UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Integração de Séries Temporais Financeiras e Informação Textual na Previsão do Mercado de Commodities Agrícola

Ivan José dos Reis Filho

Tese de Doutorado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP SÃO CARLOS

Ivan José dos Reis Filho

Integração de Séries Temporais Financeiras e Informação Textual na Previsão do Mercado de Commodities Agrícola

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientadora: Profa. Dra. Solange Oliveira Rezende
Coorientador: Prof. Dr. Ricardo Marcondes Marcacini

USP – São Carlos
Junho de 2024

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

José dos Reis Filho, Ivan
J375i Integração de Séries Temporais Financeiras e Informação Textual na Previsão do Mercado de Commodities Agrícola / Ivan José dos Reis Filho; orientador Solange Oliveira Rezende; coorientador Ricardo Marcondes Marcacini. -- São Carlos, 2024.
148 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. Séries temporais. 2. Mineração de textos. 3. Multimodalidade. 4. Commodities agrícola. I. Oliveira Rezende, Solange, orient. II. Marcondes Marcacini, Ricardo, coorient. III. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Ivan José dos Reis Filho

Integration of Financial Time Series and Textual Information in Agricultural Commodity Market Forecasting

Doctoral dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP, in partial fulfillment of the requirements for the degree of the Doctorate Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Profa. Dra. Solange Oliveira Rezende
Co-advisor: Prof. Dr. Ricardo Marcondes Marcacini

USP – São Carlos
June 2024

Este trabalho é dedicado aos meus três filhos,
Isabella Ferreira, Jorge Henrique e João Guilherme,
a quem diariamente me mostram exemplos de amor e carinho.

AGRADECIMENTOS

Os agradecimentos principais são direcionados à Deus que proporcionou a dádiva da vida. Gostaria de expressar minha profunda gratidão aos meus pais, Ivan José e Nerci Aparecida que sempre estiveram ao meu lado, oferecendo apoio incondicional e encorajamento ao longo dessa jornada.

Um agradecimento especial à minha esposa, Juliana, pela sua paciência, amor e apoio durante todo o processo de elaboração desta Tese de Doutorado. Aos meus filhos, Isabella Ferreira, Jorge Henrique e João Guilherme, que são minha fonte de inspiração e motivação diária, por darem sentido à minha vida e me impulsionarem a alcançar meus objetivos.

Não poderia deixar de mencionar a professora Solange Oliveira Rezende, cuja dedicação, orientação e paciência foram fundamentais para a realização deste trabalho. Ao meu co-orientador, Ricardo Marcacini, pelo apoio técnico valioso em todas as etapas do doutorado. Agradeço também aos meus colegas do LABIC, cujas contribuições, sugestões e diferentes pontos de vista enriqueceram significativamente meu trabalho. Em especial, agradeço ao Marcos Gôlo por sua colaboração essencial no desenvolvimento de algumas das pesquisas.

Agradeço ao Instituto de Ciências Matemáticas e de Computação e ao Laboratório de Inteligência Computacional, por me oferecerem um ambiente de estudos e trocas inigualável. Agradeço também as agências financiadoras que me apoiaram para elaboração desta tese: Programa de Capacitação de Recursos Humanos da (PCRH) da Fundação de Amparo a Pesquisa do Estado de Minas Gerias (FAPEMIG) (Processo PCRH BPG-00054-210). Ao Centro de Inteligência Artificial (C4AI) e Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) (bolsa nº 2019/07665-4) e pela IBM Corporation. A FAPESP (Processo 2019/25010-5) e ao Centro Nacional de Desenvolvimento Científico e Tecnológico (Processo 309575/2021-4). Por fim, agradeço a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento nº 88887.928770/2023-00.

“Crie uma meta e
estabeleça um limite.”
(Autor desconhecido)

RESUMO

REIS FILHO, I. J. **Integração de Séries Temporais Financeiras e Informação Textual na Previsão do Mercado de Commodities Agrícola**. 2024. 148 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

O mercado de commodities agrícolas é conhecido por sua volatilidade e complexidade, onde fatores como condições climáticas, demanda global, políticas governamentais e eventos geopolíticos exercem uma influência significativa sobre as decisões dos especialistas e os preços nos mercados. Nos últimos anos, tem havido um crescente interesse em aplicações baseadas em aprendizado de máquina no agronegócio, visando minimizar os desafios impostos pelo cenário caótico e complexo do mercado financeiro. Essas aplicações exploram avanços tecnológicos para aprimorar a previsão de preços e tendências, fornecendo percepções estratégicas aos especialistas do domínio. Recentemente, estudos têm sido desenvolvidos utilizando técnicas de processamento de linguagem natural e dados de séries temporais em diversas estratégias de fusão, gerando representações multimodais alternativas para modelos de previsão. A integração de dados de múltiplas fontes visa proporcionar previsões que considerem fatores não explícitos em dados de séries temporais. No entanto, propor modelos e representações multimodais é desafiador devido ao alinhamento temporal entre dados de textos e séries temporais. Além disso, a disponibilidade de documentos rotulados para o domínio do agronegócio é escassa, o que dificulta a aplicação direta de modelos multimodais. Diante desse cenário, esta tese busca desenvolver e avaliar representações de séries temporais integradas com representações semânticas de textos, explorando abordagens inovadoras para aprimorar previsões de séries temporais enriquecidas com textos e classificações automáticas de notícias com padrões extraídos das séries temporais. As abordagens propostas consideram estratégias que podem ser aplicadas em cenários reais de mercado. Os resultados demonstram que as abordagens propostas podem ser uma alternativa real para melhorar a precisão das previsões em mercados complexos e voláteis, oferecendo uma perspectiva inovadora na integração de dados textuais e séries temporais no contexto do agronegócio.

Palavras-chave: Séries temporais; Mineração de textos; multimodalidade; Commodities agrícola.

ABSTRACT

REIS FILHO, I. J. **Integration of Financial Time Series and Textual Information in Agricultural Commodity Market Forecasting**. 2024. 148 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

The agricultural commodities market is known for its volatility and complexity, where factors such as weather conditions, global demand, government policies, and geopolitical events significantly influence experts' decisions and prices in the markets. In recent years, there has been a growing interest in machine learning applications in agribusiness, aiming to minimize the challenges posed by the chaotic and complex landscape of the financial market. These applications explore technological advances to improve price and trend forecasting, providing strategic insights to domain experts. Recently, studies have been developed using natural language processing techniques and time series data in various fusion strategies, generating alternative multimodal representations for prediction models. Integrating data from multiple sources aims to provide forecasts considering factors not explicitly present in time series data. However, proposing multimodal models and representations is challenging due to the temporal alignment between text data and time series. Additionally, the availability of labeled documents for the agribusiness domain is scarce, making the direct application of multimodal models difficult. Given this scenario, this thesis aims to develop and evaluate time series representations integrated with semantic text representations, exploring innovative approaches to enhance time series predictions enriched with texts and automatic news classifications with patterns extracted from time series. The proposed approaches consider strategies that can be applied in real market scenarios. The results demonstrate that the proposed approaches can be an alternative to improve forecast accuracy in complex and volatile markets, offering an innovative perspective on integrating textual data and time series in the context of agribusiness.

