dados de *microblogging* para prever as variáveis do mercado de ações	Metodologia que utiliza indicadores de sentimento extraídos de *microblogs* e utilizando o Filtro *Kalman*

[7] Utilizada para ver a ligação entre o sentimento e os retornos das ações no mercado
[8] Diz que os mercados de capitais abrangem todas as informações disponíveis sobre uma empresa
[9] Para avaliar a força que a primeira metodologia tem no preço das ações da empresa

Tabela 2.7: Análise e metodologia dos estudos (continuação)

Autor	Análise feita	Metodologia de estudo
Yang et al., (2015)	identifica a quantidade de investimento correspondente e verifica os principais influenciadores nas redes sociais e fez validação cruzada	Utilizado um método de maneira a criar uma comunicação financeira e depois procuram entender as características da comunidade financeira
Ranco et al., (2015)	Analisa a relação entre os retornos dos preços das ações e o sentimento expresso nos *tweets* e também validação cruzada.	Metodologia de aprendizagem supervisionada e de aprendizagem de máquinas[10], e metodologia de estudo de eventos;
Jena & Majhi, (2023)	Análise do impacto que a Covid-19 teve nos valores das ações em diversos setores, na Índia e nos EUA	Recolha dos dados, realização dos modelos[11], treino dos modelos e avaliação do desempenho destes
Katsafados et al., (2023)	Análise textual[12]	Seleção da amostra e das variáveis e um pré-processamento dos dados
Albahli et al., (2022)	Analisa os *tweets* e os dados do *Google Finance* para prever o futuro comportamento das bolsas de valores num período de tempo	Recolha dos dados, cálculo do sentimento e treino do modelo e criação do modelo de previsão do mercado de ações
Porshnev & Redkin, (2014)	Analisa o humor dos utilizadores do *Twitter* para melhorar a precisão das previsões para os preços de mercado do ouro e da prata	Algoritmo baseado no cálculo da quantidade de palavras para cada emoção. Como o mesmo foi muito preciso, ajudou a criar uma abordagem baseada em léxico

Para além dos artigos anteriormente referidos, existem outros que abordam apenas o tipo de análise feita ou a metodologia, como é o caso de Hamraoui & Boubaker, (2022) onde se verifica uma análise da dependência entre o retorno do preço das ações e o sentimento dos *tweets*, Souza & Martins, (2022) identificou o sentimento dos *tweets* e a sua relação com o retorno do mercado de ações brasileiro e o volume negociado e Tanaltay et al., (2021) classificou os sentimentos dos *tweets* e

[10] De modo a classificar os sentimentos e os métodos utilizados para fazer a análise da correlação
[11] Que são divididos em dados de treino (80%) e dados de teste (20%)
[12] Estudada a relação entre os sentimentos das redes sociais e os mercados de ações

previu o retorno no preço das ações dos quatro clubes. Sóti et al., (2020) fez o teste de causalidade de *Granger* para determinar se as séries temporais são uteis para prever outras séries temporais, Michalak, (2020) fez uma abordagem de aprendizagem e uma abordagem baseada em léxico e Ruan et al., (2018) estudou uma correlação entre os sentimentos do *Twitter* e os retornos anormais das ações e fez validação cruzada. Nofer & Hinz, (2015) estudou a interdependência entre a análise de sentimentos e o retorno das ações, para o modelo *VAR* e Bollen et al., (2011) analisou se os estados de humor dos *tweets* estão correlacionados com o valor do *DJIA* ao longo dos anos e fez validação cruzada. Achyutha et al., (2022) irá fazer uma classificação dos utilizadores em várias categorias (ouro, prata, platina e bronze) com base no rendimento bruto anual, no seguro de saúde e em alguns outros fatores financeiros. Chahine & Malhotra, (2018) faz uma análise do histórico de eventos, ou seja, o estudo de eventos, que é utilizada para analisar os acontecimentos históricos nas empresas da *Fortune 500* que têm uma conta no *Twitter* e, para além disso, também faz uma validação cruzada e Pai & Liu, (2018) previu as vendas totais mensais de veículos, utilizando sete abordagens que estão presentes na Tabela 7.11 do Anexo B.

Em relação à origem dos dados financeiros, vários têm como origem o *Yahoo!Finance* (Achyutha et al., 2022; Bollen et al., 2011; Boylan et al., 2020; Michalak, 2020; Nisar & Yeung, 2018; Pai & Liu, 2018; Ruan et al., 2018; Sóti et al., 2020; Tanaltay et al., 2021; Valle-Cruz et al., 2022). A origem deste dados financeiros pode também vir: (i) Bolsa de Valores do Zimbábue (Nyakurukwa, 2022); (ii) *Investing.com* (Valle-Cruz et al., 2022); (iii) *DataStream* (Brans & Scholtens, 2020; Chahine & Malhotra, 2018; Oliveira et al., 2017); (iv) *NASDAQ* (Ranco et al., 2015; Schmidt et al., 2020); (v) *Thomson Reuters DataStream*(Hazem et al., 2019; Oliveira et al., 2017).

O *software* utilizado na parte financeira é: (i) o *Eviews* para Hamraoui & Boubaker, (2022); (ii) o *R Ecosystem* para as análises das tarefas de previsão para Tanaltay et al., (2021); (iii) o *Federal Deposit Insurance Company*, o *Federal Reserve Bank of Chicago* e a *Consumer Federation of America* para Boylan et al., (2020). Por fim, Katsafados et al., (2023) foi buscar os dados ao *Scott Baker*, *Nick Bloom* e *Steven Davis* (para o índice de volatilidade implícita e para a incerteza da política económica) e ao Banco de dados *Refinitiv Eikon* (para o preço do *bitcoin* e para o preço do petróleo).

Por fim, o tipo de variáveis e a sua relação entre si: (i) Hamraoui & Boubaker, (2022) são utilizadas variáveis de séries temporais estacionárias com resoluções diárias; (ii) Souza & Martins, (2022) são utilizadas variáveis heterogêneas, não normais e macroeconómicas para ver se há uma relação entre o sentimento e o retorno ou volume de negociação; (iii) Tanaltay et al., (2021) utiliza variáveis dependentes e utiliza a Bolsa *Istanbul BIST 100* como forma de comparação para ver as mudanças no mercado; (iv) Boylan et al., (2020) existem 2 variáveis independentes e uma dependente, em que essas

variáveis independentes são comparadas com a variável dependente do preço das ações; (v) Dhar & Bose, (2020) tem variável dependente, variáveis independentes e variáveis de controle, em que é necessário perceber a influência que a Covid-19 teve na variável dependente, utilizando as variáveis independentes; (vi) Hazem et al., (2019) existe uma variável dependente (preço das ações) e uma variável independente (anúncios de eventos) e este estudo vai analisar o efeito da variável independente na variável dependente e visualizar a relação entre estas últimas; (vii) Katsafados et al., (2023) tem variáveis independentes (sentimento do *Twitter* em relação ao Covid-19) e variáveis de controlo, que vão ser utilizadas para estudar a relação entre o sentimento do *Twitter* e o valor da ação; (viii) Nisar & Yeung, (2018) tem variáveis dependentes e independentes e vai haver uma comparação entre ambas, através da análise de correlação baseada no volume; (ix) Yang et al., (2015) tem variável independente e variável dependente (retorno diário do índice de mercado) e vai ser analisado se a variável independente tem uma relação significativa no retorno do mercado e na volatilidade.

2.4.3. Contextualização dos artigos para a análise de sentimentos

Verifica-se que alguns artigos utilizam, como *software*: (i) a *API* do *Twitter* (Abdulsattar et al., 2018; Albahli et al., 2022; Hazem et al., 2019; Katsafados et al., 2023; Michalak, 2020; Nisar & Yeung, 2018; Nofer & Hinz, 2015; Porshnev & Redkin, 2014; Ranco et al., 2015; Sóti et al., 2020; Tanaltay et al., 2021; Yang et al., 2015); (ii) a *API OAuth* é utilizada em Achyutha et al., (2022) e Abdulsattar et al., (2018); (iii) o *Bing Liu*, o *Sentiment 140*, o *NRC*, o *Afim* e o *SenticNet* são utilizados em Valle-Cruz et al., (2022); (iv) *API Google Cloud Natural Language* e a biblioteca *Tweepy* para a recolha dos *tweets* são utilizados por Souza & Martins, (2022); (v) os *softwares Soper Sentiment Analyzer* e *Sentiment32* são utilizados por Schmidt et al., (2020); (vi) *API text2emotion* e o Léxico VADER (*Valence Aware Dictionary and sentiment Reasoner*) são utilizados por Dhar & Bose, (2020). Para além disso, Katsafados et al., (2023) para além da *API* do *Twitter*, também utilizou o Léxico VADER, o (Albahli et al., 2022), para além da *API* do *Twitter*, utilizou o *SentiWordnet;* o *Sentiment 140; SenticNet*; *AFINN* e *SO-CAL*, o *software SentiStrength* é utilizado por Brans & Scholtens, (2020) e por Pai & Liu, (2018) e por fim, Bollen et al., (2011) utiliza o *software OpinionFinder* e *GPOMS (Google-Profile of Mood States)*.

Relativamente ao índice: (i) a grande maioria dos artigos utiliza o *S&P500* (Chahine & Malhotra, 2018; Nofer & Hinz, 2015; Oliveira et al., 2017; Pai & Liu, 2018; Porshnev & Redkin, 2014; Ranco et al., 2015; Ruan et al., 2018; Schmidt et al., 2020; Valle-Cruz et al., 2022; Yang et al., 2015); (ii) vários utilizam o índice *Dow Jones Industrial Average (DJIA)* (Bollen et al., 2011; Jena & Majhi, 2023; Oliveira et al., 2017; Pai & Liu, 2018; Ranco et al., 2015; Sóti et al., 2020; Yang et al., 2015); (iii) também são utilizados índices relacionados com a polaridade dos sentimentos (Hamraoui & Boubaker, 2022;Albahli

et al., 2022); (iv) outros estudos também usam índice de sentimentos (Achyutha et al., 2022; Michalak, 2020; Schmidt et al., 2020; Souza & Martins, 2022); (v) e existe um estudo que aborda o índice *Nifty* (Jena & Majhi, 2023).

Por tipo de sentimento, a maioria dos artigos são divididos em sentimentos positivo, negativo e neutro, à exceção de: (i) (Achyutha et al., 2022; Albahli et al., 2022; Hazem et al., 2019; Katsafados et al., 2023; Pai & Liu, 2018; Ruan et al., 2018) que apenas divide em sentimento positivo e sentimento negativo; (ii) Valle-Cruz et al., (2022) em que cada um dos *softwares* tem uma forma de classificar os sentimentos, tal como está evidenciado na Tabela 7.15 tabela presente no Anexo C; (iii) Dhar & Bose, (2020) em que o sentimento pode ser felicidade, raiva, medo ou tristeza; (iv) Nofer & Hinz, (2015) que divide em tristeza, desespero, cansaço, raiva e humor positivo; (v) Porshnev & Redkin, (2014) que divide em feliz, apaixonado, calmo, enérgico, com medo, com raiva, cansado e triste; (vi) Jena & Majhi, (2023) divide o sentimento em três formas que são, aptidão, entusiasmo e particularidade (que variam entre -1 e +1); (vii) Bollen et al., (2011) que divide em Positivo e negativo para *OpinionFinder* e Calma, alerta, certo, vital, gentil e feliz para *GPOMS*.

Por fim, relativamente à utilização dos *tweets*, na maioria dos casos, realiza-se um pré-processamento dos dados de modo possibilitar a concretização das análises. Todas as fases na utilização dos dados estão presentes nas tabelas presentes no Anexo C.

2.4.4. Avaliação dos Resultados dos artigos

Vários artigos mencionam diversas limitações com os dados. Para Nisar & Yeung (2018), existem diversos problemas nos dados dado que, ao utilizar apenas três *hashtags* outros *tweets* que são feitos sem *hashtags* ou com uma hashtag não muito utilizada, acabam por ser ignorados e, se forem usadas mais do que uma das *hashtags* analisadas, o *tweet* irá ser contado mais do que uma vez.

O tamanho da amostra e o período temporal também é limitativo por ser muito curto. Tanto para Ruan et al. (2018) como para Abdulsattar et al. (2018) o número de *tweets* é muito reduzido para a análise. Para Bollen et al. (2011) a maior limitação são os dados, isto é, apenas existem dados dos EUA, o que para este é o maior problema do artigo. Para Valle-Cruz et al. (2022) esta também é a maior limitação, uma vez que é necessário existir uma análise mais profunda nas reações dos investidores, de modo a obter maior número de dados. Para Tanaltay et al., (2021) uma das maiores limitações é o facto de não serem estudados os retornos semanais e mensais do preço das ações e também por diversos fatores influentes dos mercados de ações, como por exemplo as taxas de juro, não serem levadas em consideração na construção dos modelos de previsão.

