Programa de Pós Graduação em Instrumentação, Controle e Automação de Processos de Mineração - PROFICAM
Universidade Federal de Ouro Preto - Escola de Minas
Associação Instituto Tecnológico Vale - ITV

Dissertação

SISTEMA HÍBRIDO COM AGREGAÇÃO DE ANÁLISE DE SENTIMENTOS E SÉRIES TEMPORAIS NEBULOSAS PARA PREVISÃO DE PREÇOS DE MINÉRIO DE FERRO

Flavio Mauricio da Cunha Souza

Ouro Preto
Outubro de 2023

Flavio Mauricio da Cunha Souza

SISTEMA HÍBRIDO COM AGREGAÇÃO DE ANÁLISE DE SENTIMENTOS E SÉRIES TEMPORAIS NEBULOSAS PARA PREVISÃO DE PREÇOS DE MINÉRIO DE FERRO

Dissertação apresentada ao Programa de Pós-Graduação em Instrumentação, Controle e Automação de Processos de Mineração da Universidade Federal de Ouro Preto e do Instituto Tecnológico Vale, como parte dos requisitos para obtenção do título de Mestre em Engenharia de Controle e Automação.

Linha de Pesquisa: Tecnologia da Informação, Comunicação e Automação Industrial

Orientador: Prof. D.Sc. Gustavo Pessin
Coorientador: Prof. D.Sc. Geraldo Pereira Rocha Filho
Coorientador: Prof. D.Sc. Frederico Gadelha Guimarães

Ouro Preto, MG – Brasil
Outubro de 2023

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

S729s Souza, Flavio Mauricio da Cunha.
Sistema híbrido com agregação de análise de sentimentos e séries temporais nebulosas para previsão de preços de minério de ferro. [manuscrito] / Flavio Mauricio da Cunha Souza. - 2023.
39 f.: il.: color., gráf., tab..

Orientador: Dr. Gustavo Pessin.
Coorientadores: Prof. Dr. Frederico Gadelha Guimarães, Prof. Dr. Geraldo Pereira Rocha Filho.
Dissertação (Mestrado Profissional). Universidade Federal de Ouro Preto. Programa de Mestrado Profissional em Instrumentação, Controle e Automação de Processos de Mineração. Programa de Pós-Graduação em Instrumentação, Controle e Automação de Processos de Mineração.
Área de Concentração: Engenharia de Controle e Automação de Processos Minerais.

1. Aprendizado do computador. 2. Análise de séries temporais. 3. Processamento de linguagem natural (Computação). 4. Minérios de ferro. I. Pessin, Gustavo. II. Guimarães, Frederico Gadelha. III. Rocha Filho, Geraldo Pereira. IV. Universidade Federal de Ouro Preto. V. Título.

CDU 681.5:622.2

Bibliotecário(a) Responsável: Maristela Sanches Lima Mesquita - CRB-1716

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
ESCOLA DE MINAS
PROGR. POS GRAD. PROF. INST. CONT. E AUT.
PROCESSOS DE MIN.

FOLHA DE APROVAÇÃO

Flávio Mauricio da Cunha Souza

Sistema híbrido com agregação de análise de sentimentos e séries temporais nebulosas para previsão de preços de minério de ferro

Dissertação apresentada ao Programa de Pós-Graduação em Instrumentação, Controle e Automação de Processos de Mineração (PROFICAM), Convênio Universidade Federal de Ouro Preto/Associação Instituto Tecnológico Vale - UFOP/ITV, como requisito parcial para obtenção do título de Mestre em Engenharia de Controle e Automação na área de concentração em Instrumentação, Controle e Automação de Processos de Mineração

Aprovada em 04 de outubro de 2023

Membros da banca
Doutor - Gustavo Pessin - Orientador - Instituto Tecnológico Vale
Doutor - Geraldo Pereira Rocha Filho - Coorientador - Universidade Estadual do Sudoeste da Bahia
Doutor - Frederico Gadelha Guimarães - Coorientador - Universidade Federal de Minas Gerais
Doutora - Andrea Gomes Campos Bianchi - Universidade Federal de Ouro Preto
Doutor - Alan Demétrius Baria Valejo - Universidade Federal de São Carlos

Gustavo Pessin, orientador do trabalho, aprovou a versão final e autorizou seu depósito no Repositório Institucional da UFOP em 01/12/2023

Documento assinado eletronicamente por **Bruno Nazário Coelho**, **COORDENADOR(A) DE CURSO DE PÓS-GRADUAÇÃO EM INST. CONTROLE AUTOMAÇÃO DE PROCESSOS DE MINERAÇÃO**, em 15/12/2023, às 10:11, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0642645** e o código CRC **078B6166**.

Referência: Caso responda este documento, indicar expressamente o Processo nº 23109.016906/2023-38 SEI nº 0642645

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: (31)3552-7352 - www.ufop.br

"O coração do homem traça o seu caminho, mas o Senhor lhe dirige os passos."
Provérbios 16:9

Agradecimentos

Agradeço, primeiramente, à Deus pela vida e vitalidade para desenvolver este trabalho.

Agradeço aos meus pais, Maurício e Adriana pela minha criação e formação de valores que levo por toda a minha vida.

Dedico este trabalho à minha esposa, Fabiana, e minha filha Sophia, fontes de amor diário e motivação para sempre seguir adiante.

Ao meu enteado, Thiago, para que eu continue a lhe ser uma boa referência.

Agradeço aos meus sogros, Aloízio e Lú, que sem o suporte diário dado por eles seria impossível desenvolver e concluir este trabalho.

Ao meu orientador, Prof. Dr. Gustavo Pessin, e meus coorientadores, Prof. Dr. Geraldo P. R. Filho e Frederico G. Guimarães, pelas contribuições e conhecimentos transmitidos.

Aos meus amigos do CBMMG, literalmente "Amigos certos nas horas incertas", aos quais confio a minha vida a cada serviço. Aos amigos e colegas da UFMG, UFV, UFOP e ITV, que fizeram parte de toda a minha jornada acadêmica até então.

À UFOP e ao ITV, que desenvolvem o PROFICAM, formando mestres que certamente contribuirão muito à sociedade com os conhecimentos adquiridos.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001, do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ), do Instituto Tecnológico Vale (ITV) e da Universidade Federal de Ouro Preto (UFOP).

Resumo da Dissertação apresentada à Escola de Minas/UFOP e ao ITV como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

SISTEMA HÍBRIDO COM AGREGAÇÃO DE ANÁLISE DE SENTIMENTOS E SÉRIES TEMPORAIS NEBULOSAS PARA PREVISÃO DE PREÇOS DE MINÉRIO DE FERRO

Flavio Mauricio da Cunha Souza

Outubro/2023

Orientadores: Gustavo Pessin
Geraldo Pereira Rocha Filho
Frederico Gadelha Guimarães

O preço global do minério de ferro é determinado por um número elevado de parâmetros efetivos e uma relação complexa entre eles. A soma das expectativas dos participantes deste mercado como um todo, ao longo do tempo, definem variações e tendências numa série temporal de preços. Desenvolver um modelo de previsão confiável para a volatilidade do preço do minério de ferro e, por consequência, demais ativos ligados à esta *commodity*, que analise o mercado de forma ampla, não é uma tarefa trivial e é fundamental na definição de investimentos futuros e decisões para projetos de mineração em empresas relacionadas. Este trabalho avalia um sistema preditivo híbrido, que utiliza um índice obtido a partir da agregação de sentimentos extraídos de resumos de notícias relacionadas ao minério de ferro, baseado em conjuntos nebulosos hesitantes, e o número de notícias como variáveis exógenas para um modelo multivariado Weighted Multivariate Fuzzy Time Series (WMVFTS). Neste contexto, a aplicação do Índice de Agregação de Análise de Sentimentos com Conjuntos Nebulosos Hesitantes para Previsão de Preços de Minério de Ferro combina métodos de aprendizado de máquina que abrangem tanto análises técnicas quanto fundamentais, obtendo resultados significativos para suporte à decisão especializada em ativos de minério de ferro. Os resultados indicam a viabilidade de utilização das variáveis propostas em um conjunto de variáveis exógenas do WMVFTS com precisão superior a 80% na previsão de tendências e oscilações da variável de referência.

Palavras-chave: Aprendizado de Máquina, Séries Temporais, Processamento de Linguagem Natural, Minério de Ferro.

Macrotema: Logística.

Tema: Redução de Incerteza no Planejamento da Mina.

Linha de Pesquisa: Tecnologias da Informação, Comunicações e Automação Industrial.

Abstract of Dissertation presented to Escola de Minas/UFOP and ITV as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

HYBRID SYSTEM WITH SENTIMENT ANALYSIS AGGREGATION AND FUZZY TIME SERIES FOR IRON ORE PRICE FORECASTING

Flavio Mauricio da Cunha Souza

October/2023

Advisors: Gustavo Pessin
Geraldo Pereira Rocha Filho
Frederico Gadelha Guimarães

The global price of iron ore is determined by a large number of effective parameters and a complex relationship between them. The sum of the expectations of the participants of this market as a whole, over time, defines variations and trends in a time series of prices. Developing a reliable forecast model for the volatility of iron ore prices and, consequently, other assets linked to this commodity, which analyzes the market in a broad way, is fundamental in defining future investments and decisions for mining projects in companies related. This work evaluates a hybrid predictive system, which uses an index obtained from the aggregation of sentiments extracted from news related to iron ore, based on hesitant fuzzy sets, and the number of news as exogenous variables for a multivariate model Weighted Multivariate Fuzzy Time Series (WMVFTS). In this context, the application of the Aggregation Index of Sentiment Analysis with Hesitant Fuzzy Sets for Iron Ore Price Forecasting combines machine learning methods that encompass both technical and fundamental analysis, obtaining significant results for expert decision support specialist in iron ore assets. The results indicate the feasibility of using the proposed variables in a set of exogenous WMVFTS variables with an accuracy greater than 80% in predicting trends and oscillations of the reference variable.

Keywords: Machine Learning, Time Series, Natural Language Processing, Iron Ore.
Macrotheme: Logistics.
Theme: Reducing Uncertainty in Mine Planning.
Research Line: Information Technologies, Communications and Industrial Automation.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

BERT Bidirectional Encoder Representations from Transformers

CGOA Chaotic Grasshopper Optmization Algotithm

EEMD Ensemble Empirical Mode Decomposition

FDT Fuzzy Decision Trees

FTS Fuzzy Time Series

GOA Grasshopper Optmization Algotithm

GORU Gated Orthogonal Recurrent Unit

HFS Hesitant Fuzzy Sets

HFWA Hesitant Fuzzy Weighted Averageing

IQR InterQuartile Range

LSTM Long Short-Term Memory

MAPE Mean Absolute Percentage Error

MDA Mean Directional Accuracy

MLM Masked Language Modeling

MLP MultiLayer Perceptron

NLP Natural Language Processing

NSP Next Sentence Prediction

PWFTS Probabilistic Weighted Fuzzy Time Series

RDEU Rank Dependent Expected Utility

RMSE Root Mean Square Error

RNA Redes Neurais Artificiais

RNN Recurrent Neural Network

WMVFTS Weighted Multivariate Fuzzy Time Series

Sumário

Capítulo 1

Introdução

1.1 Motivação

O ferro é um dos metais mais aplicáveis no mundo (MENDES, 2000). O preço global do minério de ferro é determinado com base na demanda e oferta, sendo esse aspecto afetado por vários parâmetros quantitativos tais como preço do aço, produção, preço do petróleo, preço do ouro, taxa de juros, taxa de inflação, produção de ferro e preço do alumínio (LI *et al.*, 2020). Como o minério de ferro é o recurso fundamental para a produção do aço, prever seu preço é estrategicamente importante para a gestão de risco em empreendimentos e projetos relacionados (TUO; ZHANG, 2020). As projeções dos preços do minério de ferro, para empresas mineradoras como a VALE, podem ajudar na definição da viabilidade operacional de determinadas minas e controle do excedente produtivo.

O minério de ferro não é comercializado tal como outras commodities, sendo normalmente cotado através de duas fontes. Numa delas, chamada Índice de Referência, o preço geralmente é definido uma vez por dia por várias grandes empresas de referência. A transparência é um problema potencial, pois as negociações contratuais podem ocorrer a portas fechadas, limitando a previsibilidade de preços e quantidades do mercado. A outra fonte de cotação é o Mercado Futuro, onde a negociação pode ocorrer quase 24 horas por dia e todas as transações são transparentes, ou seja, a quantidade de contratos de compra e venda, os potenciais compradores e vendedores, assim como os preços que cada entidade está disposta a pagar ou vender seus contratos são apresentados de forma explícita enquanto o mercado estiver aberto (INDEX, 2022).

O processo de digitalização aplicado na governança das empresas permitiu o acesso à uma quantidade enorme de dados quantitativos e de natureza subjetiva dos mais variados mercados. Nos dias atuais, com o aumento da capacidade computacional de processamento desses dados, ferramentas que utilizam métodos de inteligência artificial têm se tornado cada vez mais viáveis e eficientes em problemas de predição (ARIAS; ARRATIA; XURIGUERA, 2014).

Aproveitando a abundância de dados disponíveis, nos estudos sobre preços futuros do minério de ferro, métodos híbridos de previsão apresentam bons resultados e se mostram promissores. Os métodos propostos por Li *et al.* (2020) e Ewees *et al.* (2020) trabalharam unindo técnicas de otimização com redes neurais artificiais. Tuo & Zhang (2020) propuseram um modelo híbrido baseado numa tecnologia de decomposição de sinal e uma rede neural artificial.

Porém, os métodos citados, por trabalharem com dados para análise técnica, tiveram dificuldades na previsão de mudanças no mercado de minério de ferro motivadas por questões essencialmente fundamentalistas. Os impactos de uma pandemia como a da COVID-19, analisados por Jowitt (2020), e de políticas de redução de emissões de carbono, relatados por Ma & Wang (2021), no mercado de minério de ferro são exemplos da grande influência de variáveis de caráter fundamentalista sobre os preços.

