INSTITUTO
UNIVERSITÁRIO
DE LISBOA

A Utilização da Aprendizagem Automática e do Text Mining na Previsão do Mercado de Ações

Alexandre Costa Pereira

Mestrado em Ciência de Dados

Orientadora:
Doutora Diana Aldea Mendes, Professora Associada,
ISCTE – Instituto Universitário de Lisboa

Co-Orientador:
Doutor Fernando Manuel Marques Batista, Professor Associado,
ISCTE – Instituto Universitário de Lisboa

Outubro, 2023

Departamento de Métodos Quantitativos para Gestão e Economia

Departamento de Ciências e Tecnologia da Informação

A Utilização da Aprendizagem Automática e do Text Mining na Previsão do Mercado de Ações

Alexandre Costa Pereira

Mestrado em Ciência de Dados

Orientadora:

Doutora Diana Aldea Mendes, Professora Associada,

ISCTE – Instituto Universitário de Lisboa

Co-Orientador:

Doutor Fernando Manuel Marques Batista, Professor Associado,

ISCTE – Instituto Universitário de Lisboa

Outubro, 2023

Agradecimentos

Terminada a elaboração desta dissertação, gostaria de agradecer a todos os que tiveram um papel essencial durante a mesma, assim como no decorrer de todo o Mestrado.

Em primeiro lugar, gostaria de agradecer aos meus orientadores, professora Diana Mendes e professor Fernando Batista, por todas as sugestões e críticas construtivas que me apresentaram durante as inúmeras reuniões realizadas no decorrer desta dissertação. O culminar deste trabalho não seria possível sem o incansável auxílio que me prestaram nos últimos meses.

Aos meus pais, por me encorajarem a inscrever neste mestrado e por sempre me motivarem no decorrer do mesmo. Estou certo de que, sem o vosso apoio, a conclusão deste trabalho não teria sido possível. O entusiasmo, seriedade e empenho que colocam em tudo o que fazem servem e sempre servirão de estímulo para eu fazer mais e melhor.

Aos meus amigos, por estarem sempre ao meu lado, e pela compreensão que demonstraram nos últimos meses quando não pude comparecer a alguns eventos, a fim de me dedicar à conclusão da dissertação.

Resumo

É inquestionável que a Inteligência Artificial e os algoritmos de *Machine Learning* estão, e continuarão nos próximos anos, a mudar o mundo e a forma como vivemos o quotidiano. Diversas áreas alteram regularmente o seu *modus operandi*, em conformidade com as novas melhorias que vão surgindo diariamente nos diferentes sectores que envolvem *Machine Learning*; a área de negócio da bolsa e da previsão das respetivas ações não é exceção.

No decorrer desta dissertação procurou-se perceber o impacto que os modelos de *Machine Learning* conseguem atingir na previsão do valor de ações, quando aplicados em conjunto com técnicas de *Text Mining* e *Web Scraping*. Os modelos de *Deep Learning* escolhidos para esta investigação foram o LSTM e o BiLSTM, por se considerar que eram aqueles que ofereciam uma melhor resposta ao âmbito proposto. Para além disso, os modelos utilizados adequam-se bem ao conjunto de dados utilizados.

Os principais conjuntos de dados utilizados no decorrer da dissertação foram os valores de fecho e de volume de diferentes índices, assim como notícias financeiras e relatórios do *Federal Reserve System*. A partir destes, foi possível calcular diversas variáveis que se utilizaram como *input* nos modelos de *Deep Learning*.

Apesar da considerável quantidade de técnicas que se podem aplicar, importa também realçar as naturais limitações que estas ainda apresentam, oferecendo assim espaço para esta área se auto superar.

Keywords: *Machine Learning*; *Text Mining*; *Web Scraping*; Bolsa de Valores

Abstract

It is unquestionable that Artificial Intelligence and Machine Learning algorithms are, and will continue to be in the coming years, changing the world and the way we live our daily lives. Various fields regularly adjust their modus operandi in accordance with the new improvements that emerge daily in the different sectors involving Machine Learning; the business area of stock market and the prediction of respective actions is no exception.

Throughout this dissertation, we sought to understand the impact that Machine Learning models can achieve in predicting stock values when applied in conjunction with Text Mining and Web Scraping techniques. The Deep Learning models chosen for this research were LSTM and BiLSTM, as they were considered to provide a better response within the proposed scope. Furthermore, the models used were well-suited to the dataset employed.

The primary datasets used during the dissertation consisted of closing values and volume data from various financial indices, as well as financial news and reports from the Federal Reserve System. From these, it was possible to calculate various variables that were used as inputs in the Deep Learning models.

Despite the considerable array of techniques that can be applied, it is also important to highlight the inherent limitations that still exist, thus providing room for this field to continually self-improve.

Keywords: *Machine Learning*; *Text Mining*; *Web Scraping*; Stock Exchange

Índice

Índice de Tabelas

Índice de Figuras

Glossário

BERT – Bidirectional Encoder Representations from Transformers

NLP – Natural Language Processing

RMSE – Root Mean Square Error

MSE – Mean Square Error

MAE – Mean Absolute Error

MAPE – Mean Absolute Percentage Error

1. Introdução

O mercado de ações desempenha um papel fundamental na economia global, sendo um dos principais motores de crescimento económico e uma fonte de investimento e financiamento, quer para empresas como indivíduos.

A volatilidade e imprevisibilidade inerentes ao mercado de ações apresentam desafios significativos para investidores e analistas financeiros. Tendo isso em consideração, a capacidade de prever as tendências do mercado e tomar decisões de investimento informadas é do maior interesse para os diferentes atores do mercado financeiro.

1.1 Background

1.1.1 Machine Learning

Machine Learning é um ramo que dá aos computadores a capacidade de aprender sem serem explicitamente programados, baseando-se na ideia de que os sistemas podem aprender a partir dos dados, identificando padrões e tomando decisões com a mínima intervenção humana. Os dois principais tipos de *Machine Learning* são a aprendizagem supervisionada e a aprendizagem não supervisionada. Na primeira, o computador é treinado com um conjunto de dados "rotulados" (*labeled*), onde o objetivo é construir um modelo que possa fazer previsões em exemplos novos, que ainda não tenham sido analisados pelo modelo, e que sejam tirados da mesma distribuição que o conjunto de treino. Aprendizagem supervisionada é tipicamente utilizado em tarefas como classificação de imagens, previsão do preço de uma casa, processamento de linguagem natural, entre outras. Na aprendizagem não supervisionada, não são fornecidos dados rotulados ao computador, pelo que este deve descobrir a estrutura dos dados através de técnicas como o *cluster* de dados. As técnicas de aprendizagem não supervisionada são normalmente utilizadas para descobrir padrões ocultos nos dados ou para aprender mais sobre a estrutura subjacente dos dados.

Existem diferentes tipos de modelos de *Machine Learning* que podem ser usados para uma panóplia de tarefas. Alguns dos modelos mais comuns são:

- *Linear Regression*: modelo utilizado para prever um valor contínuo, onde se assume que existe uma relação linear entre as variáveis independentes e a variável dependente.

- *Decision tree*: este modelo é uma estrutura de árvore semelhante a um fluxograma, usada para tomar decisões com base em determinadas condições. Cada nó interno na árvore representa um atributo, e cada nó folha representa uma "*label*" de uma classe.
- *Random Forest*: modelo que consiste em múltiplas árvores de decisão treinadas com conjuntos de dados diferentes, onde a previsão final é feita através da agregação das previsões individuais de cada uma das árvores.
- *Neural network*: este é um modelo mais complexo, inspirado na estrutura e na função do cérebro humano, e que consiste em múltiplas camadas de neurónios conectados, que podem aprender e se adaptar a novos dados.

Os artigos analisados utilizam diferentes modelos de *Machine Learning*, consoante o conjunto de dados que têm à sua disposição, ou das necessidades do problema que procuram resolver. Em alguns casos, utilizam até mais do que um modelo, de forma a conseguirem perceber qual é que se adequa mais às necessidades do problema em questão. Desta forma, não seria correto afirmar que um determinado modelo é sempre mais eficiente que os demais, uma vez que a precisão de cada modelo depende do conjunto de dados e do problema que se propõe resolver.

1.1.2 Text Mining

Text Mining consiste no processo de extração de informação útil de quantidades consideráveis de texto não estruturado. Neste processo, incluem-se tarefas como identificar padrões, encontrar relações entre palavras ou frases, e extrair dados específicos de um texto. *Text Mining* é amplamente utilizado em aplicações como análise de sentimento, classificação de documentos, deteção de spam, entre outras tarefas.

Esta área é um campo em constante evolução, pelo que existem diversas técnicas que podem ser utilizadas, dependendo do âmbito em questão, e certamente muitas outras técnicas serão criadas ou melhoradas num futuro próximo. Algumas das mais conhecidas e utilizadas atualmente são:

- *Bag of words* (BoW): é um modelo de *Text Mining* relativamente simples, onde o texto é representado como um conjunto de palavras e onde é tida em consideração a cardinalidade das palavras, isto é, onde é contabilizado o número de vezes que cada

palavra aparece no texto. Este modelo não considera a gramática nem a ordem das palavras.

- *Term Frequency-Inverse Document Frequency* (TF-IDF): o propósito deste modelo passa por refletir o quão importante é uma palavra num documento. Como o nome indica, este modelo é o produto de outras duas aplicações: *term frequency*, que consiste na frequência relativa de um termo dentro de um documento, e *inverse document frequency*, que mede quanta informação uma palavra fornece, através da análise do quão raro é esta aparecer ao longo de todos os documentos em análise. Desta forma, o valor do TF-IDF aumenta proporcionalmente ao número de vezes que uma palavra aparece num documento.
- *Bidirectional Encoder Representations from Transformers* (BERT): o principal objetivo deste modelo passa por ajudar a perceber o significado de linguagem ambígua em texto, utilizando o seu conteúdo para estabelecer contexto.

No decorrer desta revisão, serão analisadas diferentes aplicações destes modelos, para diferentes conjuntos de dados e diferentes âmbitos. Pretende-se assim perceber qual das técnicas é mais indicada para cada caso.

1.1.3 Web Scraping

Web Scraping consiste numa técnica para extrair informações de um *website* de forma automatizada, onde tipicamente se utiliza um programa para simular a ação de um utilizador. A informação extraída é normalmente utilizada para diferentes fins, tais como recolha de dados, monitorização de preços, entre outras.

Importa destacar que *Web Scraping* pode ser uma atividade ilegal, ou pelo menos eticamente questionável, dependendo do *website* que se está a analisar, ou da forma como se estão a utilizar os dados. Alguns *websites* proíbem explicitamente a utilização de técnicas de *web scraping* no mesmo, podendo mesmo agir judicialmente contra quem desrespeitar essa norma. HiQ Labs, Inc v. LinkedIn Corp é um caso judicial que retrata esta situação: o conhecido grupo LinkedIn colocou em tribunal a empresa HiQ Labs, alegando que estes extraíam dados do seu *website* e os utilizavam de forma inadequada, indo assim contra os interesses do LinkedIn, a quem o tribunal eventualmente deu razão em novembro de 2022. Tendo tudo isso em consideração, é de elevada importância aplicar estas técnicas sempre com a máxima precaução.

À imagem do que acontece com as técnicas de *Text Mining*, também as de *Web Scraping* estão em constante desenvolvimento e mudança, dado o quão recente esta é, pelo que existem algumas opções que se podem utilizar para recolher dados de um *website*. No entanto, importa destacar a biblioteca de Python "*Beautiful Soup*", dada a sua regular utilização nos artigos analisados.

1.2 Definição do Problema de Investigação.

O problema de investigação desta dissertação consiste em explorar a possibilidade de prever o mercado de ações, utilizando para isso técnicas avançadas de *Machine Learning*, *Web Scraping*, *Text Mining* e Análise Técnica. O objetivo passa por desenvolver um modelo capaz de analisar dados extraídos de diferentes fontes, como os valores de abertura, fecho e volume de ações transacionadas do mercado de ações, a análise de sentimento de notícias financeiras e relatórios financeiros do *Federal Reserve System*, assim como o cálculo de algumas métricas de análise técnica, que ajudem a perceber as oscilações da bolsa de valores.

Desta forma, pretende-se prever as tendências futuras do mercado de ações com maior precisão, precisão essa que pode fornecer valiosos *insights* para investidores e analistas financeiros, auxiliando-os na tomada de decisões informadas e a obter melhores resultados financeiros.

A previsão do mercado de ações é um tema de enorme interesse para analistas financeiros, investidores e instituições financeiras. A capacidade de antecipar os movimentos do mercado pode proporcionar vantagens competitivas, auxiliar na tomada de decisões de investimento e gerar retornos financeiros significativos. Importa, no entanto, realçar que a previsão exata do mercado de ações é uma tarefa árdua, devido à complexidade, volatilidade e imprevisibilidade do mesmo. Todos nos recordamos da pandemia COVID-19 que, sendo completamente imprevisível, impactou o mercado financeiro de formas que ninguém conseguiria antecipar.

As abordagens habituais e mais comuns de previsão do mercado de ações apresentam algumas limitações, e é com isso em mente que as técnicas de *Machine Learning*, *Web Scraping* e *Text Mining* se destacam como uma alternativa promissora.

Através da utilização de diferentes técnicas e fontes de dados, esta dissertação procura contribuir para uma maior precisão na previsão do mercado de ações. A aplicação de técnicas de *Machine Learning*, *Web Scraping* e *Text Mining* na previsão do mercado de ações tem o

potencial de melhorar a precisão das mesmas, fornecer *insights* valiosos sobre o comportamento do mercado, assim como apoiar investidores e analistas financeiros na tomada de melhores decisões. Além disso, a combinação integrada de diferentes fontes de dados e técnicas permite obter uma visão abrangente e multidimensional do mercado de ações, ajudando a proporcionar uma compreensão mais precisa e completa dos fatores que influenciam o seu comportamento.

1.3 Questões de Investigação e Objetivos da Investigação

Com a realização desta dissertação pretende-se contribuir para o avanço do conhecimento na área da previsão da bolsa de valores, fornecendo *insights* valiosos sobre a utilização de diferentes técnicas que podem contribuir para tal, onde se destacam *Machine Learning*, *Web Scraping*, *Text Mining* e análise técnica. A integração de diferentes fontes de dados e técnicas permite uma compreensão mais abrangente e precisa do mercado de ações, auxiliando investidores e analistas financeiros na tomada de decisões informadas e na sua procura por melhores resultados financeiros.

Espera-se também que os resultados e conclusões obtidos possam ser aplicados no campo da tomada de decisões financeiras, fornecendo informações importantes para qualquer pessoa ou instituição que se aventure na imprevisibilidade dos mercados financeiros. A pesquisa e trabalho aqui apresentados procuram contribuir para o desenvolvimento de metodologias mais precisas e eficazes, que superem as limitações das abordagens tradicionais de previsão do mercado de ações, utilizando para isso diversas fontes de dados e técnicas avançadas. Compreender e antecipar o comportamento do mercado de ações é de fulcral importância para aprimorar o desempenho financeiro e maximizar os retornos dos investimentos.

1.3.1 Questões de Investigação

De forma a alcançar o objetivo geral desta dissertação, foram formuladas as seguintes questões de investigação:

- Quais são as fontes de dados com maior relevância na previsão do mercado financeiro?
- De que forma é possível utilizar técnicas de *Web Scraping* para recolher e extrair informações relevantes das diferentes fontes de dados?

- É viável efetuar análise de sentimento em textos financeiros com o propósito de prever as oscilações da bolsa de valores? E se sim, de que forma é que esta se pode aplicar de forma a obter uma maior perceção do mercado financeiro?
- As técnicas de análise técnica têm sido utilizadas nas últimas décadas para prever o mercado financeiro. É possível utilizar esses resultados como *input* de um modelo de *Machine Learning*?

1.3.2 Objetivos de Investigação

Os objetivos de investigação desta dissertação são os seguintes:

- Identificar e recolher dados relevantes de diferentes fontes, onde se inclui os valores de abertura, fecho e volume de ações transacionadas no mercado de ações, notícias financeiras de fontes credíveis e relatórios financeiros do *Federal Reserve System*.
- Aplicação de técnicas de *Web Scraping* para recolher os dados mencionados no ponto anterior, assim como permitir que estes sejam armazenados no formato indicado para posterior utilização.
- Efetuar análise de sentimento às notícias e relatórios financeiros com recurso a técnicas de *Text Mining*, de forma a perceber se os resultados obtidos se relacionam de alguma forma com a oscilações da bolsa de valores.
- Aplicar algumas das mais utilizadas abordagens de análise técnica, onde se destacam as médias móveis, procurando relacionar o *output* destas com os resultados obtidos nos pontos *supra*.

1.4 Estrutura da Dissertação

Este trabalho está estruturado em cinco capítulos principais. No primeiro capítulo, a Introdução, é apresentada a contextualização e definição do problema de investigação, assim como a pertinência do estudo. No segundo capítulo é possível encontrar uma revisão da literatura relevante e que maior impacto teve na realização deste trabalho, abrangendo os principais conceitos e técnicas relacionados com a previsão do mercado de ações, *Web Scraping*, *Text Mining*, Análise Técnica e *Machine Learning*. O terceiro capítulo descreve a metodologia adotada, onde se detalha o procedimento de recolha de dados com recurso a *Web Scraping*, as técnicas de análise de sentimento utilizadas nas notícias e relatórios financeiros, a metodologia

aplicada para calcular as médias móveis da Análise Técnica, assim como de que forma tudo isto foi utilizado em diferentes modelos de *Machine Learning*, com o propósito de prever de que forma iria oscilar a bolsa de valores. O quarto capítulo retrata a visualização e análise dos dados recolhidos e utilizados nos modelos, enquanto que no quinto e último capítulo são discutidos os resultados obtidos, analisando com detalhe as diferentes abordagens utilizadas.

2. Revisão de Literatura

O investimento nos mercados financeiros é algo que desperta o interesse de imensas pessoas há várias décadas, com diferentes propósitos. Alguns querem apenas rentabilizar as suas poupanças, de modo a garantir um rendimento extra na hora de se reformarem, enquanto outras pessoas olham para a bolsa de valores como uma verdadeira mina de ouro, venda nela uma real hipótese de colecionarem fortunas e alterarem por completo o rumo das suas vidas. A abordagem de cada pessoa no mercado financeiro depende daquilo que pretende obter do mesmo: se procurar apenas aumentar de forma controlada e segura as suas poupanças, o foco será em ativos mais conservadores, como obrigações ou ações de empresas com elevado nível de liquidez (como por exemplo, as que se encontram presentes no índice S&P 500). Se por outro lado o objetivo passar por enriquecimento célere, a escolha irá cair em ativos mais voláteis, onde o risco de perder dinheiro é bastante elevado, mas onde também existe a possibilidade de acumular consideráveis quantidades de capital num curto espaço de tempo.

Existem várias técnicas que se podem utilizar de forma a procurar prever se um determinado ativo financeiro irá subir ou descer na bolsa de valores. Duas das mais utilizadas são a análise técnica e a análise fundamental: a primeira consiste num método que se baseia na análise dos gráficos históricos de preços e volumes das ações, assentando no princípio de que os preços se movem em tendências persistentes ao longo do tempo. Após identificada a tendência, é possível saber qual o melhor momento para comprar ou vender uma ação. A segunda é a análise financeira das contas de uma empresa com vista a determinar o preço justo de uma ação, e que se fundamenta na expetativa de lucros futuros. O método envolve a projeção dos *cash-flows* futuros da empresa e respetiva atualização para o momento presente. Para além destas técnicas (entre muitas outras que se aplicam nesta área), é importante também ter em consideração o impacto que as notícias têm na bolsa de valores. Quer sejam financeiras, de economia ou políticas, é inegável que as notícias impactam as oscilações dos diferentes ativos financeiros na bolsa de valores, tanto de forma negativa como positiva.