Por fim, os contributos ao longo dos artigos são variados, mas todos comprovam a teoria que os autores tentavam provar. Por exemplo, Tanaltay et al., (2021) percebeu que os sentimentos dos *tweets* conseguiam indicar quais as expetativas dos adeptos para o pré-jogo, o que fez com que, esta análise fosse utilizada para verificar as expetativas dos adeptos antes do jogo. Assim, é importante para próximos estudos perceber se estes atestam as perspetivas em relação à bibliografia já existente.

Tabela 2.8: Avaliação dos resultados dos artigos na RSL

ID	Limitações
1	N.E.
2	N.E.
3	N.E.
4	Dados
5	N.E.
6	Dados (é necessária uma análise mais profunda nas reações dos investidores)
7	N.E.
8	N.E.
9	Não foram estudados os retornos semanais e mensais do preço das ações. Necessário considerar mais fatores alternativos de importância das partidas, como o ranking das equipas. Outros fatores influentes dos mercados de ações, como taxas de juro, número de investidores e situação económica geral, não foram considerados na construção do modelo
10	Resultados limitados
11	Dados, onde as existências de eventos externos do mercado, por exemplo, poderão influenciar os resultados; e, os resultados obtidos, que mencionam um efeito no mercado das ações menor comparativamente ao estudo inicial de *Hendricks* e *Singhal*
12	O número limitado de carateres do *Twitter*, faz com que seja necessário o uso de *links*. Existe também uma limitação em termos de dados, em que o *beta* utilizado foi no momento na pesquisa e não da divulgação, o que faz com que este *beta* não esteja atualizado, e uma vez que é um estudo onde é analisado o preço da ação, eventos externos, como eventos mundiais ou corporativos, poderão influenciar esse preço da ação
13	Dados, onde existe falta de acesso a informação do mercado de alta frequência, os *tweets* são selecionados manualmente de forma a encontrar as empresas mencionadas
14	N.E.
15	N.E.
16	Dados e Período temporal
17	Dados
18	Dados e período
19	Dados e período
20	N.E.
21	N.E.
22	Dados
23	N.E.
24	N.E.
25	N.E.
26	Dados e período
27	N.E.
28	Dados

Nota: N.E. – Não Especifica

2.5. Avaliação dos artigos científicos relevantes

Para facilitar a análise dos artigos e verificar qual o melhor artigo para cada pergunta e, consequentemente, para cada critério de qualidade, vão ser avaliados os artigos para cada um dos aspetos presente na Tabela 2.5.

Para cada critério de qualidade vai ser atribuída uma pontuação a cada pergunta e para artigo:

- 0 - não responde à questão;
- 0,5 - responde parcialmente à questão;
- 1 - responde totalmente.

Esta informação está, assim presente na Tabela 2.9.

Tabela 2.9: Avaliação da qualidade dos artigos

	Objetivo		**Sentimento**		**Empresa**		**Contexto**							**Avaliação**			**Total[2]**
ID	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**	**Q10**	**Q11**	**Q12**	**Q13**	**Q14**	**Q15**	**Q16**	
1	1	0,5	1	1	0	0	1	0	1	0	1	1	1	1	0	1	10,5
2	1	0,5	1	0,5	0	0	1	0	1	1	1	1	1	1	0	1	11
3	1	1	1	1	0	0	1	0	1	1	1	1	1	1	0	1	12
4	1	1	1	0	0	1	1	0	0	0	1	1	1	1	1	0	10
5	1	1	1	0	0	0,5	1	0	0,5	1	1	1	1	1	0	0	10
6	1	1	1	1	0	0	1	0	1	1	1	0,5	0	1	1	1	11,5
7	1	1	1	0	1	0	1	0	0,5	0	0,5	1	0	1	0	1	9
8	1	1	1	0	0	1	1	0	0	1	1	1	0,5	1	0	1	10,5
9	1	0,5	1	0	0,5	1	1	0	0	1	1	1	1	1	1	0	11
10	1	0	1	1	1	0	1	0	0	0,5	1	1	0,5	1	1	0	9,5
11	1	0,5	1	0	1	1	1	0	1	0	1	1	1	1	1	0	11
12	1	0	1	1	0	0	1	0	1	1	0	0	0	1	1	1	9
13	1	1	1	0	0	0	1	0	0,5	1	1	1	0,5	1	1	1	11
14	1	1	1	1	0	0	1	0	0	1	1	1	1	1	0	1	11
15	1	0	1	0	0	0	1	0	0	1	1	0,5	0,5	1	0	1	8
16	1	1	1	0	0,5	1	0	0	1	0,5	1	1	0,5	1	1	1	11,5
17	1	0,5	1	0,5	1	1	0	0	1	0	1	1	1	1	1	1	12
18	1	1	1	0	0	0	0	0	1	1	1	1	0	1	1	1	10
19	1	1	1	0,5	0,5	0	0	0	0	1	1	1	0,5	1	1	1	10,5
20	1	0,5	1	1	0	1	1	1	1	0,5	1	1	1	1	0	1	13
21	1	0	1	0	0	0	1	0	1	1	1	0	0	1	0	0	7
22	1	1	0,5	0	0,5	1	1	0	1	1	1	0,5	1	1	1	1	12,5
23	0,5	0,5	1	1	0	0,5	1	0	0	1	1	0,5	1	1	0	0	9
24	1	0	1	0	0,5	0,5	1	0	1	0,5	1	1	1	1	0	1	10,5
25	1	1	1	0	0	0	1	0	0,5	1	1	0	1	1	0	1	9,5
26	1	1	1	1	0	0,5	0	0	0	1	1	0,5	1	1	1	0	10
27	1	0	1	0	0	0	1	0	1	1	1	0	0	1	0	0	7
28	1	1	1	0	0	0	1	0	0	1	1	0	0	1	1	1	9
Total[1]	24,5	16,5	24,5	8	5,5	10	20	1	13	19	23,5	17,5	15	25	14	16	253

Total[1]: Pontuação total de cada critério de qualidade. Total[2]: Pontuação total de cada artigo.

Dos artigos utilizados na RSL verifica-se que aquele cujo valor de qualidade é melhor, obteve um valor de 13/16 (Pai & Liu, 2018), sendo o critério de avaliação com maior qualidade o Q14, onde todos os artigos respondem a este critério. Inversamente, os artigos com piores valores de qualidade conseguiram um valor de 7/16 (Abdulsattar et al., 2018; Porshnev & Redkin, 2014), sendo o critério de avaliação com pior qualidade o Q8, apenas um artigo obtém valor.

Ao analisar as questões que apresentaram critérios com maior qualidade verifica-se que são as relativas ao objetivo e ao contexto de aplicação, uma vez que são os critérios onde, no seu conjunto, se encontram os critérios com maiores valores.

Existe também um critério que apresenta um valor de 1 (Q8). Este é o critério que questiona se os artigos são da indústria automóvel. Em virtude desta dissertação se inserir na indústria automóvel, é importante perceber se existem muitos artigos que analisam empresas da indústria automóvel ou se é um *gap* existente.

De forma a resposta à questão de investigação "Como é que os comentários na rede social *Twitter* podem prever o valor da empresa?" foi necessário serem realizados diferentes modelos com diferentes metodologias, dependendo do objetivo de cada estudo. Assim, em todos os estudos que utilizam empresas é possível perceber que existe sempre alguma associação entre os comentários na rede social *Twitter* e o valor da empresa.

2.6. Contributos da RSL

A realização desta RSL foi importante para a dissertação, dado que existe um *gap* relativamente a este tema, uma vez que é um assunto não muito abordado. Apenas foi possível encontrar um artigo sobre veículos automóveis. Assim, o mesmo será muito importante para futuras investigações sobre o tema.

A realização desta RSL possibilitou perceber que futuros estudos terão de ter em atenção o período de tempo e a quantidade de dados utilizados, uma vez que foram encontradas limitações após a realização das análises.

Capítulo 3: Metodologia

O presente capítulo aborda as fases realizadas aos dados de modo a desenvolver o estudo e obter resultados comprovando ou não a questão de investigação do estudo. Para isso, é utilizada a metodologia CRISP-DM (*Cross Industry Standard Process for Data Mining*), em virtude de ser um bom modelo para utilizar como metodologia de pesquisa, para sistematizar todos os passos que terão de ser feitos para uma boa análise (El-Din Hassanien, 2019).

A metodologia CRISP-DM é um processo *standard* utilizado em estudos de *data minin*g, focada na experiência prática, nas necessidades dos gestores, na resolução dos problemas de gestão e também pelo seu conteúdo, facilmente adaptável e por isso, flexível (Gonçalves & Jacinto, 2021).

O CRISP-DM é um processo não sequencial, ou seja, as diferentes fases podem ser realizadas mais do que uma vez dependendo dos resultados obtidos (Moreno, 2015).

O CRISP-DM é composto por seis etapas:

- Compreensão do Negócio

  Primeira fase de todos os projetos onde são definidos: (i) o problema e (ii) os objetivos propostos. É o ponto de partida para iniciar e desenvolver o projeto (Moreno, 2015). Nesta fase existe uma visão global dos recursos disponíveis e necessários de modo a aperfeiçoar o projeto (Gonçalves & Jacinto, 2021).
- Compreensão dos Dados

  Respeita à recolha e compreensão dos dados em que ocorrerá uma análise inicial dos dados verificando a sua qualidade e a obtenção dos primeiros insights (Gonçalves & Jacinto, 2021).
- Preparação dos Dados

  Fase crucial para o desenvolvimento do trabalho, onde são realizadas todas as tarefas de preparação, seleção, limpeza e tratamento dos dados. Tem como objetivo organizar o texto de forma a ser trabalhado pelos algoritmos de extração de padrões. Nesta fase haverá grande diferença entre os processos de *text mining* e de *data mining*, uma vez que em *data mining* os dados já se encontram estruturados (Moreno, 2015). Pretende ainda, esta fase, obter a informação relevante com as devidas alterações a obter, um novo *dataset* para ser utilizado nas fases posteriores (Gonçalves & Jacinto, 2021).
- Modelação

  Fase onde são analisadas as várias técnicas de aprendizagem e algoritmos que possibilitam a criação de modelos (Moreno, 2015).

- Avaliação

  Posteriormente à realização dos modelos é importante avaliá-los, de modo a perceber se os resultados obtidos são bons ou não e, se alcançam os objetivos previamente definidos (Gonçalves & Jacinto, 2021).

- Implementação

  Fase centrada na implementação, monitorização e manutenção dos dados de forma a comprovar as decisões tomadas. Esta fase de implementação pode ser demonstrada através de um relatório com todas as análises realizados e com os resultados obtidos (Gonçalves & Jacinto, 2021)

3.1. Metodologia Utilizada

Depois de analisar alguns estudos análogos, optou-se pela metodologia CRISP-DM em virtude de ser o método a mais frequentemente utilizado para estudos semelhantes (Maria et al., 2021).

3.1.1. Compreensão do Negócio

Na presente dissertação a fase da compreensão do negócio, está presente nos capítulos da Introdução e da Revisão da literatura.

3.1.2. Compreensão dos Dados

Os dados foram obtidos com base num *dataset* retirado do *Kaggle,* plataforma que inclui diversos *dataset* relacionados com a ciência de dados para problemas de negócios, contratação e para fins de pesquisa académica (Bojer & Meldgaard, 2021). Este *dataset* inclui todos os *tweets* publicados pela *Tesla*, desde 1 de janeiro de 2018 até 1 de dezembro de 2018.

Posteriormente, procurou-se o valor da ação da *Tesla* de modo a tentar comprovar ou não o estudo, isto é, se os *tweets* publicados em relação à *Tesla* influenciam o valor da ação da mesma. Assim, foi retirado o valor da ação da *Tesla* no *MarketWatch*[13] com o índice *NASDAQ*.

[13] *https://www.marketwatch.com/*

3.1.3. Preparação dos Dados

Inicialmente, foram divididos os dados por semestre no *PowerQuery*, plataforma que pertence ao *Microsoft Excel*, uma vez que, a quantidade de dados era muito grande e era necessário certificarmo-nos que era possível, posteriormente, guardar[14] esses dados. Posteriormente, e por existirem várias colunas que não são usadas no estudo, por não serem importantes para o mesmo, foram eliminadas as seguintes colunas: (i) *Source.Name*, coluna que diz qual o nome do documento, isto é, se é um dado que pertence ao documento do primeiro semestre ou do segundo semestre; (ii) *column 1*, coluna de numeração do *tweet*, ou seja, vai de 1 ao número total de *tweets* extraídos; (iii) *favorited* e *retweeted*, coluna da qual não se extrai qualquer informação[15]; (iv) *quoted_status_id_str; (v) quoted_status_short_url; (vi) quoted_status_expand_url; (vii) user_verified; (viii) in_reply_to_status_id_str; (ix) in_reply_to_user_id_str; (x) id_str, (xi) conversation_id_str; (xii) user_id_str; (xiii) user_name; (xiv) user_full_name hastags; (xv) mentions; (xvi) urls; (xvii)* media, todas estas colunas referem quem escreveu o *tweet* ou tratam de informações não relevantes para o estudo.