Keywords: time series; text mining; multimodality; Agricultural comodities..

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

1DCNN	one-Dimensional Convolutional Neural Network
ALSTM	Attention-based Long Short-Term Memory
ARIMA	*Autoregressive Integrated Moving Average*
BERT	*Bidirectional Encoder Representations from Transformers*
CBOW	*Continuous Bag-of-Words*
CHARM	*Context-Aware Hierarchical Attention Mechanism*
CNN-SC	*Convolutional Neural Network with Sentiment Check*
CoATSMP	*Collaborative Attention Transformer fusion model for Stock Movement Prediction*
DJIA	Dow Jones Industrial Average
DRL	*Deep Reinforcement Learning*
DRNN	*Deep Recurrent Neural Networks*
ELMo	*Embeddings from Language Models*
GCN	*Graph Convolutional Networks*
GELU	*Gaussian Error Linear Units*
GELU	*Gaussian Error Linear Unit*
GloVe	*Global Vectors for Word Representations*
GNN	*Graph Neural Network*
GRU	*Gated Recurrent Network*
HEM	*Efficient Market Hypothesis*
LDA	*Latent Dirichlet Allocation*
LDA	*Latent Dirichlet Allocation*
LSI	*Latent Semantic Indexing*
LSTM	*Long Short Term Memory*
MAC	*Multi-source Aggregated Classification*
MAE	Média do Erro Absoluto
MAPE	Erro Médio Percentual Absoluto
MASE	Erro Médio Absoluto Escalado
MLM	*Masked Language Model*
MLP	*Multilayer Perceptron*
MSE	Erro Quadrático Médio
MT	**Mineração de Textos**

NSP	*Next Sentence Prediction*
PIP	*Perceptually Important Points*
POS	*Part-of-speech*
PPO	*Proximal Policy Optimization*
PSO	*Particle Swarm Optimization*
RBF	Radial Base Function
RMSE	Raiz do Erro Quadrático Médio
RNR	Redes Neurais Recorrentes
S3WE	*Sequential Three-Way Decisions*
ST	**Série Temporal**
STACN	Spatial-Temporal Attention-based Convolutional Network
SVM	Support Vector Machine
TF	*Term Frequency*
TF-IDF	*Term Frequency-Inverse Document Frequency*
VADER	*Valence Aware Dictionary and sEntiment Reasoner*
VMD	*Variaton Mode Decomposition*
WCN-LSTM	*LSTM-based Weighted Categorized News*

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

Nos últimos anos, a quantidade de dados gerados e disponíveis têm crescido de maneira exponencial (JANEV *et al.*, 2020). O avanço tecnológico na área da computação permitiu a transformação desses dados em informações e conhecimentos úteis, proporcionando suporte para a tomada de decisões em diversas áreas do conhecimento. Nesse contexto, sistemas de gerenciamento de bancos de dados, aplicações com apoio em aprendizado de máquina e computação em nuvem têm impulsionado a oferta de recursos computacionais capazes de armazenar, analisar e gerenciar esse grande volume de dados (CHATFIELD; XING, 2019).

Considerando os avanços na área de Inteligência Artificial que exploram o grande volume de dados disponível, a **Contextualização e Motivação** (Seção 1.1), que formaram a base para as investigações conduzidas neste trabalho de doutorado, são apresentadas no início deste capítulo. As **Questões de Pesquisa**, que orientaram as propostas e direcionaram o estudo, são discutidas e relacionadas aos objetivos da tese na Seção 1.2. Em seguida, é apresentado o **Sumário de Contribuições** (Seção 1.3), destacando os principais resultados alcançados durante o desenvolvimento do trabalho. Por fim, a **Organização da Tese** (Seção 1.4) é detalhada, descrevendo o conteúdo de cada capítulo desta tese.

1.1 Contextualização, Motivação e Lacunas

Os dados gerados na internet podem ser classificados em três categorias: estruturados, não estruturados e semi-estruturados. Os dados estruturados consistem em valores numéricos quantitativos com uma estrutura bem definida, o que facilita seu armazenamento e processamento por aplicações de gerenciamento de dados (AGGARWAL, 2014). Em contraste, os dados não estruturados não seguem um formato específico e são mais desafiadores de processar, como mensagens de e-mail, documentos de texto, publicações em redes sociais, imagens e áudios, entre outros (HASSANI *et al.*, 2020). Os dados semi-estruturados possuem uma estrutura heterogênea, em que cada campo de dados apresenta uma descrição automática e não há imposição de formato

fixo. Alguns exemplos conhecidos de dados semi-estruturados incluem representações como XML (Extensible Markup Language), JSON (JavaScript Object Notation), RDF (Resource Description Framework) e OWL (Ontology Web Language) (GIUDICE *et al.*, 2019).

O armazenamento temporal dos dados viabiliza uma organização cronológica dos registros coletados, estrutura conhecida como **Série Temporal** (ST) (apresentada na Seção 2.1). Uma ST é definida como dados estruturados e é essencialmente uma sequência de registros numéricos, coletados em intervalos temporais específicos. A análise de séries temporais é uma área de estudo que desempenha um papel importante em diversas áreas do conhecimento, como economia, finanças, meteorologia e muitas outras (TANG *et al.*, 2022; SEZER; GUDELEK; OZBAYOGLU, 2020a). As séries temporais são formadas por dados sequenciais e que podem ser constituídas por dois tipos: estocásticas e determinísticas. As séries estocásticas são caracterizadas por possuírem componentes aleatórios, o que torna difícil prever seus padrões futuros com alta precisão. Por outro lado, as séries determinísticas apresentam padrões reconhecíveis e consistentes ao longo do tempo, permitindo a identificação de tendências, sazonalidade e comportamentos cíclicos (CHATFIELD; XING, 2019). Dessa forma, a compreensão adequada da formação e dos componentes é essencial para realizar análises precisas e eficientes em séries temporais, bem como desenvolver modelos de previsão confiáveis.