Li *et al.* (2021) combinaram Rank Dependent Expected Utility (RDEU) com teoria de jogos para analisar como mineradoras e siderúrgicas se comportam nas situações de conflito comercial. Este trabalho conclui que variáveis subjetivas, como expectativas de crescimento ou retração da atividade siderúrgica ou da capacidade de produção e disponibilidade de excedente produtivo das mineradoras, podem ser consideradas e influem de forma abrangente nas escolhas estratégicas corporativas e posicionamento dessas empresas diante do mercado. A análise de mudanças nos dados de posicionamento de atores do mercado e suas expectativas em relação a uma cadeia produtiva, permitem definir respostas comportamentais, isolar tipos específicos de padrões de fluxo e desenvolver ferramentas e sinais para melhor negociação, investimento e gerenciamento de risco (KEENAN, 2019).

Recentemente, métodos de inteligência artificial que utilizam Natural Language Processing (NLP), como Análise de Sentimentos ou Mineração de Opinião, estão alcançando bons resultados preditivos no mercado de ações com base em análises fundamentalistas, se destacando trabalhos como os de Alves (2015) e Igarashi, Valdevieso & Igarashi (2020).

1.2 Justificativa

Nesse contexto, considerando as peculiaridades do mercado de minério de ferro e do estado da arte de ferramentas preditivas de ativos similares anteriormente citados, a criação de um sistema especialista de suporte a decisão sobre ativos de minério de ferro que possa, de maneira confiável, prever o valor do minério de ferro através de uma série temporal, analisar os riscos relacionados à este ativo através de informações textuais no período da série utilizada e utilizar estes produtos para indicar a melhor opção para o gerenciamento dos recursos, seria de grande utilidade para empresas e gestores deste mercado.

A associação dos métodos de análise de sentimentos de notícias e agregação de índices com base em conjuntos nebulosos hesitantes indica a viabilidade da construção de um índice que quantifique variáveis subjetivas do mercado de minério de ferro. Além disso,

este índice pode ser utilizado como uma das variáveis de entrada em um modelo preditivo de séries temporais nebulosas, melhorando os resultados das previsões do preço do minério de ferro em relação à métodos já apresentados na literatura para este fim.

1.3 Objetivos

1.3.1 Objetivo geral

Esta pesquisa tem por objetivo estudar a aplicação de variáveis alternativas - obtidas através de notícias relacionadas ao assunto - como entradas de modelos preditivos nebulosos multivariados para estimar o comportamento futuro da série temporal de preços médios do minério de ferro refinado 62%. Para isto, um índice será construído a partir da agregação nebulosa hesitante de sentimentos extraídos de notícias referentes ao minério de ferro, por análise de sentimentos. Além deste índice, a quantidade de notícias também será avaliada como variável exógena de um modelo preditivo que séries temporais nebulosas no intuito de aumentar a robustez, em relação aos métodos já presentes na literatura para a previsão dos preços do minério de ferro.

1.3.2 Objetivos específicos

Especificamente, os objetivos deste projeto são:

a) Revisar a bibliografia referente à predição de preços de minério de ferro 62%, que utilizem diferentes técnicas de Inteligência Artificial.

b) Apurar dados e variáveis que influenciam na variação do preço do minério de ferro para selecionar as fontes de notícias adequadas;

c) Construir um índice, baseado na Análise de Sentimentos das notícias selecionadas, que possa ser utilizado como variável exógena correlacionada à variação do preço do minério de ferro;

d) Verificar, estatisticamente, a relação de causa e efeito entre o índice construído e o preço do minério de ferro;

e) Desenvolver um sistema híbrido de suporte a decisão especialista em ativos de minério de ferro que utilize técnicas de análises de sentimentos e Fuzzy Time Series para a previsão de preços;

f) Verificar os resultados obtidos aplicando métricas de acuracidade que permitam a comparação com outros modelos preditivos já propostos por outros autores.

1.4 Estrutura do trabalho

A estrutura deste trabalho está organizada da seguinte forma: No capítulo 1, é feita uma introdução sobre o mercado de minério de ferro e a importância da previsão dos preços desta commodity para a indústria mineradora, discorrendo sobre a motivação, justificativa e os objetivos do trabalho. O capítulo 2 trata do referencial teórico, abordando o comportamento do mercado de minério de ferro nos dias atuais, os conceitos de Séries Temporais e ferramentas de análise destas, Fuzzy Time Series e sua variação Weighted Multivariate Fuzzy Time Series, Hesitant Fuzzy Sets, Análise de Sentimentos e modelos de aprendizado de reforço profundo. Este capítulo também menciona os trabalhos relacionados à previsão dos preços do mercado de minério de ferro, utilização de NLP na predição de mercado de ações e sistemas autônomos de tomadas de decisões e negociação de ativos no mercado financeiro. O capítulo 3 contém a metodologia utilizada, descrevendo os dados utilizados, os processos de análise e obtenção dos sentimentos, agregação dos sentimentos e previsão de preços do minério de ferro. No capítulo 4 estão os experimentos executados, os resultados alcançados e uma discussão acerca destes. A conclusão de todo o trabalho é apresentada no capítulo 5, trazendo os pontos positivos e negativos desta proposta, sus dificuldades e limitações. Por fim, o capítulo 6 sugere como trabalhos futuros relacionados a este tema podem ser conduzidos para obter melhores resultados.

Capítulo 2

Revisão bibliográfica

2.1 Fundamentação teórica

2.1.1 Mercado de Minério Ferro

O preço do minério de ferro pode afetar significativamente o custo e o preço de venda dos produtos siderúrgicos e influenciar, ainda, outras indústrias, bem como políticas econômicas em diferentes países, através de um efeito cascata (FU, 2018). O minério de ferro é a principal matéria-prima utilizada na produção do aço. Segundo Ewees *et al.* (2020), cerca de 10 a 20 por cento do custo total do aço é atribuído ao minério de ferro, de modo que a identificação dos fatores que afetam o preço do minério de ferro podem ajudar no controle do preço do aço.

O preço de outras commodities, o frete e volume de demanda por produtos e bens de consumo estão correlacionadas ao aumento ou queda do preço do minério de ferro (PUSTOV; MALANICHEV; KHOBOTILOV, 2013).

O minério de ferro tem sofrido grandes variações de preço a partir dos anos 2000. As maiores influências sobre este mercado pode ser enumeradas como:

- A demanda de minério de ferro pela siderurgia, em especial na China;
- A oferta caracterizada pelo excedente produtivo das mineradoras e estoques nos polos siderúrgicos consumidores;
- Os custos operacionais de produção e logística de distribuição do minério de ferro;
- A dinâmica do poder de mercado internacional do minério de ferro exercida pelas indústrias mineradoras e siderúrgicas.

A distribuição global de minério de ferro é extremamente desequilibrada, promovendo a internacionalização e globalização do comércio de minério de ferro (PUSTOV; MALANICHEV; KHOBOTILOV, 2013). No sistema de comércio global, fornecedores ricos em recursos de

minério de ferro, por exemplo, Austrália, Brasil e Índia, podem satisfazer a demanda por minério de ferro doméstico e exportar para outros países em grandes quantidades. No entanto, os países demandantes, por exemplo, China, Japão e Coréia do Sul, precisam importar grandes quantidades devido à oferta interna insuficiente de minério de ferro ou à falta de recursos de minério de ferro de alta qualidade.

Zhu *et al.* (2019) verificaram que existe uma dinâmica do poder de mercado internacional do minério de ferro exercido, principalmente pela Austrália e Brasil, nas exportações para a China e mudanças no regime de preços no mercado internacional de minério de ferro têm um impacto profundo nesta dinâmica. A dinâmica do poder de mercado internacional do minério de ferro gera conflitos comerciais entre empresas siderúrgicas de países importadores e mineradoras de países exportadores. Li *et al.* (2021) modelaram, por meio de teoria de jogos, o conflito comercial entre empresas siderúrgicas chinesas e empresas mineradoras australianas, analisando o impacto dos benefícios corporativos, custos de conflitos, "fatores emocionais" e "fatores assimétricos" nas escolhas estratégicas corporativas.

2.1.2 Aprendizado de Máquinas

Aprendizado de máquinas são métodos computacionais que utilizam a experiência para fazer previsões ou melhorar o desempenho de processos. Neste contexto, experiência refere-se às informações anteriores do sistema já conhecidas, que geralmente estão disponíveis na forma de dados eletrônicos coletados e prontos para análise. Os dados podem ser provenientes de conjuntos de treinamento digitalizados e rotulados por humanos, histórico medição de variáveis de processos ou outras informações obtidas através da interação com o ambiente. Genericamente, são métodos fundamentados na combinação de dados que possuem conceitos de ciência da computação, estatística, probabilidade e otimização (MOHRI; ROSTAMIZADEH; TALWALKAR, 2018).

O aprendizado de máquina tem se mostrado muito útil em diversas aplicações práticas, podendo citar: classificação de textos e documentos, Natural Language Processing (NLP), processamento de voz, visão computacional, detecção de fraudes e problemas de previsão de valores reais.

De acordo com o tipo de dado de treinamento disponível para o aprendizado e a aplicação a ser estudada, o aprendizado de máquina pode ser classificado como aprendizagem supervisionada, aprendizagem não supervisionada e aprendizagem por reforço alpaydin2020introduction.

Na aprendizagem supervisionada, o sistema recebe, como dados para treinamento, um conjunto de exemplos rotulados e faz previsões para um horizonte de eventos futuros. É comum a aplicação deste tipo de aprendizagem a problemas que envolvem regressão e classificação.

Já, na aprendizagem não supervisionada, o sistema recebe um conjunto de exemplos

não rotulados para o treinamento e faz previsões para um horizonte de eventos futuros. Uma vez que nenhum exemplo rotulado está disponível nesse caso, pode ser difícil avaliar quantitativamente o desempenho desse sistema. *Clustering* e redução de dimensionalidade são exemplos de problemas de aprendizagem não supervisionada.

Finalmente, na aprendizagem por reforço, as fases de treinamento e teste também são misturadas. Na coleta das informações, o sistema interage ativamente com o ambiente e pode afeta-lo, recebendo uma recompensa imediata por cada ação. O objetivo deste sistema é maximizar sua recompensa ao longo de uma sequência de ações e iterações com o ambiente. No entanto, nenhuma informação de recompensa de longo prazo é fornecido pelo ambiente, e o sistema deve escolher entre explorar ações desconhecidas para obter mais informações ou utilizar as informações já coletadas.

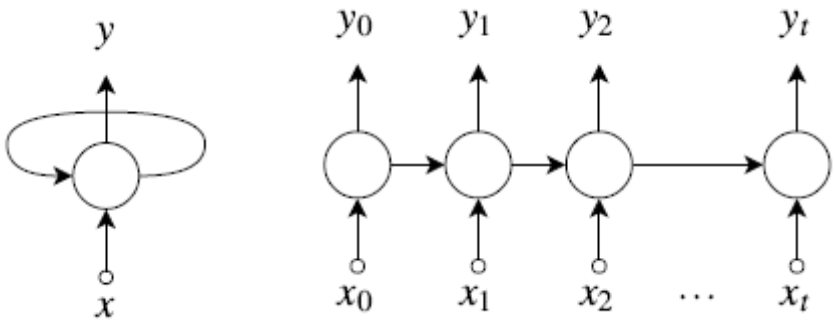
Uma estrutura importante, bastante conhecida, presente numa infinidade de métodos de aprendizado de máquinas, e que exemplifica didaticamente como é possível este tipo de aprendizado são as Redes Neurais Artificiais (RNAs).

Redes Neurais Artificiais

Uma Rede Neural Artificial (RNA) pode ser definida como uma estrutura matemática não-linear capaz de representar arbitrariamente processos não-lineares que relacionam entradas e saídas de um sistema. Em suma, deve ser uma estrutura com potencial de aplicação em situações que requeiram a classificação de padrões, aproximação de funções, aprendizado em áreas de difíceis modelagens, de previsão, e com frequentes mudanças de ambiente (HAYKIN, 2009)

Em problemas que tratam de séries temporais, um tipo de rede neural que tem ganhado campo e sendo bastante utilizada são as Recurrent Neural Network (RNN). Uma RNN possui pelo menos um laço fechado, como mostrada na Figura 2.1a com um único neurônio. Seu modelo equivalente, diretamente alimentado, é obtido pela sequência de entradas de dados de iterações passadas (Fig. 2.1b). Essa arquitetura permite, de maneira mais simplificada, processar dados sequenciais, de forma que os estados das entradas anteriores são propagados, como um tipo de memória. A ideia principal é que uma RNN pode ser alimentada reversamente com dados relevantes ao modelo. A retropropagação dos erros em RNNs é chamada retropropagação através do tempo (ou *backpropagation through time*) (WERBOS, 1990).

Porém, a habilidade de RNNs convencionais trabalharem com informações das entradas passadas é restrita, pois dificilmente consegue treinar relações de longo prazo. Por exemplo, a rede da Figura 2.1b pode facilmente aprender uma dependência entre as entradas x_2 e x_3 da sequência temporal $(x_0, x_1, x_2, \ldots, x_t)$, mas seria muito difícil treiná-la para aprender uma dependência entre as entradas x_2 e x_{12}. Esta dificuldade se dá porque o método do gradiente diminui sua eficiência conforme aumenta o período entre as dependências (BENGIO; SIMARD; FRASCONI, 1994). O gradiente propagado por muitos estágios tende



(a) (b)

Figura 2.1: (a) Rede neural recorrente de um único neurônio; (b) rede neural diretamente alimentada equivalente, obtida pelo desdobramento no tempo.

Fonte: Silva (2019).

a se perder ou divergir (*Vanish Gradient Problem*)(GOODFELLOW; BENGIO; COURVILLE, 2016).

Para evitar esse tipo de problema com o gradiente são usadas arquiteturas especiais de redes recorrentes, como a Long Short-Term Memory (LSTM).

LSTM

LSTM é um tipo especial de RNN proposto por Hochreiter & Schmidhuber (1997) para solucionar o problema do treinamento de dependências de longo prazo. O funcionamento de uma célula lstm é ilustrado na Figura 2.2. O estado da célula $\mathbf{C}_t$ acumula ou descarta ao longo do tempo informações das entradas passadas e atual. Os mecanismos que regulam quais informações serão esquecidas (*forget gate*) e quais serão agregadas (*input gate*) são unidades sigmoides, que recebem como entrada a saída anterior realimentada $\mathbf{h}_{t-1}$ e a entrada atual $\mathbf{x}_t$.