Assim, as colunas a utilizar e seguindo a ordem referida neste estudo serão: (a) *created_at; (b) full_text; (c) lang; (d) retweet_count; (e) favorite_count; (f) reply_count; (g) quote_count; (h) pos; (i) neg; (j) neu; (l) compound*.

Uma vez que os ficheiros já tinham sido divididos no *PowerQuery*, todos os ficheiros foram colocados no *software* escolhido[16] para analisar a quantidade de sentimento positivo, negativo, neutro e o *compound*, que descreve o sentimento predominante num *tweet*.

Para fazer esta análise no *software Orange3,* foi necessário colocar um nó *Corpus* que inclui os dados em análise. Seguidamente foi colocado o nó *Select Columns*, onde foram selecionadas as colunas necessárias para a análise. Posteriormente foi selecionado o nó *Sentiment Analysis* onde foi selecionado o método de análise dos sentimentos selecionados. No estudo em análise, a escolha recaiu sobre o método VADER, método fundado em regras e léxicos para a análise de sentimentos e utilizado para fazer análise de sentimentos em conteúdos do *Twitter* e outras redes sociais que não necessitam de treinamento nos dados (Borg & Boldt, 2020). Em seguida foi colocado o nó *Data Table*, onde está presente a tabela que inclui, para além do *tweet*, o valor de positivo, negativo, neutro e *compound* que constitui cada *tweet*, sendo estes valores de 0 a 1. Por último, foi utilizado o nó *Save Data*, nó que serve para guardar os dados para, se necessário, passar para outro *software* e para analisá-los. O *dataset* do que foi feita neste programa encontra-se evidenciado na Figura 3.1.

[14] o *Microsoft Excel* tem um número de linhas limitado
[15] só refere *FALSE*
[16] *Software Orange3*

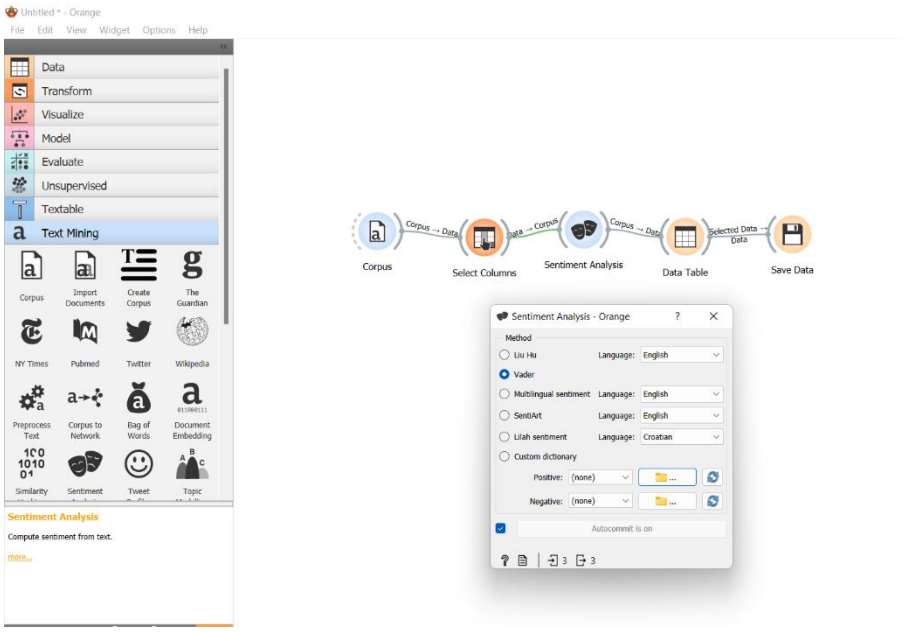

Figura 3.1: Procedimento de análise de dados no *Orange3*

Posteriormente à realização desta análise entendeu-se que seria necessário ter as colunas criadas em valor e não numa escala de 0 a 1, ou seja, foi essencial analisar em cada *tweet* qual era o sentimento predominante de modo a possibilitar a realização de uma análise clara. O objetivo era ter apenas um *tweet* em cada data, ou seja, juntar todos os *tweets* com datas iguais. Assim, para cada data verificou-se quantos *tweets* tinham um sentimento positivo, um sentimento negativo e um sentimento neutro[17].

Depois, foi realizada uma tabela dinâmica com as informações necessárias para a análise, isto é, o número de *tweets* positivos e negativos em cada dia do mês, o número de *retweets*, gostos, respostas e *quotes*. Foi também feito um filtro no *Microsoft Excel* quer pelos positivos quer pelos negativos de modo a fazer uma nuvem de palavras, para os dois semestres. Com a realização deste filtro verificou-se que existiam alguns *tweets* que, apesar de terem sentimento, não tinham qualquer informação na coluna do *full_text*. Assim foi necessário eliminá-los, o que originou uma mudança no total da divisão dos sentimentos, conforme se explica detalhadamente no Resultados e Discussão.

Foi colocado no *Orange3* o ficheiro dos *tweets* negativos e feito um pré-processamento marcando-se no *Transformation* todas as opções, no *Tokenization* foi marcada a opção *Regexp*, no *Filtering* foram marcadas as opções *Stopwords* e *Numbers*. Também foi colocado o ficheiro dos *tweets* positivos. Como não se conseguiu juntar os dois semestres no *Microsoft-Excel*, foi necessário utilizar o nó *Concatenate* e, aí, colocado o nó de *Preprocess Text* com os mesmos parâmetros do utilizado para

[17] análise efetuada no subcapítulo seguinte

os *tweets* negativos. Relativamente às *Stopwords* de cada um dos ficheiros, inseriu-se uma vasta lista incluindo *stopwords* como: “https”; “rt”; “ev(s)”; “e”; “c”; “Facebook” e “etc”.

Uma vez que apenas se analisou uma ação em diferentes períodos, trata-se de uma série temporal. Realizou-se, assim, uma análise do sentimento em relação ao valor da ação, e para isso, utilizou-se um ficheiro de *Microsoft Excel* onde estão inseridas as colunas data, negativo, positivo, soma_*retweet*, soma_*favorite*, soma_*reply*, soma_*quote*, *tweet_tone*, abertura e fecho. Estas últimas referem-se ao valor de abertura e fecho da ação no mercado de valores. Posteriormente, foi alterada a formatação da data pois, a mesma, estava numa formatação que não possibilitava a criação do gráfico. De forma a conseguir fazer a comparação do sentimento dos *tweets* com o valor da ação, foi necessário fazer a variação do valor da ação do dia, isto é, o valor de fecho do dia em comparação ao dia de fecho do dia anterior (Brealey, 2006). Assim, foi realizado o gráfico que irá ser incluído no capítulo seguinte.

3.1.4. Modelação

Foi efetuada, nesta fase, uma análise de sentimentos aos dados extraídos com o objetivo de perceber como o sentimento dos mesmos estavam divididos. Para além disso, foi também realizada uma regressão linear aos dados, que pode ser: (i) simples, quando existe apenas uma variável explicativa, ou (ii) múltipla, quando existe uma relação linear entre várias variáveis independentes e uma dependente (Oliveira et al., 2017).

A análise de sentimentos implicou dividir os comentários em sentimentos positivo, negativo e neutro. Neste processo, conclui-se pela existência de um enviesamento do estudo na medida em que a maioria dos comentários estavam associados a sentimento neutro, como se mostra na Tabela *3.1* e Tabela *3.2*. Foi necessário dividir o estudo apenas em sentimento positivo e sentimento negativo, divisão essa que irá ser analisado no capítulo seguinte.

Tabela 3.1: Divisão dos sentimentos, 1º semestre

Tipo de sentimento	Nº comentários	%
Negativo	4.189	0,55%
Neutro	752.895	98,54%
Positivo	6.984	0,91%
Total	764.068	100,00%

Tabela 3.2: Divisão dos sentimentos, 2º semestre

Tipo de sentimento	Nº comentários	%
Negativo	1.990	0,27%
Neutro	716.868	98,99%
Positivo	5.327	0,74%
Total	724.185	100,00%

A regressão linear[18] é uma regressão linear simples, por isso, tem apenas uma variável dependente. Esta tem como variável dependente: o sentimento do *tweet* e, a variável independente: a variação do valor da ação.

Foi utilizada a variação do valor da ação de modo a que conseguisse ter em conta o dia anterior, pois um *tweet* pode não influenciar o valor da ação no dia, mas também nos dias seguintes.

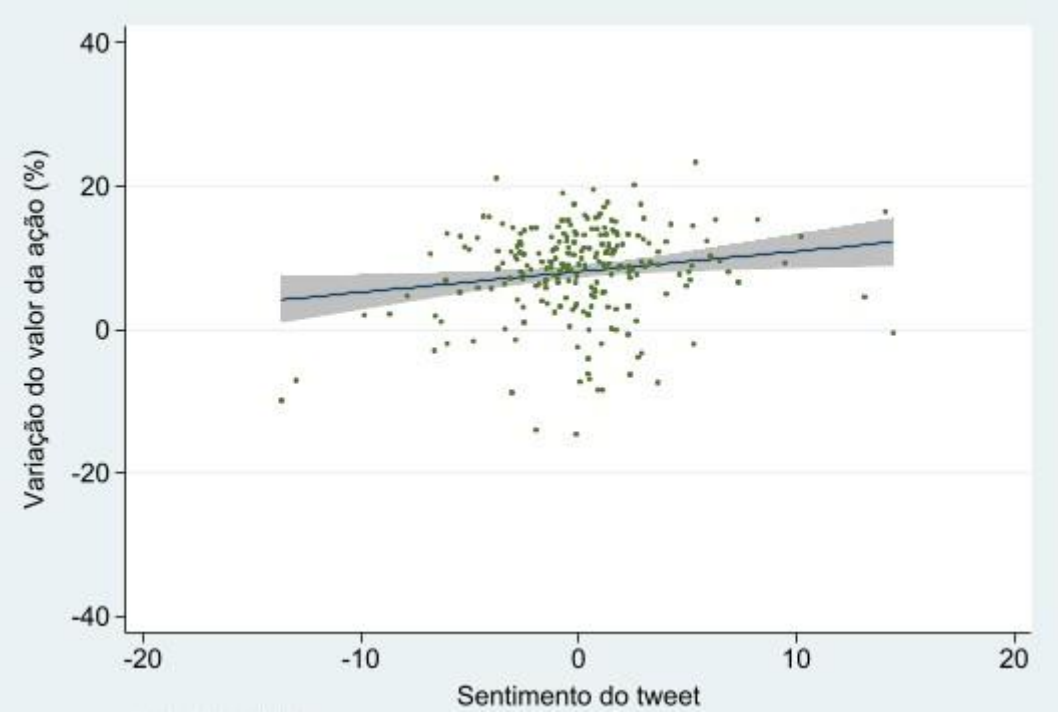

Figura 3.2: Gráfico Regressão Linear[19]

De acordo com o gráfico[20] constata-se que existe claramente uma relação positiva entre a variação do valor da ação da *Tesla* e o sentimento dos *tweets* extraídos.

[18] Realizado no *software Stata 17.0* com o comando *Tsline*
[19] Gráfico de dispersão
[20] Realizado no *software Stata 17.0* com o comando *Tsline*

3.1.5. Avaliação

Para a avaliação do modelo foi utilizado o R^2, que é uma medida estatística que serve para perceber o quanto próximo os dados estão da linha de regressão. Quanto maior o R^2 maior a realidade explicada pelo modelo.

O gráfico demonstra que o R^2 é reduzido, i.e. cerca de 2,7%, no entanto, a relação entre as duas variáveis (variação do valor da ação e o sentimento do *tweet*) é estatisticamente significativa a 5%.

Para a avaliação da análise de sentimentos foi feita uma análise manual aos sentimentos, comparando com o *ChatGPT*, *chatbot* de inteligência artificial desenvolvido pelo *OpenAI*[21], e consequentemente, compará-lo com os sentimentos extraídos do *software Orange3*. Esta análise foi feita a uma amostra de 10 *tweets*.

Depois de analisar a amostra de 10 *tweets* verifica-se que o sentimento estava bem definido nos *tweets*, i.e., através da observação dessas três variáveis (*tweets*, *ChatGPT* e *software)* foram retiradas as mesmas conclusões para o sentimento dos *tweets*.

3.1.6. Implementação

Na presente dissertação a implementação é o relatório realizado com todas as análises realizadas e os resultados encontrados.

[21] Laboratório de pesquisa de Inteligência Artificial

Capítulo 4: Resultados e Discussão

Neste capítulo vão ser demonstrados os resultados das análises feitas no capítulo anterior, i.e., na fase de Preparação dos dados e na fase de Modelação.

4.1. Análise de Sentimentos

Tal como referido na fase da Modelação dos dados, do capítulo da Metodologia, o estudo dividiu-se em: (i) sentimento positivo; e, (ii) sentimento negativo; quando se dividia em sentimento positivo, negativo e neutro, o estudo ficava enviesado. Analisando o número de comentários e a percentagem, percebe-se que existe um número muito superior de comentários com um sentimento positivo, havendo assim, uma percentagem superior, tanto para o primeiro semestre como para o segundo semestre (75,41% para o primeiro semestre e 76,31% para o segundo semestre).