Os **Modelos preditivos** (Seção 2.2) de séries temporais são construídos com base na ideia de que valores futuros podem ser estimados por funções matemáticas que são parametrizadas por meio de observações passadas (LIU *et al.*, 2021). Tradicionalmente, modelos paramétricos e lineares, como o *Autoregressive Integrated Moving Average* (ARIMA) (KENDALL, 1971), têm sido amplamente utilizados em diversas variações e mostrado bom desempenho para previsão de séries que apresentam comportamentos determinísticos (ZOU *et al.*, 2007; ADANACIOGLU; YERCAN *et al.*, 2012; AHUMADA; CORNEJO, 2016). No entanto, esses modelos lineares podem falhar em capturar padrões estocásticos e não lineares (MONDAL; SHIT; GOSWAMI, 2014). Para superar essas limitações, modelos não paramétricos, como Redes Neurais Artificiais, Aprendizado Profundo e Vetores de Suporte de Regressão, têm sido propostos e apresentado resultados promissores (LIM; ZOHREN, 2021). Paralelamente, modelos com base na arquitetura *Transformers* e *Graph Neural Network* (GNN) têm ganhado destaque nas tarefas de previsão de séries temporais (SANTOS; MARCACINI; REZENDE, 2021; CHENG *et al.*, 2022). No entanto, é importante observar que, esses modelos não paramétricos têm focado principalmente em séries temporais com comportamentos de tendências, sazonalidades e movimentos cíclicos (WEN *et al.*, 2022).

A natureza volátil dos mercados financeiros, juntamente com a influência de eventos externos imprevisíveis, torna a previsão uma tarefa complexa e sujeita a incertezas (VENTER; STRYDOM; GROVÉ, 2013). Mesmo os modelos não paramétricos, que foram propostos como alternativas para lidar com séries não lineares, também podem encontrar dificuldades para realizar análises precisas e eficientes em tais dados (XU *et al.*, 2020; ZHONG; HITCHCOCK,

2021). Nesse sentido, técnicas de **Mineração de Textos** (MT) (Seção 2.3) têm sido empregadas em estudos para selecionar dados de textos relevantes e incorporá-los nas séries temporais em outros domínios (PICASSO *et al.*, 2019) (RODRIGUES; MARKOU; PEREIRA, 2019). A ideia geral envolvida nessas abordagens é utilizar uma representação vetorial dos textos para capturar comportamentos estocásticos e informações implícitas que podem não estar diretamente presentes nas séries financeiras.

Outra área que têm ganhado destaque nos últimos anos é a análise de sentimento da notícia para compreender o impacto das mídias sociais nas tendências de mercado (FARIMANI *et al.*, 2021; DARADKEH, 2022; JI *et al.*, 2023). Alguns trabalhos visam discernir como os relatórios técnicos influenciam a dinâmica das séries de preços, enfatizando como esses relatórios moldam o comportamento de mercado (ZHONG; HITCHCOCK, 2021) (LI; WU; WANG, 2020). Recentemente, abordagens têm sido propostas para integrar múltiplas fonte de dados, como séries temporais, notícias, mídia social e relatórios técnicos para predizer os movimentos ou tendência do mercado financeiro (ZHOU *et al.*, 2020; NTI; ADEKOYA; WEYORI, 2021; WINDSOR; CAO, 2022). Nesse campo de pesquisa, muitos estudos concentram-se na classificação de notícias como positiva ou negativa, sendo que o interesse do mercado vai além da polaridade das notícias. Para os especialistas do domínio, uma notícia pode ser significativa para o mercado independente da sua polaridade. Por exemplo, uma notícia classificada como positiva pode apenas retratar uma situação do passado, e que potencialmente não influencia decisões futuras de investimento e não impacta o mercado no momento da publicação da notícia.

Como tendência de pesquisa, novas abordagens têm sido apresentadas para combinar múltiplas fontes de dados para tarefas preditivas. Na literatura, a combinação de diferentes dados é conhecida como modelos multimodais ou **Fusão de Informação** (Seção 2.4) (HUANG *et al.*, 2020). Esses modelos combinam dados provenientes do mesmo domínio em uma representação consistente e clara. As técnicas de *Early Fusion*, *Joint Fusion* e *Late Fusion* têm sido exploradas em diversas áreas, como medicina (EL-SAPPAGH *et al.*, 2021), sistemas avançados de assistência ao motorista (LIM *et al.*, 2019), sistemas baseados em blockchain (LIANG *et al.*, 2021) e no domínio financeiro (ZHONG; HITCHCOCK, 2021). Essas abordagens têm sido aplicadas para melhorar a precisão dos modelos preditivos ou para ajudar a interpretar os resultados. Entretanto, desafios ainda persistem quando se trata de lidar com séries financeiras, caracterizadas por comportamentos estocásticos, volatilidade e influência de eventos externos imprevisíveis. Nesse sentido, combinar informações textuais com dados estruturados de séries temporais têm se apresentado uma estratégia promissora para aprimorar a capacidade de previsão nesse setor financeiro volátil e complexo (VERMA; SAHU; SAHU, 2023).

Considerando as perspectivas apresentadas, a presente tese de Doutorado têm como objetivo de avaliar duas propostas com diferentes abordagens: i) abordagens para desenvolver modelos de representações de séries temporais enriquecidas com dados extraídos de textos, apresentadas no Capítulo 4; e, ii) abordagens de classificação de textos usando séries temporais,

no Capítulo 5. As propostas abrem novas possibilidades para obter previsões mais informadas ao dinâmico domínio do mercado financeiro, em específico o de commodities da soja e do milho. Ao considerar tanto os dados estruturados como os não estruturados, buscou-se validar a hipótese de que as séries financeiras podem ser consideradas como uma fonte de informação adicional para modelos de previsão multimodais.

1.2 Questões de Pesquisa e Objetivos

Esta tese de doutorado visa avançar pesquisas na área de previsão de séries temporais e de classificação de notícias, em especial apresentar novas abordagens de integração entre dados de notícias e séries temporais relacionados ao agronegócio brasileiro. O desenvolvimento deste projeto é guiado por Questões de Pesquisas (QP) separadas em duas abordagens **QP1** e **QP2**:

QP1: Em relação ao enriquecimento de séries temporais com técnicas de Mineração de Textos, em especial, o uso de representação textual para enriquecer séries temporais:

> **QP1.1** Quais são os modelos de representação textual mais utilizados para enriquecer séries temporais?
>
> **QP1.2** Como as representações (textos e séries) podem impactar as estratégias dos modelos de fusão nas tarefas preditivas?
>
> **QP1.3** Quais são os modelos amplamente utilizados e que resultam em melhores resultados para predição de séries temporais?

QP2: Em relação ao uso de séries temporais para classificação automática de textos:

> **QP2.1**: Como utilizar dados de séries financeiras para classificar notícias?
>
> **QP2.2**: Séries temporais financeiras contribuem significativamente para classificação de textos relacionados ao mercado de commodities?
>
> **QP2.3**: Quais são os modelos de aprendizado mais adequados para modelos multimodais?

De acordo com as questões de pesquisa identificadas, são definidos os seguintes objetivos para o desenvolvimento deste projeto:

- Realizar um mapeamento sistemático na literatura e analisar os trabalhos semelhantes a esta temática, assim como, explorar as abordagens recentes e avaliar como os modelos multimodais podem ser empregados para série temporal e informação textual (**QP1.1** e **QP1.3**).

- Propor, desenvolver e avaliar diferentes modelos multimodais com dados de textos e séries temporais (**QP1.2 e QP2.2**).