A formulação matemática é dada pelas equações abaixo:

$$\mathbf{f}_t = \sigma(\mathbf{W}_f \cdot \mathbf{x}_t + \mathbf{U}_f \cdot \mathbf{h}_{t-1} + \mathbf{b}_f) \tag{2.1}$$

$$\mathbf{i}_t = \sigma(\mathbf{W}_i \cdot \mathbf{x}_t + \mathbf{U}_i \cdot \mathbf{h}_{t-1} + \mathbf{b}_i) \tag{2.2}$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_o \cdot \mathbf{x}_t + \mathbf{U}_o \cdot \mathbf{h}_{t-1} + \mathbf{b}_o) \tag{2.3}$$

$$\mathbf{C}_t = \mathbf{f}_t \odot \mathbf{C}_t + \mathbf{i}_t \odot \tanh(\mathbf{W}_c \cdot \mathbf{x}_t + \mathbf{U}_c \cdot \mathbf{h}_{t-1} + \mathbf{b}_c) \tag{2.4}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{C}_t) \tag{2.5}$$

onde o operador $\odot$ é o produto de Hadamard (multiplicação elemento à elemento); $\mathbf{W}_f$, $\mathbf{W}_i$, $\mathbf{W}_o$ e $\mathbf{W}_c$ são as matrizes dos pesos que ligam as entradas $\mathbf{x}_t$ respectivamente às unidades sigmoides do *forget gate*, *input gate* e *output gate*, e à unidade tangente hiperbólica da entrada; $\mathbf{U}_f$, $\mathbf{U}_i$, $\mathbf{U}_o$ e $\mathbf{W}_c$ são os pesos ligados à realimentação h_{t-1}; e $\mathbf{b}_f$, $\mathbf{b}_i$, $\mathbf{b}_o$, $\mathbf{b}_c$ os

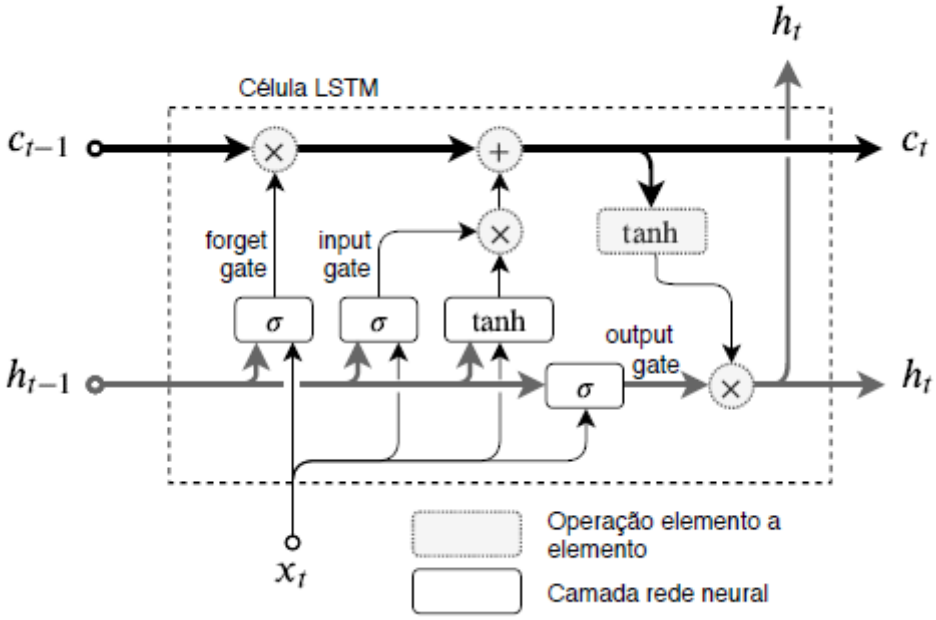


Figura 2.2: Diagrama de uma célula LSTM.

Fonte: Silva (2019).

bias correspondentes.

As LSTM são modelos de aprendizado presentes na estrutura dos mais sofisticados métodos de NLP, como por exemplo o BERT.

2.1.3 Mineração de Séries temporais

Séries temporais são observações de variáveis feitas de forma sequencial ao longo do tempo. Devido a sua estrutura, uma série temporal apresenta várias características como tendência, estacionariedade, sazonalidade e autocorrelação que, através delas, é possível encontrar padrões que podem facilitar previsões (MORETTIN; TOLOI, 2004).

Tendência é uma variação, ascendente ou descendente, que se mantém como um padrão durante um certo período de tempo. Uma série estacionária pode ser identificada quando suas propriedades estatísticas, como a média e a variância, são constantes ao longo do tempo. Uma série temporal é sazonal, ou periódica, quando os fenômenos se repetem a cada período idêntico de tempo, afetando o comportamento da série consistentemente ao longo de sua duração. Por fim, a autocorrelação exprime a influência de valores anteriores de uma série temporal (*lags*) no valor atual desta série (MORETTIN; TOLOI, 2004).

Análise

Em problemas de previsão, a análise preditiva busca por padrões não aleatórios que se repetem e que poderão se repetir no futuro indicados pela presença das características

descritas em 2.1.3.

Após a análise preditiva, para se definir a viabilidade dos métodos a serem utilizados, é necessário determinar um horizonte de previsão, ou seja, classificar se a previsão é de curto, médio ou longo prazo (BEASLEY, 2022). A maioria das decisões tomadas no ambiente de negócios podem ser classificadas conforme a Tabela 2.1:

Tabela 2.1: Classificação de previsões

Classificação	**Tempo**	**Tipo de decisão**	**Exemplos**
Curto prazo	Até 6 meses	Operacional	Controle de inventário; planejamento da produção; distribuição
Médio prazo	De 6 meses a 2 anos	Tática	Locação de instalações; contratação de funcionários
Longo prazo	Mais de 2 anos	Estratégica	Pesquisa e desenvolvimento; aquisições e fusões; mudanças de produto

Fonte: Beasley (2022).

O fundamento da classificação mostrada pela Tabela 2.1 é que diferentes métodos de previsão se aplicam em cada situação a ser estudada. Particularmente, pode-se observar que a quantidade de dados à qual as técnicas quantitativas são aplicadas normalmente varia de muito alta para previsão de curto prazo a muito baixa para previsão de longo prazo quando estamos lidando com situações de negócios (BEASLEY, 2022).

Previsão

Segundo Beasley (2022), os métodos de previsão podem ser classificados em categorias diferentes:

a) Qualitativos - onde não há modelo matemático formal, muitas vezes porque os dados disponíveis não são considerados representativos do futuro (previsão de longo prazo).

b) Regressão - uma extensão da regressão linear em que uma variável é considerada linearmente relacionada a várias outras variáveis independentes.

c) Equações múltiplas - onde há uma série de variáveis dependentes que interagem entre si através de uma série de equações (como nos modelos econômicos).

d) Séries temporais - onde temos uma única variável que muda com o tempo e cujos valores futuros estão relacionados de alguma forma com seus valores passados.

Com o aumento do número de fontes e da disponibilidade de dados históricos de diversos processos sujeitos a previsões, atualmente, os modelos baseados em aprendizado de máquinas, que podem associar métodos de previsão diferentes, se tornaram altamente viáveis e fortes concorrentes aos modelos estatísticos clássicos.

2.1.4 Fuzzy Time Series

As séries temporais nebulosas, ou Fuzzy Time Series (FTS)), propostas por Song & Chissom (1993) são métodos de previsão baseados na teoria dos conjuntos nebulosos (*fuzzy sets*). As FTS têm implementação flexível, capazes de trabalhar com dados numéricos e não numéricos, sendo utilizadas na previsão de variáveis em diversas áreas do conhecimento, como o mercado financeiro, por exemplo (CHENG; CHEN, 2018). Os principais componentes de uma FTS, conforme Lima *et al.* (2019) são:

1. Pré-processamento: Primeira etapa para reduzir a quantidade de dados espúrios ou insignificantes, pode ser aplicado ao conjunto de dados da série temporal Y a ser prevista.

2. Particionamento: Divisão do universo de discurso U em k conjuntos nebulosos, para a criação da variável linguística denominada $\tilde{A}$.

3. Fuzzificação: A partir da variável $\tilde{A}$, é criada a representação linguística F dos dados Y.

4. Extração e Representação de Regras: O modelo de conhecimento $\mathcal{M}$ reconhece padrões de F através da observação de padrões temporais em uma quantidade de defasagens Ω.

Após as regras geradas, a previsão ocorre da seguinte maneira:

1. Pré-processamento: A amostra de entrada $y(t)$ pode sofrer ações de pré-processamento que sejam necessárias.

2. Fuzzificação: A partir da variável $\tilde{A}$, é criada a representação linguística F dos dados Y.

3. Inferência: O modelo $\mathcal{M}$ utiliza elementos de Ω de F para estimar $f(t+1)$.

4. Defuzzificação: $f(t+1)$ toma o valor numérico $\hat{y}(t+1)$

5. Pós-processamento: A saída prevista $\hat{y}(t+1)$ pode sofrer outras transformações de dados de pós-processamento.

As séries temporais multivariadas são matrizes $Y \in \mathbb{R}^n$ onde $n = |\mathcal{V}|$ e $\mathcal{V}$ é o conjunto de atributos de Y. Cada vetor $y(t) \in Y$ contém todos os atributos $\mathcal{V}_i \in \mathcal{V}$ e existe uma dependência temporal entre esses pontos de dados tal que sua ordenação temporal, conforme dado pelo índice de tempo $t \in T$, deve ser respeitada.

Weighted Multivariate FTS

O método *Weighted Multivariate Fuzzy Time Series* (WMVFTS), apresentado por Silva *et al.* (2020), é um previsor de ponto de primeira ordem do tipo Multiple Input/Single Output (MISO), onde para o conjunto de variáveis $\mathcal{V}$, uma delas é escolhida como variável alvo, endógena, e as demais são referidas como variáveis explanatórias exógenas. Diferenciaremos esta variável alvo das demais por um asterisco, como $*\mathcal{V}$.

O procedimento de treinamento é um processo de três estágios responsável por criar um modelo FTS ponderado multivariado $\mathcal{M}$. O modelo WMVFTS final $\mathcal{M}$ consiste em um conjunto de variáveis $\mathcal{V}$, uma variável linguística nebulosa $\tilde{\mathcal{V}}_i$ para cada $\mathcal{V}_i \in \mathcal{V}$ e um conjunto de regras ponderadas nebulosas sobre $\tilde{\mathcal{V}}_i$. As entradas do procedimento de treinamento são os dados de treinamento da série temporal Y e o conjunto de hiperparâmetros para cada $\mathcal{V}_i \in \mathcal{V}$.

O procedimento de previsão visa produzir uma estimativa pontual $\hat{y}(t+1)$ para a variável alvo $*\mathcal{V}$, dada uma amostra de entrada Y, usando as variáveis linguísticas $\tilde{\mathcal{V}}_i$ e as regras difusas induzidas no modelo $\mathcal{M}$.

O método WMVFTS é um modelo determinístico que sempre fornecerá as mesmas previsões com os mesmos dados de entrada e hiperparâmetros, garantindo alta reprodutibilidade dos resultados, ao contrário dos modelos de rede neural. A seleção dos hiperparâmetros k_i e α_i tem um impacto direto na precisão e parcimônia do modelo. O número de partições de cada variável influencia diretamente o número de regras do modelo, pois o número máximo de regras é o produto cartesiano entre os conjuntos nebulosos $A_j^{\mathcal{V}_i} \in \widetilde{\mathcal{V}}_i$ para cada $\mathcal{V}_i \in \mathcal{V}$.

2.1.5 Hesitant Fuzzy Sets

Os *Hesitant Fuzzy Sets* (HFS) são conjuntos nebulosos utilizados em situações em que a incerteza entre diferentes valores dificulta a determinação da pertinência de um elemento a um determinado conjunto, como em problemas de tomada de decisão. Torra (2010). Seja um conjunto de referência X, a definição de um HFS em X em termos de uma função h que, quando aplicada a X, retorna um subconjunto de $[0, 1]$ Torra (2010).

Uma variante de HFS, muito utilizada em problemas que demandam um método de agregação é a *Hesitant Fuzzy Weighted Averageing* (HFWA), definida pela Equação 2.6 Xia & Xu (2011).

$$HFWA = 1 - \prod_{i=1}^{n} (1 - x_i)^{w_i} \tag{2.6}$$

Onde n é o número de elementos no subconjunto de $[0, 1]$ e w_i é o peso de cada elemento x_i, com $i = 1, 2, \ldots, n$.

Torra (2010), verificou que a maior dificuldade em problemas de tomada de decisão é determinar um grau de adesão devido á grande possibilidade de valores, que podem gerar dúvidas e incertezas. Portanto, o HFS é um bom agregador para diferentes sentimentos de notícias relacionadas a um mesmo assunto.

2.1.6 Análise de Sentimentos

Segundo a explicação fornecida por Liu (2012) em sua pesquisa, a análise de sentimento, também conhecida como mineração de opinião, é uma disciplina que se dedica ao estudo das opiniões, sentimentos, avaliações, atitudes e emoções expressas pelas pessoas em relação a diversas entidades, abrangendo produtos, serviços, organizações, indivíduos, questões, eventos, tópicos e seus respectivos atributos.

Empresas e organizações estão sempre interessadas nas opiniões de consumidores ou do público sobre seus produtos e serviços, consumidores também querem analisar as avaliações sobre um produto antes de comprá-lo e as pessoas, em geral, querem saber as opiniões umas das outras sobre candidatos políticos antes de tomar uma decisão de voto numa eleição (LIU, 2012).

As redes sociais online são excelentes fontes de dados para diversas aplicações de análise de sentimentos. Porém, a tarefa de capturar opiniões com o objetivo de observar a dinâmica do pensamento humano em redes sociais é complexa.O número de usuários inscritos e a quantidade de mensagens compartilhadas nesses sistemas aumenta, o volume de dados cresce continuamente e explorar esses dados em tempo hábil, considerando a diversidade e ausência de estrutura formal nas construções dos textos, dificulta a extração de informações úteis (ARIAS; ARRATIA; XURIGUERA, 2014). Os fatores citados tornam o uso de automação imprescindível na análise e obtenção de conhecimento através destas fontes (Alves, 2015).