Tabela 4.1: Divisão dos sentimentos sem sentimento neutro, 1º semestre

Tipo de Sentimento	Nº comentários	%
Negativo	186.848	24,59%
Positivo	573.081	75,41%
Total	759.929	100,00%

Tabela 4.2: Divisão dos sentimentos sem sentimento neutro, 2º semestre

Tipo de sentimento	Nº comentários	%
Negativo	170.541	23,69%
Positivo	549.360	76,31%
Total	719.901	100,00%

4.2. Caracterização do *dataset* por países

Ao analisar a origem dos *tweets*, no 1º semestre houve uma variedade de 24 países diferentes e, no 2º semestre 25 países diferentes. Destacando-se com o maior número de comentários os da língua inglesa, francesa, alemã, espanhola e holandesa, apesar da língua que mais se destaca ser o inglês, em ambos os semestres (758.824 no 1º semestre e 718.922 no 2º semestre), Tabela 4.3 e Tabela 4.4.

Tabela 4.3: Lista das línguas dos comentários, 1º semestre

Língua	Nº comentários
en	758.824
fr	266
de	166
es	115
nl	111
ca	90
et	56
tl	52
ht	51
it	33
tr	32
eu	25
pt	19
in	19
sv	18
da	18
pl	10
cs	7
no	4
hu	4
cy	3
lv	3
vi	2
ja	1
Total	759.929

Tabela 4.4: Lista das línguas dos comentários, 2º semestre

Língua	Nº comentários
en	718.922
fr	320
nl	124
de	109
es	69
ht	52
et	49
tl	48
ca	40
tr	35
da	28
it	24
pt	20
sv	12
in	10
no	9
ja	7
lv	6
vi	5
pl	4
eu	4
und	1
lt	1
hu	1
cs	1
Total	719.901

4.3. Principais palavras presentes nos comentários

Foram realizadas duas nuvens de palavras para os sentimentos positivos e negativos[22].

No caso da nuvem de palavras para o sentimento positivo, Figura 4.1, as *stopwords* mais utilizadas são: (i) *tesla*; (ii) *Elon Musk* (CEO da *Tesla*); (iii) *model*; (iv) *new car*; (v) *electric*; (vi) *spacex* (Uma das empresas de *Elon Musk*).

No caso da nuvem de palavras para o sentimento negativo, Figura 4.2, as *stopwords* mais utilizadas são: (i) *tesla;* (ii) *Elon Musk* (CEO da *Tesla*); (iii) *model*; (iv) *car*; (v) crash; (vi) autopilot; (vii) *new*.

Assim, observando ambas as nuvens de palavras, verifica-se que existe um padrão de *stopwords* utilizadas, sendo, em ambos os casos, semelhantes as mais utilizadas.



Figura 4.1: Nuvem de palavras para sentimento Positivo

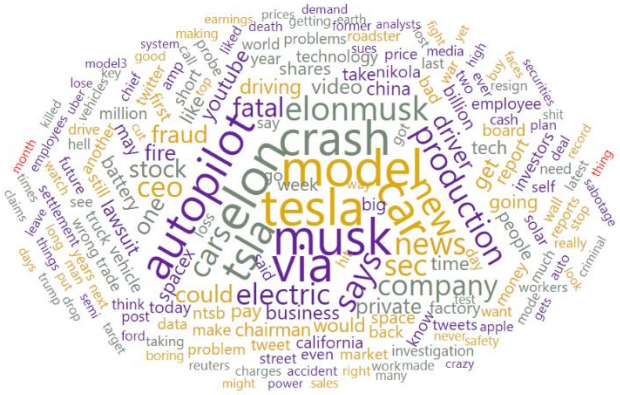

Figura 4.2: Nuvem de palavras para sentimento Negativo

[22] Realizadas no *Software Orange3*

4.4. Séries temporais da variação do valor da ação e sentimento

O gráfico final, Figura 4.3 , demonstra a existência de uma relação entre a variação do valor da ação e o *tweet tone*, que estuda o sentimento predominante de um *tweet* e que varia entre [-1;1]. Utilizou-se a variação do valor da ação, pois a mesma, compara o preço de fecho do dia com o preço de fecho do dia anterior. Essa é uma análise importante, uma vez que, para a previsão do valor da ação de uma empresa, um *tweet* pode não condicionar esse valor no próprio dia e condicioná-lo nos dias posteriores.

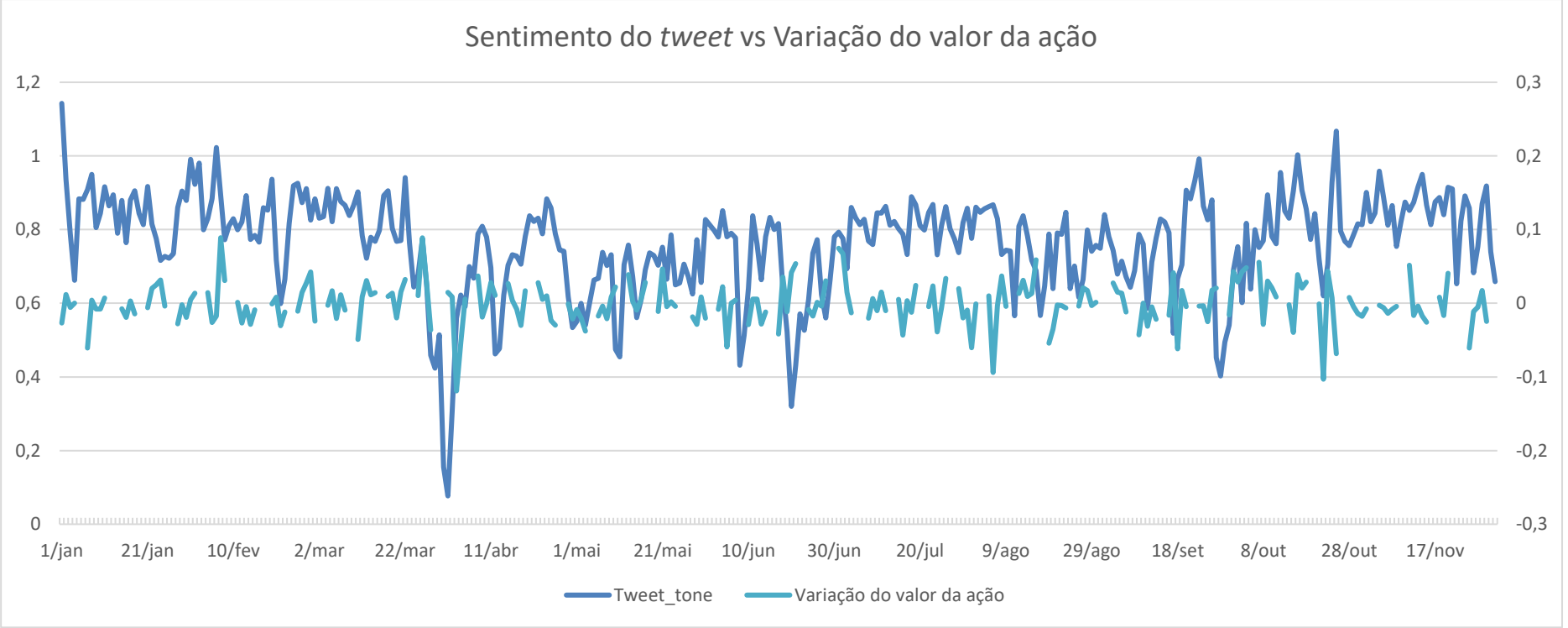

Figura 4.3: Gráfico com a comparação entre *Tweet tone* e Variação do valor da ação

Pela análise do gráfico verifica-se a existência de alguns picos. Tanto negativos como positivos de sentimento dos *tweets*. Assim, foram analisadas algumas situações, conforme Figura 4.4:

- 1º Situação: dia 1 de abril de 2018 - saiu uma notícia, em que o *Elon Musk* anunciava que a *Tesla* tinha acabado de entrar num processo de proteção contra credores, i.e., numa insolvência, o que levou a que houvesse um aumento significativo de *tweets* com um sentimento negativo, e consequentemente, a existir um decréscimo no valor da ação.
- 2º Situação: dia 17 de agosto de 2018 - saiu uma notícia sobre o *Elon Musk*, dizendo que tinha sido o ano mais difícil e doloroso da sua carreira, pois nesta altura, a *Tesla* enfrentava uma falta de liquidez e era investigada pela bolsa norte-americana. Existia instabilidade na empresa agravada pelos *tweets* colocados pelo mesmo, justificando-o por "fruto de exaustão". Com esta situação e através do gráfico que compara o sentimento dos *tweets* com a variação do valor da ação, é possível verificar que, apesar do sentimento dos *tweets* e a variação do valor da ação, não terem chegado a valores muito baixos, percebe-se que a variação do valor da ação estava a subir e, posteriormente a esta noticia, o valor da variação do valor da ação parte de um ponto mais baixo, o que leva a perceber que houve uma descida abrupta do valor da variação do valor da ação da *Tesla.*
- 3º situação: 30 de setembro de 2018 - saiu uma notícia que informava que o *Elon Musk*, apesar de continuar com o cargo de presidente executivo da *Tesla*, ia renunciar à presidência do conselho de administração, durante 3 anos. Para além disso, também tinha sido punido com uma multa de 20 milhões de dólares. As ocorrências originaram uma grande queda no sentimento dos *tweets* e consequentemente, do valor da variação do valor da ação, mas não tão significativa.

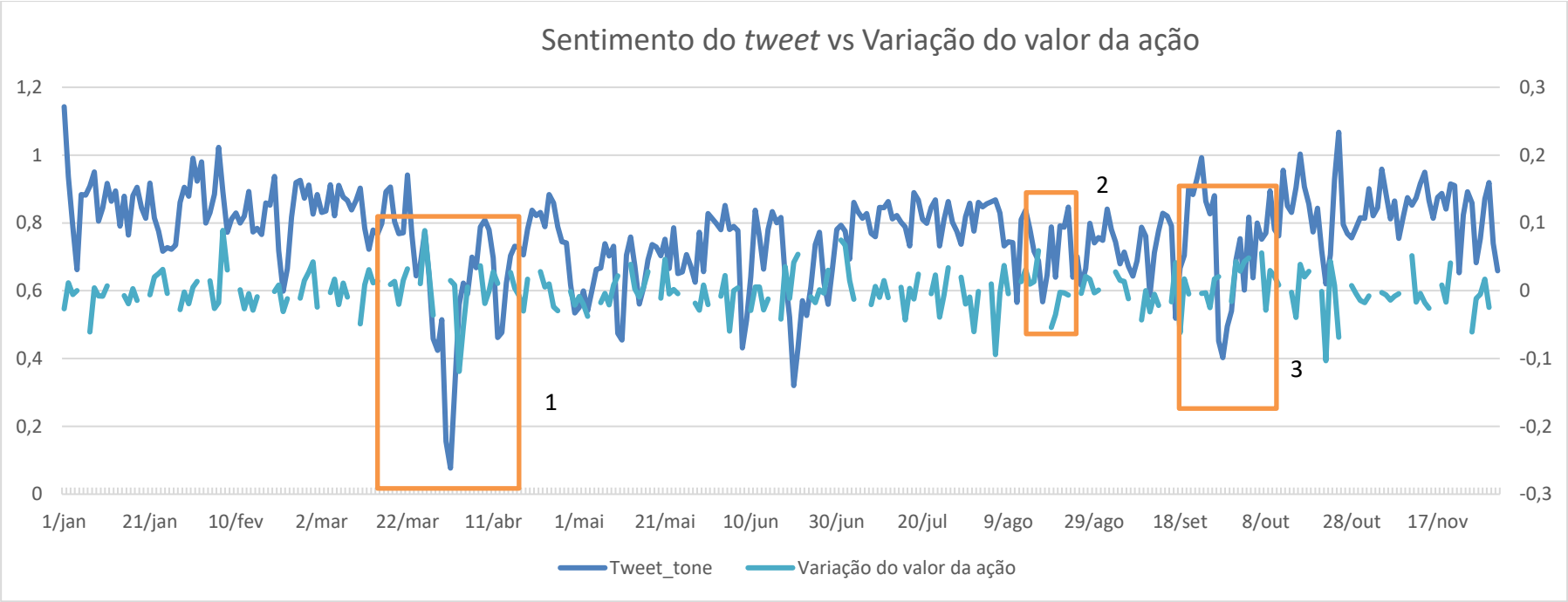

Figura 4.4: Gráfico com a comparação entre *Tweet tone* e Variação do valor da ação (3 situações)

Capítulo 5: Conclusões e Recomendações

5.1. Conclusão

Com o desenvolvimento das novas tecnologias, as redes sociais tornaram-se um importante canal de informação em todo o mundo e definem-se como um conjunto de aplicativos baseados na Internet que se apoiam na *Web 2.0* e que permitem a criação e a troca de conteúdos entre os utilizadores, conteúdos esses que ajudam as empresas a tomar decisões (Yang et al., 2015).