- Propor, desenvolver e avaliar diferentes estratégias de previsão de mercado financeiro para um cenário real. (**QP2.1**, **QP2.2** e **QP2.3**).

Após abordar as questões de pesquisa e atingir os objetivos propostos, a próxima seção destaca os trabalhos publicados na literatura.

1.3 Sumário de Contribuições

Os principais resultados obtidos durante o desenvolvimento do trabalho de doutorado e que foram publicados e submetidos:

REIS FILHO, I. J.; CORRÊA, G. B.; FREIRE, G. M.; REZENDE, S. O. Forecasting future corn and soybean prices: an analysis of the use of textual information to enrich time-series. In: **Anais do VIII Symposium on Knowledge Discovery, Mining and Learning**. SBC, p. 113-120. 2020.

CARMO, P.; REIS FILHO, I. J.; MARCACINI, R. Commodities trend link prediction on heterogeneous information networks. In: **Anais do IX Symposium on Knowledge Discovery, Mining and Learning**. SBC, p. 81-88. 2021.

REIS FILHO, I. J.; MARCACINI, Ricardo M.; REZENDE, Solange O. Previsão do preço futuro de commodities agrícolas: um estudo para enriquecer séries temporais. In: **Simpósio Brasileiro de Automação Inteligente-SBAI**. 2021.

REIS FILHO, I. J.; MARCACINI, R. M.; REZENDE, S. O.. On the enrichment of time series with textual data for forecasting agricultural commodity prices. **MethodsX**, v. 9, p. 101758, 2022.

CARMO, P.; REIS FILHO, I. J.; MARCACINI, R. TRENCHANT: TRENd PrediCtion on Heterogeneous InformAtion NeTworks. **Journal of Information and Data Management**, v. 13, n. 6, 2022.

REIS FILHO, I. J.; MARTINS, L. H. D.; PARMEZAN, A. R. S.; MARCACINI, R. M.; REZENDE, S. O.. Sequential short-text classification from multiple textual representations with weak supervision. In: **Brazilian Conference on Intelligent Systems. Cham: Springer International Publishing**, 2022. p. 165-179.

TRINDADE, R. N, MARTINS, L. H., CORREA, G. N., REIS FILHO, I. J.. Using a labeling function for automatic classification of agribusiness news: A weak supervisory approach. In: **Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional**. SBC, p. 73-82. 2022.

MARTINS, L. H. D., TRINDADE, R. N., CORREA, G. N., HEITOR, C. C. C., REIS FILHO , I. J. Avaliação Do TD-Bert Com Diferentes Modelos De Representação Textual para Tarefas de Classificação De Textos. **RETEC-Revista de Tecnologias**, v. 16, n. 1, p. 40-52, 2023.

REIS FILHO, I. J. R.; COLETI, J. C.; MARCACINI, R. M.; REZENDE, S. O. Dataset: Annotated Soybean Market News Articles. **Data in Brief**, p. 110545, 2024.

REIS FILHO, I. J. R.; GOLO, M. P. S.; MARCACINI, R. M.; REZENDE, S. O. How do financial time series enhance the detection of news significance in market movements? a study using graph neural networks with heterogeneous representations. **Neural Computing and Applications**, Springer (em revisão - submetido em 20/02/2024).

1.4 Organização da Tese

No Capítulo 1 é apresentado a Introdução da Tese, assim como os conceitos gerais sobre séries temporais, modelos paramétricos e não paramétricos, estratégias de fusão de informação e as limitações existentes na previsão no mercado financeiro. Como propostas, alternativas para a combinação de dados de séries temporais e informação textual foram apresentadas, assim como questões de pesquisa, objetivos e o sumário de contribuições.

No Capítulo 2 são apresentados os métodos que fundamentam esta tese de Doutorado. Primeiro, uma definição, técnicas e métodos de séries temporais são apresentadas a fim de abordar as limitações de cada tipo de previsão. Técnicas de mineração de textos para representação vetorial de textos, bem como os modelos de linguagens disponíveis na literatura são apresentadas para elucidar as especificações de cada representação.

Um mapeamento sistemático foi realizado e detalhado no Capítulo 3. O objetivo é apresentar os estudos relacionados a esta tese, destacando as contribuições significativas e delineando o estado atual na área de pesquisa. Além disso, busca-se elucidar os desafios enfrentados nessa área, fornecendo uma visão sobre o enriquecimento de séries temporais com informações textuais. Essa análise é fundamental para situar o presente trabalho no contexto mais amplo da pesquisa em previsão de commodities agrícolas, evidenciando lacunas de conhecimento e oportunidades para inovação.

As propostas são apresentada em dois Capítulos. No primeiro Capítulo 4 é apresentada a proposta de enriquecer séries temporais com dados extraídos de textos, sendo discutidas três abordagens diferentes. No Capítulo 5 é detalhado a segunda proposta de classificar textos usando dados de séries temporais, sendo discutidas duas abordagens distintas. Por fim, no Capítulo 6 são discutidos as conclusões obtidas durante o trabalho de Doutorado, apresentados os trabalho futuros .

CAPÍTULO

2

FUNDAMENTOS

Os fundamentos que serviram como base para as investigações realizadas no âmbito deste trabalho de Doutorado são apresentados no presente capítulo. De início, os conceitos **de Séries Temporais** (Seção 2.1), uma área essencial para a compreensão de seus tipos, são abordados quanto as classificações e composição dos dados temporais. Nessa exploração, os conceitos fundamentais relativos a séries estocásticas e determinísticas, estacionárias e não estacionárias, além de examinarmos os componentes de tendência, sazonalidade, cíclico e ruídos, são elucidados a fim de proporcionar uma compreensão fundamentada dos aspectos que moldam as características das séries temporais.

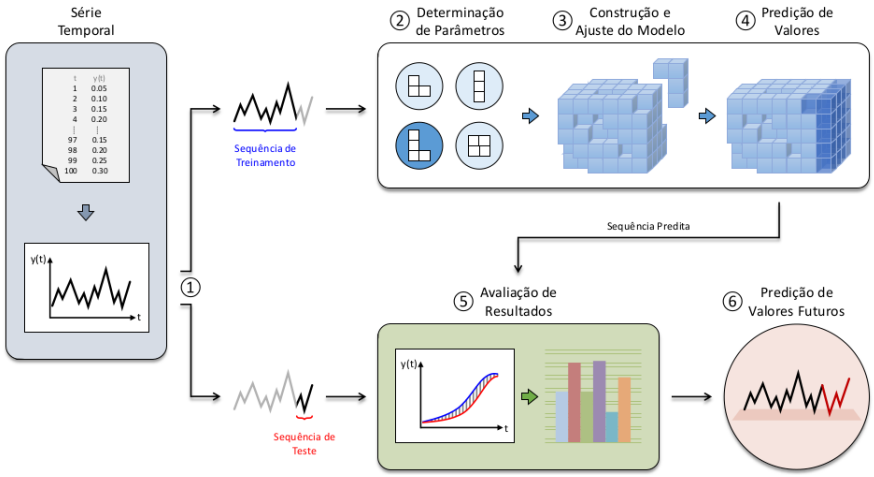
Em seguida, são apresentados os **Modelos Preditivos** (Seção 2.2), abrangendo tanto abordagens paramétricas quanto não paramétricas. Técnicas tradicionais de **Mineração de Textos** (Seção 2.3) são discutidas com objetivo de apresentar as representações vetoriais a partir de dados textuais. Conceitos são explicitados sobre as técnicas de processamento de linguagem natural, abordagens que desempenham um papel fundamental na captação de significados contextuais e relações semânticas presentes nos textos. Como ponto importante neste trabalho, técnicas de **Fusão de Informação** (Seção 2.4) são exploradas para examinar os modelos de arquiteturas para diferentes estratégias de fusão, como Fusão inicial (*early*), conjunta (*join*) e tardia (*late*). O entendimento das estratégias de fusão visa propor novos modelos de representações integradas de séries temporais e dados de textos.