Com o crescimento notável que a Inteligência Artificial e os modelos preditivos de *Machine Learning* têm alcançado nos últimos anos, é bastante compreensível que atualmente se olhe para estes como uma poderosa ferramenta para prever a oscilação dos ativos na bolsa de valores. Diferentes estudos procuraram alcançar esta previsão com a maior precisão possível, utilizando para isso diferentes técnicas, *datasets* e modelos de *Machine Learning*, obtendo diferentes níveis de sucesso conforme as variáveis utilizadas.

O propósito deste trabalho passa pela construção de um algoritmo de *Machine Learning* que consiga prever as oscilações de diferentes ações. Para isso, serão recolhidas notícias financeiras de diferentes fontes, onde se incluem a Reuters e a Bloomberg. Utilizar-se-ão diferentes técnicas de *Text Mining*, de forma a se efetuar uma análise de sentimento às mesmas, procurando assim perceber se estas induzem notícias positivas ou negativas, ajudando assim a indicar que trajeto seguirão as ações. Para além disto, será também aplicado análise técnica às respetivas ações, de forma a ter outras ferramentas que permitam tirar conclusões. Para isso, será utilizado *Web Scraping* com o propósito de recolher dados históricos e atuais relativamente ao valor das ações, para efetuar análise técnica.

Com a aplicação deste modelo e de todas as técnicas mencionadas *supra*, procurar-se-á responder à derradeira questão: é possível construir um modelo de *Machine Learning* que consiga prever a oscilação de diferentes ações na bolsa de valores? E se sim, que nível de precisão se pode alcançar?

2.1 Metodologia da Revisão de literatura

Uma revisão de literatura consiste num processo de pesquisa que procura identificar, analisar e sintetizar os principais trabalhos publicados sobre um determinado tópico. O seu principal propósito passa por fornecer um panorama da situação atual do conhecimento sobre o tema em questão, sendo que a sua importância se divide em múltiplas razões.

A principal razão passa por permitir à pessoa que está a pesquisar entender o contexto e o estado atual da pesquisa sobre um determinado tópico, onde se inclui os principais trabalhos publicados, as teorias existentes e debatidas atualmente, e as principais lacunas no estudo que se está a efetuar. Ajuda também a identificar os métodos e técnicas utilizadas em pesquisas anteriores, algo que pode ser útil para orientar a metodologia de uma futura pesquisa. Importa também destacar que uma revisão de literatura permite à pessoa que está a efetuar a pesquisa avaliar a qualidade e a credibilidade dos trabalhos publicados, de forma a assegurar que a pesquisa está baseada em evidências sólidas. Para além de tudo isto, uma revisão de literatura fornece um conjunto de referências que podem ser usadas como base para o seu próprio trabalho, permitindo citar trabalhos relevantes e estabelecer uma linha de continuidade com o conhecimento previamente adquirido.

Como será comprovado no decorrer desta revisão de literatura, os temas mencionados e detalhados *supra* (*Machine Learning*, *Text Mining* e *Web Scraping*) foram aqueles a que se deu mais destaque aquando da pesquisa de artigos, procurando sempre interconectar esses temas com a previsão do valor das ações na bolsa de valores. Os métodos e as técnicas utilizadas para obter os artigos mais indicados para efetuar a revisão de literatura serão analisados no próximo ponto.

No decorrer da revisão de literatura, foram utilizadas diferentes combinações de palavras, de forma a obter diferentes resultados e artigos para analisar. Numa primeira fase, optou-se por utilizar a seguinte combinação:

(Machine Learning) AND (Stock Market) AND (Algorithm) AND (SP500) AND (Trading) AND (Forecast)

Que apenas devolveu dois artigos. Apesar de interessantes, estavam longe de corresponder ao esperado, e certamente não seriam suficientes para realizar uma revisão de literatura. Desta forma, optou-se por reduzir ligeiramente as métricas aplicadas na pesquisa, recorrendo às seguintes palavras:

(Machine Learning) AND (Stock Market) AND (Algorithm)

Esta combinação resultou na disponibilização de mais de mil artigos. Apesar de alguns parecerem adequados para aquilo que se pretendia, a grande maioria dos artigos devolvidos não se enquadrava no âmbito mencionado anteriormente, o que obrigou a que a combinação de palavras aplicadas na pesquisa fosse alterada para a seguinte:

(Machine Learning) AND (Stock Market) AND (Financial News)

Esta pesquisa, apesar de devolver menos de 250 artigos, apresentou resultados com um nível superior de qualidade, e muito mais próximos do âmbito ambicionado. A expressão "Financial News" acabou por se revelar essencial para se obterem os artigos mais adequados para realizar esta revisão de literatura, uma vez que foi assim que se obtiveram a grande maioria dos artigos referentes a *Text Mining* e a análise de sentimento.

Para além disso, como foi visto anteriormente, o tema de *Web Scraping* assume uma importância considerável na realização deste trabalho. Tendo isso em consideração, foram também efetuadas pesquisas que incluíssem a palavra "*Web Scraping*", de forma a que se obtivessem artigos que retratassem a aplicação desta técnica na previsão do valor das ações na bolsa de valores.

2.2 Resultados

No decorrer da pesquisa por artigos foram tomadas em consideração diferentes métricas, com o objetivo de assegurar que os resultados obtidos apresentavam a maior diversidade possível. Nesse sentido, foram recolhidos artigos de diferentes fontes, com *datasets* de diferentes países, e que tenham sido publicados em anos distintos. Desta forma, pretendeu-se que os resultados recolhidos abrangessem diferentes perspetivas, onde fosse evidente que eram tratadas realidades distintas, mas sem nunca descurar o ano em que o artigo tinha sido publicado. Na área de estudo em questão, é imperativo que se tome em alta consideração quando é que o artigo foi publicado, uma vez que isso pode impactar de forma bastante significativa o estudo que se irá realizar, inclusivamente a nível da veracidade dos temas que estamos a estudar. Como já foi mencionado anteriormente, os principais temas abordados no decorrer deste documento são *Machine Learning*, *Text Mining* e *Web Scraping*. Ora, tudo isto são temas que mudam recorrentemente, dado serem relativamente recentes e alvo de estudo de imensas pessoas por todo o mundo. Tendo isso em consideração, é de elevada importância que se dê prioridade aos artigos mais recentes. Nesse sentido, foram priorizados os artigos do ano de 2022.

Foram recolhidos um total de 38 artigos, sendo que o principal critério para definir o nível de interesse e de potencial importância de cada um foi a análise do respetivo *abstract*. Seguiu-se uma análise mais detalhada de cada artigo, onde se procurou perceber efetivamente o que é que retratava, de forma a procurar entender se se encontrava próximo do âmbito que se pretendia, chegando-se assim a uma conclusão expectável: nem todos os artigos recolhidos correspondiam às expetativas criadas pelo seu *abstract*, pelo que a sua utilidade era reduzida. Ainda assim, todos os artigos recolhidos foram contabilizados na análise efetuada, uma vez que todos tiveram impacto no desenvolvimento desta revisão de literatura.

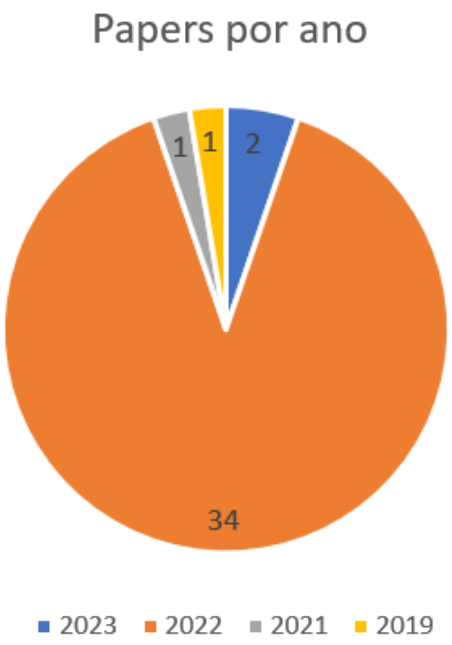

Figura 1 - Distribuição dos artigos analisados por ano de publicação

Como tinha sido mencionado anteriormente, é imperativo que os artigos analisados sejam recentes, devido às rápidas e constantes alterações que são efetuadas nas áreas alvo de estudo. Como se pode ver na Figura 1, a pesquisa de artigos foi efetuada em conformidade com essa filosofia: aproximadamente 95% dos documentos considerados para análise foram publicados nos últimos 12 meses (aquando da realização da revisão de literatura), assegurando assim que são tidas em consideração as técnicas e recursos mais recentes.

Para além de serem recentes, os artigos escolhidos também se destacam por serem originários de diferentes países. Os artigos selecionados são originários de 10 países diferentes, como se pode ver pela Figura 2.

Figura 2 - Distribuição dos artigos analisados por país

Não só se destaca uma distribuição razoável dos artigos por diferentes países, como também é evidente a multiplicidade e discrepância entre os países escolhidos. Este fator ajuda não só a assegurar a diversidade entre *datasets* utilizados (uma vez que muitos destes artigos trabalham com dados do próprio país), como também permite perceber a variação de técnicas utilizadas em diferentes zonas do globo.

A pesquisa por artigos que relacionassem as previsões no mercado financeiro através de modelos preditivos de *Machine Learning* e de técnicas de *Web Scraping* não foi muito frutífera, pelo menos a nível de quantidade, uma vez que não se obtiveram muitos resultados que relacionassem estes três temas. Ainda assim, existem algumas conclusões que se podem extrair dos (poucos) artigos analisados.

2.3 Discussão de resultados

A técnica que mais se destaca aquando da pesquisa por artigos que utilizem *Web Scraping* é, sem dúvida, *Beautiful Soup*. *Beautiful Soup* é uma biblioteca de Python que facilita a extração de dados de arquivos *HTML* e *XML*, permitindo navegar, pesquisar e modificar documentos de forma mais eficiente do que se faria manualmente, tendo também a capacidade de transformar documentos complexos de *HTML* numa árvore de objetos Python. Esta ferramenta não só

auxilia a efetuar o "*scrape*" das páginas *web*, como também pode ser utilizada para limpar os dados. Tendo tudo isto em consideração, torna-se fácil perceber o porquê de tantos estudos recorrerem a esta ferramenta quando necessitam de efetuar *Web Scraping*. Em [29], os dados são recolhidos da *National Stock Exchange* (NSE), com recurso a *Web Scraping*, utilizando a ferramenta *Beautiful Soup*. De seguida, é utilizado um modelo de classificação *Random Forest* e recorre-se a análise técnica de forma a conseguir prever o valor das ações, usando para isso as técnicas MACD (*Mobile Average Convergence-Divergence*), P/R e *Moving Averages*. Em [33], é utilizado *Web Scraping* com dois propósitos distintos: numa primeira fase, são recolhidas noticias financeiras sobre a Bolsa de Valores de Tehran (que é o alvo de estudo do artigo em questão), de forma a ser possível efetuar análise de sentimento sobre as mesmas. São também recolhidos os valores do ouro, dólar e petróleo, com o objetivo de se realizar análise fundamental sobre estes valores. Em [35], para além de se aplicar a técnica de *Beautiful Soup* para efetuar *Web Scraping*, outras ferramentas são utilizadas, tais como o *Selenium*, o *Pandas*, ou as bibliotecas de Python *Time Library* e *Sys Library*. Tal como acontece na maioria dos artigos analisados, neste também se efetua uma análise de sentimento a notícias financeiras e a publicações do Twitter e do Reddit. Por fim, este artigo destaca-se também pela utilização de um modelo preditivo *Support Vector Machine* (SVM), um modelo probabilístico com capacidades generativas (*Topic Modeling – Latent Dirichlet Allocation*) e a utilização de SARIMA (*Seasonal Autoregressive Integrated Moving Average*). Em [17], é novamente utilizado a ferramenta de Python, *Beautiful Soup*, de forma a extrair informação do *website* Google Finance, que contem o valor atual das ações e índices em diferentes mercados bolsistas. Para extrair a informação passada sobre estes temas, recorre-se à API Yahoo! Finance. Este artigo destaca-se também pela utilização de *dashboards* para efetuar a análise técnica e pelo uso de *Long Short-Term Memory* (LSTM), com 3 algoritmos de *Machine Learning*: *Linear Regression*, *Ridge Regression* e *Lasso Regression*. Em [5] o autor procura efetuar *Web Scraping* com técnicas de *Machine Learning* e *Natural Language Processing* (NPL) em conjunto com a ferramenta *Named Entity Recognition* (NER), que o autor considera que garante melhores resultados em dados não estruturados.

Foi também efetuada uma pesquisa semelhante à anterior, mas onde se procuravam artigos que mencionassem a utilização de técnicas de *Text Mining*, ao invés de *Web Scraping*. Esta pesquisa obteve mais resultados, mas observou-se que a maioria deles utilizavam abordagens e técnicas semelhantes. Em [15] é utilizado FINBERT, um modelo de processamento de linguagem natural baseado na arquitetura BERT, mas que foi treinado com um vasto *dataset* de

notícias financeiras, pelo que o seu principal propósito é trabalhar com este tipo de documentos. Este artigo destaca-se também pela utilização do *dataset* Financial Phrase Bank, que contêm 5000 frases financeiras próprias para a análise de sentimento, e pela utilização do modelo *Random Forest*. Em [39], é utilizado o algoritmo *Stanford NLP*, com o objetivo de realizar análise de sentimento de notícias financeiras. Importa também destacar que neste artigo apenas se recorre aos títulos de cada notícia, ao contrário do que acontecia no artigo anterior, onde se analisava o título e o conteúdo de cada notícia. No artigo [22] são utilizados 3 modelos para efetuar a análise de sentimento: *Vader*, *the Sentiment Roberta Large English* (SRLE) e *Twitter Roberta Base Sentiment* (TRBS) (este último não é relevante para o âmbito pretendido, uma vez que apenas se pretende realizar análise de sentimento a notícias financeiras, e não a *tweets*). Em [32], o objetivo passa por prever o valor do índice turco BIST30, pelo que o *dataset* é composto por notícias financeiras e *tweets* publicados na Turquia. Para além disso, utilizam também o modelo de análise de sentimento BERTurk, direcionado para funcionar com a linguagem do próprio país. Em [19], são utilizadas as técnicas *Vader* e de *Word Embedding*, onde se incluem a *Bag of Words* (BoW), TF-IDF, BERT e ROBERTA. No artigo [10] é aplicada a técnica *TextBlob* para efetuar análise de sentimento. É também aplicado o método *Sliding Window*, de forma a extrair as *features* mais representativas das notícias financeiras. Foram ainda analisados outros artigos que aplicam outras técnicas, como o [27] que recorre ao *Word2vec* e ao *ELMo*, ou o [23], que utiliza *Natural Language Toolkit* (Nltk).

A grande maioria dos artigos menciona os modelos de *Machine Learning* utilizados, pelo que se considerou relevante mencionar e listar os modelos mais utilizados, que se podem analisar na Figura 3.

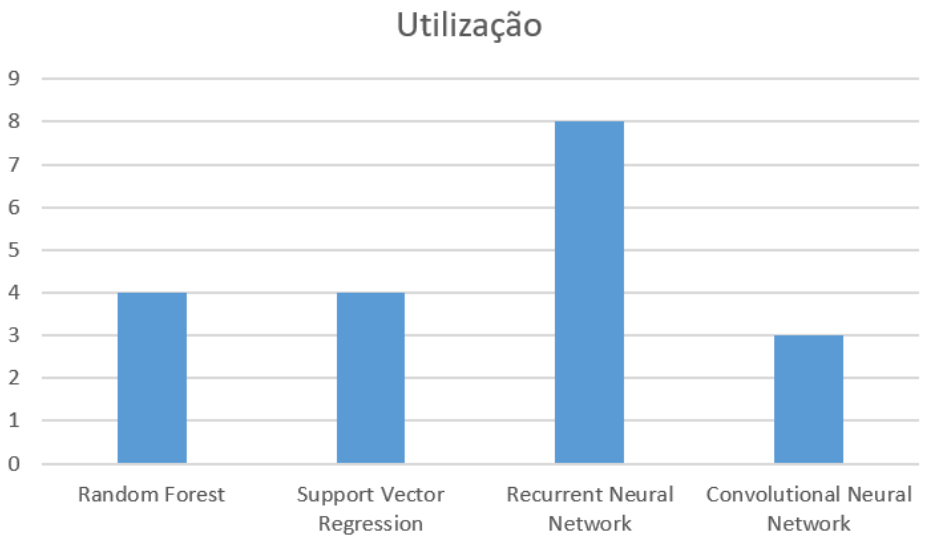

Figura 3 - Distribuição dos modelos de Machine Learning utilizados nos artigos analisados

Como se pode concluir através da análise do gráfico na Figura 3, existe uma forte utilização de modelos RNN (*Recurrent Neural Network*), com especial destaque para a rede LSTM (*Long Short-Term Memory*). A considerável utilização deste tipo de modelos para a previsão do valor das ações justifica-se pela forte adequação destes algoritmos em lidar com séries temporais. Importa também destacar a notável utilização dos modelos *Random Forest* e *Support Vector Regression* (SVM).

Para terminar o processo de análise de artigos, foi também tomada em consideração os *datasets* utilizados, assim como a origem dos mesmos. No decorrer da revisão de literatura foram analisados certos artigos que retratavam um estudo específico de um determinado país, sendo que para isso utilizam documentos locais, do próprio país; recorde-se por exemplo o artigo [32], cujo alvo de estudo é o índice BIST30, que representa a *performance* das 30 empresas com maior capital na Turquia, pelo que eram utilizadas noticias financeiras do próprio país. Naturalmente, este tipo de *datasets* não são relevantes para o âmbito pretendido, pelo que não devem ser tomados em consideração. Em contrapartida, existem fontes de informação que se destacam pela forte utilização no decorrer dos artigos, como é o caso do *website* Yahoo! Finance. Nesta página é possível encontrar o valor atual de diversas ações e índices, pelo que se deve claramente equacionar a utilização deste *website* para extrair este tipo de dados, com recurso a *Web Scraping*. Apesar da notável utilização que se dá a esta página, importa também ter em consideração outro tipo de fontes de dados. Em [17], por exemplo, recorre-se ao *website* Google Finance para extrair este tipo de informação, pelo que se deve analisar detalhadamente

as melhores opções à nossa disposição para recolher dados relativos ao valor das ações nas diversas bolsas de valores. Relativamente às fontes de dados para efetuar *Text Mining*, a diversidade de opções é maior. Em alguns artigos analisados, a análise de sentimento era realizada com recurso a outras fontes, nomeadamente o Twitter ou o Reddit. Estes dois *websites* não se aplicam ao âmbito pretendido, uma vez que, como já foi mencionado, apenas é do interesse deste trabalho a análise de notícias financeiras. Nesse sentido, foram encontradas diversas fontes de notícias financeiras que devem ser consideradas, onde se destacam a Reuters, a Bloomberg e a Financial Times. Neste grupo, deve-se prestar especial atenção à primeira mencionada, uma vez que é a fonte de notícias mais utilizada nos artigos analisados, sem nunca descurar outros *media* que possam ser relevantes e adequados para a análise de sentimento pretendida.