Esta dissertação tem como objetivo estudar a possibilidade de existir uma relação entre o valor da ação da *Tesla* e os sentimentos presentes nos *tweets*. Para isso, foi necessário a recolha dos dados de modo a compará-los e analisá-los para retirar conclusões.

A *Tesla* é uma empresa especializada na produção de veículos elétricos, que utilizam baterias de ião lítio para o armazenamento de energia e auxiliar na instalação de painéis solares (Liu, 2021). A análise de sentimentos ajuda uma empresa a analisar as opiniões e *feedback*, dos seus clientes. Utilizar esta análise, será benéfico para a empresa, uma vez que faz com que haja uma melhoria dos serviços e da qualidade dos produtos no futuro, ajudando a atrair novos clientes, mas também a fidelizar os atuais (Abdulsattar et al., 2018).

Assim, em primeiro lugar, realizou-se uma Revisão Sistemática da Literatura, onde foram analisados 28 artigos científicos, de um total de 142 documentos repartidos em duas bases de dados científicas, que abordavam as diversas áreas.

Ao analisar estes artigos entendeu-se que a maioria eram da área financeira, existindo um *gap* relativamente à área automóvel, pois, apenas um artigo abordava este tema, concluindo-se do interesse desta dissertação em virtude de não tratar um assunto muito estudado.

Posteriormente, foi iniciada a metodologia do trabalho. Foi utilizada a Metodologia CRISP-DM. Metodologia não sequencial, que possibilita "voltar atrás" na fase, dependendo dos resultados obtidos e é focada na experiência prática, nas necessidades dos gestores, na resolução dos problemas de gestão e também pelo seu conteúdo, facilmente adaptável e por isso, flexível (Gonçalves & Jacinto, 2021).

Na metodologia, primeiro houve uma compreensão do negócio e dos dados, posteriormente passou-se a uma preparação dos mesmos. Estes foram extraídos, limpos, filtrados e preparados, de modo a que fosse possível modelá-los e avaliá-los, para que, posteriormente, possibilitasse a comprovação ou não do objetivo da dissertação[23].

[23] Verificar a existência de uma relação entre o valor da ação da *Tesla* e os sentimentos presentes nos *tweets*

Seguidamente passou-se à fase da modelação, tendo sido feito um estudo à análise de sentimentos. Estes foram divididos por diferentes tipos de sentimento, de modo a perceber se existia alguma classe de sentimentos que fosse predominante. Percebeu-se assim, que existia um enviesamento da classificação dos sentimentos, dado que a maioria dos sentimentos conduzia para um sentimento neutro. Por isso, foi necessário eliminar a divisão de sentimento neutro e dividir os *tweets* apenas entre sentimento positivo e negativo. Análise realizada no capítulo seguinte.

Também na fase da modelação, foi realizada uma regressão linear aos dados, utilizando-se como variável dependente: o sentimento do *tweet* e como variável independente: a variação do valor da ação. A variação do valor da ação foi utilizada de modo a ter em conta o dia anterior. Um *tweet* pode condicionar o valor da ação no dia e também nos dias seguintes. Com a regressão linear foi possível verificar que existe uma relação positiva entre a variação do valor da ação da *Tesla* e o sentimento dos *tweets*.

De modo a avaliar o modelo de Regressão Linear foi realizado um estudo ao R^2, que apesar de ser reduzido, 2,7%, confirma a existência de uma relação entre as duas variáveis (variação do valor da ação e sentimento do *tweet*) significativa, a 5%.

Também para avaliar o modelo, mas de análise de sentimentos, foram comparados os sentimentos expressos no *software* com uma análise manual e com o *ChatGPT*, utilizando uma amostra de 10 *tweets*, aleatoriamente. Depois de analisados estes 10 *tweets*, conclui-se que: pela observação das três variáveis (*tweets*, *chatGPT* e *software*) estas estavam todas de acordo e era possível retirar as mesmas conclusões.

Por fim, foram analisados e discutidos os resultados. No caso da análise de sentimentos, com a eliminação do sentimento neutro, conclui-se que existia um número superior de *tweets* com um sentimento positivo, com uma percentagem de 75,41% no primeiro semestre e de 76,31% no segundo semestre. Seguidamente foi realizada uma análise por países, em ambos os semestres e, verificou-se que a língua dominante era a língua inglesa.

Foi também realizada uma análise às nuvens de palavras, verificando-se que as *stopwords* mais utilizadas eram “tesla”; “elon musk” (CEO da *Tesla*); “electric” e “new”, para ambos os sentimentos. Por fim, foi feita a análise do gráfico final, onde se encontra o gráfico da série que evidencia a comparação entre o sentimento dos *tweets* com a variação do valor da ação. Neste gráfico são demonstradas algumas variações, que mostram alguns picos e alguns fossos bem como a explicação dos mesmos, justificando algumas variações.

Assim, com a ajuda do gráfico, verifica-se que com a utilização do sentimento dos *tweets* ajuda à previsão da variação do valor da ação.

5.2. Limitações

As limitações sentidas ao longo da dissertação foram as seguintes:

1. A análise ter apenas em conta os *tweets*, não considerando as opiniões dos utilizadores das outras redes sociais;
2. Não serem consideradas outras variáveis, que de alguma forma podem influenciar o valor da ação ou os *tweets* dos utilizadores, que podem influenciar a que a análise tivesse diferentes resultados;
3. A não utilização de outros anos, de modo a possibilitar a comparação entre os diferentes anos e/ou se percebesse se este ano foi um ano anómalo, ou se o sentimento dos *tweets* influencia sempre o valor da ação.

5.3. Trabalho futuro

Com esta dissertação abrem-se portas para trabalhos futuros, tendo em atenção as limitações encontradas.

Seria importante em futuros estudos, a utilização de outros anos, i.e., em 2018, foi possível perceber que os *tweets* influenciam o valor da ação da *Tesla*, mas será que nos anos anteriores ou posteriores também influenciaram? Será que apenas acontece no ano de 2018? Se acontecesse nos anos posteriores, também seria bom e importante a comparação entre os anos, por exemplo, comparar os resultados obtidos em 2018 e em 2019, para perceber se esta influência foi mais notória ou não.

Para trabalho futuro também seria importante a inclusão de outras variáveis que ajudassem a perceber se existem fatores externos que poderão influenciar o valor da ação da empresa, sem ser os *tweets*, pois a variação da ação poderá não ser causa-efeito dos *tweets,* mas sim, de outros fatores externos.

Capítulo 6: Referências Bibliográficas

Abdulsattar, G., Alkubaisi, A. J., Sakira Kamaruddin, S., & Husni, H. (2018). Conceptual Framework for Stock Market Classification Model Using Sentiment Analysis on Twitter Based on Hybrid Naïve Bayes Classifiers. In *International Journal of Engineering & Technology* (Vol. 7, Issue 2).

Abrantes, C. (2020). *Os modelos preditivos do sucesso de candidaturas a fundos europeus: o papel da manipulação de resultados*.

Achyutha, P. N., Chaudhury, S., Bose, S. C., Kler, R., Surve, J., & Kaliyaperumal, K. (2022). User Classification and Stock Market-Based Recommendation Engine Based on Machine Learning and Twitter Analysis. *Mathematical Problems in Engineering*, *2022*. https://doi.org/10.1155/2022/4644855

Albahli, S., Irtaza, A., Nazir, T., Mehmood, A., Alkhalifah, A., & Albattah, W. (2022). A Machine Learning Method for Prediction of Stock Market Using Real-Time Twitter Data. *Electronics (Switzerland)*, *11*(20). https://doi.org/10.3390/electronics11203414

Azar, P. D., & Lo, A. W. (2016). *The Journal of Portfolio Management The Wisdom of Twitter Crowds: Predicting Stock Market Reactions to FOMC Meetings via Twitter Feeds*. www.iijournals.com

Baker, K. A., & Weeks, S. M. (2014). An overview of systematic review. *Journal of Perianesthesia Nursing*, *29*(6), 454–458. https://doi.org/10.1016/j.jopan.2014.07.002

Bojer, C. S., & Meldgaard, J. P. (2021). Kaggle forecasting competitions: An overlooked learning opportunity. *International Journal of Forecasting*, *37*(2), 587–603. https://doi.org/10.1016/j.ijforecast.2020.07.007

Bollen, J., Mao, H., & Zeng, X. (2011a). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8. https://doi.org/10.1016/j.jocs.2010.12.007

Bollen, J., Mao, H., & Zeng, X. (2011b). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1–8. https://doi.org/10.1016/j.jocs.2010.12.007

Borg, A., & Boldt, M. (2020). Using VADER sentiment and SVM for predicting customer response sentiment. *Expert Systems with Applications*, *162*. https://doi.org/10.1016/j.eswa.2020.113746

Boylan, D. H., Thomas, M., Matangi, E., & Brown, M. (2020). Semantic analysis: Can MD&A word choice increase stock price of american publicly traded banks disseminating using twitter? *International Journal of Digital Accounting Research*, *20*, 45–73. https://doi.org/10.4192/1577-8517-v20_3

Brans, H., & Scholtens, B. (2020). Under his thumb the effect of president Donald Trump's Twitter messages on the US stock market. *PLoS ONE*, *15*(3). https://doi.org/10.1371/journal.pone.0229931

Chahine, S., & Malhotra, N. K. (2018). Impact of social media strategies on stock price: the case of Twitter. *European Journal of Marketing*, *52*(7–8), 1526–1549. https://doi.org/10.1108/EJM-10-2017-0718

Dhar, S., & Bose, I. (2020). Emotions in Twitter communication and stock prices of firms: the impact of Covid-19 pandemic. *DECISION*, *47*(4), 385–399. https://doi.org/10.1007/s40622-020-00264-4

El-Din Hassanien, H. (2019). Web scraping scientific repositories for augmented relevant literature search using crisp-dm. *Applied System Innovation*, *2*(4), 1–22. https://doi.org/10.3390/asi2040037

Gonçalves, M., & Jacinto, P. (2021). *O Impacto do COVID-19 na Satisfação dos Clientes nos Hotéis: uma Análise de Text Mining*.

Hamraoui, I., & Boubaker, A. (2022). Impact of Twitter sentiment on stock price returns. *Social Network Analysis and Mining*, *12*(1). https://doi.org/10.1007/s13278-021-00856-7

Hazem, S.-M., Mohamed, E. K. A., & Ali, H. (2019). WHAT TWITTER CAN TELL US ABOUT THE STOCK MARKET. *Anduli*, *18*, 219–242. https://doi.org/10.12795/anduli.2019.i18.10

Jawid, R. F. (2022). Study of using applications of artificial intelligence in performance of financial markets. *Journal of Cases on Information Technology*, *24*(2). https://doi.org/10.4018/JCIT.20220401.oa4

Jena, P. R., & Majhi, R. (2023). Are Twitter sentiments during COVID-19 pandemic a critical determinant to predict stock market movements? A machine learning approach. *Scientific African*, *19*. https://doi.org/10.1016/j.sciaf.2022.e01480

Katsafados, A. G., Nikoloutsopoulos, S., & Leledakis, G. N. (2023). Twitter sentiment and stock market: a COVID-19 analysis. *Journal of Economic Studies*. https://doi.org/10.1108/JES-09-2022-0486

Liu, S. (2021). Competition and Valuation: A Case Study of Tesla Motors. *IOP Conference Series: Earth and Environmental Science*, *692*(2). https://doi.org/10.1088/1755-1315/692/2/022103

Maria, J., Ferreira, G., & do Sacramento, F. (2021). *Sentiment Analysis in the Stock Market Based on Twitter Data*.