A utilização de NLP na análise de sentimentos visa o desenvolvimento de técnicas automáticas de extração de informações subjetivas de textos, como opiniões e sentimentos, subsidiando a tomada de decisão (BENEVENUTO; RIBEIRO; ARAÚJO, 2015).

Alessia *et al.* (2015) expõem três abordagens na classificação de sentimentos em fontes textuais: supervisionada, não supervisionada e híbrida.

A abordagem supervisionada usa algoritmos de aprendizado de máquina que requerem, normalmente, um conjunto de dados de treinamento e um conjunto de dados de teste. O conjunto de treinamento aprende características de um texto e o conjunto de teste é usado

para validar este aprendizado.

A abordagem não supervisionada, ou baseada em léxico, usa um dicionário de sentimento com palavras de opinião e as combina com os dados para determinar a polaridade e não necessita de sentenças previamente classificadas para criar o modelo.

Finalmente, na abordagem híbrida, combina as duas anteriores, tendo o potencial de melhorar o desempenho da classificação de sentimentos.

Embeddings

Embeddings são representações que têm por objetivo garantir significado semântico ao texto por meio de vetores densos de tamanho fixo (CARVALHO, 2020). A técnica em questão foi introduzida por Mikolov *et al.* (2013) por meio do algoritmo *word2vec*, que cria representações distribuídas (*embeddings*) ao treinar um modelo neural de forma não supervisionada em um extenso texto corporal. Os *embeddings* são capazes de capturar o significado das palavras com base no contexto em que são usadas. Cada dimensão de um vetor de *embeddings* representa uma característica associada à palavra. A Figura 2.3 representa este comportamento.

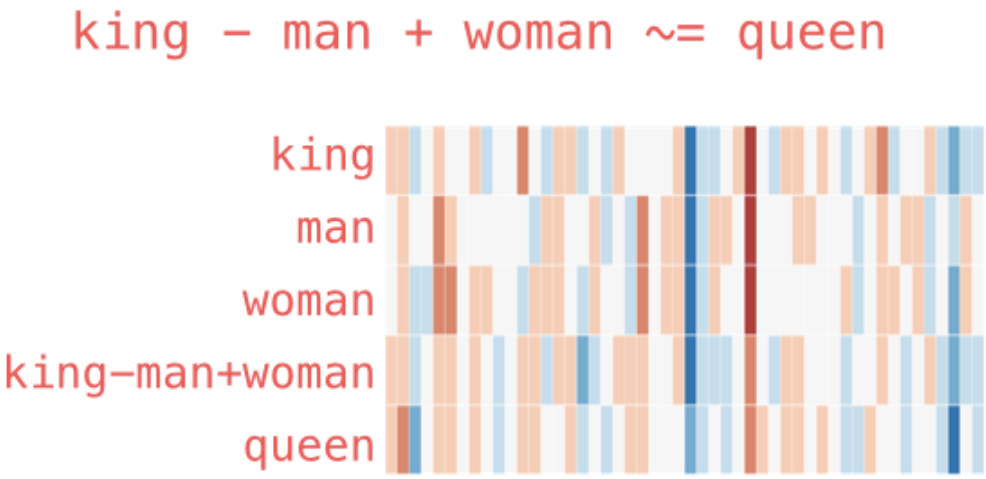


Figura 2.3: Representação de vetores de *embeddings* e analogia das composições de características

Fonte: ⟨https://jalammar.github.io/illustrated-word2vec/⟩

Por serem vetores num espaço N dimensional, os *embeddings* possuem propriedades podem ser obtidas através do cálculo de distancias ou ângulos entre estes vetores. Um meio de se verificar isso é através da similaridade cosseno, que pode ser obtida a partir do cálculo do cosseno do angulo entre dois vetores A e B, conforme Equação 2.7. Vetores com pequena diferença de angulo entre si são correlacionados, enquanto vetores que são ortogonais são não correlacionados.

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}, \tag{2.7}$$

Um exemplo, para visualizar como funcionam os *embeddings*, é de que o vetor que representa a diferença entre rei e rainha, deve ter uma similaridade cosseno alta com o vetor que representa a diferença entre homem e mulher Mikolov *et al.* (2013).

word2vec

O word2vec gera os embeddings tentando prever a probabilidade de uma certa palavra aparecer próxima à outras palavras num texto. Uma técnica chamada *skip-gram with negative sampling* é utilizada para escolher quais serão estas outras palavras dentro do modelo (MIKOLOV *et al.*, 2013). O modelo de *skip-gram with negative sampling* é composto por duas partes importantes. O *skip-gram model* trata uma palavra e uma janela de palavras em volta dela como exemplos positivos para um classificador e usa o restante das palavras do texto como exemplos negativos. Em seguida, um modelo logístico é treinado e os pesos obtidos através deste modelo são usados como *embeddings*, que sumarizam o aprendizado da rede a respeito das palavras usadas no processo de treinamento do modelo. Já a *negative sampling*, escolhe k palavras de maneira aleatória para cada palavra a ser treinada ao invés de serem escolhidas todas as demais palavras do corpos de palavras como exemplos negativos, gerando um exemplo positivo, como pode ser visto na Figura 2.4.

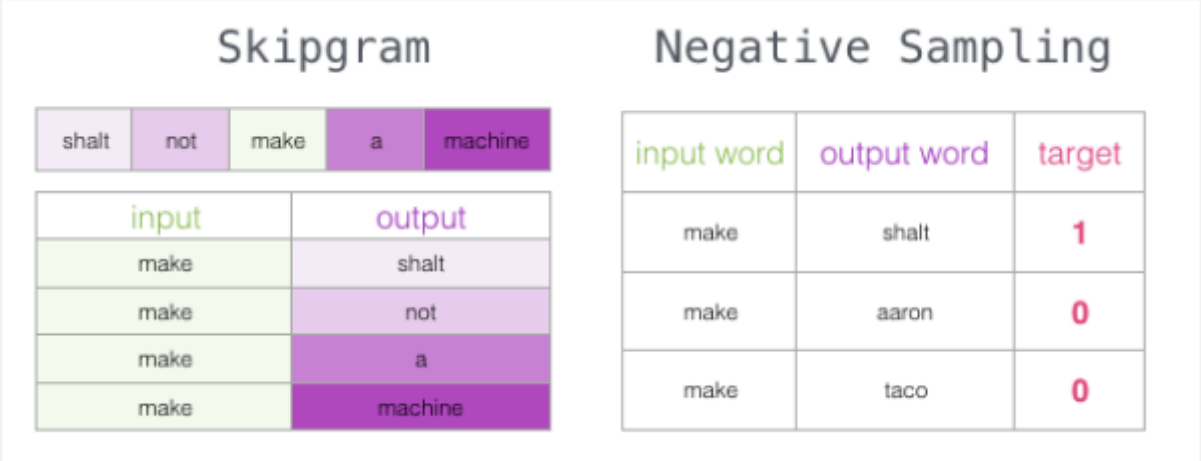


Figura 2.4: Representação da técnica *skip-gram with negative sampling*

Fonte: ⟨https://jalammar.github.io/illustrated-word2vec/⟩

BERT

O Bidirectional Encoder Representations from Transformers (BERT), apresentado por Devlin *et al.* (2018) é uma combinação de *embeddings*, estratégias bidirecionais e

transformers. O *transformer*, representado pela Figura 2.5 é uma arquitetura de sequência a sequência baseada exclusivamente em mecanismos de atenção para *encoders* (VASWANI *et al.*, 2017).

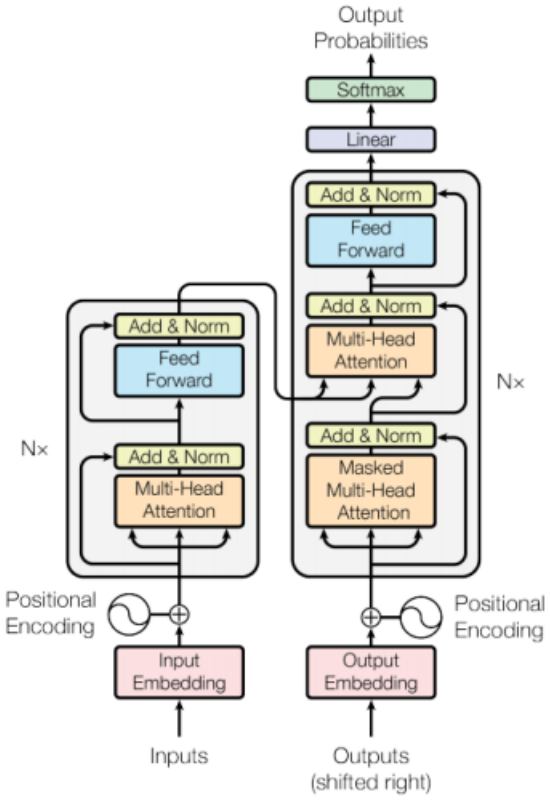


Figura 2.5: Arquitetura *transformer*

Fonte: Vaswani *et al.* (2017)

Um dos diferenciais do BERT é o método de pré-treinamento Masked Language Modeling (MLM). O MLM mascara alguns *tokens* de entrada, substituindo-os por um *token* chamado [MASK]. Em seguida, tenta prever esses *tokens* utilizando o contexto em que ele se encontra. Com isso, o BERT gera um modelo de linguagem robusto bidirecional, ou seja, capaz de enxergar os *tokens* antecedentes e procedentes. Além do MLM, o BERT realiza um procedimento de Next Sentence Prediction (NSP). Neste procedimento, o BERT tenta prever se uma determinada sentença A é seguida de uma determinada sentença B. A Figura 2.6 mostra as tarefas de MLM e de NSP no pré-treinamento do BERT.

Após o pré-treinamento, o BERT passa pela etapa de *fine-tuning* conforme a tarefa e os dados a serem trabalhados. A Figura 2.7 ilustra o funcionamento de algumas tarefas. A tarefa de classificação de texto pode ser observada na Figura 2.7b. Nesta tarefa, o *token* [CLS] representado como C na saída é utilizado como *feature* na classificação da sentença. Isso é feito através da adição de uma camada de classificação contendo uma função *softmax* tendo uma função de custo. Com isso, os parâmetros da rede são ajustados para se adequarem ao problema (DEVLIN *et al.*, 2018).

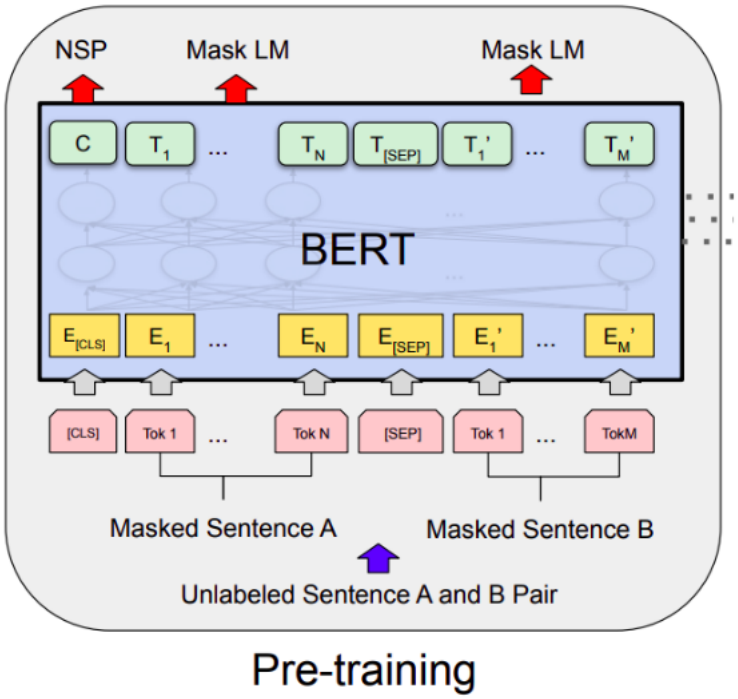


Figura 2.6: Pré-treinamento BERT

Fonte: Devlin *et al.* (2018)

2.2 Trabalhos relacionados

Iniciando pelos métodos que utilizam NLP, Igarashi, Valdevieso & Igarashi (2020) estudaram a influência de notícias e editoriais,de veículos de comunicação especializados no mercado financeiro, na movimentação do mercado. Este estudo comparou a polaridade dos sentimentos das notícias analisadas com a tendência apontada por técnicas de análise como as médias móveis, concluindo que existe uma correlação predominantemente moderada entre os sentimentos e a movimentação do preço das ações. No intuito de agilizar a analise de textos das últimas notícias do mercado financeiro e auxiliar investidores a tomar decisões rápidas, Sousa *et al.* (2019) propuseram o uso de Bidirectional Encoder Representations from Transformers (BERT) para a realizar análise de sentimento dessas notícias, chegando a alcançar 72,5% do F-score.

No sentido da agregação de métodos de inteligência artificial para fundamentar o desenvolvimento de sistemas previsores híbridos no mercado de ações, Dias *et al.* (2021) adotaram valores linguísticos e seus conjuntos hesitantes correspondentes, os tweets publicados pela Bloomberg e os valores de fechamento do Standard & Poor's 500 Index e do Nasdaq Composite Index, para representar preços e sentimentos do mercado de ações americano, empregando a *Weighted Multivariate Fuzzy Time Series* (WMVFTS) como modelo de aprendizado de máquina, conseguindo melhorar resultados do método FTS.