O capítulo é encerrado explorando as **Métricas de Avaliação** (Seção 2.5). O objetivo é apresentar as bases teóricas e práticas necessárias para a concepção de modelos de previsão sólidos e eficazes. Ao contemplar os aspectos apresentados anteriormente, o objetivo é de embasar as etapas subsequentes, as quais visam aprofundar e aplicar esses conhecimentos para enriquecer a análise de séries temporais financeiras por meio de informações textuais, contribuindo assim para o aprimoramento na capacidade de previsão e tomada de decisões no cenário dinâmico e desafiador.

2.1 Séries Temporais

Série Temporal (ST) é uma sequência de pontos de dados coletados em intervalos de tempo uniformes ou irregulares, representando a evolução de uma variável ao longo do tempo (MONTGOMERY; JENNINGS; KULAHCI, 2015). Ao desenvolver modelos de previsão para séries temporais, é possível utilizar dados históricos de diferentes horizontes temporais, incluindo curto, médio e longo prazo. Independentemente do modelo escolhido, o processo de previsão em séries temporais geralmente compreende seis etapas fundamentais, como ilustrado na Figura 1. Essas etapas envolvem: 1) a coleta e **Análises das Séries Temporais** (Subseção 2.1.1); 2) Determinação de parâmetros; 3) Construção e ajuste do modelo; 4) Predição de valores; 5) Avaliação dos resultados; e, 5) Predição de valores futuros (PARMEZAN; SOUZA; BATISTA, 2019).

Figura 1 – Processo de predição de valores em ST.



Fonte: (PARMEZAN; SOUZA; BATISTA, 2019).

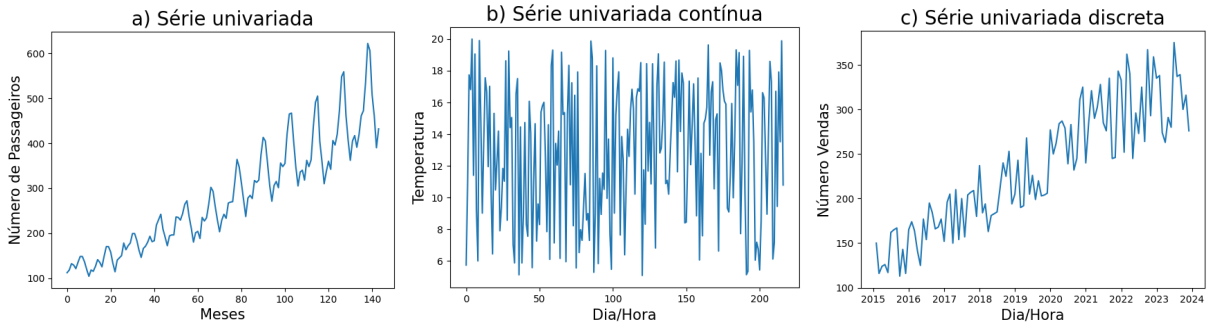
As séries temporais podem ser analisadas de diversas maneiras, levando em consideração as particularidades de suas variações, tais como a presença de tendências, sazonalidades, ciclos e comportamentos irregulares. Essas categorizações desempenham um papel fundamental na orientação das abordagens de análise e previsão, permitindo que os especialistas escolham os métodos adequados para cada contexto. Além disso, a definição dos parâmetros, a construção e o ajuste do modelo, bem como a previsão dos valores das séries temporais, estão intrinsecamente relacionados aos **Tipos de Previsões** (subseção 2.1.2) desejado para o determinado domínio de aplicação (PARMEZAN; SOUZA; BATISTA, 2019). Nesse contexto, os diferentes tipos de previsão, como a previsão de multi-etapas, previsão direta e previsão recursiva, oferecem uma gama de abordagens para antecipar eventos futuros, adaptando-se às necessidades específicas de cada situação.

2.1.1 Análises de Séries Temporais

Uma ST de tamanho m pode ser definida como uma sequência de dados $S = (s_1, s_2, ...s_m)$, em que $s_t \in \mathbb{R}^d$ representa um ponto coletado s no tempo t. A estrutura da série temporal é ser considerada **univariada** quando $d = 1$ ou **multivariada** quando $d > 1$. As observações univariadas são analisadas em função do tempo e consiste de registros únicos, enquanto que nas observações multivariadas são incluídas outras variáveis independentes na estrutura do modelo. Outra característica das séries temporais é quanto a temporalidade dos registros, definidas como **contínuas** ou **discretas** (HAMILTON, 2020). As observações contínuas são coletadas de modo que não há intervalos fixos entre pontos coletados, ou seja, as observações são registradas em intervalos indefinidos. As séries discretas são observações registradas em intervalos fixos e bem definidos, isto é, cada ponto é registrado em um período específico no tempo.

A Figura 2 demonstra exemplos de séries univariada, univariada contínua e univariada discreta. A Figura 2 (a) ilustra uma série temporal univariada discreta que registra mensalmente o número de passageiros de companhias aéreas na década de 50 nos Estados Unidos. A Figura (b) exemplifica o registro de um termômetro em um dado local, situação em que as leituras podem ocorrer a qualquer momento sem um intervalo definido. Outro exemplo é a Figura (c), uma série discreta referente a quantidade de vendas mensais de um produto de uma determinada loja. As vendas são registradas em intervalos regulares independentemente se houve ou não a venda do produto.

Figura 2 – Exemplos de séries temporais: a) Dados Passageiros de companhias áreas nos Estados Unidos; b) Registro de um termômetro; e c) Vendas mensais de um produto.