É importante também mencionar que os diversos artigos analisados apresentam fórmulas matemáticas que poderão ser úteis no decorrer do trabalho, nomeadamente ao nível de cálculos de análise técnica.

No início desta revisão de literatura foi proposto procurar perceber se é possível prever o valor e as oscilações das ações na bolsa de valores, recorrendo para isso a modelos de *Machine Learning*, e a tecnologias como *Text Mining* e *Web Scraping*.

Nesse sentido, no decorrer desta revisão de literatura procurou-se explorar diferentes artigos que apresentassem diferentes abordagens para alcançar este tipo de previsões. Utilizando a plataforma *Scopus*, foi possível encontrar documentação de diversos países, onde são utilizados diferentes algoritmos e *datasets*, mas sempre com um objetivo comum: procurar alcançar a combinação que obtivesse melhores resultados na previsão do valor das ações. Os resultados obtidos foram bastante satisfatórios, uma vez que permitiram não só perceber quais são os modelos e conjunto de *datasets* que tipicamente obtém melhores resultados, como também contribuíram para descobrir novas técnicas que podem ser utilizadas na realização deste trabalho, nomeadamente a nível de *Text Mining* e *Web Scraping*.

Apesar da considerável quantidade de informação que se pôde extrair através da análise destes artigos, tornou-se evidente que esta é uma área onde ainda existem diversas melhorias que se podem realizar, nomeadamente ao nível das técnicas que se aplicam para realizar análise de sentimento, pelo que seria interessante explorar e aprofundar estas metodologias no futuro.

3. Metodologia

Neste capítulo será apresentada a metodologia utilizada no decorrer do trabalho, que poderá ser visualizada de forma esquemática no fluxograma presente na Figura 4:

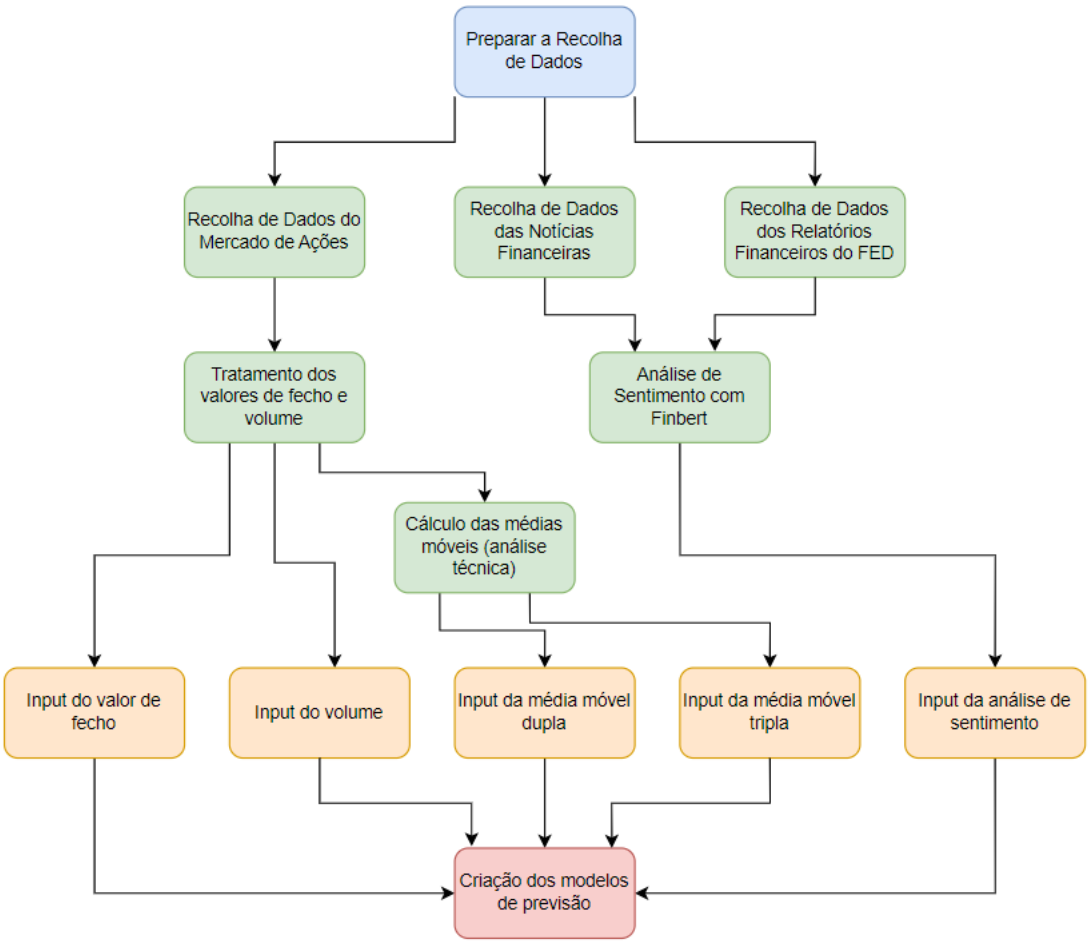

Figura 4 - Metodologia utilizada no trabalho

3.1 Recolha de Dados

A recolha de dados desempenha um papel crucial nesta dissertação, na medida em que fornece as informações necessárias para a análise e previsão do mercado de ações. Neste estudo, a recolha de dados é realizada através de técnicas de *Web Scraping*, que permitem extrair informações relevantes de fontes *online*. Os dados recolhidos são de diversas naturezas: informações do mercado de ações, notícias financeiras de algumas das mais prestigiadas fontes internacionais e relatórios financeiros do *Federal Reserve System*.

Em particular, os modelos de previsão desenvolvidos neste estudo utilizam diversos *inputs* para realizar as previsões do valor de fecho dos diferentes casos de estudo. Os *inputs* utilizados são os seguintes:

1. Valores de abertura e fecho do mercado de ações para os índices S&P 500, Nikkei, Dow Jones, Russell, Nasdaq, ouro e petróleo.
2. Volume de ações transacionadas para os índices mencionados no ponto 1.
3. Análise de sentimento das notícias financeiros e dos relatórios financeiros do *Federal Reserve System*,
4. Os valores das médias móveis duplas e médias móveis triplas, calculados a partir dos valores de fecho do mercado de ações.

Estes *inputs* são combinados e processados pelos modelos de *Machine Learning*, com o propósito de gerar previsões para as tendências futuras do mercado de ações.

3.1.1 Recolha de Dados do Mercado de Ações

De forma a obter dados do mercado de ações, são utilizadas técnicas de *Web Scraping* para extrair informações dos valores de abertura, fecho e volume de ações transacionadas de cinco índices diferentes: S&P 500, Nikkei, Dow Jones, Russel e Nasdaq. São também recolhidos os valores de abertura, fecho e volume do ouro e petróleo. Estas informações são essenciais para a análise do comportamento do mercado de ações, e serão utilizadas como um dos *inputs* para o modelo de previsão.

Os dados relativos aos mercados financeiros foram todos extraídos do *website* https://finance.yahoo.com/ . Nesta página, os valores de fecho, abertura, volume de ações, entre outros, são diariamente atualizados para os sete casos de estudo utilizados. Para além disso, a estrutura *HTML* da página permite que os dados sejam facilmente extraídos com recurso a *Web Scraping*, pelo que se considerou que esta era uma ótima fonte para efetuar a recolha de dados. As técnicas de *Web Scraping* permitem automatizar o processo da recolha de dados, percorrendo as páginas relevantes, identificando os elementos de interesse e extraindo os dados necessários. O facto de todo o processo de recolha de dados estar automatizado permite também que qualquer utilizador consiga obter todos os dados necessários, sem despender tempo nessa tarefa. As variáveis são depois armazenadas num formato adequado para análise futura.

3.1.2 Recolha de Notícias Financeiras

As notícias financeiras desempenham um papel importante na análise e previsão do mercado de ações. De forma a obter as mesmas, são também utilizadas técnicas de *Web Scraping* para extrair informações relevantes de *websites* de notícias financeiras, onde se inclui a Bloomberg, a Reuters, a Zacks, entre outras. À imagem do que se sucede com a extração dos valores de abertura, fecho e volume das ações, também este processo foi construído de forma que possa ser utilizado por qualquer pessoa, independentemente dos seus conhecimentos técnicos, e de forma extremamente célere. As notícias são guardadas numa base de dados SQLite – onde se garante que não é adicionado conteúdo duplicado às tabelas, assegurando assim a unicidade dos dados – para posterior análise.

3.1.3 Recolha de Relatórios Financeiros do *Federal Reserve System*

Os relatórios financeiros emitidos pelo *Federal Reserve System* são também fontes importantes de informação para a análise e previsão do mercado de ações. Tal como acontece nos pontos anteriores, também aqui são utilizadas técnicas de *Web Scraping* para extrair informações relevantes desses relatórios.

O tratamento dos dados dos relatórios financeiros é bastante semelhante ao que é efetuado com os das notícias financeiras: ambos são registados em bases de dados SQLite, e é efetuada uma análise de sentimento nos dois casos, procurando perceber se o seu conteúdo indica notícias positivas, negativas ou neutras. No próximo ponto, é explicado com maior detalhe como funciona este procedimento.

3.2 Análise de Sentimento com Finbert

Parte dos dados utilizados para a análise de sentimento foram também retirados da mesma página *web* (https://finance.yahoo.com/topic/stock-market-news/), onde se encontravam diversas notícias financeiras referentes à bolsa de valores. Os restantes dados utilizados para estudar a análise de sentimento foram retirados de blogues e páginas oficias do *Federal Reserve System*. Estes dados foram registados numa base de dados *SQLite*, com o intuito de serem posteriormente utilizados como *input* de algoritmos de *Text Mining*, que indicam se as notícias são positivas ou negativas, com o valor da análise de sentimento a variar entre 0 e 1.

A análise de sentimento é uma técnica utilizada para avaliar e perceber o tom emocional de um texto. Neste estudo, a análise de sentimento é aplicada às notícias financeiras recolhidas através de técnicas de *Web Scraping*. Para realizar a análise de sentimento, é utilizada a biblioteca FinBERT, (derivada da conhecida *framework* de *Text Mining* BERT) um modelo de linguagem pré-treinado desenvolvido especificamente para análise de sentimento de conteúdo financeiro.

A biblioteca recebe um *input* – o texto financeiro que se pretende analisar – e devolve três *outputs*: o valor "positivo", o valor "neutro" e o valor "negativo". A soma destes três valores corresponde a 1, e procura indicar a probabilidade de o texto transmitir um sentimento positivo, negativo ou neutro. Utilizando o seguinte caso como exemplo:

	content	contentSentimentAnalysisPos	contentSentimentAnalysisNeg	contentSentimentAnalysisNeu
1	In the latest trading session, Docebo Inc. (DCBO) closed at $38.18, marking a -1.75% ...	0.92939060926437	0.04747396707535	0.02313548512757
2	Exact Sciences (EXAS) closed the most recent trading day at $66.28, moving +1.33% ...	0.94859826564789	0.03180510550737	0.01959661208093
3	In the latest trading session, Perion Network (PERI) closed at $38.28, marking a +0.18...	0.92889833450317	0.05105701833963	0.02004463784397

Figura 5 - Exemplo de notícias positivas

Na Figura 5 é possível encontrar três exemplos de notícias onde a probabilidade de a notícia transmitir sentimentos positivos é elevadíssima (superior a 92%). Por outro lado, a hipótese da notícia transmitir sentimentos negativos é praticamente residual (inferior a 5%).

	content	contentSentimentAnalysisPos	contentSentimentAnalysisNeg	contentSentimentAnalysisNeu
1	Investors are quickly dividing corporate borrowers into the haves and the have-nots. ...	0.01774548180401	0.92633843421936	0.05591610819101
2	Peter Tuchman, one of the most recognizable stock brokers on Wall Street, has been ...	0.03049922734499	0.2411572933197	0.72834348678589
3	Commercial real estate is probably the next pain point for regional banks and the sto...	0.01352495793253	0.95826184749603	0.02821323089302

Figura 6 - Exemplo de notícias negativas e neutras

Naturalmente, o oposto também pode acontecer, como se pode concluir através da análise da Figura 6. Neste exemplo, é possível encontrar duas notícias que claramente transitem um sentimento negativo, como se pode concluir através dos valores de 92.63% e 95.82%. Por fim, temos também a representação da terceira opção possível, espelhada na segunda linha da Figura 6: o texto não é identificado como sendo positivo ou negativo, mas sim como neutro.

A análise de sentimento é efetuada tanto para as notícias financeiras, como para os relatórios financeiros do *Federal Reserve System*. O objetivo passa por relacionar o respetivo *output* com a oscilação do valor dos índices mencionados na secção 3.1.1, com recurso a um modelo de *Machine Learning*, de forma a perceber se é possível encontrar uma correlação entre os mesmos.

3.3 Análise Técnica

Os dois elementos da análise técnica (média móvel dupla e média móvel tripla) definem os terceiro e quarto *inputs* aplicados nos modelos de *Machine Learning*, juntamento com o valor de fecho das ações e o volume das ações transacionadas. A análise técnica concentra-se no estudo dos padrões e tendências de preços passados, de forma a prever o comportamento futuro do mercado de ações. Neste estudo, são consideradas duas abordagens de análise técnica: médias móveis duplas e médias móveis triplas. A média móvel dupla envolve o cálculo de duas médias móveis simples (*Simple Moving Average* – SMA) num conjunto de dados, onde uma das médias deverá ser de curto prazo, enquanto a segunda será de longo prazo. No presente trabalho, optou-se por utilizar uma média de 5 dias (curto prazo) e outra de 20 dias (longo prazo). O cálculo das SMA é realizado da seguinte forma:

1. **Cálculo da SMA de curto prazo**: Para cada elemento do *dataset*, somar os valores de fecho dos 5 dias anteriores, e dividir o valor obtido por 5 para obter a primeira média.
2. **Cálculo da SMA de longo prazo**: O processo utilizado neste cálculo é o mesmo do anterior, mas de forma a obter a segunda média, devemos somar os valores de fecho dos 20 dias anteriores, e dividir o valor obtido por 20.
3. **Gerar sinais de compra ou venda**: Quando a SMA de curto prazo cruza a SMA de longo prazo no sentido ascendente, pode ser interpretado como um sinal de compra. No sentido inverso, quando a SMA de curto prazo cruza a SMA de longo prazo no sentido ascendente, pode ser interpretado como um sinal de venda.

O cálculo das médias móveis triplas é realizado seguindo os mesmos critérios, mas ao invés de se utilizarem apenas duas SMA – uma de curto prazo e outra de longo prazo – é utilizada uma terceira SMA, de médio prazo.

De forma a obter os valores das médias móveis duplas e triplas, são utilizados os valores de fecho do mercado de ações recolhidos anteriormente, pelo que o seu alcance é bastante semelhante – aproximadamente cinco anos. Estes valores serão também utilizados como *input* dos modelos de previsão.

3.4 Modelo de *Machine Learning*

O desenvolvimento do modelo de *Machine Learning* é uma etapa crucial desta dissertação. Com base nos dados recolhidos e nas informações obtidas através da análise de sentimento e análise técnica, serão utilizados diferentes modelos com a capacidade de prever as tendências futuras do mercado de ações. A principal diferença nos modelos utilizados prende-se com o número de *inputs* utilizados em cada um deles. Serão construídos modelos com uma diferente combinação de *inputs*, de forma a tentar perceber qual combinação de variáveis é que assegura os melhores resultados.

Para efeitos de modelação, são utilizados algoritmos de *Deep Learning*, onde se destacam as redes neuronais artificiais recorrentes, através da utilização de modelos LSTM. Os dados recolhidos (limpos e pré-processados) definem os *inputs* que são utilizados para treinar e ajustar os modelos, procurando assim maximizar a sua capacidade de previsão. Será também efetuada uma avaliação dos resultados dos modelos, de forma a determinar a sua eficácia e precisão.

4. Análise exploratória dos dados

Serão realizadas análises aos dados recolhidos, com o intuito de observar as séries temporais dos valores de fecho, volume de transações e médias móveis. Da mesma forma, serão também examinadas as distribuições e estatísticas descritivas dos dados, a fim de compreender melhor as características e comportamentos do mercado.

O conjunto de dados, que compreende todas as variáveis financeiras, engloba observações diárias (*Business Day*), para os últimos cinco anos (entre agosto de 2018 e agosto de 2023), o que garante um intervalo de dados adequado para o objetivo do estudo, que se reflete num total de cerca 1300 linhas para cada um dos casos de estudo. Na Figura 7 é possível encontrar a representação gráfica da série temporal dos preços de fecho do índice S&P 500, para os quase 1300 dias úteis entre agosto de 2018 e agosto de 2023 (representados no eixo 'Date'). No Anexo A encontram-se as representações gráficas de outras variáveis referentes ao mesmo índice.

Figura 7 - Dataset para os valores de fecho do índice S&P 500

Na Figura 7 pode-se identificar uma queda abrupta na série temporal quando o valor da variável '*Date'* é próximo de 400, pelo que é possível que os valores durante esse período sejam identificados como atípicos devido à sua significativa distância em relação aos valores médios. A série dos preços de fecho do índice S&P 500 tem um comportamento irregular, volátil, sem indícios de uma tendência global ou de sazonalidade.

Analisando os gráficos de linha das restantes séries temporais, percebe-se que esta queda é comum a todos os *datasets* utilizados na dissertação, e que ocorreu devido ao início da pandemia COVID-19. Desta forma, é importante notar que essa queda não se deveu a erros nos dados ou a anomalias no mercado, mas sim a um evento global extraordinário, pelo que estes dados não foram considerados um *outlier*, nem foram removidos dos *datasets*. Ainda assim, para alguns *datasets*, foram encontrados valores omissos, onde não estava representado o valor de fecho e/ou o volume. Para estes casos, optou-se por fazer imputação do valor de fecho e/ou volume com o valor não nulo anterior.

Nos dados utilizados para realizar análise de sentimento não foram detetados *outliers* nem valores omissos. Na Figura 8 pode-se observar a distribuição das notícias pelos diferentes valores de análise de sentimento. Como se pode concluir através da análise dos histogramas, a distribuição dos resultados não é uniforme, uma vez que para os ambos os casos, os resultados mais frequentes se encontram nas caudas de distribuição. O valor de sentimento varia entre 0 e 1, onde um valor próximo de 0 representa um valor de sentimento fraco, enquanto que um valor de sentimento próximo de 1 representa um valor de sentimento forte.