No caso do minério de ferro, a grande maioria dos trabalhos recentes de previsão de

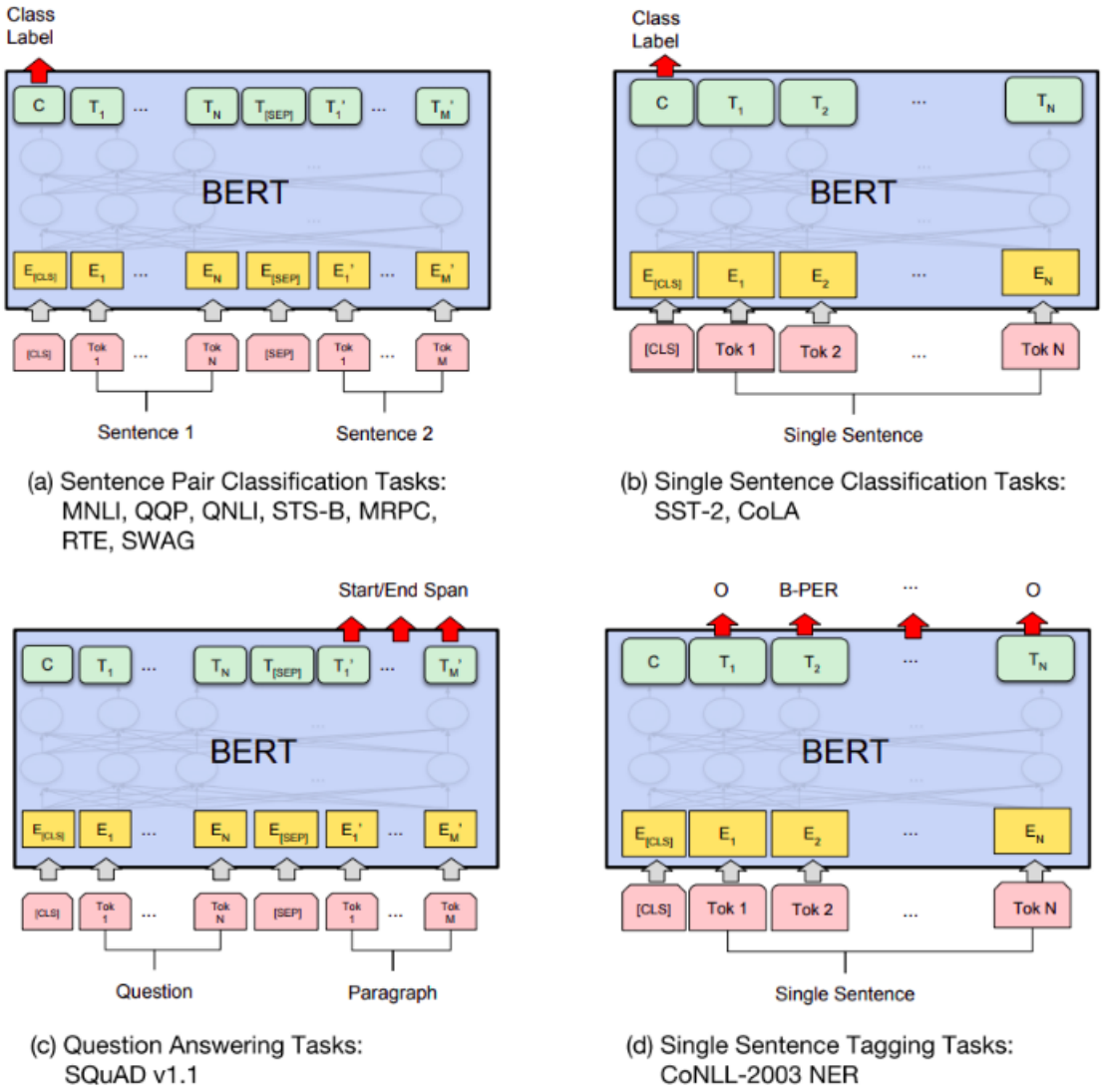


Figura 2.7: *Fine-tuning* BERT

Fonte: Devlin *et al.* (2018)

preço lida com métodos de inteligência artificial alimentados por dados essencialmente quantitativos numa análise técnica. O modelo de Ewees *et al.* (2020) integra o comportamento caótico em um método meta-heurístico recente do *Grasshopper Optmization Algotithm* (GOA) para formar um novo algoritmo chamado *Chaotic Grasshopper Optmization Algotithm* (CGOA), que é usado no treinamento de uma rede *MultiLayer Perceptron* (MLP). Tuo & Zhang (2020) propuseram um modelo híbrido chamado *Ensemble Empirical Mode Decomposition - Gated Orthogonal Recurrent Unit* (EEMD-GORU), baseado numa tecnologia de decomposição de sinal e uma rede neural artificial, e um novo método de reconstrução de dados para explorar o risco de preço e correlações de flutuação entre os futuros de minério de ferro da China e os mercados à vista. Recentemente, Li *et al.* (2021) utilizaram uma combinação entre *Rank Dependent Expected Utility* (RDEU) com o modelo de jogo Falcão-Pombo para analisar o conflito comercial do minério de ferro entre

empresas siderúrgicas chinesas e mineradoras australianas. Em um estudo mais recente, Jr & Guimarães (2022) realizaram uma avaliação da precisão de modelos nebulosos, incluindo os *Probabilistic Weighted Fuzzy Time Series* (PWFTS) e *Fuzzy Decision Trees* (FDT), em comparação com os modelos preditivos ARIMA, Multilayer Perceptron (MLP) e Xgboost, para prever os preços do minério de ferro.

Como contribuição para a inovação científica, a pesquisa a ser desenvolvida nesta peça reúne técnicas utilizadas nos trabalhos acima relacionados para apresentar uma nova variável, que expresse o caráter subjetivo do mercado, seja utilizada para obter previsões mais robustas sobre os preços do minério de ferro. Para tal, são utilizadas NLP para a análise de sentimentos de dados textuais (notícias sobre minério de ferro), agregação baseada em conjuntos nebulosos hesitantes para construir o índice de sentimentos agregados e uma série temporal nebulosa na composição de um sistema híbrido, com diferentes métodos de aprendizagem de máquina, capaz de processar tipos diversos de variáveis de entrada e retornar previsões sobre uma série de preços do minério de ferro mais robusta do que as soluções já apresentadas.

A Tabela 2.2 enumera os principais assuntos abordados pelos trabalhos relacionados e a abrangência desta pesquisa em sua última linha.

Tabela 2.2: Comparativo de Trabalhos Relacionados

Trabalho	NLP	Sistema Hibrido	Minério de Ferro
(IGARASHI et al., 2020)	✓		
(SOUSA et. al., 2019)	✓		
(DOLABELA DIAS et. al., 2021)	✓	✓	
(EWEES et al., 2020)		✓	✓
(TUO and ZHANG, 2020)		✓	✓
(LI et al., 2021)		✓	✓
(TONIDANDEL Jr. et al., 2022)		✓	✓
(SOUZA, 2023)	✓	✓	✓

Fonte: o autor.

Capítulo 3

Metodologia

Neste capítulo, é descrita cada etapa da metodologia para o desenvolvimento de um sistema que utiliza um índice de sentimento agregado em uma FTS para a previsão de preços de minério de ferro. Primeiro, é relatado como se deram a escolha, coleta e processamento de dados. Em seguida, é apresentado o método utilizado para a análise de sentimentos dos tweets. Depois, é descrito o método usado para agregar os sentimentos de diferentes tweets pertencentes a um mesmo período de tempo considerado. Na sequência, é explicado o método WMVFTS para a previsão. Além do desenvolvimento do sistema, para avaliação da sua eficiência, serão procedidos teste de configurações experimentais e a comparação entre o sistema desenvolvido com outras soluções presentes na literatura. A Figura 3.1 ilustra resumidamente a metodologia do trabalho.

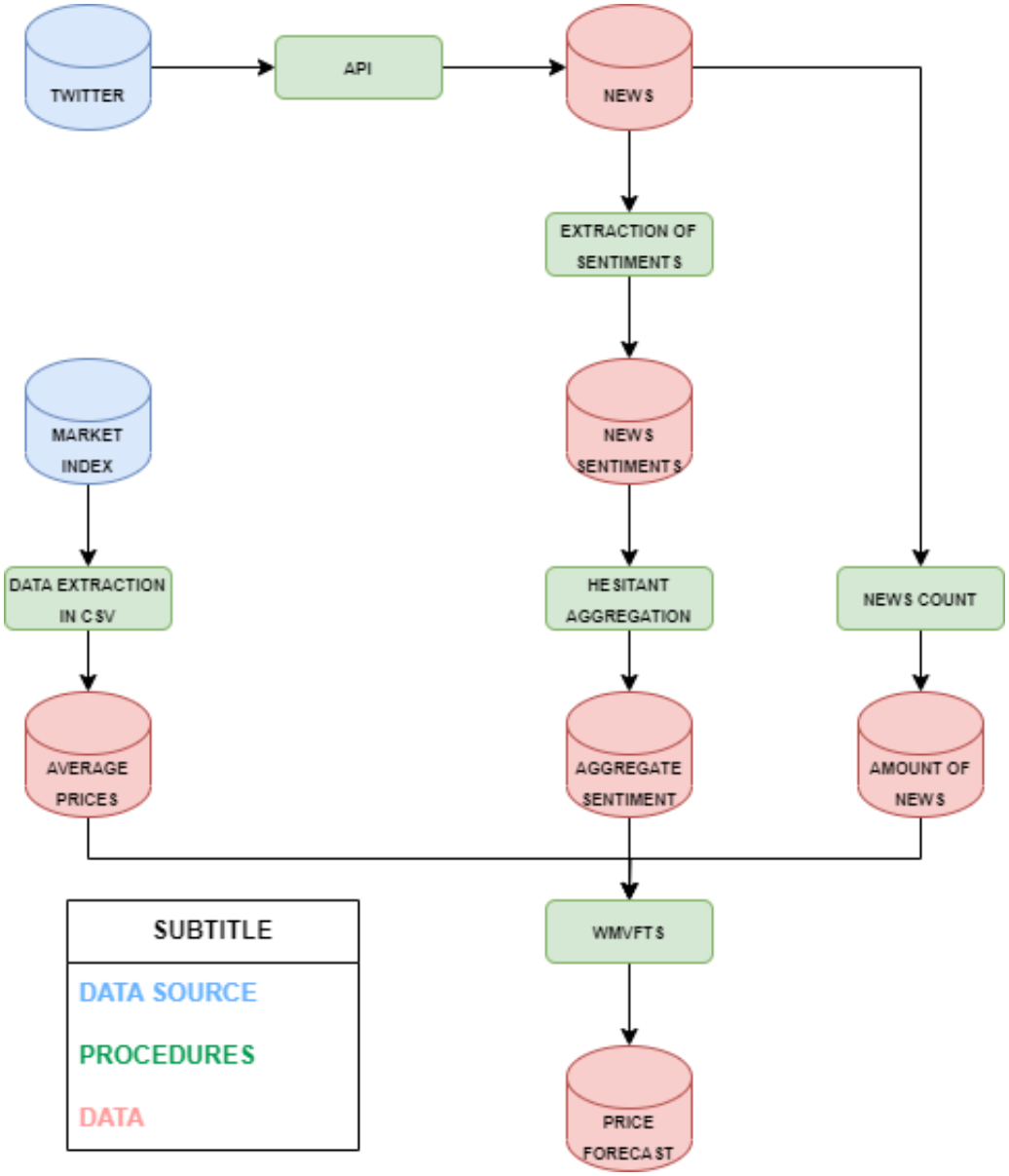


Figura 3.1: Fluxograma da proposta de metodologia

Fonte: O autor.

3.1 Dados

Foram analisados e previsto o comportamento de uma série temporal mensal de preços médios de contratos futuros de minério de ferro refinado com teor de 62% de ferro, exemplificada pela Figura 3.2, considerando este o principal ativo relacionado ao minério de ferro.

Nesta série, foi considerado o período compreendido entre julho de 2015 e janeiro de 2022, uma vez que é o mesmo período estudado por Jr & Guimarães (2022), facilitando a comparação entre seus métodos e o proposto por este trabalho.

Além do conjunto de dados da série principal, foi feita a análise de sentimentos dos

Figura 3.2: Série temporal de preços médios mensais do minério de ferro 62%

Fonte: Index (2022)

tweets postados por Bloomberg[1], referentes ao mercado de minério de ferro ao longo da série temporal considerada. Uma amostra dos tweets a serem analisados pode ser observada na Tabela 3.1.

Tabela 3.1: Amostra dos tweets coletados

Data	**Tweet**
2019-07-25	The mining industry is starting to split on who bears responsibility for all the carbon emissions caused by smeltin
2019-07-25	Anglo American plans to buy back up to billion of shares after the diversified miner reaped bumper profits from
2019-07-20	Vale's second quarter production due next week may offer clues on an end to shortages
2019-07-19	BHP forecasts iron ore production will rise as much as this fiscal year after output slumped to a first annual d
2019-07-12	Forget about oil bonds and tech. This tiny ETF has gained more than so far in July
2019-07-12	The world s largest mining company says it could build more iron ore mines over the next to years in nort

Fonte: o autor.

3.2 Análise preditiva da série de preços de minério de ferro

Tonidandel Junior (2022) confirmou que a série de preços médios de contratos futuros de minério de ferro refinado com teor de 62%, em relação ao período compreendido entre

[1]https://twitter.com/business

julho de 2015 e janeiro de 2022, é não-estacionária evidenciada pelo padrão de tendência. Este fato foi constatado através do teste de raiz unitária de *Dickey-Fuller* aumentado (teste ADF), que retornou *p-valor* igual a 0,93.

Nesta série com periodicidade mensal, observou-se forte autocorrelação com os *lags* vizinhos (até o décimo segundo mês de atraso), com baixo decaimento dos valores de autocorrelação ao longo do tempo. Isto indica que os *lags* vizinhos são altamente significativos para explicar o comportamento da série em um dado instante.

3.3 Obtenção dos sentimentos

A obtenção dos sentimentos será através da aplicação do BERT sobre o conteúdo dos tweets. A utilização do BERT se dá por ser um modelo pré-treinado que pode ser ajustado para uma pontuação de sentimento para cada tweet foi calculada, variando de 0 como a pontuação mais negativa, a 1 como a pontuação mais positiva. A Tabela 3.2 contém a amostra de tweets e seus respectivos sentimentos obtidos pelo BERT.

Tabela 3.2: Tweets e sentimentos analisados pelo BERT

Tweet	**Sentimento BERT**
The mining industry is starting to split on who bears responsibility for all the carbon emissions caused by smeltin	0.30201486
Anglo American plans to buy back up to billion of shares after the diversified miner reaped bumper profits from	0.61859
Vale's second quarter production due next week may offer clues on an end to shortages	0.19472954
BHP forecasts iron ore production will rise as much as this fiscal year after output slumped to a first annual d	0.33379474
Forget about oil bonds and tech. This tiny ETF has gained more than so far in July	0.20920986
The world s largest mining company says it could build more iron ore mines over the next to years in nort	0.35044497

Fonte: o autor.