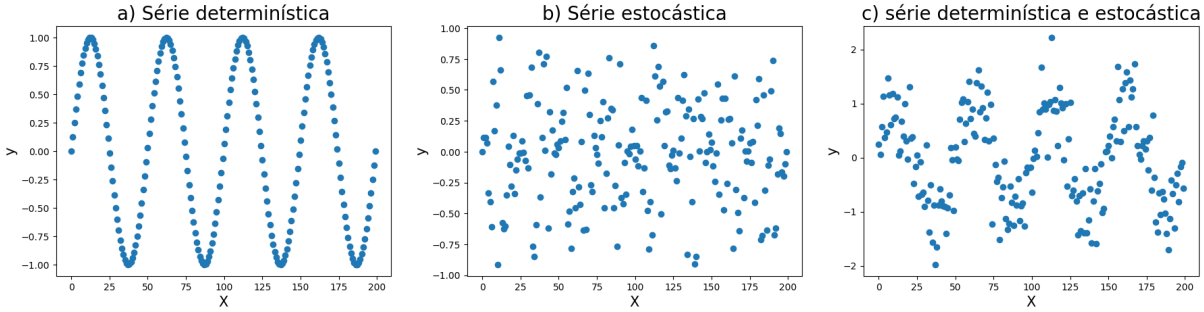
Fonte: Elaborado pelo Autor.

Além das características, a ST pode ser tipificada de acordo com seu determinismo ou estocasticidade; classificada quanto a linearidade ou estacionariedade; e, decomposta em componentes de Tendência, Sazonalidade, Ciclos e Ruídos (MONTGOMERY; JENNINGS; KULAHCI, 2015). Formada por observações $S = (s_1, s_2, ...s_m)$, uma ST é definida como **determinística** se um registro s_t depende única e exclusivamente de uma combinação de observações anteriores. Ou seja, os valores de uma ST são definidas por meio de uma função matemática $s_t = f(x)$. Por outro lado, se um registro s_t sofre influências de alguma variável aleatória ε_t, a série é definida como **estocástica** (Figura 3). Nesse caso, além do modelo matemático que considera o tempo,

um termo aleatório ε_t é considerado na função $s_t = f(x, \varepsilon_t)$.

A Figura 3 ilustra exemplos de uma série puramente determinística (a), uma série estocástica (b) e ambas tipificações (c), sendo: a) a série determinística é formada $f(x) = sen(s_t)$, em que cada s_t é um valor regularmente espaçado $(-4\pi, 4\pi)$; b) a série estocástica possui a influência de uma variável aleatória $\varepsilon(-1, 1)$; e, c) a série determinística estocástica é composta por uma distribuição normal para cada s_t, em que $\mu = 0$ e $\sigma = 0.5$. Um exemplo real de série determinística é o movimento de um pêndulo simples em um relógio de parede, em que ocorre uma trajetória previsível e consistente sob a influência da gravidade. Uma série estocástica pode ser exemplificada por variações diária das taxas de câmbio de moedas estrangeiras em relação a uma moeda nacional. Nesse caso, as flutuações das taxas de câmbio ocorrem devido a uma combinação aleatória de fatores econômicos, políticos, sociais e entre outros. Um exemplo de uma série determinística e estocástica é o ciclo de temperaturas ao longo do ano em uma região com quatro estações distintas. Nesse caso, as variações nas temperaturas seguem um padrão previsível e consistente à medida que as estações mudam.

Figura 3 – Tipificações de séries temporais. a) série determinística; b) série estocástica; e, c) série determinística e estocástica.



Fonte: Elaborado pelo Autor.

Independentemente do tipo, as séries temporais podem ser categorizadas de acordo com sua **estacionariedade**. Essa organização utiliza as relações implícitas entre as observações (ISHII; RIOS; MELLO, 2011). Uma ST é estacionária quando se desenvolve no tempo, aleatoriamente ao redor de uma média e variância constante, refletindo alguma forma de equilíbrio estável. A Figura 3 (a) exemplifica um cenário de série temporal **estacionária** com média constante que oscila no intervalo de 1 a -1. Em outros cenários, como por exemplo aplicações econômicas e financeiras, apresentam tendências positivas ou negativas e são conhecidas como **não estacionárias** com tendência linear ou não linear Figura 3 (c). Essas séries são caracterizadas pelo processo de Passeio Aleatório (do inglês, *Randon Walk*) (CRYER; CHAN; CHAN, 2008).

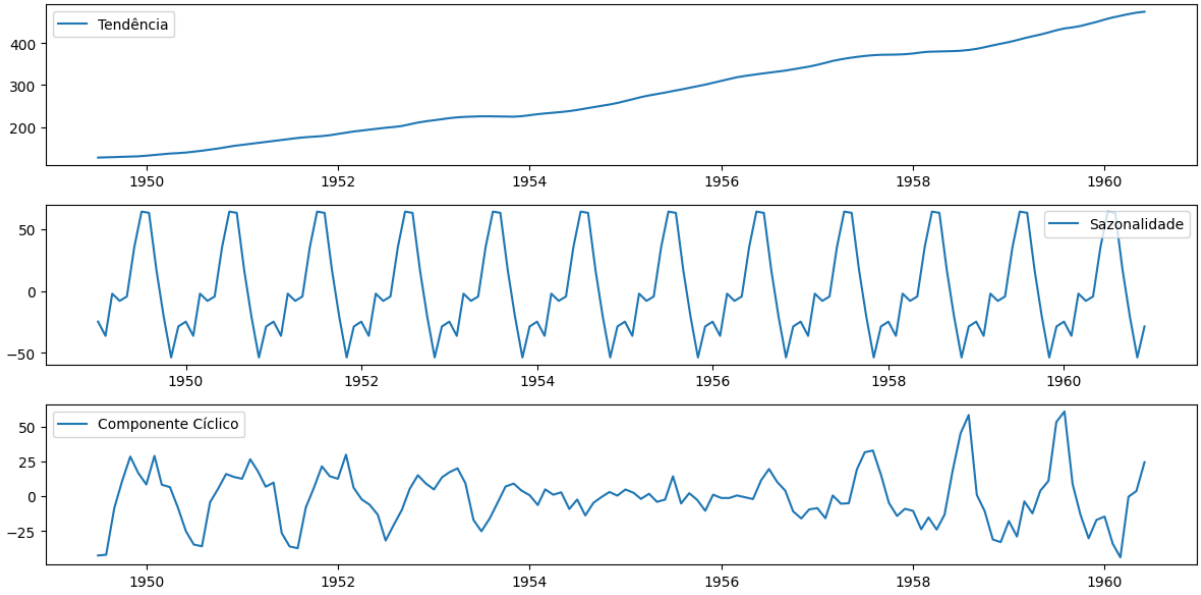
Uma série temporal pode ser genericamente decomposta em Tendência, Sazonalidade, Ciclo e Ruído. A série pode ser fundamentada de modelo aditivo, multiplicativo ou misto (PARMEZAN; SOUZA; BATISTA, 2019). No modelo aditivo, a série é vista como a soma dos diferentes componentes (tendência, sazonalidade, ciclo e ruído). O modelo multiplicativo é

visto como o produto dos componentes e o misto pode ter características tanto aditiva quanto multiplicativas. Por exemplo, o modelo aditivo pode ser caracterizada como