Md Khudzari, J., Kurian, J., Tartakovsky, B., & Raghavan, G. S. V. (2018). Bibliometric analysis of global research trends on microbial fuel cells using Scopus database. *Biochemical Engineering Journal*, *136*, 51–60. https://doi.org/10.1016/j.bej.2018.05.002

Michalak, J. (2020). Does pre-processing affect the correlation indicator between Twitter message volume and stock market trading volume? *Ekonomia i Prawo*, *19*(4), 739. https://doi.org/10.12775/eip.2020.048

Moreno, Á. C. (2015). *Análise de Sentimentos na Classificação de Comentários Online Aplicando Técnicas de Text Mining* [ISCTE-IUL]. https://repositorio.iscte-iul.pt/bitstream/10071/11504/1/ACMoreno_MSIAD_F.pdf

Mourão, A. I. M. (2018). *"O Google como medida de sentimentos nos mercados financeiros."*

Nisar, T. M., & Yeung, M. (2018). Twitter as a tool for forecasting stock market movements: A short-window event study. *Journal of Finance and Data Science*, *4*(2), 101–119. https://doi.org/10.1016/j.jfds.2017.11.002

Nofer, M., & Hinz, O. (2015). Using Twitter to Predict the Stock Market: Where is the Mood Effect? *Business and Information Systems Engineering*, *57*(4), 229–242. https://doi.org/10.1007/s12599-015-0390-4

Nyakurukwa, K. (2022). *The wisdom of the Twitter crowd in the stock market: Evidence from a fragile state*. https://www.theguardian.com/technology/2021/nov/08/will-elon-musk-abide-by-twitter-poll-

Oliveira, N., Cortez, P., & Areal, N. (2017). The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems with Applications*, *73*, 125–144. https://doi.org/10.1016/j.eswa.2016.12.036

Pai, P. F., & Liu, C. H. (2018). Predicting vehicle sales by sentiment analysis of twitter data and stock market values. *IEEE Access*, *6*, 57655–57662. https://doi.org/10.1109/ACCESS.2018.2873730

Porshnev, A., & Redkin, I. (2014). Analysis of twitter users' mood for prediction of gold and silver prices in the stock market. *Communications in Computer and Information Science*, *436*, 190–197. https://doi.org/10.1007/978-3-319-12580-0_19

Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M., & Mozetič, I. (2015). The effects of twitter sentiment on stock price returns. *PLoS ONE*, *10*(9). https://doi.org/10.1371/journal.pone.0138441

Ruan, Y., Durresi, A., & Alfantoukh, L. (2018). Using Twitter trust network for stock market analysis. *Knowledge-Based Systems*, *145*, 207–218. https://doi.org/10.1016/j.knosys.2018.01.016

Schmidt, C. G., Wuttke, D. A., Ball, G. P., & Heese, H. S. (2020). Does social media elevate supply chain importance? An empirical examination of supply chain glitches, Twitter reactions, and stock market returns. *Journal of Operations Management*, *66*(6), 646–669. https://doi.org/10.1002/joom.1087

Sóti, A., Ancarani, A., Stéger, J., & Vattay, G. (2020). Influence of Twitter activity on the stock price of soccer clubs. *Social Network Analysis and Mining*, *10*(1). https://doi.org/10.1007/s13278-020-00691-2

Souza, D. M. S. de, & Martins, O. S. (2022). Brazilian stock market performance and investor sentiment on Twitter. *Revista de Gestao*. https://doi.org/10.1108/REGE-07-2021-0145

Tanaltay, A., Langroudi, A. S., Akhavan-Tabatabaei, R., & Kasap, N. (2021). Can Social Media Predict Soccer Clubs' Stock Prices? The Case of Turkish Teams and Twitter. *SAGE Open*, *11*(2). https://doi.org/10.1177/21582440211004153

Valle-Cruz, D., Fernandez-Cortez, V., López-Chau, A., & Sandoval-Almazán, R. (2022). Does Twitter Affect Stock Market Decisions? Financial Sentiment Analysis During Pandemics: A Comparative

Study of the H1N1 and the COVID-19 Periods. *Cognitive Computation*, *14*(1), 372–387. https://doi.org/10.1007/s12559-021-09819-8

Yang, S. Y., Mo, S. Y. K., & Liu, A. (2015). Twitter financial community sentiment and its predictive relationship to stock market movement. *Quantitative Finance*, *15*(10), 1637–1656. https://doi.org/10.1080/14697688.2015.1071078

Capítulo 7: Anexos

7.1. Anexo A

Tabela 7.1: Contextualização dos artigos da RSL

ID	Área de estudo	Nº de empresas	Nº de utilizadores	Nº de *tweets*	Período	País da amostra
1	Covid-19 e Finanças	N.E.	N.E.	Cerca de 1 milhão	5 meses (janeiro a maio de 2020)	EUA e India
2	Covid-19	N.E.	N.E.	1.616.007	janeiro a junho de 2021	7 países (Bélgica, Alemanha, França, Itália, Espanha, Reino e EUA)
3	Finanças	N.E.	N.E.	224,22 mil[24]	1/03/2021 a 21/03/2021	N.E.
4	Finanças	22 empresas listadas no *Tunindex*	N.E.	856	12 meses (1/10/2016 a 30/09/2017)	Tunísia
5	Finanças	N.E.	N.E.	N.E.	N.E.	N.E.
6	Finanças	N.E.	N.E.	N.E.	junho a julho 2009 (H1N1); janeiro a maio 2020 (Covid)	China
7	Finanças	1	N.E.	10.997	24/02/2019 a 22/06/2020	Zimbábue
8	Finanças	N.E.	N.E.	314.864	01/01/2017 a 31/12/2018	Brasil
9	Área desportiva	4 clubes de futebol turcos	N.E.	12.814.581	dezembro de 2017 a março de 2019	Turquia
10	Área desportiva	3 clubes de futebol	N.E.	N.E.	01/01/2012 a 28/12/2014	Itália, Inglaterra, Holanda
11	Finanças	150 empresas de capital aberto e 213 falhas na cadeia de suprimentos	N.E.	2,05 bilhões	janeiro de 2013 a dezembro de 2017	Estados Unidos da América (EUA)
12	Finanças	34 bancos americanos de capital aberto[25]	N.E.	N.E.	4 trimestres	Estados Unidos da América (EUA)
13	Momentos Eleitorais	1 de capital aberto	1[26]	100 *tweets*[27]	8/11/2016 a 8/11/2018 [28]	EUA

[24] 1,6 milhões e depois de filtrar ficaram 224,22 mil
[25] 17 usam o *twitter* e 17 não usam o *twitter*
[26] Conta de *Donald Trump (*@realDonaldTrump)
[27] Mais de 5600, mas foram eliminados *retweets*, *tweets* de outras empresas, etc, e sobraram 100 *tweets*
[28] Desde o dia das eleições a essa data 2 anos depois

Tabela 7.2: Contextualização dos artigos da RSL (continuação)

ID	Área de estudo	Nº de empresas	Nº de utilizadores	Nº de *tweets*	Período	País da amostra
14	Covid-19	105 listadas na NYSE da lista da *Fortune 1000*	N.E.	189.303	01/01/2019 a 30/06/2020 (6 trimestres)	N.E.
15	Finanças	N.E.	N.E.	N.E.	01/01/2016 a 31/12/2017	N.E.
16	Finanças	67 empresas[29]	N.E.	833 *tweets*[30]	01/01/2014 a 16/02/2016	N.E.
17	Finanças	312 empresas da *Fortune 500*	N.E.	N.E.	N.E.	N.E.
18	Momentos eleitorais	N.E.	N.E.	60.944	6 dias (04/05 a 9/05/2016)	Reino Unido
19	Finanças	8	3.929.933	775.928.121	8 meses (1/1/2015 a 31/8/2015)	N.E.
20	Venda de veículos	4	N.E.	6 milhões	janeiro de 2008 a julho de 2017	Coreia
21	Tecnologias / Finanças	N.E.	N.E.	N.E.	N.E.	N.E.
22	Finanças	3800	N.E.	31 milhões	22/12/2012 a 29/10/2015	EUA
23	Finanças	N.E.	N.E.	N.E.	2009 a 2014	N.E.
24	Finanças	50	154.203	1.606.104	15/02/2014 a 15/06/2014	EUA e Reino Unido
25	Finanças	30	N.E.	mais de 1,5 milhões	15 meses (01/06/2013 a 18/09/2014)	N.E.
26	Finanças	N.E.	N.E.	100 milhões	janeiro de 2011 a novembro de 2013[31]	Alemanha
27	Finanças	N.E.	N.E.	755 milhões	13/02/2013 a 29/09/2013	N.E.
28	Finanças	N.E.	2,7 milhões	9.853.498	28/02/2008 a 19/12/2008	EUA

Nota: N.E. – Não Especifica

[29] 100 maiores empresas, com base no tamanho da empresa e na capitalização do mercado, listadas na FTSE 100. Mas foram excluídas as empresas que não tinham *twitter*, restando 67 empresas

[30] 75.154 *tweets,* mas como só interessavam financeiros ficaram 1.747. Como só queriam um tweet de cada dia ficaram 854 e como ainda haviam *tweets* com falta de disponibilidade de dados de preços de ações à data do estudo, sobraram 833 tweets

[31] Dividido em dois períodos: janeiro/2011 a 17/março/2012 e dezembro/2012 a 30/novembro/2013

7.2. Anexo B

Tabela 7.3: Contextualização dos artigos da RSL para a parte financeira

ID	Modelo utilizada	Métrica de avaliação	Tipo de análise feita	Metodologia feita	Origem dos dados financeiros	Tipo de variáveis	Análise da relação entre as variações
1	LSTM; Modelos de séries temporais (modelo de médias moveis autorregressivos e modelo de regressão linear)	Correlação de *Pearson*; RMSE; MSE; R^2	Impacto da Covid-19 nos valores das ações dos diversos setores, nos EUA e na India	Recolha dos dados; Realização dos modelos selecionados; Divisão dos dados em dados de treino (80%) e dados de teste (20%); Treino dos modelos, ajustando os parâmetros necessários e, por fim, avaliação do desempenho do modelo	N.E.	N.E.	N.E.
2	Modelos Painel VAR; Painel ARDL	*P-value*; Teste de cointegração de *Pedroni*; Teste de causalidade de *Granger*; Teste de *Hausman*; Teste de robustez	Análise textual, onde é estudada a relação entre os sentimentos nas redes sociais e os mercados de ações	Seleção de amostra e variáveis; Pré-processamento dos dados	Para as variáveis de controlo: Banco de dados *Refinitiv Eikon* (para as duas primeiras); *Scott Baker, Nick Bloom* e *Steven Davis* (para as restantes duas)	Variáveis independentes[32] ; Variáveis de controlo[33]	Utilização das variáveis para estudar a relação entre o sentimento do *twitter* e o valor da ação
3	Modelo SSWN[34]; Modelo de espaço vetorial	Retorno do mercado; *Accuracy*; Previsão; *Recall*; *F-measure*; Curva ROC	Analisar os *tweets* e os dados do *Google Finance* para prever o futuro comportamento das bolsas de valores num período de tempo	Recolha dos dados; Cálculo do sentimento e treino do modelo; Criação do modelo de previsão do mercado de ações	N.E.	N.E.	N.E.

[32] Sentimento do *twitter* em relação ao Covid-19

[33] Preço do *bitcoin*; preço do petróleo; índice de volatilidade implícita (VIX); incerteza da política económica

[34] Para treinar máquinas de aprendizagem extrema (ELM) e rede neuronal recorrente (RNN)

Tabela 7.4: Contextualização dos artigos da RSL para a parte financeira (continuação)

ID	Modelo utilizada	Métrica de avaliação	Tipo de análise feita	Metodologia feita	Origem dos dados financeiros	Tipo de variáveis	Análise da relação entre as variações
4	Modelo VAR com AIC, SC e HQ	Correlação de *Pearson*; Teste de Causalidade *Granger*; Teste *Dickey-Fuller* Aumentado e teste Philipp Person[35]	Dependência entre o retorno de preços de ações e o sentimento de *tweets*	N.E.	N.E.	Variáveis de séries temporais estacionárias[36]	Série temporal de sentimentos expressos em *tweets*, com resoluções diárias
5	Classificação utilizando árvores de decisão, utilizando algoritmo C4.5; Algoritmo *Naive Bayes*	Previsão do preço das ações	Classificação dos utilizadores em categorias[37]	Classificação dos utilizadores em diversas categorias: ouro, prata, platina e bronze	*YahooFinance*	N.E.	N.E.
6	Detetar correlação através dos máximos valores de correlação	N.E.	Correlação entre os índices financeiros e as polaridades dos *tweets*	Metodologia baseada na análise do sentimento financeiro e correlações dos diferentes índices financeiros	*YahooFinance; Investing.com*	N.E.	N.E.

[35] Ambos para ver se as variáveis são não estacionárias
[36] Comprovado depois dos testes presentes nas métricas
[37] Com base na renda bruta anual, seguro de saúde e outros fatores financeiros

Tabela 7.5: Contextualização dos artigos da RSL para a parte financeira (continuação)

ID	Modelo utilizada	Métrica de avaliação	Tipo de análise feita	Metodologia feita	Origem dos dados financeiros	Tipo de variáveis	Análise da relação entre as variações
7	LSTM	Retornos compostos; Análise *Wavelet*; *P-value*	Relação entre o sentimento médio dos *tweets* e os retornos das ações da *Old Mutual Zimbabue Limited*	Análise dos dados e a utilização da Análise *Wavelet*. Esta análise é utilizada para ver a ligação entre o sentimento e os retornos das ações no mercado	Bolsa de Valores do Zimbábue	N.E.	N.E.
8	Regressões Quantílicas; Modelo VAR	Testes para a raiz unitária nas variáveis; Teste *Dickey-Fuller* aumentado; Teste de *Jarque-Bera;* Teste de rodustez; Matriz de correlação entre as variáveis; *AIC*; *SC*; Causalidade de *Granger*	Identificar o sentimento dos *tweets* e a sua relação com o retorno do mercado de ações brasileiro e o volume negociado	N.E.	*Refinitiv Eikon*	Variáveis heterogêneas, não normais e macroeconómicas	Ver se existe relação entre o sentimento e o retorno ou volume de negociação

Tabela 7.6: Contextualização dos artigos da RSL para a parte financeira (continuação)

ID	Modelo utilizada	Métrica de avaliação	Tipo de análise feita	Metodologia feita	Origem dos dados financeiros	Tipo de variáveis	Análise da relação entre as variações
9	3 modelos[38]; Classificadores Naïve Bayes, SVM e Regressão Logística	*Accuracy*; Matriz de confusão; *Recall*; *Precision*; *F-score*; R^2/MSE^2; *AIC*	Classificar os sentimentos dos *tweets* e prever o retorno no preço das ações dos 4 clubes	N.E.	*YahooFinance*[39]; Bolsa *Istanbul; BIST 100*	Variáveis dependentes[40]	As variáveis e o Bolsa *Istanbul BIST 100* são comparadas para ver as mudanças no mercado
10	Regressão linear	Teste F, *P-value*, Distribuição média, Desvio padrão, Valor médio do retorno anormal absoluto	Teste de causalidade de *Granger* que determina se as séries temporais são uteis para prever outras series temporais	N.E.	Bolsa italiana, *NYSE*, *Euronext Amsterdam*. *Yahoofinance*	Variáveis temporais	N.E.