3.4 Agregação dos sentimentos

A imprecisão no processo de atribuição de sentimentos à textos, devido à própria natureza desses dados, e a dificuldade de se estabelecer o grau de contribuição de cada elemento para o sentimento final, favorecem a escolha do agregador hesitante para lidar com essas características.

Para obter os sentimentos das notícias, primeiramente, os sentimentos dos tweets, calculados pelo BERT, foram agregados diariamente, devido ao aparecimento de diferentes notícias em um mesmo dia, aplicando-se a HFWA proposta por Xia & Xu (2011). Em seguida, esta mesma técnica de agregação foi aplicada para a agregação dos sentimentos no período mensal, de forma que ficassem com a mesma periodicidade dos preços médios mensais do minério de ferro. Tomando a Equação 2.6 como base, quando $w = (1/n, 1/n, \ldots, 1/n)^T$, obtem-se a Equação 3.1 como resultado, que agrega os sentimentos presentes no período determinado e entrega um índice de sentimentos com valor numérico que também varia entre 0 e 1, sendo 0 o valor para o sentimento mais negativo e 1 o valor mais positivo.

$$HFA = 1 - \prod_{i=1}^{n} (1 - x_i)^{\frac{1}{n}} \tag{3.1}$$

Onde n é o número de elementos no subconjunto de $[0, 1]$ e $i = 1, 2, \ldots, n$.

A Tabela 3.3 mostra como ficaria a agregação dos sentimentos de tweets de um mesmo período.

Tabela 3.3: Agregação dos sentimentos

Data	**Tweet**	**Sentimento BERT**	**Sentimento Agregado**
2019-07	The mining industry is starting to split on who bears responsibility for all the carbon emissions caused by smeltin	0.30201486	0.35299
	Anglo American plans to buy back up to billion of shares after the diversified miner reaped bumper profits from	0.61859	
	Vale's second quarter production due next week may offer clues on an end to shortages	0.19472954	
	BHP forecasts iron ore production will rise as much as this fiscal year after output slumped to a first annual	0.33379474	
	Forget about oil bonds and tech. This tiny ETF has gained more than so far in July	0.20920986	
	The world s largest mining company says it could build more iron ore mines over the next to years in nort	0.35044497	

Fonte: o autor.

3.5 Previsão de preços do minério de ferro

Os valores históricos do preço médio mensal do minério de ferro, obtido no Market Index[2], e do índice de sentimentos agregados e a quantidade mensal de tweets foram utilizados para prever o preço do minério de ferro por meio do método WMVFTS, como descrito em 2.1.4. Este procedimento pode ser visualizado pela Figura 3.3.

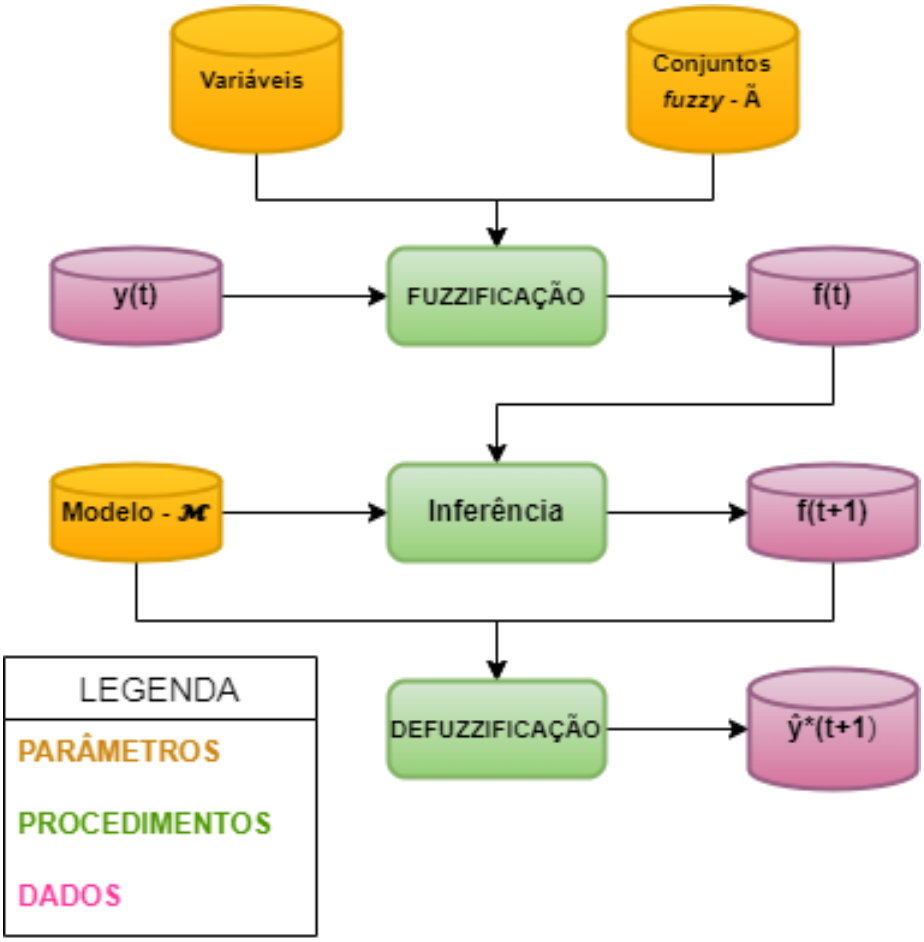


Figura 3.3: Fluxograma do procedimento de previsão

Fonte: O autor.

Neste trabalho, as variáveis exógenas são: a própria série de preços médios mensais do minério de ferro série de sentimentos agregados e a quantidade de tweets postados à cada mês. A variável endógena é a série de preços médios mensais do minério de ferro.

Na WMVFTS cada variável foi particionada em conjuntos nebulosos definidos empiricamente através do processo de otimização de hiperparâmetros, chegando ao resultado ótimo de 40 partições para o preço médio mensal do minério de ferro, 10 para o índice de sentimentos agregados e 15 para a quantidade mensal de tweets.

[2]https://www.marketindex.com.au/iron-ore

3.6 Métricas de desempenho

Após a etapa de previsão, foi feita uma análise comparativa da acurácia deste modelo com os modelos já presentes na literatura para a previsão de preços do minério de ferro. A grande maioria dos trabalhos citados na seção 2.2 utilizam o *Root Mean Square Error* (RMSE) e o *Mean Absolute Percentage Error* (MAPE) como métricas de avaliação de desempenho dos modelos e, por isto, também foram adotados neste para facilitar as comparações.

O RMSE informa o erro médio na mesma unidade que é dado o preço do minério de ferro. O MAPE informa um erro percentual médio dos dados.

Também foi calculada a *Mean Directional Accuracy* (MDA), que descreve a taxa percentual de acerto da tendência, de alta ou queda, da variável predita.

O RMSE e o MAPE são calculados da seguinte forma:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2} \tag{3.2}$$

$$\text{MAPE} = \frac{1}{n}\sum_{j=1}^{n}\frac{|y_j - \hat{y}_j|}{y_j}.100\% \tag{3.3}$$

sendo y_j o j-ésimo valor real da série e $\hat{y}_j$ o j-ésimo valor previsto, considerando um vetor de tamanho n.

Já o MDA é calculado pela equação 3.4.

$$\frac{1}{N}\sum_{t} 1_{\text{sgn}(A_t - A_{t-1}) = \text{sgn}(F_t - A_{t-1})} \tag{3.4}$$

onde A_t, é o valor real no tempo t e F_t, é o valor previsto no tempo t. A variável N representa o número de pontos de previsão. A função sgn$(\cdot)$ é a função de sinal e $\mathbf{1}$ é a função de indicador.

Capítulo 4

Experimentos e resultados

Este capítulo traz o detalhamento do experimento proposto, conforme metodologia descrita no capítulo anterior, e os resultados obtidos. De início, é apresentado um passo-a-passo do experimento, detalhando o pré-processamento dos dados extraídos do Twitter para a análise de sentimentos e os parâmetros utilizados nesta análise. Na sequência, é colocada a forma como foi feita a agregação dos sentimentos extraídos dos tweets que resultou no índice aplicado como entrada no modelo preditor. Em seguida, o modelo preditivo é descrito e suas variações de configurações definidas. Por fim, os resultados são apresentados, discutidos e comparados com outros estudos.

4.1 Experimentos

Os códigos[1] dos algoritmos utilizados neste projeto foram desenvolvidos na linguagem *Python*, aproveitando os diversas bibliotecas e ferramentas dedicadas à inteligência computacional e à previsão de séries temporais disponíveis para tal.

Para a realização dos experimentos com algoritmos de análise de sentimentos, foram coletados, através da API do Twitter, todos os tweets com a *tag "iron ore"* postados por Bloomberg (@business) de 01/07/2015 a 31/01/2022, num total de 504. O pré-processamento aplicado nestes dados foram a remoção de caracteres especiais, links e números da parte textual. Além das informações textuais, foram mantidas e aproveitadas apenas as informações temporais das postagens.

Após a obtenção e agregação dos sentimentos dos tweets, como descritos em 3.3 e 3.4, foram tratados os casos de dados faltantes, observados nos meses em que não houve tweet publicado com os filtros pré-determinados. Nestes casos, considerando que este fato indica uma neutralidade de sentimento, o valor de sentimento agregado inserido foi de 0,5; uma vez que o sentimento mais negativo tem valor 0,0 e o mais positivo 1,0. O resultado deste tratamento foram 79 dados de sentimentos agregados e quantidade de tweets referentes

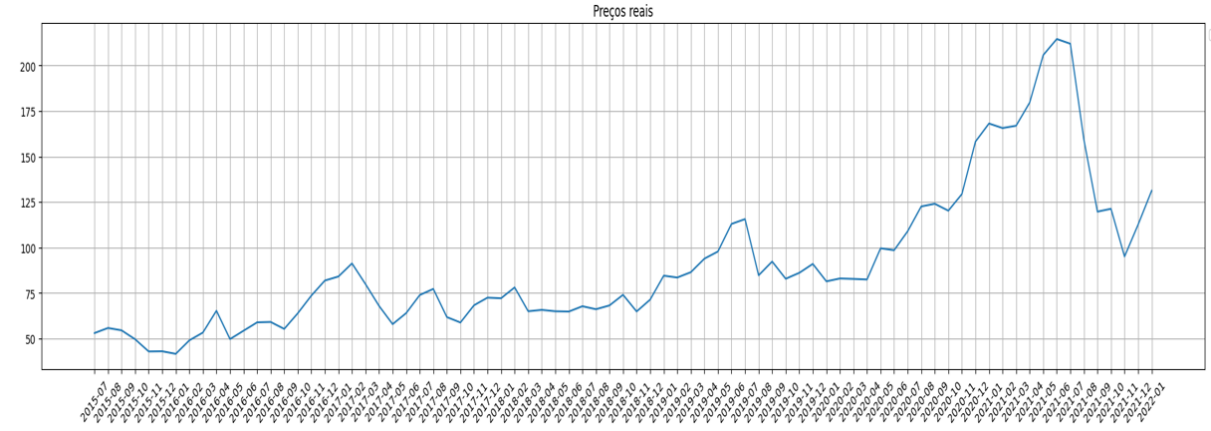
[1]Disponível em https://github.com/flaviomcs/wmvfts_minerio

aos meses compreendidos no período determinado já citado.

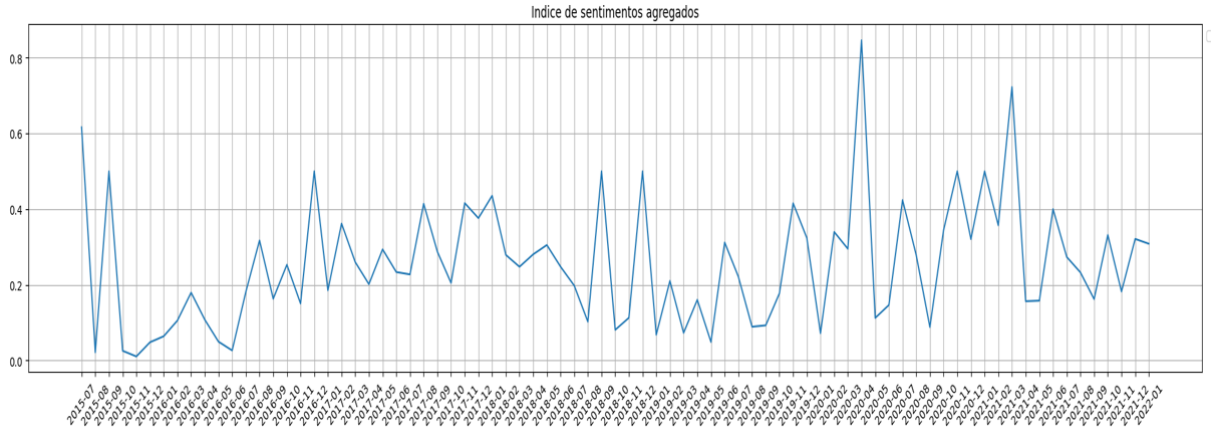
Os experimentos seguiram para a parte preditiva com a modelagem da WMVFTS, usando como variáveis de entrada a série de preços reais do minério de ferro, o índice dos sentimentos agregados mensalmente e a contagem mensal de tweets. Para esta modelagem foi utilizada a biblioteca PyFTS[2].

A série temporal destas variáveis é ilustrada pela Figura 4.1.

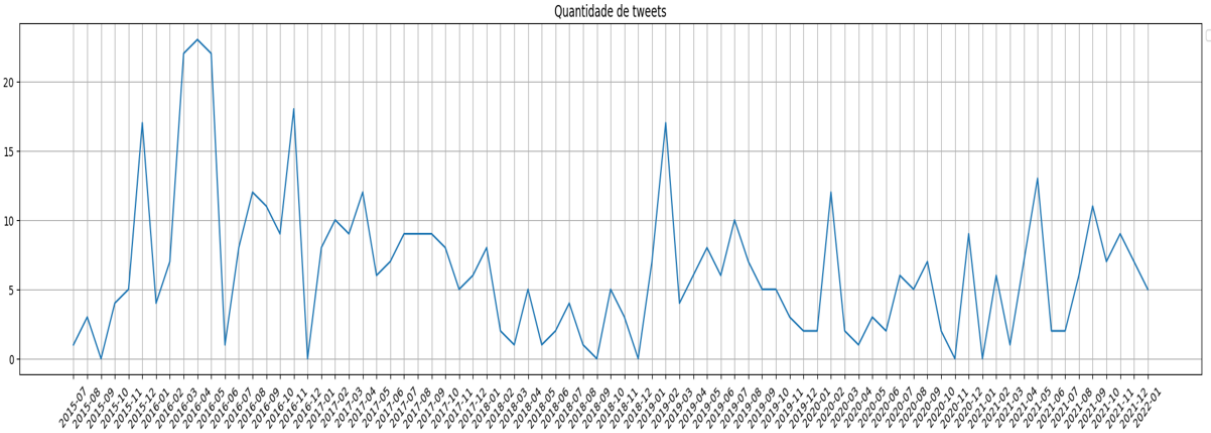
[2]https://github.com/PYFTS



(a) Preços reais do minério de ferro



(b) Sentimentos agregados



(c) Quantidade de tweets

Figura 4.1: Série temporal das variáveis utilizadas nos experimentos

Fonte: O autor.