$$Y_t = T_t + S_t + C_t + E_t \tag{2.1}$$

em que Y_t é a série temporal, T_t captura a tendência, S_t a Sazonalidade e E_t uma irregularidade em função do tempo t. A **tendência** refere-se à direção em que série está evoluindo ao longo do tempo, podendo ser: crescente, decrescente ou constante. A **sazonalidade** são padrões que se repetem em intervalos fixos e conhecidos. Esses intervalos podem ser diário, semanal, mensal ou qualquer outra frequência previsível. O componente **cíclico** representa padrões repetitivos e não fixos que ocorrem em intervalos maiores do que a sazonalidade. Os ciclos não têm uma periodicidade fixa como a sazonalidade, e sua duração pode variar. O **ruído** é a variação aleatória presente na série que não pode ser atribuída a nenhuma das outras componentes (MONTGOMERY; JENNINGS; KULAHCI, 2015). Pode ser resultado de fatores imprevisíveis e externos. A Figura 4 apresenta a decomposição aditiva da série temporal referente aos dados ilustrados na Figura 2 (a)).

Figura 4 – Decomposição aditiva da série temporal de passageiros de companhias aéreas na década de 50.



Fonte: Elaborado pelo Autor.

A Figura 4, a linha do componente de tendência mostra uma tendência crescente ao longo dos anos. Isso sugere que, em média, o número de passageiros de companhias aéreas está aumentando consistentemente com o tempo. Quanto à sazonalidade, as flutuações periódicas nos dados mostram que existem padrões ao longo dos anos. Os picos positivos, que ocorrem aproximadamente no meio do ano, indicam que sazonalidade é anual e que o número de passageiros atinge seu ponto máximo durante esse período no ano. Em relação ao componente cíclico, pode-se observar que as variações ao longo do tempo não se encaixam nos padrões de

tendência ou sazonalidade. Isso pode ser influenciado por fatores econômicos, eventos atípicos ou outras influências que afetam o número de passageiros de forma não sazonal. O componente ruído não é exemplificado devido fatores imprevisíveis que não são decompostos nas séries.

2.1.2 Tipos de previsões

Uma das complexidades enfrentadas pelos pesquisadores na tarefa de previsão de séries temporais está na determinação dos parâmetros e na configuração experimental para a realização das previsões. A configuração dos experimentos pode abranger tanto situações do mundo real quanto cenários concebidos exclusivamente para avaliar o desempenho dos modelos preditivos. Nesse contexto, as previsões podem ser categorizadas como multi-etapas **recursiva**, **direta** e **direta recursiva** (SORJAMAA *et al.*, 2007). Cada uma dessas abordagens tem suas vantagens e desvantagens, e a seleção entre elas depende das características dos dados e das metas de previsão.

As previsões multi-etapas são aquelas em que as saídas previstas podem retroalimentar as entradas futuras do modelo. Essa retroalimentação das previsões para as entradas subsequentes é uma característica distintiva das previsões multi-etapas . Essa abordagem é mais frequentemente aplicada a séries temporais univariadas, uma vez que, no caso de séries multivariadas, nem todas as séries presentes nas entradas (X) são previstas para os períodos futuros de previsão. Isso ocorre devido à natureza da retroalimentação, que pode limitar a capacidade de prever múltiplas séries ao mesmo tempo, uma vez que as previsões de uma série podem influenciar as previsões de outras séries de maneira complexa (TAIEB *et al.*, 2012).

A previsão multi-etapas em séries temporais univariadas, com T observações, tem como principal objetivo estimar as observações futuras $y_1, y_2, ..., y_{T+h}$, em que h representa o horizonte da previsão. O valor de h geralmente é categorizado como curto, médio ou longo prazo, dependendo do contexto. Enfrentando esse desafio de previsão multi-etapas, a escolha da estratégia de previsão mais adequada para uma situação específica requer uma consideração cuidadosa (SORJAMAA *et al.*, 2007). A abordagem **recursiva** se baseia na criação de um único modelo (f), visando minimizar a variância do erro em h passos no futuro. Para ilustrar essa estratégia, considere o exemplo de previsão a 3 passos à frente ($h = 3$), que pode ser demonstrado por meio das equações a seguir:

$$\begin{aligned} \hat{y}(t+1) &= f(y(t), y(t-1), y(t-2), y(t-3)) \\ \hat{y}(t+2) &= f(\hat{y}(t+1), y(t), y(t-1), y(t-2)) \\ \hat{y}(t+3) &= f(\hat{y}(t+2), \hat{y}(t+1), y(t), y(t-1)) \end{aligned} \tag{2.2}$$

em que $\hat{y}(t+h)$ representa a variável a ser prevista. Nos passos subsequentes, a variável $\hat{y}(t+h)$ é retroalimentada e utilizada como entrada para a próxima previsão. Uma vantagem notável dessa estratégia é que somente um modelo é necessário, resultando em economia de tempo

computacional (TAIEB; HYNDMAN, 2014). Entretanto, uma desvantagem significativa é a sensibilidade a erros acumulados ao longo do horizonte de previsão. Um exemplo de aplicação para a estratégia recursiva é a previsão de demanda de energia elétrica. Nesse cenário, um único modelo é construído para prever os valores futuros da demanda de energia em várias etapas à frente. As previsões anteriores são retroalimentadas como entradas para prever as etapas subsequentes, considerando a dependência temporal da série. Isso pode ajudar as empresas de energia a planejar de forma mais eficiente a geração e distribuição de energia, otimizando recursos e minimizando custos.

A estratégia **direta** se diferencia da estratégia recursiva por não considerar a retroalimentação das previsões anteriores para as etapas futuras. Na estratégia direta, sempre são utilizados os dados reais como entrada, o que impede a acumulação de erros ao longo do horizonte de previsão (TAIEB; HYNDMAN, 2014). No entanto, essa abordagem exige a adaptação de diferentes modelos de previsão (f_i) para cada etapa futura. Um exemplo da estratégia direta para previsão de 3 passos ($h = 3$) à frente pode ser ilustrado pelas seguintes equações:

$$\begin{aligned}
\hat{y}(t+1) &= f_1(y(t), y(t-1), y(t-2), y(t-3)) \\
\hat{y}(t+2) &= f_2(y(t), y(t-1), y(t-2), y(t-3)) \\
\hat{y}(t+3) &= f_3(y(t), y(t-1), y(t-2), y(t-3))
\end{aligned} \tag{2.3}$$

A estratégia direta é frequentemente utilizada em cenários em que a acumulação de erros ao longo do horizonte de previsão é uma preocupação (SORJAMAA *et al.*, 2007). Um exemplo de aplicação para essa estratégia é a previsão de vendas em varejo. Nesse contexto, cada modelo f_i é projetado para prever diretamente o valor futuro das vendas em um determinado passo à frente. Isso permite que a previsão seja feita sem a necessidade de retroalimentação das previsões anteriores, reduzindo o potencial de acumulação de erros. Cada modelo é projetado para lidar com a relação específica entre as variáveis de entrada e a variável de saída em uma etapa de previsão específica. Isso proporciona uma abordagem mais adaptativa, permitindo que diferentes modelos sejam usados para diferentes horizontes de previsão.