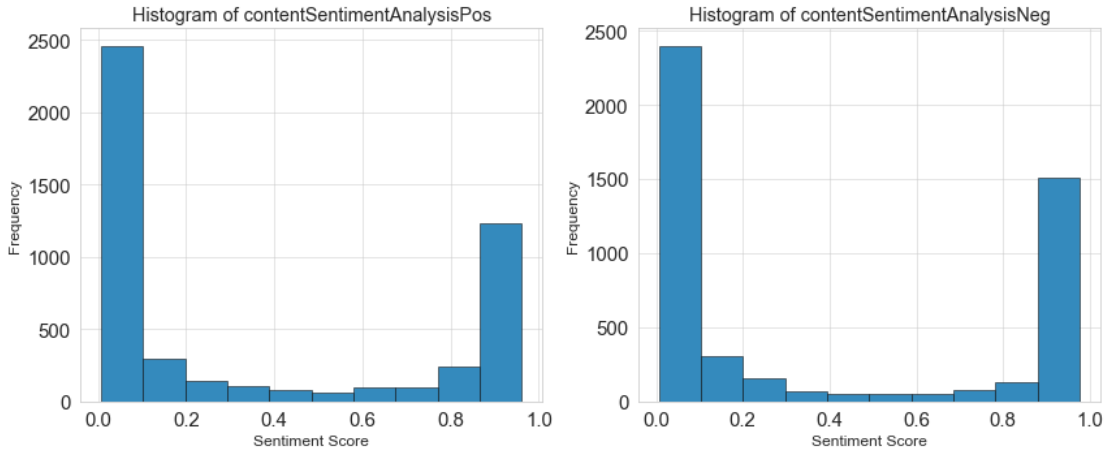

Figura 8 - Distribuição das notícias analisadas por análise de sentimento

4.1 Estatística descritiva das variáveis

De seguida, foi efetuada uma análise da estatística descritiva das variáveis. Nesta fase, apenas serão analisados os resultados obtidos para o índice S&P 500, que se encontram nas Tabelas 1 e 2.

	Close	Volume	Double Moving Average	Triple Moving Average
count	1271.00	1271.00	1271.00	1271.00
mean	3625.40	4292914579.07	0.07	0.03
std	652.18	1068224281.28	0.34	0.24
min	2237.40	1296530000.00	0.00	0.00
25%	2968.43	3632945000.00	0.00	0.00
50%	3749.63	4057340000.00	0.00	0.00
75%	4175.34	4692150000.00	0.00	0.00
max	4796.56	9976520000.00	2.00	2.00

Tabela 1 - Estatística Descritiva das variáveis de Valor de Fecho, Volume, Média Móvel Dupla e Média Móvel Tripla

i	Sentiment Analysis	sentimentAnalysis OneWeekAfter	sentimentAnalysis TwoWeekAfter	sentimentAnalysis FewDaysAfter
count	1271.00	1271.00	1271.00	1271.00
mean	-0.07	-0.06	-0.05	-0.17
std	0.33	0.19	0.29	0.30
min	-0.71	-0.71	-0.71	-0.71
25%	-0.27	-0.15	-0.22	-0.31
50%	-0.05	0.00	-0.15	-0.15
75%	0.05	0.00	0.13	0.00
max	0.77	0.77	0.77	0.77

Tabela 2 - Estatística Descritiva das variáveis de análise de sentimento

Relativamente à Tabela 1, que compreende informações sobre o valor de fecho (Close), volume de ações (Volume), valor da média móvel dupla (Double Moving Average) e valor da média móvel tripla (Triple Moving Average):

- O número total de observações é 1271.
- O valor médio de fecho é de 3625.40, com um desvio padrão de 652.18. O valor mínimo registado foi de 2237.40, enquanto o valor máximo alcançado foi de 4796.56.
- O volume médio de ações transacionadas é de 4.292.914.579,07, com um desvio padrão de 1.068.224.281,28. O volume mínimo é de 1.296.530.000,00, e o volume máximo atinge 9.976.520.000,00.
- A média móvel dupla tem uma média de 0.07, com um desvio padrão de 0.34, variando de 0.00 a 2.00.
- A média móvel tripla tem uma média de 0.03, com um desvio padrão de 0.24, variando de 0.00 a 2.00

Na Tabela 2 encontram-se as médias da análise de sentimento em diferentes intervalos de tempo:

- A média da análise de sentimento é de -0.07, com um desvio padrão de 0.33, variando de -0.71 a 0.77.
- A média da análise de sentimento uma semana antes (sentimentAnalysisOneWeekAfterAverage) é de -0.06, com um desvio padrão de 0.19, variando de -0.71 a 0.77.
- A média da análise de sentimento duas semanas antes (sentimentAnalysisTwoWeekAfterAverage) é de -0.05, com um desvio padrão de 0.29, variando de -0.71 a 0.77.
- A média da análise de sentimento alguns dias antes (sentimentAnalysisFewDaysAfterAverage) é de -0.17, com um desvio padrão de 0.30, variando de -0.71 a 0.77.

4.2 Correlação entre as variáveis

Terminada esta análise, procurou-se perceber a correlação para as variáveis, recorrendo assim a uma matriz de correlação linear (de Pearson) que se encontra na Figura 9.

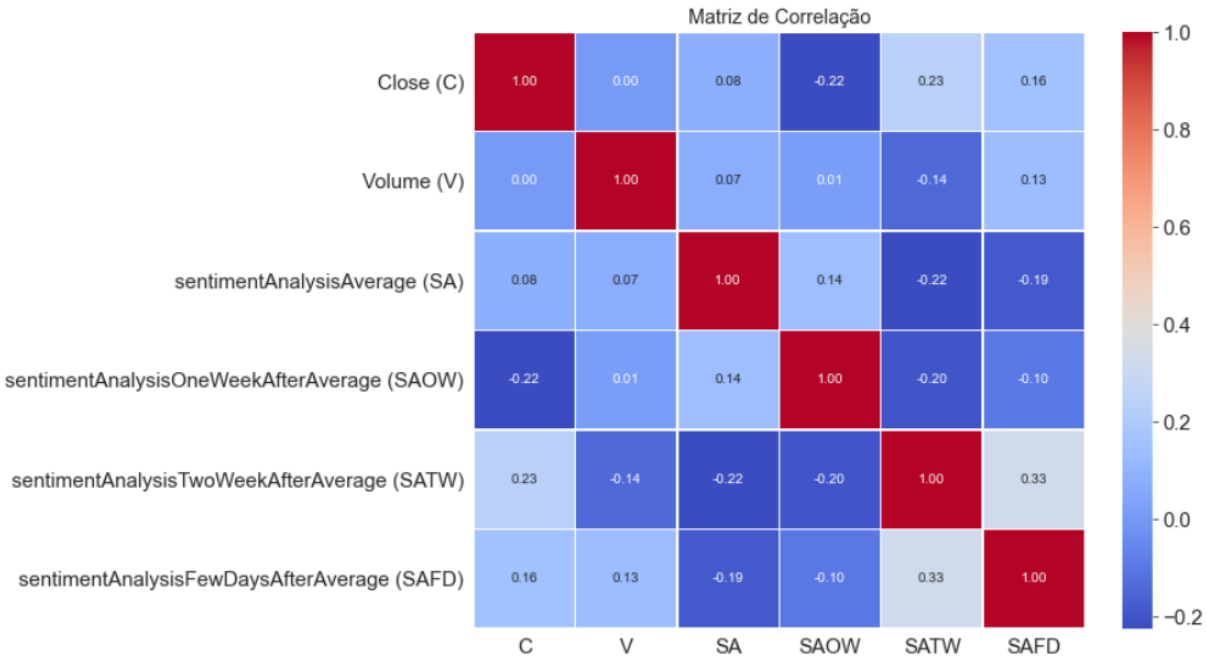

Figura 9 - Matriz de correlação do dataset S&P 500

- A correlação entre o valor de fecho e a análise de sentimento (sentimentAnalysisAverage) é de 0.08, indicando ausência de correlação linear.
- A correlação entre o valor de fecho e a análise de sentimento uma semana antes (sentimentAnalysisOneWeekAfterAverage) é de -0.22, sugerindo uma correlação negativa moderada-fraca.
- A correlação entre o valor de fecho e a análise de sentimento duas semanas antes (sentimentAnalysisTwoWeekAfterAverage) é de 0.23, indicando uma correlação positiva moderada-fraca.
- A correlação entre o valor de fecho e a análise de sentimento alguns dias antes (sentimentAnalysisFewDaysAfterAverage) é de 0.16, sugerindo uma correlação positiva fraca.

Estas correlações fornecem indicações sobre como as variáveis se relacionam umas com as outras. Correlações próximas a 1 indicam uma forte correlação positiva, correlações próximas a -1 indicam uma forte correlação negativa, enquanto correlações próximas a 0 indicam ausência de correlação.

Os valores possíveis para as duas variáveis de médias móveis são um conjunto fixo e limitado, que varia entre 0 (manter as ações), 1 (comprar mais ações) e 2 (vender as ações). Devido à natureza nominal das variáveis Média Móvel Dupla e Média Móvel Tripla, não é apropriado incluí-las na matriz de Pearson, uma vez que esta é destinada a medir correlações entre variáveis numéricas, sendo essa a razão pela qual estas duas variáveis não se encontram representadas na Figura 9.

É possível consultar as matrizes de correlação dos índices Dow Jones, Russell e do valor do ouro, nas Figuras 13, 14 e 15, presentes no anexo B da dissertação.

5. Modelação e Resultados

5.1 Normalização dos dados

De forma a melhor preparar os dados que serão utilizados nos modelos, foi aplicada uma técnica conhecida como Min-Max Scaler. Esta técnica é utilizada para normalizar os dados, ou seja, para ajustar os valores das variáveis dentro de um intervalo específico. Neste caso, o objetivo foi escalar os dados para o intervalo padrão de [0,1]. Uma vez que a média móvel dupla e a média móvel tripla são variáveis de natureza nominal, cujos valores possíveis variam entre 0, 1 e 2, a normalização das mesmas não faria sentido estatístico, pelo que a técnica Min-Max Scaler não foi aplicada para essas duas variáveis.

5.2 Conjuntos de treino e teste

A escolha da proporção entre o conjunto de treino e o conjunto de teste é uma decisão importante. Optou-se por aplicar uma divisão de 80% para treino e 20% para teste, pelas seguintes razões:

- **Tamanho Suficiente**: o conjunto de treino com 80% dos dados é suficiente para treinar modelos complexos de forma eficaz.
- **Teste Significativo**: o conjunto de teste com 20% dos dados oferece uma amostra representativa para avaliação e verificação do desempenho do modelo.
- **Redução do Risco de *Overfitting***: uma proporção razoável de dados no conjunto de treino ajuda a mitigar o risco de *Overfitting*, permitindo que o modelo aprenda padrões úteis sem se ajustar em excesso aos dados de treino.

Essa divisão estratégica e a escolha da proporção 80/20 visam garantir que o modelo resultante seja preciso e capaz de generalizar e fornecer previsões fiáveis para novos dados.

5.3 Introdução

No decorrer do trabalho, foram desenvolvidos diversos modelos com o intuito de procurar prever as oscilações do mercado de ações, de forma a explorar e perceber a eficácia de diferentes abordagens. De forma a atingir esse objetivo, recorreu-se a dois tipos de modelos de *Machine Learning*: LSTM e BiLSTM. Para além disso, foram testados diferentes cenários onde se variavam as variáveis de *input* utilizadas, com o propósito de perceber de que forma as suas diferentes combinações afetavam os resultados obtidos.

Inicialmente, foram utilizados modelos LSTM e BiLSTM com apenas uma variável de *input* – o valor de fecho das ações. Esses modelos iniciais serviram como ponto de partida para a análise comparativa com as variações posteriores.

Como foi mencionado anteriormente, no decorrer deste estudo, foram considerados sete casos de estudo: os índices S&P 500, Dow Jones, Nasdaq, Russell e Nikkei, assim como os valores do ouro e do petróleo. A dimensão temporal considerada para a recolha destes dados foi de 5 anos, entre agosto de 2018 e agosto de 2023. Relativamente à recolha de notícias e relatórios financeiros, a dimensão temporal foi de apenas 6 meses, entre março e agosto de 2023.

5.4 Modelos com um input

Numa primeira abordagem, onde foi utilizado um modelo LSTM com apenas uma variável de *input* - o valor de fecho das ações (o momento presente, assim como os últimos dois momentos históricos – utilizando assim um *lag* de 2 *time steps*) - os resultados foram, de modo geral, satisfatórios. A correlação linear entre os valores de fecho e os valores previstos era bastante elevada para todos os casos, oscilando entre 0.931 (para o índice Russell) e 0.988 (para o valor do ouro). Os valores de correlação para os sete casos de estudo foram calculados com recurso ao conjunto de teste de cada *dataset*, contabilizando assim aproximadamente 254 observações para cada conjunto de dados. Os coeficientes de correlação de Pearson, para cada série temporal em estudo e os seus valores preditos, encontram-se na Tabela 3.

Índice	Valor de Correlação
S&P 500	0.975
Dow Jones	0.966
Nasdaq	0.983
Russell	0.931
Nikkei	0.988
Gold	0.987
Crude Oil	0.938

Tabela 3 - Correlação entre os valores reais e preditos – modelo LSTM com 1 input

De forma a avaliar os resultados, foram utilizadas diferentes métricas para calcular a *performance* do modelo, onde se destacam a *Root Mean Square Error* (RMSE), a *Mean Absolute Error* (MAE) e o coeficiente de determinação R^2. Antes de proceder à análise dos resultados obtidos, importa explicar brevemente o que representa cada uma das métricas utilizadas:

- MAE (Erro Médio Absoluto): Esta métrica representa o valor médio dos erros absolutos entre as previsões e os valores reais.
- RMSE (Raiz do Erro Quadrático Médio): O RMSE mensura a raiz quadrada do erro quadrático médio entre as previsões e os valores reais
- Mean: A média dos valores reais é uma métrica importante para compreender a escala das previsões.
- MAPE (Erro Percentual Absoluto Médio): O MAPE calcula a média dos erros percentuais absolutos entre as previsões e os valores reais.
- R^2 (Coeficiente de Determinação): Medida estatística que representa a proporção da variabilidade na variável dependente que é explicada pelo modelo, variando de 0 a 1, onde 1 indica uma explicação perfeita da variabilidade.

Portanto, quanto menor forem os valores do RMSE, MAE e MAPE, melhor é a adequação do modelo ao conjunto de dados (treino) e a sua capacidade de previsão (teste), uma vez que quanto menor forem estes valores, menor é a diferença entre os valores reais e as previsões. O modelo é tanto melhor quanto maior for o coeficiente de determinação, uma vez que este demonstra o quão bem os dados se adequam ao modelo.

As métricas são calculadas em diferentes momentos, de forma a perceber a evolução do modelo no decorrer do processo. O primeiro cálculo das métricas é efetuado imediatamente após serem geradas as previsões por parte do modelo (primeira abordagem). Numa segunda fase, é utilizado o RandomizedSearchCV (um método da *scikit-learn*, que é uma biblioteca de *Python*) com o objetivo de obter os melhores hiperparâmetros, que serão então utilizados para a criação de um novo modelo (segunda abordagem). Após treinar este novo modelo, que contém os hiperparâmetros otimizados, é efetuada uma nova previsão com recurso ao conjunto de teste. Utilizando as novas previsões, são novamente calculadas as métricas de avaliação de *performance* do modelo. Numa terceira e última fase, é utilizado um modelo BiLSTM (Bidirectional LSTM), que consiste num modelo com dois LSTM: um que leva o input adiante, e outro que o leva na direção oposta (terceira abordagem). Desta forma, o modelo BiLSTM incrementa a quantidade de informação disponível para a rede, melhorando assim o contexto disponível para o algoritmo. À imagem do que acontece na segunda abordagem, os modelos BiLSTM também contêm hiperparâmetros otimizados. Após o modelo BiLSTM ser treinado e serem geradas as previsões do mesmo, são novamente calculadas as métricas de avaliação. Este processo foi seguido para os sete casos de estudo. Os resultados de *performance* obtidos estão disponíveis nas Tabelas 4, 5 e 6.

Primeira Abordagem				
Índice	**MAE**	**RMSE**	**R^2**	**MAPE**
S&P 500	42.94	53.43	0.94	1.06
Dow Jones	333.55	424.86	0.56	1.02
Nasdaq	167.97	207.35	0.96	1.42
Russell	25.85	32.33	0.86	1.42
Nikkei	294.51	370.98	0.92	1.01
Gold	24.12	29.13	0.97	1.28
Crude Oil	3.04	3.55145	0.68	3.89

Tabela 4 - Resultados das métricas para o modelo LSTM com 1 input após serem geradas as previsões

Segunda Abordagem				
Índice	**RMSE**	**MAE**	**R^2**	**MAPE**
S&P 500	51.71	40.88	0.95	1.01
Dow Jones	338.28	255.92	0.94	0.78
Nasdaq	185.62	148.78	0.97	1.27
Russell	31.29	25.07	0.84	1.38
Nikkei	328.10	253.03	0.97	0.87
Gold	18.71	14.55	0.97	0.79
Crude Oil	2.33	1.86	0.86	2.38

Tabela 5 - Resultados das métricas para o modelo LSTM com 1 input após a otimização de parâmetros

Terceira Abordagem				
Índice	**MAE**	**RMSE**	**MAPE**	**R^2**
S&P 500	39.29	48.97	0.98	0.95
Dow Jones	268.75	344.18	0.82	0.94
Nasdaq	152.29	187.84	1.29	0.97
Russell	21.78	27.09	1.19	0.88
Nikkei	254.11	330.87	0.87	0.97
Gold	12.92	17.27	0.69	0.97
Crude Oil	1.69	2.09	2.15	0.89

Tabela 6 - Resultados das métricas para o modelo BiLSTM com 1 input

Após efetuar a análise dos resultados, existem algumas conclusões que se podem imediatamente tirar. Começando pelo Coeficiente de Determinação, que é calculado apenas na segunda e terceira abordagem, percebe-se que o seu valor não varia de forma significativa após ser utilizado o modelo BiLSTM. Dos sete casos de estudo apresentados, o valor do R^2 melhora em quatro, quando se recorre ao modelo BiLSTM, piorando nos restantes três. Ainda assim, as diferenças apresentadas entre as duas abordagens são residuais (com exceção do índice Russell e do petróleo, onde as diferenças são de 0.038 e 0.026, respetivamente), pelo que esta métrica não nos permite obter uma conclusão precisa relativamente à utilização do modelo BiLSTM ao invés do modelo LSTM com os hiperparâmetros otimizados.

Relativamente às métricas RMSE e MAE, os resultados obtidos entre a primeira e segunda abordagem vão de encontro ao esperado. Uma vez que a principal diferença entre estas duas abordagens reside no facto de se ter utilizado um modelo com parâmetros otimizados na segunda, seria de esperar que os resultados melhorassem, o que de facto se acaba por cumprir, como se pode analisar através da análise das tabelas: os valores do RMSE e do MAE diminuíram para todos os exemplos utilizados, após se ter utilizado o modelo com os hiperparâmetros otimizados. No entanto, os resultados não são lineares quando comparamos a segunda abordagem com a terceira; à imagem do que acontece com o Coeficiente de Determinação, também aqui os resultados são bastante díspares. Em quatro casos (S&P 500, Russell, ouro e petróleo) os valores melhoram quando se utiliza o modelo BiLSTM, uma vez que se calculam valores inferiores para estas duas métricas, enquanto que para os restantes três (Dow Jones, Nasdaq e Nikkei) os valores obtidos são melhores quando se utiliza o modelo LSTM com os parâmetros otimizados. À semelhança do que acontece com o R^2, quando o modelo apenas conta com uma variável de *input*, nenhuma destas duas abordagens é indiscutivelmente melhor que a outra. Importa ainda destacar que os resultados são melhores na terceira abordagem, quando comparando com os obtidos na primeira, pelo que se conclui que um modelo BiLSTM com hiperparâmetros otimizados apresenta melhores resultados do que um modelo LSTM sem os parâmetros otimizados.

A métrica do MAPE é calculada apenas na primeira e terceira abordagem. À imagem do que acontece com o RMSE e o MAE, também aqui os resultados são melhores quando se utiliza o modelo BiLSTM, algo que também já era expectável, e que ajuda a reforçar a utilização desta técnica ao invés de um modelo LSTM simples sem os hiperparâmetros otimizados.