Foram executadas variações do conjunto destas variáveis de entrada, de modo que foram considerados os resultados para o conjunto que associa as três variáveis (identificado como WMVFTS (P+S+C)), para a associação da série de preços com os sentimentos agregados (identificado como WMVFTS (P+S)) e para a associação da série de preços

com a contagem mensal de tweets (identificado como WMVFTS (P+C)).

4.2 Resultados

Para os três conjuntos de variáveis de entrada, foram utilizados 90% dos dados para treinamento e 10% para testes do modelo proposto, chegando aos resultados apresentados pela Figura 4.2 e Tabela 4.1.

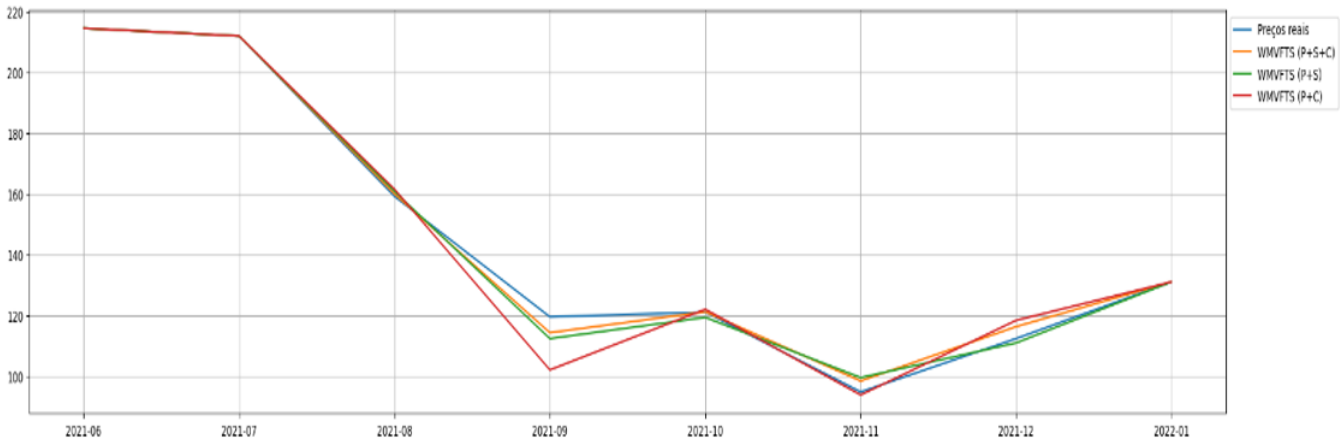


Figura 4.2: Previsões para 7 meses com o conjunto total dos dados

Fonte: o autor.

Tabela 4.1: Resultados estatísticos das previsões para 7 meses com o conjunto total dos dados

Modelo	RMSE	MAPE	MDA
WMVFTS (P+S+C)	28.21	17.86	1.0
WMVFTS (P+S)	28.62	18.36	1.0
WMVFTS (P+C)	29.65	19.98	1.0

Fonte: o autor.

Para validação do modelo, seguindo o método de janelas deslizantes realizado por Jr & Guimarães (2022). Neste método, os dados são divididos em 23 subconjuntos que deslisam um mês subsequente à cada rodada de previsão, sendo os dados de testes correspondentes à previsão dos três últimos meses de cada subconjunto. Os principais resultados deste método de validação são exibidos pela Tabela 4.2.

A precisão do modelo pode ser avaliado a cada janela de previsão através dos valores presentes no conjunto de RMSE e MAPE e da análise da mediana, intervalo interquartílico ou *InterQuartile Range* (IQR) e *outliers* de cada conjunto.

Os resultados obtidos na validação constam nas Tabelas 4.3 e 4.4, com os respectivos *boxplots* exibidos pelas Figuras 4.3 e 4.4.

Tabela 4.2: Resultados estatísticos das previsões para 3 meses com os conjuntos de janelas deslizantes

Modelo	RMSE médio	MAPE médio	MDA médio
WMVFTS (P+S+C)	1.08	0.74	0.96
WMVFTS (P+S)	3.98	2.96	0.8
WMVFTS (P+C)	3.02	2.12	0.96

Fonte: o autor.

Tabela 4.3: Resultados RMSE das previsões para 3 meses com os conjuntos de janelas deslizantes

JANELAS	(P+S+C)	(P+S)	(P+C)
1	3.775928e-01	9.476497e+00	7.422818e+00
2	3.517391e+00	5.782077e+00	8.055296e+00
3	3.497065e+00	3.764110e+00	4.463958e+00
4	3.510000e+00	4.323221e+00	2.233889e+00
5	3.093199e-01	4.463076e+00	2.280640e+00
6	3.093199e-01	4.463076e+00	3.372684e+00
7	7.098718e-02	8.917855e+00	3.150359e+00
8	8.204641e-15	8.480083e+00	2.628264e+00
9	1.834613e-14	1.190689e+01	1.160311e-14
10	1.834613e-14	8.358358e+00	1.160311e-14
11	7.186586e-02	3.961987e+00	1.911513e-01
12	0.000000e+00	1.640928e-14	1.640928e-14
13	0.000000e+00	1.640928e-14	1.640928e-14
14	0.000000e+00	0.000000e+00	0.000000e+00
15	1.640928e-14	1.640928e-14	1.640928e-14
16	2.320623e-14	2.320623e-14	2.772159e-01
17	2.842171e-14	2.320623e-14	2.320623e-14
18	2.842171e-14	1.640928e-14	1.640928e-14
19	4.480830e-01	9.498228e-01	1.367489e+00
20	3.040797e+00	4.239225e+00	1.016279e+01
21	3.041069e+00	4.357888e+00	1.017627e+01
22	3.652402e+00	5.043679e+00	1.010045e+01
23	3.072750e+00	3.008876e+00	3.540239e+00
Mínimo	0.00	0.00	0.00
Mediana	1.53	3.59	3.64
Máximo	3.65	11.91	10.18
IQR	3.04	5.41	4.00

Fonte: o autor.

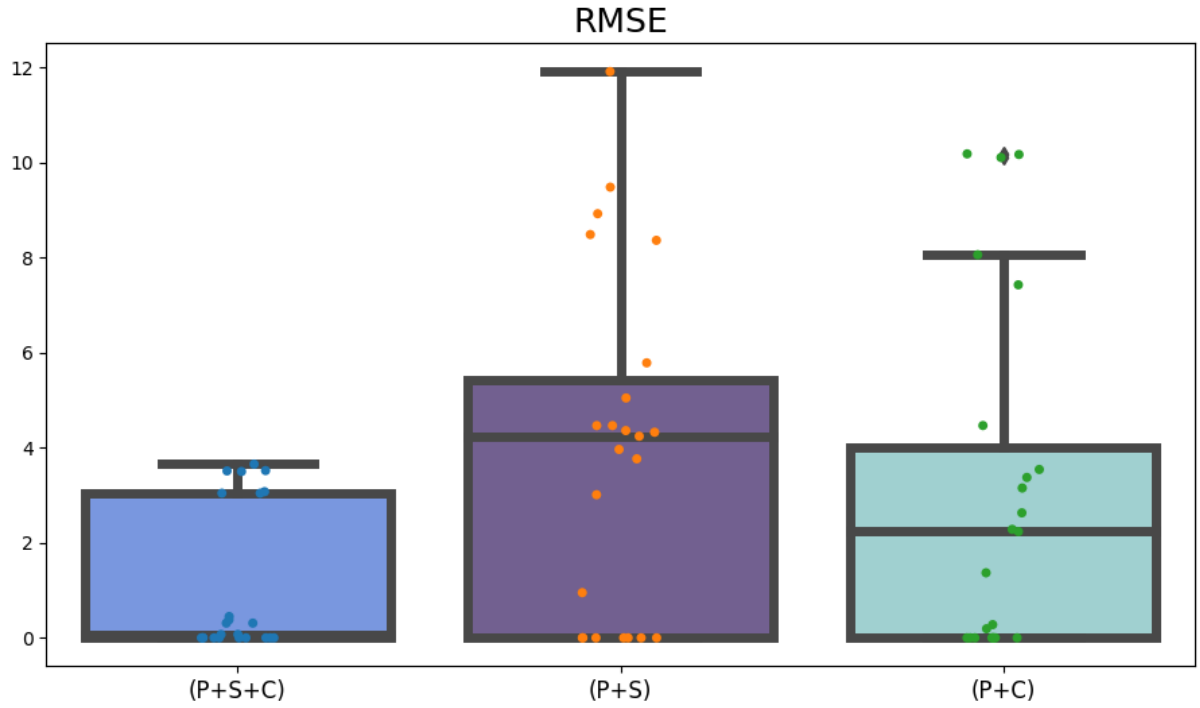


Figura 4.3: *Boxplots* de RMSE dos experimentos

Fonte: o autor.

Tabela 4.4: Resultados MAPE das previsões para 3 meses com os conjuntos de janelas deslizantes

JANELAS	(P+S+C)	(P+S)	(P+C)
1	2.679819e-01	1.072199e+01	5.487901e+00
2	2.709669e+00	6.821579e+00	7.703170e+00
3	2.441688e+00	3.706609e+00	4.281940e+00
4	2.616380e+00	3.799871e+00	2.468884e+00
5	2.163226e-01	3.653685e+00	2.317934e+00
6	2.163226e-01	3.653685e+00	3.110803e+00
7	4.162973e-02	5.614178e+00	2.412031e+00
8	3.820738e-15	3.995738e+00	1.393415e+00
9	1.170318e-14	8.010798e+00	7.761958e-15
10	1.170318e-14	4.015060e+00	7.761958e-15
11	3.452182e-02	1.903198e+00	9.182231e-02
12	0.000000e+00	5.990454e-15	5.990454e-15
13	0.000000e+00	5.990454e-15	5.990454e-15
14	0.000000e+00	0.000000e+00	0.000000e+00
15	5.274121e-15	5.274121e-15	5.274121e-15
16	9.879139e-15	9.879139e-15	9.589615e-02
17	1.429485e-14	9.879139e-15	9.879139e-15
18	1.348976e-14	4.605018e-15	4.605018e-15
19	1.624495e-01	3.443519e-01	4.957739e-01
20	1.613715e+00	2.337911e+00	5.355056e+00
21	1.633100e+00	2.818931e+00	5.604335e+00
22	2.730212e+00	4.122692e+00	5.459222e+00
23	2.443290e+00	2.553418e+00	2.372273e+00
Mínimo	0.00	0.00	0.00
Mediana	1.09	2.89	2.43
Máximo	2.73	10.72	7.70
IQR	1.62	4.01	3.70

Fonte: o autor.

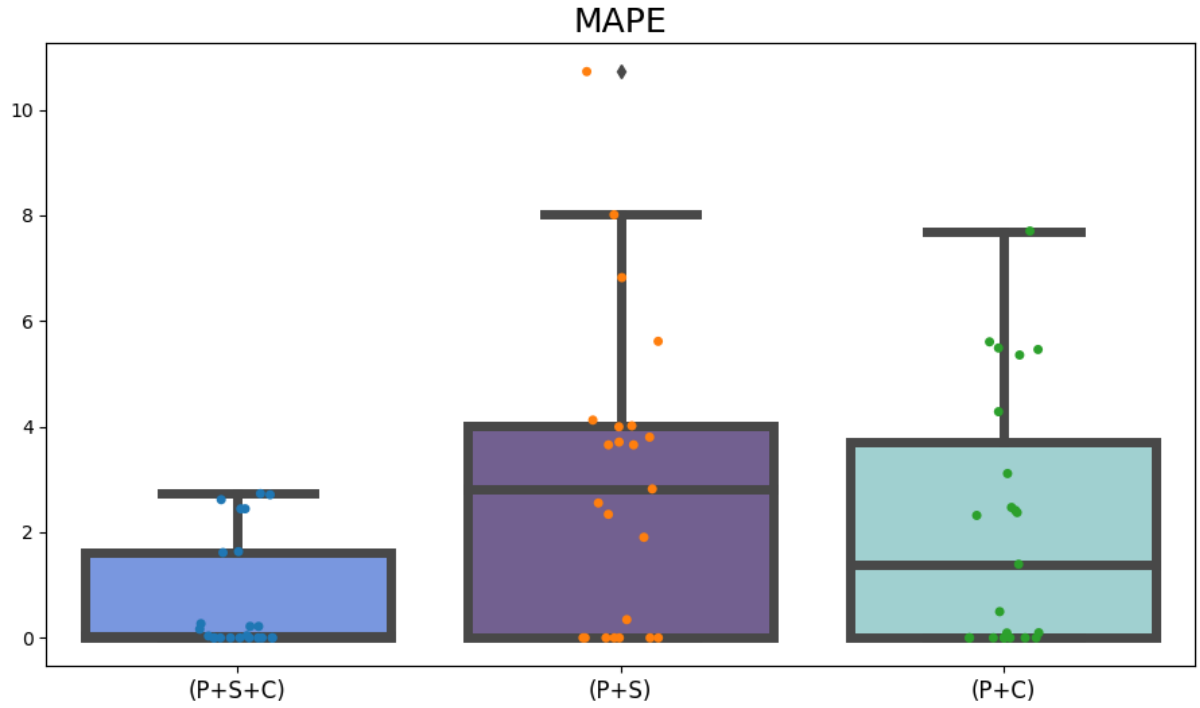


Figura 4.4: *Boxplots* de MAPE dos experimentos

Fonte: o autor.