[38] 1º baseado no desempenho da partida e probabilidade de apostar como indicador de sentimento; 2º Utilliza os dados do *twitter* como indicador alternativo dos sentimentos; 3º Combina os sentimentos do *twitter* com o desempenho da partida. São usados modelos de regressão e classificação

[39] Ações até março de 2019

[40] Data da transação, os preços de abertura e fecho da ação, os preços mais altos e mais baixos e o volume da ação vendida em uma determinada data

Tabela 7.7: Contextualização dos artigos da RSL para a parte financeira (continuação)

ID	Modelo utilizada	Métrica de avaliação	Tipo de análise feita	Metodologia feita	Origem dos dados financeiros	Tipo de variáveis	Análise da relação entre as variações
11	Modelo de 4 fatores; OLS; Modelo ajustado ao mercado e ajustado à média[41]; Modelo de Regressão de efeitos aleatórios	Retorno anual médio e acumulado; Significância estatística dos retornos anuais médios e acumulados; Teste *t* e dois testes não paramétricos[42]; R^2 ajustado e R^2; Teste de *PSM*	Volume de *tweets* e sentimento do tweet (relação linear)	Metodologia de estudo de eventos	Bolsa de valores de Nova York e *NASDAQ*[43]; *Thomson Reuters*, para os dados financeiros da empresa	N.E.	Relação entre a variável dependente e o retorno de ações, o volume de tweets e o sentimento do tweet
12	N.E.	Retorno anormal acumulado (CARs). Retorno diário (R_d). Teste t; F-valor; Distribuição normal	Entender se o sentimento expresso nos *tweets* irá ter influência no preço das ações da empresa	Metodologia da teoria financeira[44] e a metodologia de estudo de eventos[45]	*Measuringworth.com* *Yahoo!Finance.com*	Variável dependente[46] e 2 variáveis independentes[47]	As variáveis independentes são comparadas com a variável dependente do preço das ações

[41] Para estimar os retornos esperados do mercado de ações
[42] Teste de *Wilcoxon* e teste de sinais binomiais
[43] As duas maiores bolsas de valores dos EUA
[44] Menciona que os mercados de capitais abrangem todas as informações disponíveis sobre uma empresa
[45] Para avaliar a força que tem a metodologia da teoria financeira no preço das ações da empresa
[46] Variáveis no valor do acionista, consoante o aumento ou queda no preço das ações
[47] Uso de redes sociais e análise das *stopwords* com a ajuda de um *software* de análise semântica

Tabela 7.8: Contextualização dos artigos da RSL para a parte financeira (continuação)

ID	Modelo utilizada	Métrica de avaliação	Tipo de análise feita	Metodologia feita	Origem dos dados financeiros	Tipo de variáveis	Análise da relação entre as variações
13	Modelo de retornos ajustados ao mercado e ao risco	Retorno anormal médio e anormal cumulativo; EDA; Teste t (*p-value*)	Análise textual	Metodologia de estudo de eventos	*DataStream*	N.E.	N.E.
14	Regressão linear	Teste t; R^2; *P-value*	N.E.	N.E.	N.E.	Variável dependente[48]; Variáveis independentes[49]; Variáveis de controlo[50]	Perceber a influência que a covid 19 teve na variável dependente, utilizando as variáveis independentes
15	Classificador de Naive Bayes	Correlação de *Pearson*; Precisão; *Recall*; *F-mensure*	Abordagem de aprendizagem e Abordagem baseada em Léxico	N.E.	*Yahoo! Finance*	Variáveis independentes	N.E.

[48] Preço da ação de fechamento escalado min-max (*ScaledClose*)
[49] *HappyScore, AngryScore, FearScore, SadScore, SurpriseScore* e *Crisis* (variável *dummy* para a crise)
[50] Tamanho da organização (nº de funcionarios), volume de *tweets* e idade do autor do *tweet*

Tabela 7.9: Contextualização dos artigos da RSL para a parte financeira (continuação)

ID	Modelo utilizada	Métrica de avaliação	Tipo de análise feita	Metodologia feita	Origem dos dados financeiros	Tipo de variáveis	Análise da relação entre as variações
16	Modelo de mercado simples; Modelo de mercado ajustado ao risco	Retorno esperado das ações (no 1º modelo); Retorno esperado das ações com base no nível de risco (no 2º modelo); Retorno anormal médio; Retorno anormal médio acumulado; *T-value*	Avaliar o efeito que os anúncios de eventos financeiros têm sobre o rendimento das ações do *Financial Times Stock Exchange 100 Index*	Metodologia de estudo de eventos, em que existe um interesse nos eventos relativos a eventos financeiros anunciados nas contas de *Twitter* das empresas em estudo	*Thomson Reuters DataStream*	Variável independente (anúncios de eventos); Variável dependente (preço das ações)	Analisar o efeito que eventos financeiros partilhados no *Twitter* pelas empresas do *FTSE 100* no retorno das ações e visualizar a relação entre as duas variáveis
17	*Probit regression*; Modelo de *Heckman*	Retorno anual acumulado (CAR); Retorno anormal buy-and-hold (BHAR); Desvio padrão anualizado anormal; *F-statistics*	Análise do histórico de eventos (EHA), mais conhecida como estudo de eventos[51] Validação cruzada	Análise do histórico de eventos (EHA), mais conhecida como estudo de eventos, usada para analisar acontecimentos históricos nas empresas da Fortune 500 que têm uma conta de *Twitter*	*DataStream*	Variáveis independentes (*dummys*) e variáveis de controlo, como por exemplo: percentual de *dividend yield*	N.E.

[51] Usada para analisar acontecimentos históricos nas empresas da Fortune 500 que têm uma conta de *Twitter*

Tabela 7.10: Contextualização dos artigos da RSL para a parte financeira (continuação)

ID	Modelo utilizada	Métrica de avaliação	Tipo de análise feita	Metodologia feita	Origem dos dados financeiros	Tipo de variáveis	Análise da relação entre as variações
18	Regressão linear	Correlação de *Pearson*; R; R^2; R^2 ajustado; F-ratio	Análises de correlação e regressão para comparar a análise de sentimentos com as variações no preço do *FTSE 100*	Seleção da amostra; Método de cálculo do sentimentos e métodos de análise estatística	*Yahoo! Finance*	Variáveis dependentes e variáveis independentes	Comparação entre as variáveis dependentes e as variáveis independentes, através da análise de correlação baseada no volume[52].
19	Regressão linear	Covariância; Retornos esperados das ações; MAE, R^2 Ajustado, *P-value*; Correlação de *Pearson*	Correlação entre a valência do sentimento do *Twitter* e os retornos anormais das ações; Validação cruzada	N.E.	*Yahoo! Finance*	Variável independente (*RSP*), α é a interceção e β é o coeficiente linear; variáveis independentes e variáveis de controle	N.E.

[52] Cada variável necessita de ser normalizada usando o z-scores com a média e o desvio-padrão como base

Tabela 7.11: Contextualização dos artigos da RSL para a parte financeira (continuação)

ID	Modelo utilizada	Métrica de avaliação	Tipo de análise feita	Metodologia feita	Origem dos dados financeiros	Tipo de variáveis	Análise da relação entre as variações
20	LSSVR Modelos de séries temporais[53]	*MAPE (%),* *WAPE (%);* *NMAE*	Prever as vendas totais mensais de veículos	7 abordagens: método de Naive, ES, ARIMA, SARIMA, BPNN, regressão vetorial de suporte de mínimos quadrados com modelos de séries temporais e a regressão de vetor de suporte de mínimos quadrados	*Bureau of Economic Analysis;* *US Department of Commerce;* *Yahoo! Finance*	Variáveis independentes	N.E.
21	Classificadores Naive Bayes (NBC)[54]	*Recall,* Precisão, *F-Measure,* *Accuracy*	Análise temporal (data/hora do tweet) e espacial (localização geográfica do tweet)	Colheita dos dados; Pré-processamento; Caracterização dos sentimentos; Classificação e desempenho e avaliação	N.E.	N.E.	N.E.

[53] *Naïve model, the exponential smoothing (ES) model, the autoregressive integrated moving average (ARIMA) model, the seasonal autoregressive integrated moving average (SARIMA) model, backpropagation neural networks (BPNN)* e *least squares support vector regression with time series (LSSVRTS)*

[54] Dividido em NB Gaussiano, NB Multinomial, NB Bernoulli e BN Semi-Supervisionado

Tabela 7.12: Contextualização dos artigos da RSL para a parte financeira (continuação)

ID	Modelo utilizada	Métrica de avaliação	Tipo de análise feita	Metodologia feita	Origem dos dados financeiros	Tipo de variáveis	Análise da relação entre as variações
22	SVM; Árvore de decisão; Rede neuronal; Regressão linear múltipla	*MAE* e *NMAE;* Retornos diários; Teste de *Diebold-Mariano (DM)*	Avaliar o valor dos dados de microblogging para prever variáveis do mercado de ações: retornos, volatilidade e volume de negociação de diversos índices e carteiras	Metodologia para avaliar o valor dos dados de *microblogging* para prever as variáveis do mercado de ações[55]	*DataStream;* *Thompson Reuters Datastream*	Variáveis independentes e uma variável dependente (Reg LM)	N.E.
23	Regressão linear	Teste *Dickey-Fuller* aumentado para a polaridade média e polaridade média ponderada; *Drawdown sharpe*; Valor de alavancagem	N.E.	N.E.	*Wharton Research Data Services (WRDS)*	Variáveis independentes	N.E.

[55] Utilizados indicadores de sentimento e atenção extraídos de *microblogs* e para uso de um Filtro Kalman para misturar o *microblog* e as fontes de pesquisa. São utilizados vários métodos de aprendizado de máquina e o teste *Diebold-Mariano* para validar se as previsões baseadas em sentimento e atenção são valiosas quando comparadas com uma linha de base autorregressiva

Tabela 7.13: Contextualização dos artigos da RSL para a parte financeira (continuação)

ID	Modelo utilizada	Métrica de avaliação	Tipo de análise feita	Metodologia feita	Origem dos dados financeiros	Tipo de variáveis	Análise da relação entre as variações
24	Regressão linear; Regressão Multivariada	*P-value;* Retorno diário do mercado e volatilidade*;* *MSE*	Identificar a quantidade de investimento correspondente e verificar os principais influenciadores nas redes sociais; Validação cruzada	É utilizado um método de maneira a criar uma comunicação financeira[56] e depois tentam compreender as características destas comunidades financeiras no *Twitter* através da observação direta. Na análise de sentimentos é realizada uma metodologia à parte que será dividida em 4 componentes[57]	N.E.	Variável dependente (retorno diário do índice de mercado) e variável independente	Relação entre a variável dependente e a variável independente, isto é, vai ser analisado se a variável independente tem uma relação significativa no retorno do mercado e na volatilidade
25	Método de aprendizagem supervisionada[58]; Modelo *SVM*; Modelo de Retorno Normal; *Naive Bayes*	Correlação de *Pearson;* Causalidade de *Granger*; Teste *Dickey-Fuller;* *Teste Ljung-box*; Teste *F*; Retornos anormais, reais e normais esperados; *recall*	Relação entre os retornos dos preços das ações e o sentimento expresso nos *tweets*; Validação cruzada	Aprendizagem supervisionada e aprendizagem de máquina[59]; depois os métodos utilizados para fazer a análise da correlação e por fim, é apresentada a metodologia do estudo de eventos	Site da *NASDAQ*	N.E.	N.E.