Em suma, podemos concluir que os resultados obtidos através da análise dos modelos onde se utilizou apenas um *input* são bastante positivos. Os resultados das métricas RMSE, MAE e MAPE revelaram-se relativamente baixos, principalmente quando temos em consideração o valor da média dos dados, o que nos assegura que as previsões obtidas não estão muito distantes dos valores reais. Os valores do Coeficiente de Determinação (em alguns casos, próximo dos 98%) ajudam-nos também a concluir que o modelo se ajusta bem aos dados. Ficou também claro que os resultados obtidos pela segunda e terceira abordagem (modelo LSTM com parâmetros otimizados e BiLSTM, respetivamente) superam com alguma margem os resultados obtidos na primeira abordagem, onde as métricas foram calculadas imediatamente após ser gerado o modelo. No entanto, não foi possível obter uma conclusão clara relativamente a se devemos preterir o modelo BiLSTM ao modelo LSTM com parâmetros otimizados, uma vez

que nenhum destes demonstrou ser nitidamente melhor que o outro, como foi possível concluir através da análise das métricas. A interpretabilidade de ambos os modelos é bastante semelhante, pelo que a utilização deste critério também não é suficiente para se optar por um dos modelos em detrimento do outro. O tempo de execução do modelo BiLSTM é, no entanto, inferior ao tempo de execução do modelo LSTM, sendo que esse é o único critério onde o modelo BiLSTM se destaca quando comparando com o LSTM. Contudo, a diferença temporal entre a execução dos dois modelos não é elevada o suficiente para que se possa dizer que o modelo BiLSTM é claramente superior. Na próxima fase, onde serão avaliados modelos com duas variáveis de *input*, procurar-se-á perceber se esta situação se altera.

5.5 Modelos com dois inputs

No cenário seguinte foram considerados duas variáveis de *input* nos modelos utilizados: o valor de fecho de mercado e o volume de ações transacionadas. À imagem do que acontece no cenário anterior, também neste modelo se utiliza um *lag* de 2 *time steps*. Um volume de transações elevado que seja acompanhado por um aumento do valor das ações pode significar um interesse forte por parte dos compradores nessas ações. Por outro lado, volumes elevados de transações acompanhados por uma queda de preços pode demonstrar um sinal de preocupação por parte dos compradores. No presente cenário, à imagem do que aconteceu no modelo com apenas um *input*, os valores de correlação entre o valor previsto e os valores de fecho foram bastante elevados, variando entre 0.943 e 0.990. Os valores podem ser consultados na Tabela 7.

Índice	Correlação Previsão vs Valor de Fecho
S&P 500	0.979
Dow Jones	0.971
Nasdaq	0.986
Russell	0.943
Nikkei	0.990
Gold	0.989
Crude Oil	0.948

Tabela 7 - Correlação entre os valores reais e preditos – modelo LSTM com 2 inputs

No decorrer do presente trabalho, entre as diferentes pesquisas efetuadas, procurou-se apenas variar o número de variáveis *input* utilizadas, mantendo a restante estrutura dos modelos. Dessa forma, independentemente do número de *inputs* presentes em cada estudo, as técnicas utilizadas foram sempre as mesmas. Significa isto que as três abordagens apresentadas no cenário anterior - com uma variável *input* - se mantiveram neste cenário, onde existem duas variáveis *input*, assim como nos cenários que serão apresentados posteriormente. Os resultados obtidos no presente cenário encontram-se presentes nas Tabelas 8, 9 e 10:

Primeira Abordagem				
Índice	**MAE**	**RMSE**	**R^2**	**MAPE**
S&P 500	50.26	61.65	0.93	1.25
Dow Jones	348.07	424.30	0.87	1.32
Nasdaq	156.33	196.76	0.96	1.32
Russell	22.30	27.73	0.88	1.22
Nikkei	265.53	336.05	0.95	0.91
Gold	28.27	32.52	0.98	1.50
Crude Oil	2.16	2.82	0.61	2.74

Tabela 8 - Resultados das métricas para o modelo LSTM com 2 inputs após serem geradas as previsões

Segunda Abordagem				
Índice	**RMSE**	**MAE**	**R^2**	**MAPE**
S&P 500	47.72	36.76	0.96	0.92
Dow Jones	424.30	348.07	0.91	1.05
Nasdaq	178.37	141.82	0.97	1.20
Russell	30.09	24.10	0.85	1.32
Nikkei	371.05	288.38	0.97	0.99
Gold	19.22	15.23	0.97	0.83
Crude Oil	2.41	1.88	0.85	2.41

Tabela 9 - Resultados das métricas para o modelo LSTM com 2 inputs após a otimização de parâmetros

Terceira Abordagem				
Índice	**MAE**	**RMSE**	**MAPE**	**R^2**
S&P 500	38.74	48.81	0.96	0.95
Dow Jones	268.00	342.56	0.81	0.94
Nasdaq	147.13	182.31	1.25	0.97
Russell	22.18	27.66	1.21	0.88
Nikkei	237.41	305.53	0.82	0.98
Gold	12.77	17.11	0.69	0.97
Crude Oil	1.71	2.15	2.18	0.88

Tabela 10 - Resultados das métricas para o modelo BiLSTM com 2 inputs

Começando por analisar a variação dos valores obtidos na primeira abordagem, quando comparados com os valores analisados no cenário anterior, percebe-se que não é possível obter uma conclusão clara, uma vez que dos sete casos de estudo apresentados, quatro (Nasdaq, Russell, Nikkei e petróleo) melhoram quando é introduzido no modelo a segunda variável *input*, enquanto que os restantes três (S&P 500, Dow Jones e ouro) obtêm melhores resultados quando se utiliza apenas uma variável *input*. O modelo com dois *inputs* apresenta resultados surpreendentes na segunda abordagem, onde se utiliza o modelo com hiperparâmetros otimizados, quando comparando com os obtidos na primeira abordagem. Ao contrário do que acontece no cenário anterior, onde todas as métricas melhoram quando se utiliza o modelo com os parâmetros otimizados, no modelo com dois *inputs* é possível encontrar um caso de estudo onde os valores se mantêm constantes (os valores do RMSE e do MAE são os mesmos para o índice Dow Jones), e dois casos onde o valor das métricas pioram (nos índices Russell e Nikkei). À imagem do que aconteceu na análise da primeira abordagem entre os dois cenários, também na segunda abordagem os resultados oscilam, não sendo assim possível estabelecer um padrão que identifique em qual dos modelos os resultados são claramente melhores. Enquanto que os índices S&P 500, Nasdaq e Russell obtêm melhores resultados quando se utilizam duas variáveis *input* no modelo com parâmetros otimizados, os resultados são melhores para os restantes quatro casos de estudo (Dow Jones, Nikkei, ouro e petróleo) quando se utiliza apenas uma variável *input*.

No entanto, quando comparamos os resultados das métricas da terceira abordagem – o modelo BiLSTM com duas variáveis *input* – com o mesmo modelo do primeiro cenário, são

obtidos melhores resultados em cinco dos sete casos de estudo, sendo que nos dois casos onde isso não acontece (Russell e petróleo) as diferenças entre os resultados são pouco significativas. Na comparação das métricas de performance obtidas pelos modelos BiLSTM e LSTM com hiperparâmetros otimizados, torna-se evidente uma diferença significativa nos resultados: os modelos LSTM apenas são melhores em dois índices (S&P 500 e Nasdaq), sendo que nos restantes cinco casos de estudo, os resultados dos modelos BiLSTM são superiores.

O modelo BiLSTM com hiperparâmetros otimizados e duas variáveis *input* (valor de fecho e volume) demonstra assim ser aquele que apresenta melhores resultados, superando os valores das métricas obtidos pelos modelos LSTM com duas variáveis *input*, assim como os resultados obtidos pelo modelo BiLSTM com apenas uma variável *input*. Na próxima fase do estudo serão estudados e analisados os resultados obtidos por modelos com quatro variáveis *input*, pelo que será interessante perceber se algum deles consegue superar os resultados obtidos pelo modelo BiLSTM que utiliza duas variáveis *input*.

5.6 Modelos com quatro inputs

Nesta fase, para além dos dois *inputs* utilizados no exemplo anterior, são adicionados dois novos: o valor das médias móveis duplas e das médias móveis triplas. Tal como foi explicado previamente, estes são dois dos indicadores mais utilizados em análise técnica, que auxiliam os investidores a procurar perceber em que sentido se irá movimentar o valor das ações. Neste modelo, ao contrário dos dois anteriores, apenas foi utilizado um *lag* de *time step*.

Através da análise da Tabela 11, pode-se concluir que os coeficientes de correlação linear entre os valores previstos e os valores de fecho aumentaram ligeiramente para todos os índices, assim como para o valor do ouro, sendo que a única exceção se registou no caso do petróleo.

Índice	Correlação: Previsão vs. Valor de Fecho
S&P 500	0.980
Dow Jones	0.974
Nasdaq	0.987
Russell	0.947
Nikkei	0.990
Gold	0.989
Crude Oil	0.944

Tabela 11 - Correlação entre os valores reais e preditos – modelo LSTM com 4 inputs

Como foi explicado previamente, as únicas alterações significativas entre cada modelo são o número de variáveis de *input* utilizadas. Para o modelo atual, onde foram utilizadas quatro variáveis *input*, os resultados obtidos foram os seguintes:

Primeira Abordagem				
Índice	**MAE**	**RMSE**	**R^2**	**MAPE**
S&P 500	37.94	49.47	0.95	0.94
Dow Jones	300.61	385.94	0.94	0.91
Nasdaq	203.78	245.93	0.95	1.72
Russell	26.28	32.87	0.87	1.43
Nikkei	246.83	320.31	0.96	0.68
Gold	12.66	16.86	0.96	1.50
Crude Oil	3.73	4.39	0.82	4.68

Tabela 12 - Resultados das métricas para o modelo LSTM com 4 inputs após serem geradas as previsões

Segunda Abordagem				
Índice	**RMSE**	**MAE**	**R^2**	**MAPE**
S&P 500	47.72	37.32	0.96	0.93
Dow Jones	336.94	258.76	0.94	0.78
Nasdaq	219.02	178.15	0.96	1.51
Russell	29.42	23.51	0.86	1.29
Nikkei	369.09	290.88	0.97	1.01
Gold	20.16	15.97	0.97	0.85
Crude Oil	2.12	1.75	0.88	2.22

Tabela 13 - Resultados das métricas para o modelo LSTM com 4 inputs após a otimização de parâmetros

Terceira Abordagem				
Índice	**MAE**	**RMSE**	**MAPE**	**R^2**
S&P 500	38.43	52.08	0.95	0.95
Dow Jones	254.60	327.39	0.77	0.94
Nasdaq	137.07	175.63	1.16	0.97
Russell	20.73	26.10	1.13	0.89
Nikkei	259.59	330.77	0.89	0.97
Gold	12.88	17.11	0.69	0.97
Crude Oil	1.64	2.13	2.10	0.88

Tabela 14 - Resultados das métricas para o modelo BiLSTM com 4 inputs

Após analisar os resultados da primeira abordagem (imediatamente após se calcularem os valores previstos) com um modelo de quatro variáveis de *input*, percebe-se que não existe uma melhoria considerável, quando comparando com os resultados obtidos na primeira abordagem do modelo anterior, onde se utilizaram duas variáveis de *input*. No presente cenário, os valores do MAE e do RMSE melhoraram para os índices S&P 500, Dow Jones, Nikkei e ouro, mas pioraram para os restantes (Nasdaq, Russel e petróleo). Importa também destacar que, para o mesmo modelo com quatro variáveis de *input*, os resultados entre a primeira e a segunda abordagem (onde se utilizam parâmetros otimizados) não tendem a apresentar melhorias

significativas (com exceção do petróleo, onde se identificam melhorias consideráveis nos resultados). Na verdade, os resultados inclusive pioram para os resultados do índice Nikkei e do ouro, entre as duas abordagens.

Comparando os resultados da terceira abordagem, onde se utiliza um modelo BiLSTM com hiperparâmetros otimizados e quatro variáveis de *input*, com os resultados da segunda abordagem, é possível concluir que os resultados tendem a melhorar, sendo que a única exceção é o índice S&P 500, onde os resultados obtidos são ligeiramente melhores quando se utiliza o modelo LSTM com os hiperparâmetros otimizados, assim como a métrica RMSE para o petróleo. No entanto, importa recordar que no cenário previamente analisado, onde se utilizaram apenas duas variáveis de *input*, se chegou à conclusão de que o modelo BiLSTM com duas variáveis de *input* apresentava os melhores resultados para as métricas utilizadas. Será que esse cenário se mantém? Quando se utilizam quatro variáveis de *input*, o modelo BiLSTM também é aquele que, de um modo geral, apresenta os melhores resultados, pelo que é interessante comparar os resultados obtidos entre os modelos BiLSTM, quando se utilizam duas e quatro variáveis de *input*.

Após comparar os dois modelos BiLSTM com algum detalhe, conclui-se que aquele que recorre a quatro variáveis de *input* tende a apresentar melhores resultados. O modelo com quatro variáveis de *input* apresenta melhores resultados para os índices S&P 500, Dow Jones, Nasdaq e Russell, assim como para os valores do petróleo (importa apenas destacar que, para o índice S&P 500, a métrica RMSE apresenta melhores resultados quando se utilizam duas variáveis de *inputs* no modelo BiLSTM, sendo esta a única exceção nos resultados analisados). O modelo com duas variáveis de *input* apenas apresenta melhores resultados para o índice Nikkei e para o valor de fecho do ouro, ainda que para este último, a diferença de resultados seja mínima. Tendo tudo isto em consideração, é justo afirmar que, de todos os modelos que já foram estudados e avaliados, o modelo BiLSTM com hiperparâmetros otimizados e quatro variáveis de *input* aparenta ser aquele que apresenta melhores resultados, pelo que este modelo será a referência para as próximas análises.

Apesar do modelo BiLSTM com quatro variáveis de *input* ser aquele que, de modo geral, aparenta apresentar os melhores resultados, estamos na presença de um fenómeno que merece destaque: a relação entre os resultados obtidos na primeira abordagem (onde as métricas são calculadas imediatamente após se calcularem as previsões - sem otimização de hiperparâmetros) e na terceira abordagem (onde se utiliza o modelo BiLSTM com otimização de hiperparâmetros). No primeiro cenário estudado, onde foi utilizado apenas uma variável de

input, a diferença entre os resultados obtidos na primeira e na terceira abordagem é notória, uma vez que a melhoria nos resultados se torna evidente quando se utiliza o modelo BiLSTM. No segundo cenário, onde se utilizam duas variáveis de *input*, o modelo BiLSTM continua a apresentar melhores resultados do que aqueles obtidos na primeira abordagem, apesar de existir uma maior aproximação nos valores para alguns índices (note-se, por exemplo, o índice Russell, onde a diferença de resultados entre a primeira e a terceira abordagem é praticamente residual). No presente cenário, onde se recorrem a modelos com quatro variáveis de *input*, esta aproximação de resultados é ainda maior, sendo que para dois dos casos de estudo utilizados (Nikkei e ouro) a primeira abordagem (modelo LSTM) obtém melhores resultados do que a terceira (modelo BiLSTM). Este fenómeno pode-se explicar parcialmente devido ao facto de a variação de *inputs* afetar mais os resultados na primeira abordagem do que o modelo BiLSTM.

Na próxima fase, onde serão utilizadas mais variáveis de *input*, será interessante perceber se algum dos modelos irá obter melhores resultados do que o modelo BiLSTM com hiperparâmetros otimizados e quatro variáveis de *input*.

5.7 Modelos com oito inputs

No presente cenário, foi inicialmente adicionado apenas mais uma variável de *input* às quatro utilizadas anteriormente: o valor da análise de sentimento. Este valor foi obtido com recurso à aplicação de técnicas de *Text Mining* (mais especificamente, a técnica FinBERT) em diversas notícias e relatórios financeiros, recolhidos e analisados ao longo de diversos meses. Deste modo, foi possível construir uma base de dados onde se registavam todas estas notícias e relatórios, o seu conteúdo, o valor que o modelo FinBERT atribuía ao respetivo conteúdo, assim como a que índice(s) é que esse conteúdo (e respetivo valor de análise de sentimento) podia ser associado. Desta forma, foi possível associar um valor de análise de sentimento a cada um dos sete casos de estudo (o valor dos cinco índices, do ouro e do petróleo). De forma a conseguir adicionar esta variável de *input* ao modelo, foi necessário efetuar alguns ajustes. Os quatro *inputs* usados previamente têm algo em comum: o valor de fecho, o volume, e as médias móveis duplas e triplas estão associadas a apenas um dia, isto é, para um qualquer dia, temos apenas um e um só valor destes *inputs*. Existe, portanto, uma relação de um-para-um entre estas variáveis de *input*, o que facilita a sua organização nos *datasets* utilizados nos modelos. Os valores da análise de sentimento não estão distribuídos da mesma forma, uma vez que para cada dia é possível (e altamente provável) que exista mais do que uma notícia sobre um determinado

índice. É, assim, uma relação de zero (para um qualquer dia, é possível que não existam notícias sobre um determinado índice) para muitos, o que não permite que este *input* se adicione da mesma forma aos *datasets* utilizados nos modelos. A solução encontrada consistiu em calcular a média dos valores de análise de sentimento para cada dia, e utilizar esse resultado como *input* no *dataset*. A título de exemplo: se para o dia 20 de agosto de 2022 a tabela da base de dados tiver dez registos com o valor da análise de sentimento para o índice S&P 500, será calculada a média desses dez valores, e o *output* calculado será utilizado como *input* no *dataset* para a coluna de análise de sentimento, na linha onde a data corresponda a 20 de agosto de 2022. Desta forma, é possível transformar uma relação de zero para muitos numa relação de um-para-um, possibilitando assim que esse valor seja utilizado como *input* nos *datasets*.

Numa primeira fase, apenas foi adicionado o valor da média da análise de sentimento correspondente ao dia da ação em questão. Isto é, para cada linha do *dataset* é recolhida a respetiva data, e através dessa data, são recolhidos os respetivos valores da tabela da base de dados que guarda os dados da análise de sentimento. Procede-se a calcular a média desses valores, e a adicionar o respetivo *output* ao *dataset*. Desta forma, para cada data no *dataset*, era calculada a média da análise de sentimento das notícias publicadas no mesmo dia. Após alguma reflexão, conclui-se que utilizar os dados da análise de sentimento apenas desta forma significaria um mau aproveitamento dos mesmos, e que poderia até eventualmente transmitir informações erradas, influenciado assim negativamente os resultados do modelo. As notícias financeiras são utilizadas por diversos investidores, uma vez que através destas é possível compreender melhor o mercado financeiro, levando assim a que sejam tomadas melhores decisões. No entanto, nem todas as notícias impactam da mesma forma os mercados, nem à mesma velocidade. Algumas notícias são responsáveis por uma oscilação quase imediata nos mercados, enquanto outras demoram dias, semanas ou até meses a afetar os valores dos diversos instrumentos financeiros. Tendo isso em consideração, é interessante perceber não só o impacto imediato das notícias no mercado financeiro, mas também o impacto que terão nos dias seguintes. Com isso em mente, foram adicionados outros três *inputs*: o valor da análise de sentimento uma semana antes do valor de fecho, o valor da análise de sentimento duas semanas antes, e o valor da análise de sentimento dois a quatro dias antes (a variação dos dias neste último *input* justifica-se pelo facto de as bolsas se encontrarem encerradas durante o fim de semana, pelo que o número de dias utilizados para efetuar este cálculo deve ter isso em consideração). Assim, este *dataset* terá os seguintes *inputs*:

- Valor de fecho

- Valor do volume de transações
- Médias móveis duplas
- Médias móveis triplas
- Análise de Sentimento (próprio dia)
- Análise de Sentimento (uma semana antes)
- Análise de Sentimento (duas semanas antes)
- Análise de Sentimento (dois a quatro dias antes)

Como se irá analisar e perceber posteriormente, em muitos casos, os resultados obtidos com estas oito variáveis de *input* não foram os melhores, pelo que se optou por criar novos modelos que apenas utilizassem os *inputs* que demonstrassem ter maior impacto nos modelos, de forma a perceber se seria possível obter melhores resultados dessa forma. Os três *inputs* escolhidos para realizar esses testes foram os que apresentaram maior correlação com os valores previstos, sendo estes o valor de fecho, o valor de volume e um dos *inputs* utilizados para a análise de sentimento (a razão desta variação para o *input* da análise de sentimento será explicada posteriormente).