Para verificar a significância estatística destes resultados, foi aplicado o Teste de Friedman em que a hipótese nula (H0) indica que os testes possuem a mesma distribuição de probabilidade, e a hipótese alternativa (H1) indica a ausência de evidências estatísticas para afirmar que as distribuições são iguais (SIEGEL; CASTELLAN, 1988). Para um nível de significância de 0.05, foi obtida a estatística q = 10.0000 e um p-valor de 0.0068, o que rejeita H0 e comprova que a distribuição de probabilidade dos resultados possuem diferenças significativas.

Tomando como referência o trabalho de Tonidandel Junior (2022), que lida justamente com o mesmo tipo de previsão na mesma série temporal, pode-se visualizar e comparar as estatísticas descritivas RMSE e MAPE dos seus modelos multivariados (identificados por FDTFTS_ID3, FDTFTS_CART, FDTFTS_RF) com os resultados deste trabalho através das Tabelas 4.5 e 4.6.

Tabela 4.5: Tabela comparativa de resultados de RMSE com a referência

RESULTADOS	(P+S+C)	(P+S)	(P+C)	FDTFTS_ID3	FDTFTS_CART	FDTFTS_RF
Mínimo	0.00	0.00	0.00	7.96	7.96	4.69
Mediana	1.53	3.59	3.64	19.15	19.20	16.17
Máximo	3.65	11.91	10.18	37.30	37.73	30.30
IQR	3.04	5.41	4.00	15.51	13.28	16.07

Fonte: o autor.

Tabela 4.6: Tabela comparativa de resultados de MAPE com a referência

RESULTADOS	(P+S+C)	(P+S)	(P+C)	FDTFTS_ID3	FDTFTS_CART	FDTFTS_RF
Mínimo	0.00	0.00	0.00	3.71	3.71	1.97
Mediana	1.09	2.89	2.43	11.94	12.06	10.50
Máximo	2.73	10.72	7.70	30.25	30.47	22.17
IQR	1.62	4.01	3.70	10.63	8.73	9.04

Fonte: o autor.

Analisando as Tabelas 4.5 e 4.6, observa-se que o método proposto neste trabalho, em todas as configurações de variáveis de entrada - (P+S+C), (P+S) e (P+C) - na WMVFTS, teve melhores resultados em relação aos métodos multivariados - FDTFTS_ID3, FDTFTS_CART e FDTFTS_RF - da referência em todas as métricas, indicando melhor acurácia na previsão de preços de minério de ferro na série temporal estudada.

Capítulo 5

Conclusões

Dados os objetivos almejados em 1.3, este trabalho propôs avaliar a aplicação de variáveis alternativas como variáveis exógenas de um modelo preditivo que séries temporais nebulosas no intuito de aumentar a robustez, em relação aos métodos já presentes na literatura para a previsão dos preços do minério de ferro. Para isto, foi construído um índice através da agregação nebulosa hesitante de sentimentos extraídos de notícias referentes ao minério de ferro e considerada a utilização a quantidade de notícias como variáveis.

As conclusões desta pesquisa apontam que a viabilidade do uso de um índice construído através da agregação de sentimentos obtido das notícias relacionadas ao minério de ferro para a previsão de preços e da quantidade de notícias se mostraram positivas.

A abordagem proposta indica que o método Weighted Multivariate Fuzzy Time Series (WMVFTS), na análise dos dados em questão, obtém melhores resultados quando o conjunto de dados de entrada contém todas as variáveis correlacionadas. Este método preditivo também de mostrou superior aos multivariados utilizados no trabalho tido como referência quando comparados pelas análises das estatísticas descritivas RMSE e MAPE. Do ponto de vista de planejamento e auxílio nas tomada de decisões de analistas sobre os preços futuros do minério de ferro, ao analisar a métrica MDA com precisão acima de 80% em todos os testes, a proposta apresentada por este estudo se mostra promissora e confiável para a previsão, principalmente, das tendências e oscilações da variável de interesse no curto e médio prazo.

5.1 Dificuldades e limitações

Ao longo da pesquisa foram enfrentadas algumas dificuldades e os resultados tiveram limitações.

Determinar uma fonte de notícias que fosse imparcial, de modo que fosse compatível com os parâmetros do BERT para a entrega de uma análise de sentimentos coerente é um complicador. Notícias sobre algo que influencie na alta do preço do minério de ferro são boas para mineradoras e ruins para siderúrgicas. Essas mesmas notícias podem ter

conotações diferentes dependendo a qual setor do mercado as mídias que as divulgam estão atreladas.

A utilização de um modelo de análise de sentimentos que possua uma base de dados de treinamento com sentimentos rotulados de fontes textuais especializadas no mercado de minério de ferro e direcionadas ao segmento de interesse, mineração ou siderurgia, tende a melhorar os resultados obtidos no presente estudo. Porém, construir esta base de dados é um grande dispêndio, uma vez que só seria possível através da rotulagem manual, feita por especialistas, de um número elevado de notícias que impactam de alguma forma o mercado de minério de ferro.

Capítulo 6

Trabalhos futuros

As sugestões para trabalhos futuros, que tenham esta dissertação como referência, aprimorarem a metodologia empregada são:

- Adicionar outras variáveis, que possam ter correlação com a série de preços do minério de ferro, ao modelo preditor na tentativa de aumentar sua acurácia.
- Explorar outras fontes textuais de informação sobre o minério de ferro e outras variáveis correlacionadas, como relatórios e artigos especializados, para a extração de sentimentos.
- Acrescentar a este modelo módulos para decisão automática, que utilizem *Deep Learning*, e comparar suas decisões com decisões tomadas por especialistas humanos.
- Aplicar a metodologia proposta neste trabalho na previsão de outros tipos de variáveis.

Referências Bibliográficas

ALESSIA, D. *et al.* Approaches, tools and applications for sentiment analysis implementation. **International Journal of Computer Applications**, Citeseer, v. 125, n. 3, 2015.

Alves, D. S. **Uso de técnicas de computação social para tomada de decisão de compra e venda de ações no mercado brasileiro de bolsa de valores**. 2015. Tese de Doutorado em Engeharia de Sistemas Eletrônicos e Automação, Publicação FT.PGEA-n 102=2015, Departamento de Engenharia Elétrica, Faculdade de Tecnologia, Universidade de Brasília, Brasília, DF, 133p.

ARIAS, M.; ARRATIA, A.; XURIGUERA, R. Forecasting with twitter data. **ACM Transactions on Intelligent Systems and Technology (TIST)**, ACM New York, NY, USA, v. 5, n. 1, p. 1–24, 2014.

BEASLEY, J. E. **OR-Notes**. 2022. Brunel University London. "Acessado em 28/07/2022". Disponível em: ⟨http://people.brunel.ac.uk/~mastjjb/jeb/or/forecast.html⟩.

BENEVENUTO, F.; RIBEIRO, F.; ARAÚJO, M. Métodos para análise de sentimentos em mídias sociais. **Sociedade Brasileira de Computação**, 2015.

BENGIO, Y.; SIMARD, P.; FRASCONI, P. Learning Long-Term Dependencies with Gradient Descent is Difficult. **IEEE Transactions on Neural Networks**, 1994. ISSN 19410093.

CARVALHO, F. F. de. **Transferência de Aprendizado com Embeddings Contextuais para Classificação de Texto em Cenários de Baixo Volume de Dados**. 2020. Dissertação submetida para obtenção do Título de Mestre em Inteligência Computacional, Escola de Engenharia, Universidade Federal de Minas Gerais, Belo Horizonte, MG.

CHENG, C.-H.; CHEN, C.-H. Fuzzy time series model based on weighted association rule for financial market forecasting. **Expert Systems**, Wiley Online Library, v. 35, n. 4, p. e12271, 2018.

DEVLIN, J. *et al.* Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.

DIAS, B. C. D. *et al.* Aggregation of sentiment analysis index with hesitant fuzzy sets for financial time series forecasting. In: **2021 IEEE World AI IoT Congress (AIIoT)**. [S.l.: s.n.], 2021. p. 0433–0439.

EWEES, A. A. *et al.* Improving multilayer perceptron neural network using chaotic grasshopper optimization algorithm to forecast iron ore price volatility. **Resources Policy**, v. 65, p. 101555, 2020. ISSN 0301-4207. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/S0301420719300832⟩.

FU, Z. The mechanism of imported iron ore price in china. **Modern Economy**, n. 9, p. 1908–1931, 2018.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [S.l.]: MIT Press, 2016.

HAYKIN, S. **Neural networks and learning machines**. 3rd. ed. [S.l.]: Prentice Hall/Pearson, 2009. ISBN 978-0-13-147139-9.

HOCHREITER, S.; SCHMIDHUBER, J. Long Short-Term Memory. **Neural Computation**, 1997. ISSN 08997667.

IGARASHI, W.; VALDEVIESO, G. S.; IGARASHI, D. C. C. Análise de sentimentos e indicadores técnicos: uma análise da correlação dos preços de ativos com a polaridade de notícias do mercado de ações. **Brazilian Journals of Business**, n. 1, p. 470–486, 2020. ISSN 2596-1934.

INDEX, M. **FAQs**. 2022. Market Index. "Acessado em 28/08/2022". Disponível em: ⟨https://www.marketindex.com.au/iron-ore⟩.

JOWITT, S. M. Covid-19 and the global mining industry. **SEG Discovery**, GeoScienceWorld, n. 122, p. 33–41, 2020.

JR, H. T.; GUIMARÃES, F. G. Aplicação de modelos nebulosos univariados e multivariados na previsão de preços de minério de ferro: Um estudo comparativo. 2022.

KEENAN, M. J. S. **Advanced Positioning, Flow, and Sentiment Analysis in Commodity Markets: Bridging Fundamental and Technical Analysis**. [S.l.]: Wiley, 2019. ISBN 978-1-119-60384-9.

LI, D. *et al.* Development of a group method of data handling technique to forecast iron ore price. **Applied Sciences**, v. 10, n. 7, 2020. ISSN 2076-3417. Disponível em: ⟨https://www.mdpi.com/2076-3417/10/7/2364⟩.

LI, W. *et al.* Rdeu hawk-dove game analysis of the china-australia iron ore trade conflict. **SSRN**, 2021. Disponível em: ⟨https://ssrn.com/abstract=3982302⟩.

LIMA, P. C. de *et al.* Scalable models for probabilistic forecasting with fuzzy time series. Universidade Federal de Minas Gerais, 2019.

LIU, B. Sentiment analysis and opinion mining. **Synthesis lectures on human language technologies**, Morgan & Claypool Publishers, v. 5, n. 1, p. 1–167, 2012.

MA, Y.; WANG, J. Time-varying spillovers and dependencies between iron ore, scrap steel, carbon emission, seaborne transportation, and china's steel stock prices. **Resources Policy**, Elsevier, v. 74, p. 102254, 2021.

MENDES, J. A. O ferro na história: das artes mecânicas às belas-artes. **Gestão e Desenvolvimento**, n. 9, p. 301–318, 2000.

MIKOLOV, T. *et al.* Distributed representations of words and phrases and their compositionality. **Advances in neural information processing systems**, v. 26, 2013.

MOHRI, M.; ROSTAMIZADEH, A.; TALWALKAR, A. **Foundations of machine learning**. [S.l.]: MIT press, 2018.

MORETTIN, P. A.; TOLOI, C. M. C. **Análise de séries temporais**. São Paulo: ABE-Projeto Fisher e Editora Edgard Blucher, 2004.

PUSTOV, A.; MALANICHEV, A.; KHOBOTILOV, I. Long-term iron ore price modeling: Marginal costs vs. incentive price. **Resources Policy**, v. 38, n. 4, p. 558–567, 2013. ISSN 0301-4207. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/S0301420713000743⟩.

SIEGEL, S.; CASTELLAN, N. **Nonparametric Statistics for the Behavioral Sciences**. [S.l.]: McGraw-Hill, 1988. (McGraw-Hill international editions. Statistics series). ISBN 9780071003261.

SILVA, J. T. F. **Um Modelo Computacional de Redes Neurais para Localização de Faltas em Redes de Distribuição**. 2019. Monografia (Graduação em Engenharia de Controle e Automação), Escola de Engenharia, Universidade Federal de Minas Gerais, Belo Horizonte.

SILVA, P. C. *et al.* Distributed evolutionary hyperparameter optimization for fuzzy time series. **IEEE Transactions on Network and Service Management**, IEEE, v. 17, n. 3, p. 1309–1321, 2020.

SONG, Q.; CHISSOM, B. S. Fuzzy time series and its models. **Fuzzy sets and systems**, Elsevier, v. 54, n. 3, p. 269–277, 1993.

SOUSA, M. G. *et al.* Bert for stock market sentiment analysis. In: IEEE. **2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)**. [S.l.], 2019. p. 1597–1601.

Tonidandel Junior, H. **Previsão de preços de minério de ferro utilizando modelos de inteligência computacional**. 2022. Dissertação de Mestrado Profissional submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em Instrumentação, Controle e Automação de Processos de Mineração da Universidade Federal de Ouro Preto, Ouro Preto, MG.

TORRA, V. Hesitant fuzzy sets. **International journal of intelligent systems**, Wiley Online Library, v. 25, n. 6, p. 529–539, 2010.

TUO, J.; ZHANG, F. Modelling the iron ore price index: A new perspective from a hybrid data reconstructed eemd-goru model. **Journal of Management Science and Engineering**, v. 5, n. 3, p. 212–225, 2020. ISSN 2096-2320. Special Issue on Contemporary Mega-Project and Program Management. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/S2096232020300408⟩.

VASWANI, A. *et al.* Attention is all you need. **Advances in neural information processing systems**, v. 30, 2017.

WERBOS, P. J. Backpropagation Through Time: What It Does and How to Do It. **Proceedings of the IEEE**, 1990. ISSN 15582256.

XIA, M.; XU, Z. Hesitant fuzzy information aggregation in decision making. **International journal of approximate reasoning**, Elsevier, v. 52, n. 3, p. 395–407, 2011.

ZHU, X. *et al.* Time-varying international market power for the chinese iron ore markets. **Resources Policy**, v. 64, p. 101502, 2019. ISSN 0301-4207. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/S0301420718306858⟩.