[56] Subconjuntos de utilizadores do *Twitter* com interesses semelhantes no mercado financeiro
[57] Pré-processamento dos dados; a correspondência de entidade; caracterização do sentimento da *stopword* e cálculo da pontuação do sentimento.
[58] Para cálculo do sentimento do *Twitter*
[59] De modo a classificar os sentimentos

Tabela 7.14: Contextualização dos artigos da RSL para a parte financeira (continuação)

ID	Modelo utilizada	Métrica de avaliação	Tipo de análise feita	Metodologia feita	Origem dos dados financeiros	Tipo de variáveis	Análise da relação entre as variações
26	*OLS regression*; Modelo VAR bivariado	R^2; AIC; BIC; Teste de causalidade de *Granger*	Interdependência linear entre a análise de sentimentos e os retornos das ações (modelo VAR)	N.E.	N.E.	Variável dependente[60]; Variáveis *dummys*	N.E.
27	SVM; Redes neuronais	*Accuracy*; Correlação entre quantidade normalizada de *tweets* em cada categoria e preços da prata e ouro no mercado de ações	Analisar o humor dos utilizadores do *Twitter* para melhorar a precisão das previsões para os preços de mercado do ouro e da prata	Algoritmo baseado no cálculo da quantidade de palavras para cada emoção. Como o mesmo foi muito preciso, ajudou a criar uma abordagem baseada em léxico	*InvesterPoint*	N.E.	Influencia das variáveis (crescimento, valor de abertura e de fecho nos 3 últimos dias e valor máximo e mínimo nos 3 últimos dias) no valor da prata e do ouro
28	*Self-organizing Fuzzy Neural Network* (SOFNN)*;* Regressão Linear Multipla	*MAPE (%)*; Causalidade de *Granger;* Teste *t;* R^2 Ajustado*;* *Teste F;* *P-value*	Analisar se os estados de humor dos *tweets* estão correlacionados com o valor do *DJIA* ao longo dos anos; Validação cruzada	N.E.	*Yahoo! Finance*	N.E.	N.E.

Nota: N.E. – Não Especifica

[60] Retorno entre o ganho ou perda entre o primeiro e último preço do dia da negociação

7.3. Anexo C

Tabela 7.15: Contextualização dos artigos da RSL para a análise de sentimentos

ID	*Software* utilizado	Índice Utilizado	Tipo de sentimento	Utilização dos *tweets*
1	N.E.	*DJIA;* *Nifty*	Aptidão, entusiasmo e particularidade (todos a variar entre -1 e +1)	Os *tweets* foram divididos tendo em conta a localização dos utilizadores e foram filtrados pelas *stopwords* "Covid-19" e "Corona". Foi feito o pré-processamento dos dados, em que, por exemplo, foram substituídos os valores por o valor da média, moda ou mediana e posteriormente é feita a análise de sentimentos
2	*VADER;* *API* do *Twitter*	N.E.	Positivo e negativo	N.E.
3	SentiWordNet; *Sentiment 140*; *Tweepy*[61]; *SenticNet*; *AFINN*; *SO-CAL*	Pontuação de sentimento	Positivo e negativo (No caso do *AFFIN* e do *SO-CAL* o sentimento varia entre - 5 e + 5)	Após a extração dos *tweets*, existe uma limpeza dos mesmos, em que são removidos *tweets* que sejam *spam* ou que sejam irrelevantes. Eliminam-se também os *links*, *tweets* duplicados, etc. Posteriormente, é feita a fase da identificação do sentimento associado a cada tweet[62].
4	N.E.	Polaridade dos sentimentos	Positivo, negativo e neutro	N.E.
5	*API OAuth*	Índice de sentimento; Índice de discrepância de sentimento	Positivo e negativo	N.E.
6	*Bing Liu;* *Sentiment 140;* *Afim*	*(i) IPC; (ii) S&P 500; (iii) NASDAQ 100; (iv) DJIA; (v) FTSE 100; (vi) BOVESPA; (vii) CAC 40; (viii) DAX; (ix) Hang Seng*	Bing Liu (positivo e negativo); Sentiment 140 (0 para negativo, 2 para neutro e 4 para positivo); Afim (-5 polaridade negativa a +5 polaridade positiva);	*Tweets* foram analisados para determinar uma abordagem baseada em léxico, havendo assim um pré-processamento dos dados

[61] *API* do *Twitter*

[62] A divisão em sentimento neutro é eliminada pois estes não desempenham nenhuma polaridade ao sentimento.

Tabela 7.16: Contextualização dos artigos da RSL para a análise de sentimentos (continuação)

ID	***Software* utilizado**	**Índice Utilizado**	**Tipo de sentimento**	**Utilização dos *tweets***
7	N.E.	N.E.	Positivo, negativo e neutro	O sentimento médio dos *tweets* é extraído da *Bloomberg Inc* e existe um cálculo dos sentimentos médios e segue-se uma análise manual do conjunto dos *tweets* com a ajuda dos diferentes especialistas. Nesses *tweets* que são analisados manualmente são utilizados modelos de Aprendizagem de máquinas e posteriormente são usados para analisar novos *tweets* e onde são classificados de -1 a +1 pelo seu grau de sentimento.
8	*API Google Cloud Natural Language* utilizando a biblioteca *Tweepy* para a recolha dos *tweets*	Índice de Sentimento do Investidor (ISI)	Varia entre 1 e -1, sendo 1 o sentimento positivo, -1 negativo e 0 neutro	Os *tweets* foram recolhidos segundo alguns parâmetros que são: estarem em português e conterem termos relacionados com ações ou índices de mercado. Posteriormente, foram eliminados *tweets* que incluíam *links*, *tweets* que incluíam a rede social *youtube* e por fim, *tweets* repetidos
9	*API* pública do *Twitter* usando Logstash[63] e Elasticsearch[64] ; Apache Spark (para extração de emoções)	N.E.	Positivo, negativo e neutro	Preparação dos dados e extração dos mesmos, de forma a utilizar metodologias supervisionadas ou não supervisionadas. Após esta fase é realizada a mineração dos dados de forma a fazer a extração dos sentimentos, em que há a extração das emoções, limpeza do texto, extração das emoções e por fim, a construção dos modelos, validando-os.
10	*API* do *Twitter*	*DJIA*	N.E.	Os dados foram filtrados tendo em conta o seu conteúdo, isto é, todos os registos sobre os clubes de futebol mais *tweetados*. Devido ao tipo de informação presente, foi necessário selecionar as abordagens de estudo de caso múltiplo. Posteriormente, os *tweets* foram filtrados com base nas *stopwords* dos *tweets*.
11	N.E.	*S&P500*; Índice de volume; Índice de sentimento	Positivo, negativo e neutro	N.E.

[63] Para recolha dos *tweets*

[64] Para indexação

Tabela 7.17: Contextualização dos artigos da RSL para a análise de sentimentos (continuação)

ID	*Software* utilizado	Índice Utilizado	Tipo de sentimento	Utilização dos *tweets*
12	*Soper Sentiment Analyzer*; *Sentiment32*	N.E.	1 para positivo, 0 para neutro e -1 para negativo (*soper sentiment analyzer*) 1 para positivo, 0 para neutro e -1 para negativo (*sentiment32*)	N.E.
13	*SentiStrength*	*S&P500*	-5 (emoção ou energia negativa muito forte) a +5 (emoção ou energia positiva muito forte)	N.E.
14	*API text2emotion;* Léxico *VADER*	Índice de felicidade	Felicidade, raiva, medo e tristeza	N.E.
15	*API* do *Twitter* ou *Twitter.search.web*	*Twitter sentiment analysis*	Positivo, negativo e neutra	Existem 2 abordagens: Abordagem de aprendizagem (treina no algoritmo) e Abordagem baseada em Léxico (usa o n-gram e o TF-IDF). Há assim, um pré-processamento dos dados
16	*API* do *Twitter*	N.E.	Positivo e não positivo	São feitas diversas divisões nos *tweets* de modo a perceber se têm influência nas métricas dos resultados dos modelos
17	N.E.	*S&P500*	N.E.	N.E.
18	*Tweetcatcher* com acesso à *API* do *Twitter*	Classificador de sentimento baseado em léxico - *Umigon*	Positivo, negativo e neutro	É feita uma limpeza nos dados onde são apenas incluídos *tweets* onde aparecem as *hashtags* em estudo. Posteriormente e ainda nesta limpeza, são apenas incluídos *tweets* onde a sua localização seja no Reino Unido. Também são incluídos tanto *tweets* como *retweets*, pois quanto mais vezes um *tweet* for *retweetado* mais aceite é. Por fim, são eliminados *tweets* que sejam considerados *spam*.
19	*SentiStrength* e dicionário financeiro de *Loughran* e *McDonald's*; *API* aberta do *Twitter e* a biblioteca *twitter4j*	*S&P500*	Positivo e negativo	São extraídos três tipos de *tweets*: menções; respostas e *retweets*. O sentimento de um *tweet* é avaliado pela valência ou polaridade
20	*SentiStrength*	*DJIA;* *S&P500*	Positivo (1 a 5); negativo (-1 a -5)	Antes dos *tweets* serem analisados, foram realizados processos de pré-processamento. Estes processos de pré-processamentos incluem eliminar *tweets* porque são normalmente *tweets* publicitários e eliminar *tweets* que incluam sites ou símbolos

Tabela 7.18: Contextualização dos artigos da RSL para a análise de sentimentos (continuação)

ID	*Software* utilizado	Índice Utilizado	Tipo de sentimento	Utilização dos *tweets*
21	*API* do *Twitter* (*API REST*); *Oauth*	N.E.	Positivo, negativo e neutro	O pré-processamento dos dados é feito usando o *NLP* onde começa com uma etapa de transformações e termina com a extração dos recursos necessários. Nesta etapa de transformação existe a tokenização do texto e a eliminação de *hastags*, *URLs.*
22	*API Rest*[65]	*S&P500;* *NASDAQ100;* *DJIA;* *Russell 2000;* RMRF; SMB; *HML;* *MOM;* *Volatility Index (VIX);* *Psize*	Positivo, negativo e neutro	O nº dos *tweets* foi usado para criar indicadores. Foi utilizada a primeira diferença do volume de postagem porque há um visível crescimento do número de *tweets* durante o período analisado. Para criar os indicadores de sentimento do investidor, foram utilizadas as pontuações de sentimento produzidas pela análise de sentimento de todos os *tweets.* São realizadas tarefas de pré-processamento como a substituição de todas as *cashtags* por "tkr", todos os endereços de internet por "URL", a tokenização e identificar quais são positivos e quais são negativos.
23	*API Topsy;* *"Pattern"* do *Python*; *Senti Wordnet*	N.E.	-1 a +1 (-1 puramente negativo; +1 puramente positivo)	N.E.
24	*API* do *Twitter*; *SentiWordNet*	*DJIA;* *S&P500;* *NASDAQ*; *Russel 3000 ETF*	Positivo, negativo e neutro	Realizaram um pré-processamento dos dados que inclui a filtragem dos dados, onde serão eliminados *links*, *hatshtags*, etc; a *stemming*, em que cada palavra é transformada na palavra raiz; divisão de palavras e reconhecimento de *n-gram.* Realiza também uma correspondência com a entidade financeira, isto é, identifica os *tweets* relacionados aos temas do mercado financeiro, que inclui a extração da entidade dos principais utilizadores financeira do *twitter*, o nome da empresa e o seu símbolo e o cálculo da pontuação da entidade financeira. Depois realiza caracterização do sentimento da *stopword* onde será importante perceber se existem palavras que tornam frases positivas em frases negativas como por exemplo "não". Por fim, é feito o cálculo da pontuação diária do sentimento que é calculado para o dia de observação para um determinado utilizador.

[65] *R* no processamento e *Mongodb* no armazenamento dos dados

Tabela 7.19: Contextualização dos artigos da RSL para a análise de sentimentos (continuação)

ID	*Software* utilizado	Índice Utilizado	Tipo de sentimento	Utilização dos *tweets*
25	*API* do *Twitter*; Método de aprendizagem supervisionado (para cálculo do sentimento)	*DJIA*; *S&P500*	Positivo, negativo e neutro	É feito o pré-processamento dos dados, onde são eliminados os *URLs* e o @ do utilizador, de modo a que este não interfira na análise. Posteriormente são feitos os procedimentos de tokenização, lematização, construção de grama (n-grama e bi-grama) e o TF-IDF.
26	*API* do *Twitter*; Para análise do Humor: *ASTS* e *WATS*	*S&P500; GFK*; Índice de *Sharpe*	Tristeza, desespero, cansaço, raiva e humor positivo	Todos os *tweets* incluem o *ID* do *tweet*, a hora da publicação, número de seguidores e *retweets* e a informação do *tweet*. São eliminados todos os *tweets* que não conseguem ser classificados como positivos ou negativos pelo dicionário presente no estudo.
27	*API* do *Twitter*	*DJIA*; *S&P500; NASDAQ*	Feliz, apaixonado, calmo, enérgico, com medo, com raiva, cansado e triste	N.E.
28	*OpinionFinder* e *Google-Profile of Mood States* (GPOMS)	*DJIA*	Positivo e negativo (*OpinionFinder*); Calma, alerta, certo, vital, gentil e feliz (GPOMS)	Serão removidas todas as *stopwords* e pontuações e são juntados todos os *tweets* que foram escritos na mesma data

Nota: N.E. – Não Especifica