Antes de prosseguir para a análise dos valores de correlação, importa destacar uma última alteração que apenas foi efetuada no presente cenário: a variação do número de *lags* utilizados. Este termo refere-se a dependências temporais presentes em dados sequenciais, atrasos esses que são utilizados para perceber de que forma os valores anteriores de uma variável influenciam os seus valores futuros. Esta abordagem é particularmente relevante na análise e previsão de séries temporais, nas quais cada dado utilizado está associado a um momento específico no tempo. Os *lags* são essenciais para incorporar informações de *time steps* passados em modelos de *Machine Learning*, permitindo a identificação de padrões e dependências temporais para aprimorar a precisão das previsões.

Assim, no presente cenário, onde se utilizaram os valores da análise de sentimento, foram testadas diversas hipóteses. Primeiramente foram considerados os oito *inputs* mencionados previamente, com a utilização de apenas um *lag*, para cada um dos índices. De seguida, os mesmos modelos foram testados com uma ligeira variação no valor dos *lags*, alterando este valor entre 1, 2 e 3. Para terminar, foram utilizados apenas os três *inputs* que apresentavam o maior valor de correlação, como foi explicado anteriormente. Para estes cenários, onde se utilizavam apenas três *inputs*, também se recorreu a uma variação dos *lags* utilizados, entre 1, 2 e 3. Os resultados obtidos para cada um destes cenários encontram-se na Tabela 15.

Índice	Valor de Fecho	Volume	Média Dupla	Média Tripla	Sentimento			
					dia	1 sem.	2 sem.	2-4 dias
S&P 500_1 lag	0,9689	-0,3552	NaN	NaN	0,24	-0,62	0,48	-0,10
S&P 500_1 lag_3 inputs	0,99082	-0,2157	-	-	-	-	-	-0,06
S&P 500_2 lags	0,88736	-0,5706	NaN	NaN	0,26	-0,50	0,34	-0,09
S&P 500_2 lags_3 inputs	0,98596	-0,2142	-	-	-	-	-	-0,05
S&P 500_3 lags	0,72539	-0,78	NaN	NaN	0,19	-0,36	0,19	-0,05
S&P 500_3 lags_3 inputs	0,97878	-0,2049	-	-	-	-	-	-0,04
Dow Jones_1 lags	0,96893	-0,203	0,21566	NaN	0,36	-0,04	0,72	0,42
Dow Jones_1 lags_3 inputs	0,96287	-0,2092	-	-	-	-	-	0,41
Dow Jones_2 lags	0,93058	-0,2254	0,16949	NaN	0,30	0,03	0,68	0,47
Dow Jones_2 lags_3 inputs	0,96446	-0,2057	-	-	-	-	-	0,41
Dow Jones_3 lags	0,94016	-0,2097	0,18428	NaN	0,35	0,04	0,68	0,50
Dow Jones_3 lags_3 inputs	0,97307	-0,1961	-	-	-	-	-	0,42
Nasdaq_1 lag	0,98267	0,56316	NaN	NaN	-0,53	-0,71	0,60	0,29
Nasdaq_1 lag_3 inputs	0,99071	0,65812	-	-	-	-	-	0,39
Nasdaq_2 lags	0,98611	0,55684	NaN	NaN	-0,55	-0,73	0,59	0,31
Nasdaq_2 lags_3 inputs	0,99007	0,6673	-	-	-	-	-	0,38
Nasdaq_3 lags	0,98509	0,63306	NaN	NaN	-0,67	-0,81	0,61	0,38
Nasdaq_3 lags_3 inputs	0,98728	0,66654	-	-	-	-	-	0,38
Russell_1 lag	0,93457	-0,479	NaN	NaN	0,03	-0,37	0,50	0,16
Russell_1 lag_3 inputs	0,95677	-0,4174	-	-	-	-	-	0,10
Russell_2 lags	0,91174	-0,4681	NaN	NaN	0,02	-0,37	0,51	0,20
Russell_2 lags_3 inputs	0,92396	-0,4253	-	-	-	-	-	0,20
Russell_3 lags	0,88379	-0,4765	NaN	NaN	-0,04	-0,36	0,50	0,18
Russell_3 lags_3 inputs	0,89994	-0,423	-	-	-	-	-	0,22
Nikkei_1 lag	0,67124	-0,0951	NaN	NaN	0,23	-0,05	0,12	-0,31
Nikkei_1 lag_3 inputs	0,69154	-0,0477	-	-	-	-	-	-0,27
Nikkei_2 lags	0,58634	-0,1416	NaN	NaN	0,39	-0,05	0,20	-0,49
Nikkei_2 lags_3 inputs	0,70598	0,08351	-	-	-	-	-	-0,33
Nikkei_3 lags	0,49797	-0,0321	NaN	NaN	0,44	-0,10	0,21	-0,54
Nikkei_3 lags_3 inputs	0,97861	0,43125	-	-	-	-	-	0,56

Gold_1 lag	0,05049	0,69907	-0,1142	NaN	-0,03	0,19	0,34	0,29
Gold_1 lag_3 inputs	0,12778	0,74602	-	-	-	-	-	0,44
Gold_2 lags	-0,3341	0,61463	-0,5975	NaN	0,42	0,10	0,68	-0,06
Gold_2 lags_3 inputs	0,02371	0,71095	-	-	-	-	-	0,34
Gold_3 lags	-0,0534	0,62755	-0,1744	NaN	0,04	0,19	0,46	0,22
Gold_3 lags_3 inputs	-0,0097	0,63802	-	-	-	-	-	0,27
Crude Oil_1 lag	0,87229	-0,9316	NaN	0,19994	0,61	NaN	0,66	NaN
Crude Oil_1 lag_3 inputs	0,819	-0,9092	-	-	-	-	-	NaN
Crude Oil_2 lag	0,69319	-0,9062	NaN	-0,035	0,28	NaN	0,54	NaN
Crude Oil_2 lag_3 inputs	0,66918	-0,8999	-	-	-	-	-	NaN
Crude Oil_3 lag	0,60084	-0,8283	NaN	-0,0633	0,11	NaN	0,48	NaN
Crude Oil_3 lag_3 inputs	0,63614	-0,8626	-	-	-	-	-	NaN

Tabela 15 - Correlação entre os valores reais e preditos – modelo LSTM com 8 e 3 inputs

Uma análise célere permite tirar algumas conclusões imediatas, sendo que a primeira talvez resida nos valores de correlação obtidos para as médias móveis duplas e médias móveis triplas. Na fase anterior, onde se utilizaram apenas quatro variáveis de *input*, concluiu-se que os valores obtidos para estes dois valores foram baixos. No entanto, nesta fase, onde se utilizam oito variáveis de *input*, percebe-se que a correlação entre os valores de previsão e as médias móveis duplas e triplas corresponde a *NaN* (*Not a Number*) para a grande maioria dos casos, sendo que as exceções são os valores obtidos para o índice Dow Jones e para o ouro (para as médias móveis duplas) e para o petróleo (para as médias móveis triplas).

Outra conclusão que se pode retirar da análise da tabela é relativa aos valores de correlação entre o valor previsto e o valor de fecho sobre o conjunto de teste: para a grande maioria dos exemplos estudados, os valores de correlação melhoram quando se utilizam apenas três variáveis de *input*, ao invés das oito variáveis iniciais. Das vinte e uma combinações realizadas onde se varia apenas o número de variáveis de *input*, em apenas cinco (Dow Jones com 1 *lag*, ouro com 2 e 3 *lags* e petróleo com 1 e 2 *lags*) o valor de correlação diminui quando se utilizam três variáveis de *input*, sendo que para os restantes dezasseis casos de estudo, os resultados são piores quando se recorrem às oito variáveis de *input*. Relativamente à correlação do valor previsto com o volume, a tendência é oposta à do valor de fecho: dos vinte e um casos de estudo, os valores de correlação melhoram em apenas nove (Dow Jones com 1 *lag*, Nasdaq, Nikkei com 3 *lags*, ouro e petróleo com 3 *lags*) quando se reduz o número de variáveis de *input* para

três. Antes de terminar a análise às variáveis de valor de fecho e de volume, importa realçar que os seus valores de correlação foram extremamente fracos para alguns casos de estudo, onde se destacam o valor de correlação para o valor de fecho do ouro quando se utilizam três *lags* e três *inputs* (-0.00972), ou o valor de correlação para o valor do volume do índice Nikkei, quando se utilizam três *lags* (-0.03213).

Relativamente às quatro colunas que contemplam os valores de correlação da análise de sentimento, através de uma análise detalhada, percebe-se que optar por adicionar as três variáveis de *input* adicionais (análise de sentimento uma semana antes, duas semanas antes e alguns dias antes) foi uma decisão acertada e fundamentada, uma vez que para os vinte e um casos de estudo onde se calcularam os valores destas quatro variáveis de *input*, em nenhum dos casos o *input* original (análise de sentimento para o próprio dia) foi o que apresentou os maiores valores de correlação. A variável de *input* com a análise de sentimento duas semanas antes é aquela que consegue, de forma destacada, apresentar os melhores valores de correlação: dos vinte e um casos de estudo, apresenta o melhor valor em quinze. A variável de *input* com a análise de sentimento uma semana antes apresenta o melhor resultado de correlação para os três casos do índice S&P 500, enquanto a variável de *input* com a análise de sentimento alguns dias antes apresenta o melhor valor para os três casos do índice Nikkei. Como foi mencionado anteriormente, quando os modelos com três variáveis de *input* foram criados, dois dos *inputs* usados são constantes (valor de fecho e valor do volume), enquanto o terceiro variava entre a análise de sentimento duas semanas antes, uma semana antes ou alguns dias antes. A razão desta variação, como se pode concluir, depende do valor de correlação de cada um, uma vez que se optou por priorizar aquele que apresenta o valor mais elevado. Desta forma, a variável de *input* de análise de sentimento uma semana antes foi escolhido como terceiro *input* nos três modelos do índice S&P 500, a análise de sentimento alguns dias antes foi escolhida como a terceira variável de *input* nos modelos do índice Nikkei, enquanto a análise de sentimento duas semanas antes foi escolhida para os restantes quinze modelos que utilizam três variáveis de *input*.

Agora que os valores de correlação foram analisados, prosseguir-se-á para a interpretação das métricas calculadas através dos modelos com estas variáveis de *input*. Numa primeira fase, serão apenas analisados os resultados obtidos para os modelos que utilizam oito variáveis de *input*, para um número variado de *lags*. Os resultados das métricas encontram-se nas Tabelas 16, 17 e 18.

Primeira Abordagem				
Índice	**MAE**	**RMSE**	**R^2**	**MAPE**
S&P 500_1 lag	29.46	42.44	0.44	0.68
S&P 500_2 lags	59.79	81.61	0.48	1.38
S&P 500_3 lags	90.41	118.49	0.16	2.09
Dow Jones_1 lag	194.27	225.11	0.89	0.58
Dow Jones_2 lags	257.52	293.22	0.85	0.76
Dow Jones_3 lags	333.40	361.94	0.75	0.99
Nasdaq_1 lag	262.62	311.90	0.87	1.91
Nasdaq_2 lags	405.35	466.22	0.60	2.96
Nasdaq_3 lags	524.72	558.08	0.35	3.87
Russell_1 lag	21.60	35.25	0.84	1.15
Russell_2 lags	20.86	35.71	0.73	1.11
Russell_3 lags	23.59	41.97	0.64	1.25
Nikkei_1 lag	220.08	421.96	0.33	0.69
Nikkei_2 lags	282.01	450.80	0.33	0.89
Nikkei_3 lags	225.88	453.15	0.12	0.71
Gold_1 lag	9.81	14.74	-2.28	0.50
Gold_2 lags	20.02	24.99	-5.48	1.02
Gold_3 lags	19.20	37.53	-13.88	0.98
Crude Oil_1 lag	2.36	2.43	0.25	3.36
Crude Oil_2 lags	1.59	1.96	0.27	2.22
Crude Oil_3 lags	1.51	2.12	0.30	2.09

Tabela 16 - Resultados das métricas para o modelo LSTM com 8 inputs após serem geradas as previsões

Segunda Abordagem				
Índice	**RMSE**	**MAE**	**R^2**	**MAPE**
S&P 500_1 lag	56.08	25.47	0.75	0.49
S&P 500_2 lags	40.31	22.45	0.87	0.52
S&P 500_3 lags	43.38	26.43	0.85	0.61
Dow Jones_1 lag	284.16	187.37	0.84	0.55
Dow Jones_2 lags	233.59	193.71	0.89	0.58
Dow Jones_3 lags	281.01	245.81	0.84	0.73
Nasdaq_1 lag	126.03	96.07	0.95	0.71
Nasdaq_2 lags	253.83	172.85	0.80	1.24
Nasdaq_3 lags	146.61	108.56	0.93	0.82
Russell_1 lag	51.10	38.33	0.56	2.05
Russell_2 lags	46.84	24.72	0.63	1.31
Russell_3 lags	47.72	35.62	0.62	1.89
Nikkei_1 lag	415.11	227.31	0.27	0.71
Nikkei_2 lags	413.45	273.73	0.28	0.87
Nikkei_3 lags	370.65	227.80	0.42	0.72
Gold_1 lag	12.82	11.54	-0.83	0.59
Gold_2 lags	12.06	8.80	-0.62	0.45
Gold_3 lags	82.42	34.06	-74.81	1.74
Crude Oil_1 lag	2.07	1.81	0.33	2.57
Crude Oil_2 lags	2.14	0.95	0.29	1.28
Crude Oil_3 lags	2.16	1.46	0.27	2.01

Tabela 17 - Resultados das métricas para o modelo LSTM com 8 inputs após a otimização de parâmetros

Terceira Abordagem				
Índice	**MAE**	**RMSE**	**MAPE**	**R^2**
S&P 500_1 lag	24.47	34.99	0.56	0.90
S&P 500_2 lags	54.01	72.21	1.25	0.59
S&P 500_3 lags	60.15	77.40	1.39	0.53
Dow Jones_1 lag	193.84	268.87	0.57	0.86
Dow Jones_2 lags	194.70	300.50	0.57	0.82
Dow Jones_3 lags	300.94	381.10	0.88	0.72
Nasdaq_1 lag	249.05	283.92	1.82	0.75
Nasdaq_2 lags	472.99	544.02	3.45	0.11
Nasdaq_3 lags	422.20	497.64	3.07	0.26
Russell_1 lag	23.62	34.92	1.25	0.79
Russell_2 lags	29.03	47.57	1.52	0.62
Russell_3 lags	38.98	61.84	2.04	0.36
Nikkei_1 lag	146.93	382.22	0.45	0.38
Nikkei_2 lags	192.10	439.77	0.60	0.18
Nikkei_3 lags	358.29	537.94	1.12	-0.21
Gold_1 lag	5.20	8.59	0.26	0.17
Gold_2 lags	12.54	28.47	0.64	-8.04
Gold_3 lags	13.52	23.75	0.69	-5.29
Crude Oil_1 lag	0.69	1.57	0.95	0.61
Crude Oil_2 lags	1.18	2.29	1.60	0.19
Crude Oil_3 lags	2.79	2.88	3.95	-0.28

Tabela 18 - Resultados das métricas para o modelo BiLSTM com 8 inputs

No primeiro cenário analisado, onde se utiliza apenas uma variável de *input*, era significativa a melhoria dos valores das métricas entre a primeira e a segunda abordagem, onde se otimizavam os parâmetros do modelo. Nos segundo e terceiro cenários, onde se utilizavam duas e quatro variáveis de *input*, respetivamente, apesar de os resultados na segunda abordagem serem tendencialmente melhores, foi possível encontrar alguns casos onde a utilização de parâmetros otimizados resultava na obtenção de piores resultados, nomeadamente no índice

Nikkei e no ouro. No presente cenário, onde se utilizam oito variáveis de *input*, é possível encontrar diversos casos onde os resultados das métricas pioram na segunda abordagem, onde se utilizam parâmetros otimizados. No decorrer do estudo é possível encontrar um padrão inversamente proporcional entre o número de variáveis de *input* utilizadas e a eficácia da utilização de parâmetros otimizados, onde a última diminui à medida que o número de variáveis de *input* utilizadas aumenta.

Comparando os resultados da primeira abordagem do presente cenário com os resultados do cenário anterior, onde não se utilizaram as variáveis de *input* de análise de sentimento, percebe-se que as métricas atuais apresentam melhores resultados, quando se utiliza apenas um *lag*. No entanto, é também evidente o quanto os resultados pioram à medida que o número de *lags* aumenta. Excetuando o índice Russell e o petróleo, os resultados pioram sempre que se incrementa o número de *lags* de um para dois. Quando se aumenta o número de *lags* de dois para três, os resultados pioram em quatro casos de estudo, mas melhoram nos restantes três.

Na segunda abordagem, não é possível encontrar um padrão entre o presente cenário e o cenário anterior que permita concluir se os resultados das métricas melhoram ou pioram com o aumento do número de variáveis de *input*. A variação do número de *lags* apresenta resultados semelhantes, uma vez que a maioria dos casos de estudo não apresenta uma variação constante à medida que o número de *lags* aumenta. O índice Nikkei é o único onde os resultados melhoram sempre quando se aumenta o número de *lags*, enquanto o petróleo é o único exemplo onde aumentar o número de *lags* se reflete em piores resultados.

Por fim, a terceira abordagem. À imagem do que acontece na primeira abordagem, não é possível encontrar um padrão claro entre os resultados obtidos com e sem a utilização das variáveis de *input* de análise de sentimento, com exceção do valor do R^2, onde é evidente que este piora quando se utilizam oito variáveis de *input* (nota especial para os valores obtidos no estudo do ouro, que são extremamente fracos). No entanto, ao contrário do que acontece nas abordagens anteriores, com a utilização do modelo BiLSTM é notório que os resultados pioram quando se aumentam o número de *lags*, para praticamente todos os casos de estudo – a exceção é o índice Nasdaq, onde os resultados melhoram ligeiramente quando se incrementa o número de *lags* de dois para três.

No estudo dos resultados no cenário anterior, onde se utilizaram quatro variáveis de *input*, conclui-se que o modelo BiLSTM era o que apresentava, até então, os melhores resultados. No cenário atual, onde se utilizam oito variáveis de *input*, e quando se utiliza apenas um *lag*, os

resultados das métricas MAE, RMSE e MAPE tendem a ser melhores. Apesar disso, os resultados do R^2 são baixos para a maioria dos casos atuais (note-se, por exemplo, o valor da métrica para o Nikkei e para o ouro: 0.38614 e 0.17653, respetivamente). Não seria correto afirmar que uma das abordagens é indubitavelmente superior à outra, tendo em consideração as discrepâncias de resultados presentes nesta métrica. Dependendo do índice que se pretende procurar prever, pode-se utilizar o modelo que tendencialmente apresente melhores resultados. Analisando, por exemplo, o índice S&P 500: o resultado do R^2 é superior quando se utilizam apenas quatro variáveis de *input* (0.952 ao invés de 0.904). No entanto, os resultados das restantes métricas são claramente superiores quando se recorre à utilização de oito variáveis de *input*. Tendo isso em consideração, para este índice em particular, o mais sensato talvez fosse utilizar o modelo BiLSTM com oito *inputs* e um *lag*.

5.8 Modelos com três inputs

Terminada a análise das métricas para os modelos com oito variáveis de *input*, segue-se o estudo dos resultados obtidos para os modelos onde apenas foram utilizadas as três variáveis de *input* com maior correlação com os valores previstos.

Primeira Abordagem				
Índice	**MAE**	**RMSE**	**R^2**	**MAPE**
S&P 500_1 lag	14.37	30.85	0.96	0.32
S&P 500_2 lags	26.99	58.87	0.86	0.60
S&P 500_3 lags	38.98	89.59	0.64	0.87
Dow Jones_1 lag	130.16	179.49	0.94	0.38
Dow Jones_2 lags	185.47	226.09	0.91	0.55
Dow Jones_3 lags	262.40	291.33	0.84	0.78
Nasdaq_1 lag	287.80	411.66	0.51	2.06
Nasdaq_2 lags	306.09	419.89	0.11	2.20
Nasdaq_3 lags	647.03	878.34	-0.14	4.66
Russell_1 lag	12.27	24.95	0.91	0.65
Russell_2 lags	16.74	30.79	0.84	0.89
Russell_3 lags	19.71	35.06	0.79	1.04
Nikkei_1 lag	189.84	413.80	0.27	0.59
Nikkei_2 lags	256.39	397.41	0.35	0.80
Nikkei_3 lags	165.77	381.37	0.15	0.51
Gold_1 lag	8.38	16.02	-2.24	0.43
Gold_2 lags	17.90	36.33	-11.93	0.91
Gold_3 lags	19.42	47.60	-29.32	0.99
Crude Oil_1 lag	1.39	1.63	0.66	1.96
Crude Oil_2 lags	1.53	1.95	0.39	2.13
Crude Oil_3 lags	1.51	2.12	0.38	2.09

Tabela 19 - Resultados das métricas para o modelo LSTM com 3 inputs após serem geradas as previsões

Segunda Abordagem				
Índice	**RMSE**	**MAE**	**R^2**	**MAPE**
S&P 500_1 lag	80.36	44.60	0.49	0.34
S&P 500_2 lags	17.66	9.73	0.97	0.22
S&P 500_3 lags	32.26	13.67	0.91	0.31
Dow Jones_1 lag	200.74	110.78	0.92	0.33
Dow Jones_2 lags	220.25	167.90	0.90	0.49
Dow Jones_3 lags	239.58	208.85	0.89	0.62
Nasdaq_1 lag	143.28	98.84	0.93	0.71
Nasdaq_2 lags	353.05	255.57	0.62	1.83
Nasdaq_3 lags	457.91	338.48	0.37	2.44
Russell_1 lag	53.75	35.72	0.52	1.88
Russell_2 lags	37.87	22.43	0.76	1.18
Russell_3 lags	37.260	21.18	0.76	1.13
Nikkei_1 lag	436.41	186.06	0.19	0.58
Nikkei_2 lags	404.41	219.56	0.31	0.69
Nikkei_3 lags	401.96	247.65	0.32	0.78
Gold_1 lag	20.86	12.12	-3.85	0.62
Gold_2 lags	19.10	12.01	-3.07	0.62
Gold_3 lags	15.75	10.90	-1.77	0.56
Crude Oil_1 lag	1.56	0.83	0.62	1.14
Crude Oil_2 lags	1.80	1.26	0.50	1.75
Crude Oil_3 lags	2.16	1.46	0.27	1.87

Tabela 20 - Resultados das métricas para o modelo LSTM com 3 inputs após a otimização de parâmetros

Terceira Abordagem				
Índice	**MAE**	**RMSE**	**Mape**	**R^2**
S&P 500_1 lag	6.59	14.99	0.15	0.98
S&P 500_2 lags	30.71	43.33	0.70	0.85
S&P 500_3 lags	31.47	45.00	0.72	0.84
Dow Jones_1 lag	156.43	227.89	0.46	0.90
Dow Jones_2 lags	249.57	316.65	0.73	0.80
Dow Jones_3 lags	409.01	457.65	1.21	0.60
Nasdaq_1 lag	34.89	70.10	0.25	0.98
Nasdaq_2 lags	69.37	110.14	0.50	0.96
Nasdaq_3 lags	252.28	290.10	1.84	0.74
Russell_1 lag	14.22	25.35	0.75	0.89
Russell_2 lags	22.67	39.44	1.19	0.74
Russell_3 lags	31.36	51.91	1.65	0.55
Nikkei_1 lag	131.96	377.46	0.41	0.40
Nikkei_2 lags	225.84	372.52	0.71	0.41
Nikkei_3 lags	271.09	457.34	0.85	0.12
Gold_1 lag	3.63	7.68	0.18	0.34
Gold_2 lags	14.06	37.53	0.72	-14.71
Gold_3 lags	20.39	62.50	1.04	-42.60
Crude Oil_1 lag	0.77	1.45	1.05	0.67
Crude Oil_2 lags	1.52	2.22	2.10	0.23
Crude Oil_3 lags	2.79	2.88	3.95	-0.28

Tabela 21 - Resultados das métricas para o modelo BiLSTM com 3 inputs

Após uma breve análise às métricas obtidas na primeira abordagem, percebe-se que os modelos tendem a obter piores resultados quando o número de *lags* incrementa. Isto é algo que acontece em praticamente todos os casos de estudo da primeira abordagem, sendo que as únicas exceções se apresentam no índice Nikkei, quando o número de *lags* passa de dois para três, e em algumas métricas do petróleo, quando o número de *lags* também incrementa para três. Na terceira abordagem encontramos um cenário semelhante, onde os valores das métricas tendem

a piorar à medida que o número de *lags* aumenta. Na verdade, quando se utiliza o modelo BiLSTM, os resultados de todas as métricas pioram sempre que o número de *lags* incrementa, mostrando assim que, para os *datasets* utilizados no estudo, os modelos BiLSTM com hiperparâmetros otimizados que utilizem estas três variáveis de *input* são tanto melhores quanto menor for o número de *lags*. Na segunda abordagem, onde se utiliza o modelo LSTM com os hiperparâmetros otimizados, não é possível concluir com tanta clareza o impacto do número de *lags* utilizados, uma vez que a análise das métricas obtidas não permite identificar um padrão claro na relação entre o número de *lags* utilizados e o valor das métricas estudadas.

Ao comparar a segunda abordagem com a terceira percebe-se que é preferível recorrer ao modelo LSTM com hiperparâmetros otimizados na maioria dos casos, uma vez que para os vinte e um casos de estudo apresentados, este demonstra melhores resultados treze vezes, enquanto o modelo BiLSTM apenas apresenta as melhores métricas nos restantes oito casos de estudo. Foi também realizada uma comparação entre os resultados obtidos para as mesmas abordagens, quando se utilizam três e oito variáveis de *input*. Para o modelo BiLSTM com hiperparâmetros otimizados, pode-se concluir que a utilização de três variáveis de *input* significa quase sempre que se irão obter melhores resultados. Relativamente à segunda abordagem, onde se utilizam os modelos LSTM com hiperparâmetros otimizados, não é possível afirmar com tanta segurança que um dos modelos é indiscutivelmente melhor que o outro, devido à disparidade de resultados.

Para terminar, uma breve comparação entre os modelos BiLSTM com três e quatro variáveis de *input*. Antes de os modelos com a componente de análise de sentimento serem analisados, tinha-se concluído que o modelo BiLSTM que utilizava quatro variáveis de *input* era, de modo geral, o que apresentava melhores resultados. No cenário anterior, onde se comparou esse modelo com o modelo BiLSTM que utilizava oito variáveis de *input*, concluiu-se que não era possível identificar com totais certezas qual dos dois apresentava melhores resultados, uma vez que os resultados variavam bastante entre os diferentes casos de estudo. De modo geral, o modelo com oito variáveis de *input* apresentava melhores resultados para as métricas MAE, RMSE e MAPE, mas para certos casos de estudo, o seu valor do R^2 era extremamente fraco. O modelo com quatro variáveis de *input*, por outro lado, apresentava valores bastante elevados para o R^2, pelo que para certos índices seria mais sensato recorrer a esse modelo, não obstante os piores resultados nas métricas MAE, RMSE e MAPE. No modelo com três variáveis de *input*, quando se utiliza apenas um *lag*, as conclusões são bastante semelhantes: analisando apenas as métricas MAE, RMSE e MAPE, conclui-se que o modelo

com apenas três variáveis de *input* obtém quase sempre melhores resultados, quando comparando com o modelo com quatro variáveis de *input*. No entanto, à imagem do que acontece com o modelo com oito variáveis de *input*, também no modelo com três variáveis de *input* os valores da métrica R^2 tendem a ser piores, como se pode concluir através da observação do índice Nikkei e do ouro, onde o valor desta métrica é, respetivamente, 0.40132 e 0.34036. Importa, no entanto, realçar que para alguns casos o valor desta métrica é extremamente positivo, quando se utiliza o modelo BiLSTM com três variáveis de *input*. Na verdade, revisitando todos os casos de estudo apresentados no decorrer deste trabalho, onde se utilizaram diferentes modelos (LSTM e BiLSTM), número de variáveis de *input* e de *lags*, conclui-se que os dois melhores resultados para a métrica R^2 se obtiveram no presente modelo BiLSTM com três variáveis de *input*, para os índices S&P 500 e Nasdaq, onde os resultados para esta métrica foram, respetivamente, 0.982 e 0.985. Pode-se assim concluir que, para alguns índices, o modelo BiLSTM que utiliza estas três variáveis de *input* e um *lag* é, indubitavelmente, o que apresenta os melhores resultados, e que deve ser fortemente considerado quando se tenta prever o valor de fecho futuro destes índices.

6. Conclusões e Considerações Finais

No decorrer do presente trabalho, foi possível estudar e perceber o impacto da aplicação de técnicas de *Machine Learning*, *Text Mining* e *Web Scraping* na previsão dos mercados financeiros.

No início do trabalho, foram formuladas algumas questões de investigação, de forma a alcançar o objetivo geral da dissertação. Numa primeira fase, procurou-se perceber quais eram as fontes de dados com maior relevância na previsão do mercado financeiro. Como foi possível concluir no decorrer da análise apresentada, foram utilizados diferentes *inputs*, provenientes de diferentes fontes de dados: o valor de fecho dos índices, o volume de ações transacionados, as médias móveis duplas e triplas e a análise de sentimento efetuada às notícias financeiras e aos relatórios financeiros do *Federal Reserve System*. O valor de fecho apresentou valores de correlação consideravelmente elevados em praticamente todos os casos de estudo realizados. O volume, apesar de ter demostrando não ser tão relevante como o valor de fecho, também demonstrou ser importante para prever o valor de fecho em certos casos. Tendo isso em consideração, pode-se concluir que a informação histórica dos índices, proveniente do *website* https://finance.yahoo.com/, demonstrou ser uma fonte de dados preciosa para procurar prever os valores de fecho futuros. As médias móveis baseiam-se em cálculos realizados com os valores de fecho dos dias anteriores, relacionando-se o resultado obtido com o valor de fecho do presente dia. No final, existem três *outputs* possíveis: vender, comprar, ou manter as ações, sendo que, para a grande maioria dos casos, o *output* gerado é manter. Admite-se que a forma como estas variáveis foram representadas e calculadas possa ter impactado negativamente a importância das mesmas. Apesar de os resultados das métricas terem apresentado melhorias quando se adicionaram as variáveis de *input* das médias móveis, não se pode excluir a hipótese de as duas variáveis não estarem representadas na sua plenitude. Devido à vasta utilização de análise técnica (onde se destacam as médias móveis) por parte de investidores na previsão da bolsa de valores, não se recomenda que estas variáveis sejam excluídas de futuros trabalhos. Deve-se, no entanto, procurar métodos alternativos para as representar, de forma a assegurar que a importância das médias móveis se reflete ao máximo nos resultados obtidos.

No decorrer do trabalho procurou também perceber-se de que forma é possível utilizar técnicas de *Web Scraping* para recolher e extrair informações relevantes das diferentes fontes de dados. Como foi referido durante a dissertação, foram recolhidos dados de duas fontes de dados: do *website* https://finance.yahoo.com/, onde se extraíram os valores atualizados dos

índices e as notícias financeiras, e dos sites e blogues oficiais do *Federal Reserve System*. Estes dados serviram de base para calcular e preparar todos os *inputs* utilizados nos modelos de *Deep Learning*, pelo que se pode afirmar que as técnicas de *Web Scraping* foram essenciais para atingir todos os resultados obtidos no decorrer do trabalho. Naturalmente, todos estes dados poderiam ter sido recolhidos manualmente, mas o esforço humano necessário teria sido incomensuravelmente maior. Recorda-se que estes dados devem ser recolhidos numa base diária; o *website* https://finance.yahoo.com/ atualiza os valores dos índices todos os dias e disponibiliza diariamente dezenas ou até centenas de notícias relacionadas com os mercados financeiros. Desenvolver um *script* com poucas centenas de linhas de código é suficiente para recolher todos estes dados, mitigando assim a necessidade de os recolher manualmente todos os dias. Para além disso, uma vez que o *script* é desenvolvido com base em regras, a utilização de *Web Scraping* assegura a uniformidade dos dados, garantindo assim que toda a informação recolhida estará disposta na mesma estrutura, o que facilita o tratamento e posterior utilização da mesma.

Explorou-se também a viabilidade da utilização de análise de sentimento em textos financeiros, com o propósito de prever as oscilações da bolsa de valores. Os resultados obtidos nos modelos que utilizam *inputs* de análise de sentimento provam que, quando os dados são bem estruturados e utilizados, a incorporação destes *inputs* pode ser uma mais-valia para obter melhores resultados. Importa, no entanto, reforçar que a análise de sentimento nem sempre apresenta o mesmo impacto. No decorrer do trabalho, realçou-se que as notícias financeiras nem sempre impactam os mercados financeiros no exato momento em que são publicadas. Por vezes, passam dias, semanas ou até meses para o impacto de uma notícia se refletir nos valores da bolsa de valores. Nesse sentido, é importante considerar o impacto que as notícias poderão ter no futuro, procurando perceber com a maior exatidão possível quanto tempo é que será necessário para a publicação de uma notícia se refletir no mercado financeiro, de forma a maximizar o proveito da análise de sentimento. Na presente dissertação, estudou-se a relação do valor de fecho com a análise de sentimento do próprio dia, da semana anterior, das duas semanas anteriores, e de 2 a 4 dias antes, com o propósito de perceber qual destas variáveis apresentava melhores resultados. Concluiu-se que não existe uma resposta exata para qual das variáveis apresenta melhores resultados, uma vez que é impossível dissociar os *outputs* de análise de sentimento do conteúdo de cada notícia, e que se deverá procurar perceber qual destas quatro variáveis se melhor adequa a cada modelo.

No decorrer do trabalho, foi possível encontrar respostas satisfatórias para a maioria das questões que se pretendiam responder, assim como obter resultados que permitiram obter diversas conclusões. No entanto, como já foi mencionado na presente conclusão, certas soluções poderiam ter sido implementadas de forma diferente, pelo que se devem registar diferentes abordagens para trabalhos futuros que se foquem em problemas semelhantes. Para além da procura de métodos alternativos para representar as médias móveis, cujas motivações foram explicadas no início da presente conclusão, também se deve refletir sobre a forma como as variáveis de análise de sentimento são implementadas. As técnicas de *Text Mining* utilizadas na dissertação revelaram-se extremamente úteis, nomeadamente quando se optou por utilizar quatro variáveis de análise de sentimento, ao invés de apenas uma. No entanto, o facto de os espaços temporais considerados (uma semana antes, duas semanas antes e dois a quatro dias antes) serem definidos *à priori* poderá significar que os resultados obtidos não são necessariamente os melhores. Em trabalhos futuros, deverá ser considerada a hipótese de utilizar outros espaços temporais, de forma a incrementar o leque de resultados obtidos.

Recomenda-se também que, em trabalhos futuros, se pondere seriamente sobre a utilização de análise fundamental. Na presente dissertação apenas se recorreu a análise técnica, aquando da utilização das médias móveis duplas e triplas, ignorando uma metodologia que poderia ter sido extremamente útil. A análise fundamental consiste na análise financeira das contas de uma empresa, com vista a determinar o preço justo de uma ação, e que se fundamenta na expetativa de lucros futuros. O método envolve a projeção dos *cash-flows* futuros da empresa e respetiva atualização para o momento presente. A não utilização de análise fundamental no presente trabalho justifica-se com a complexidade em encontrar certos dados que são importantes para calcular a mesma, assim como a dificuldade de os recolher com recurso a *Web Scraping*. Contudo, recomenda-se que a utilização desta metodologia seja considerada em trabalhos futuros, não obstante a possível dificuldade em encontrar e recolher alguns dados importantes para a aplicação da mesma.

Referências Bibliográficas

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) doi:10.3390/electronics11203414

Albahli, S., Nazir, T., Mehmood, A., Irtaza, A., Alkhalifah, A., & Albattah, W. (2022). AEI-DNET: A novel DenseNet model with an autoencoder for the stock market predictions using stock technical indicators. *Electronics (Switzerland), 11*(4) doi:10.3390/electronics11040611

Al-Maadid, A., Alhazbi, S., & Al-Thelaya, K. (2022). Using machine learning to analyze the impact of coronavirus pandemic news on the stock markets in GCC countries. *Research in International Business and Finance, 61* doi:10.1016/j.ribaf.2022.101667

Beg, M. O., Awan, M. N., & Ali, S. S. (2022). Algorithmic machine learning for prediction of stock prices. *Research anthology on machine learning techniques, methods, and applications* (pp. 1271-1293) doi:10.4018/978-1-6684-6291-1.ch066 Retrieved from www.scopus.com

Bhardwaj, B., Ahmed, S. I., Jaiharie, J., Sorabh Dadhich, R., & Ganesan, M. (2021). Web scraping using summarization and named entity recognition (NER). Artigo presented at the *2021 7th International Conference on Advanced Computing and Communication Systems, ICACCS 2021,* 261-265. doi:10.1109/ICACCS51430.2021.9441888 Retrieved from www.scopus.com

Bhavsar, H., Jivani, A., Amesara, S., Shah, S., Gindani, P., & Patel, S. (2023). *Stock price prediction using sentiment analysis on news headlines* doi:10.1007/978-981-19-3571-8_4 Retrieved from www.scopus.com

Carosia, A. E. d. O. (2022). Sentiment analysis applied to news from the brazilian stock market. *IEEE Latin America Transactions, 20*(3), 512-518. doi:10.1109/TLA.2022.9667151

Chen, Q., & Robert, C. -. (2022). Graph-based learning for stock movement prediction with textual and relational data. *Journal of Financial Data Science, 4*(4), 152-166. doi:10.3905/jfds.2022.1.104

Chen, Y., Zhao, P., Zhang, Z., Bai, J., & Guo, Y. (2022). A stock price forecasting model integrating complementary ensemble empirical mode decomposition and independent component analysis. *International Journal of Computational Intelligence Systems, 15*(1) doi:10.1007/s44196-022-00140-2

Chiong, R., Fan, Z., Hu, Z., & Dhakal, S. (2022). A novel ensemble learning approach for stock market prediction based on sentiment analysis and the sliding window method. *IEEE Transactions on Computational Social Systems,* , 1-11. doi:10.1109/TCSS.2022.3182375

Deng, S., Zhu, Y., Duan, S., Fu, Z., & Liu, Z. (2022). Stock price crash warning in the chinese security market using a machine learning-based method and financial indicators. *Systems, 10*(4) doi:10.3390/systems10040108

Dogra, V., Verma, S., Kavita, Jhanjhi, N. Z., Ghosh, U., & Le, D. -. (2022). A comparative analysis of machine learning models for banking news extraction by multiclass classification with imbalanced datasets of financial news: Challenges and solutions. *International Journal of Interactive Multimedia and Artificial Intelligence, 7*(3), 35-52. doi:10.9781/ijimai.2022.02.002

Fathali, Z., Kodia, Z., & Ben Said, L. (2022). Stock market prediction of NIFTY 50 index applying machine learning techniques. *Applied Artificial Intelligence, 36*(1) doi:10.1080/08839514.2022.2111134

Fattah, D. A., Naim, A. A., Desuky, A. S., & Zaki, M. S. (2022). AutoKeras and particle swarm optimization to predict the price trend of stock exchange. *Bulletin of Electrical Engineering and Informatics, 11*(2), 1100-1109. doi:10.11591/eei.v11i2.3373

Fazlija, B., & Harder, P. (2022). Using financial news sentiment for stock price direction prediction. *Mathematics, 10*(13) doi:10.3390/math10132156

Jabeen, A., Yasir, M., Ansari, Y., Yasmin, S., Moon, J., & Rho, S. (2022). An empirical study of macroeconomic factors and stock returns in the context of economic uncertainty news sentiment using machine learning. *Complexity, 2022* doi:10.1155/2022/4646733

Jacob, S. S., Thanikaiselvan, V., Patra, S., & Kapinesh, G. (2022). Monitoring of stocks using LSTM model and prediction of stock prices. Artigo presented at the *International Conference on Edge Computing and Applications, ICECAA 2022 - Proceedings,* 1689-1695. doi:10.1109/ICECAA55415.2022.9936204 Retrieved from www.scopus.com

Joshi, P., Wang, J., & Busler, M. (2022). A study of the machine learning approach and the MGARCH-BEKK model in volatility transmission. *Journal of Risk and Financial Management, 15*(3) doi:10.3390/jrfm15030116

Kamal, S., Sharma, S., Kumar, V., Alshazly, H., Hussein, H. S., & Martinetz, T. (2022). Trading stocks based on financial news using attention mechanism. *Mathematics, 10*(12) doi:10.3390/math10122001

Khan, W., Ghazanfar, M. A., Azam, M. A., Karami, A., Alyoubi, K. H., & Alfakeeh, A. S. (2022). Stock market prediction using machine learning classifiers and social media, news. *Journal of Ambient Intelligence and Humanized Computing, 13*(7), 3433-3456. doi:10.1007/s12652-020-01839-w

Khoa, B. T., & Huynh, T. T. (2022). Forecasting stock price movement direction by machine learning algorithm. *International Journal of Electrical and Computer Engineering, 12*(6), 6625-6634. doi:10.11591/ijece.v12i6.pp6625-6634

Lakatos, R., Bogacsovics, G., & Hajdu, A. (2022). Predicting the direction of the oil price trend using sentiment analysis. Artigo presented at the *2022 IEEE 2nd Conference on Information Technology and Data Science, CITDS 2022 - Proceedings,* 177-182. doi:10.1109/CITDS54976.2022.9914158 Retrieved from www.scopus.com

Lakshya, Prateek, & Sethia, D. (2022). Stock price prediction using news sentiment analysis. Artigo presented at the *2022 2nd International Conference on Intelligent Technologies, CONIT 2022,* doi:10.1109/CONIT55038.2022.9847747 Retrieved from www.scopus.com

Lee, P., Huang, Z., & Tang, Y. (2022). Trend prediction model of asian stock market volatility dynamic relationship based on machine learning. *Security and Communication Networks, 2022* doi:10.1155/2022/5972698

Li, F., Wang, Z., & Zhou, P. (2022). Ensemble investment strategies based on reinforcement learning. *Scientific Programming, 2022* doi:10.1155/2022/7648810

Li, Y., & Pan, Y. (2022). A novel ensemble deep learning model for stock prediction based on stock prices and news. *International Journal of Data Science and Analytics, 13*(2), 139-149. doi:10.1007/s41060-021-00279-9

Lin, W. -., Tsai, C. -., & Chen, H. (2022). Factors affecting text mining based stock prediction: Text feature representations, machine learning models, and news platforms. *Applied Soft Computing, 130* doi:10.1016/j.asoc.2022.109673

López Rodríguez, F. S., & Zurita López, J. M. (2022). *Detection of buy and sell signals using technical indicators with a prediction model based on neural networks* doi:10.1007/978-3-030-97273-8_48 Retrieved from www.scopus.com

Maurya, B. B. P., Ray, A., Upadhyay, A., Gour, B., & Khan, A. U. (2019). Recursive stock price prediction with machine learning and web scrapping for specified time period. Artigo presented at the *IFIP International Conference on Wireless and Optical Communications Networks, WOCN, , 2019-December* doi:10.1109/WOCN45266.2019.8995080 Retrieved from www.scopus.com

Mohapatra, S., Mukherjee, R., Roy, A., Sengupta, A., & Puniyani, A. (2022). Can ensemble machine learning methods predict stock returns for indian banks using technical indicators? *Journal of Risk and Financial Management, 15*(8) doi:10.3390/jrfm15080350

Raubitzek, S., & Neubauer, T. (2022). An exploratory study on the complexity and machine learning predictability of stock market data. *Entropy, 24*(3) doi:10.3390/e24030332

Sariyer, M., Akil, A., Bulgurcu, F. N., Oge, F. E., & Ganiz, M. C. (2022). Individual stock price prediction by using KAP and twitter sentiments with machine learning for BIST30. Artigo presented at the *16th International Conference on INnovations in Intelligent SysTems and Applications, INISTA 2022*, doi:10.1109/INISTA55318.2022.9894172 Retrieved from www.scopus.com

Shamisavi, M., & Jahanshahi, A. (2022). Forecasting tehran stock exchange trend with time series analysis, fundamental data, and sentiment analysis in news. Artigo presented at the *2022 30th International Conference on Electrical Engineering, ICEE 2022*, 1-7. doi:10.1109/ICEE55646.2022.9827232 Retrieved from www.scopus.com

Sheng, Y., & Ma, D. (2022). Stock index Spot–Futures arbitrage prediction using machine learning models. *Entropy, 24*(10) doi:10.3390/e24101462

Sidhu, A. S., Misra, N., Kaushik, V., Shankar, A., Joshi, K., & Singh, R. (2022). Analysis of global finance using web scraping and topic modeling. Artigo presented at the *Proceedings of 3rd International Conference on Intelligent Engineering and Management, ICIEM 2022*, 747-753. doi:10.1109/ICIEM54221.2022.9853165 Retrieved from www.scopus.com

Tao, M., Gao, S., Mao, D., & Huang, H. (2022). Knowledge graph and deep learning combined with a stock price prediction network focusing on related stocks and mutation points. *Journal of King Saud University - Computer and Information Sciences, 34*(7), 4322-4334. doi:10.1016/j.jksuci.2022.05.014

Zaffar, A., & Hussain, S. M. A. (2022). Modeling and prediction of KSE – 100 index closing based on news sentiments: An applications of machine learning model and ARMA (p, q) model. *Multimedia Tools and Applications, 81*(23), 33311-33333. doi:10.1007/s11042-022-13052-2

Zaznov, I., Kunkel, J., Dufour, A., & Badii, A. (2022). Predicting stock price changes based on the limit order book: A survey. *Mathematics, 10*(8) doi:10.3390/math10081234

Khan, W., Ghazanfar, M. A., Azam, M. A., Karami, A., Alyoubi, K. H., & Alfakeeh, A. S. (2022). Stock market prediction using machine learning classifiers and social media, news. *Journal of Ambient Intelligence and Humanized Computing, 13*(7), 3433-3456. doi:10.1007/s12652-020-01839-w

Anexo A

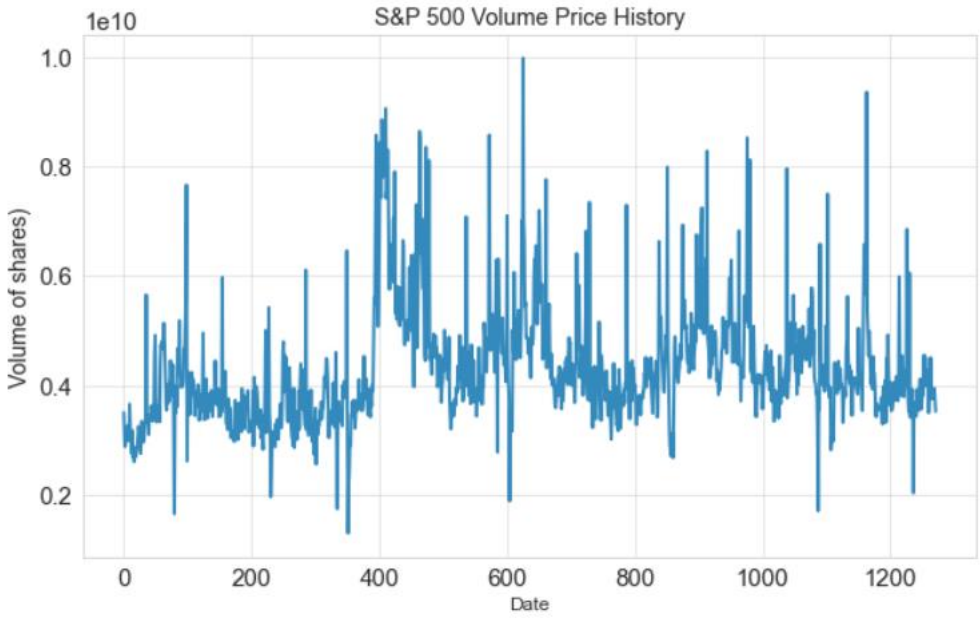

Figura 10 - Dataset para os valores de volume do índice S&P 500

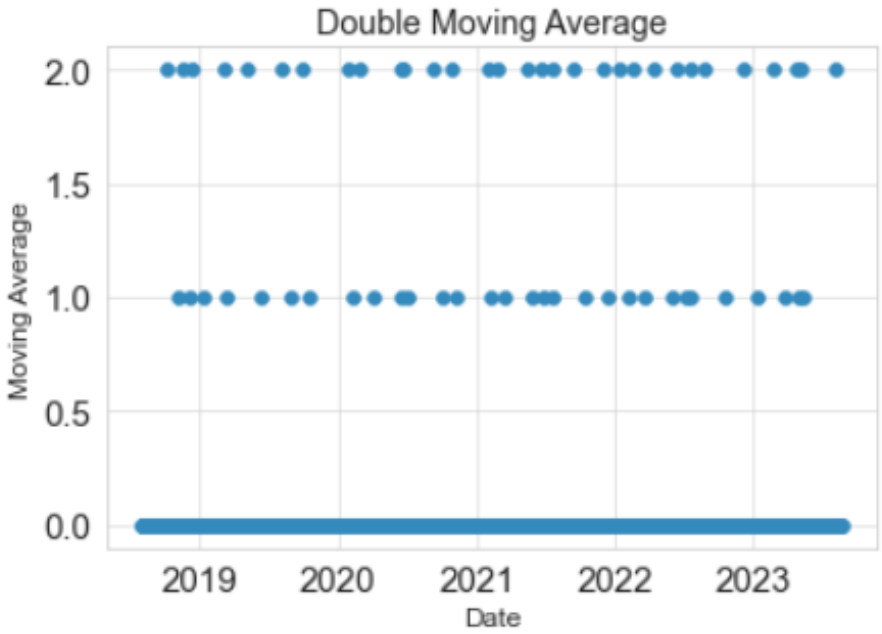

Figura 11 - Dataset para os valores das médias móveis duplas do índice S&P 500

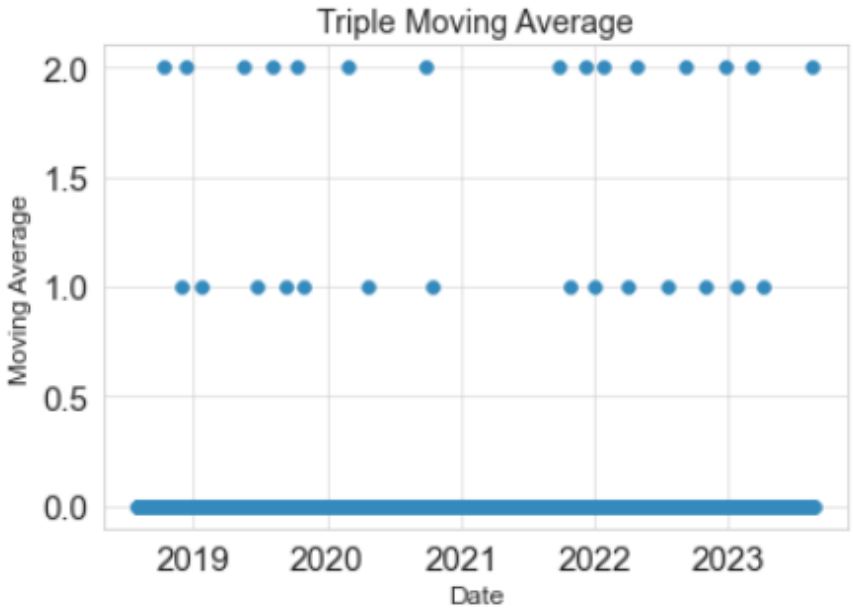

Figura 12 - Dataset para os valores das médias móveis triplas do índice S&P 500

Anexo B

Figura 13 - Matriz de correlação do dataset Dow Jones

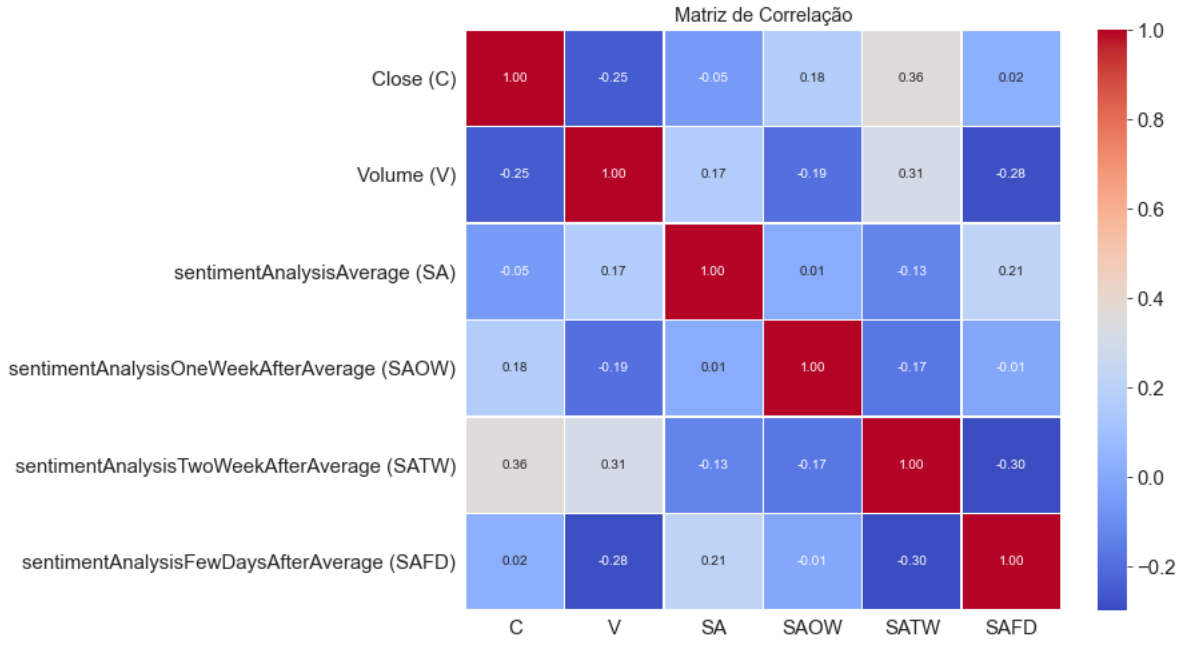

Figura 14 - Matriz de correlação do dataset Russell

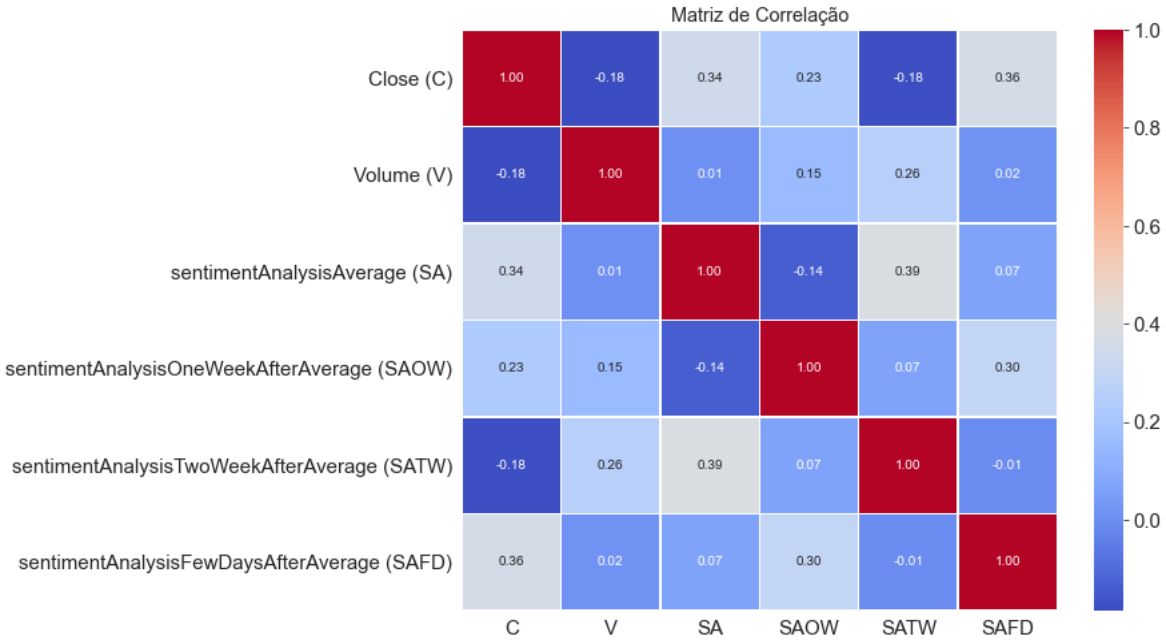

Figura 15 - Matriz de correlação do dataset Gold