A estratégia híbrida **Direta-Recursiva** é uma abordagem que combina vantagens das estratégias direta e recursiva. Nessa estratégia, são construídos modelos distintos (f_i) para cada etapa de previsão, levando em consideração as previsões realizadas nas etapas anteriores como valores de entrada (TAIEB; HYNDMAN, 2014). Isso significa que cada modelo é adaptado para estimar a previsão em um passo à frente, usando informações tanto das previsões anteriores quanto dos valores reais observados. A estratégia híbrida pode ser fundamentada pelas equações:

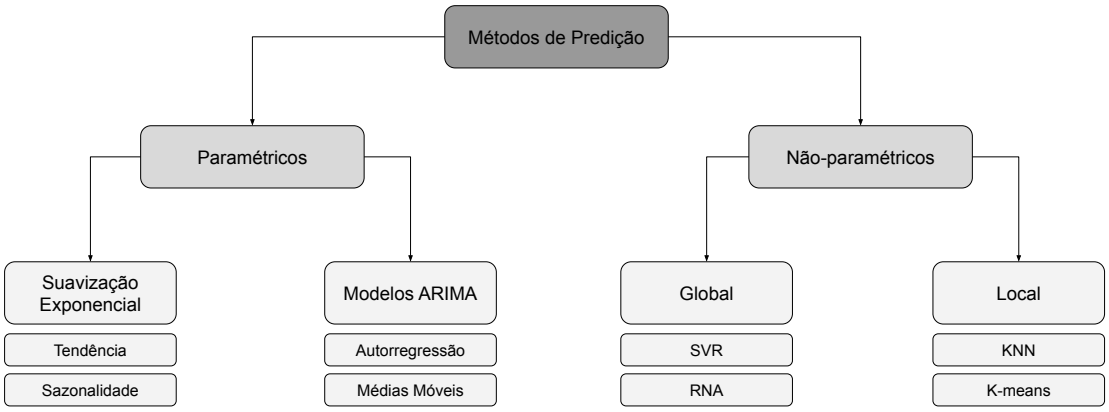
$$\begin{aligned}
\hat{y}(t+1) &= f_1(y(t), y(t-1), y(t-2), y(t-3)) \\
\hat{y}(t+2) &= f_2(\hat{y}(t+1), y(t), y(t-1), y(t-2)) \\
\hat{y}(t+3) &= f_3(\hat{y}(t+2), \hat{y}(t+1), y(t), y(t-1))
\end{aligned} \tag{2.4}$$

em que f_i é o modelo que faz previsões com valores retroalimentados ($\hat{y}_j$). Essa abordagem visa mitigar o erro acumulativo ao longo do horizonte de previsão, presente na estratégia recursiva, e aproveitar a precisão de curto prazo da estratégia direta. Essa combinação de características pode resultar em previsões mais confiáveis e precisas em séries temporais (TAIEB; HYNDMAN, 2014).

2.2 Modelos Preditivos

Métodos para predição de séries temporais apresentaram notável evolução nos últimos anos, desde técnicas mais simples de regressão até algoritmos mais sofisticados no campo da Estatística e da Inteligência Artificial. Cada modelo preditivo depende do conhecimento prévio da distribuição dos dados, os quais podem ser agrupados em abordagens paramétricas (Suavização Exponencial e modelos com base em ARIMA) e não paramétricas (Global ou Local) (CHATFIELD; XING, 2019), ilustrados na Figura 5.

Figura 5 – Hierarquia de abordagens para predição de ST.



Fonte: adaptado de Parmezan, Souza e Batista (2019).

Os modelos de **Suavização Exponencial** são caracterizados por decompor a ST em componentes (tendência, sazonalidade e nível), cujos valores são atribuídos por pesos exponencialmente decrescentes, conforme a observação se torna mais antiga. Em outras palavras, observações recentes são tomadas com mais peso relativamente na previsão do que as observações mais antigas (PARMEZAN; SOUZA; BATISTA, 2019). Os **Modelos ARIMA** (Autorregressivos Integrados de Médias Móveis) são modelos estatísticos utilizados para analisar e prever séries temporais. Eles combinam componentes de autoregressão (AR), diferenciação (I) e médias móveis (MA). O ARIMA atribui diferentes pesos às observações passadas, com base na estrutura do modelo e na autocorrelação dos dados. Esses pesos são determinados durante a etapa de estimação do modelo e podem variar de acordo com o tipo do ARIMA e as características específicas da série temporal (MONDAL; SHIT; GOSWAMI, 2014).

Os métodos não paramétricos, em que se destacam os métodos de Aprendizado de Máquina (AM), buscam descrever as propriedades dos dados sem o conhecimento prévio da distribuição dos mesmos. Por não dependerem de parâmetros específicos para modelar o comportamento do fenômeno, essas abordagens demonstram considerável desempenho mesmo quando aplicados à séries complexas e não-lineares. Segundo Tsay e Chen (2018), pelo modo como as observações da ST são aproveitadas no modelo preditivo, os métodos de AM podem ser divididos em duas abordagens: Global e Local.

Na abordagem **Global**, os métodos de AM constroem modelos a partir de um procedimento de treinamento que recebe como entrada todas as observações da série (ISLAM; SIVAKUMAR, 2002). Exemplos de modelos não paramétricos globais incluem Redes Neurais Artificiais (RNA), Support Vector Regression (SVR), Long Short-Term Memory (LSTM), entre outros. Na abordagem **Local**, os métodos dividem a série temporal original em subsequências menores, chamadas de janelas ou vizinhanças (WU; LEE, 2015). Essas janelas contêm um conjunto limitado de observações, geralmente em torno de um ponto de interesse na série temporal. Exemplos de métodos que seguem essa abordagem incluem: i) Médias Móveis: Calculam a média dos valores em uma janela específica e usam essa média como estimativa para o próximo valor da série temporal; ii) Regressão Local: Ajusta modelos de regressão simples dentro de janelas de dados específicas para prever o próximo valor da série; iii) K-Vizinhos Mais Próximos (K-Nearest Neighbors - KNN): Estima o próximo valor da série baseado nos valores das k observações mais próximas na série temporal (ISLAM; SIVAKUMAR, 2002). Esses métodos são úteis para capturar padrões locais e não lineares nos dados, especialmente quando a série temporal exibe comportamentos não estacionários ou variações complexas ao longo do tempo .

2.2.1 Modelos Paramétricos

Os modelos paramétricos desempenham um papel fundamental na previsão de séries temporais, constituindo uma abordagem amplamente utilizada para analisar e estimar padrões de comportamento ao longo do tempo. Os trabalhos mais relevantes para análises paramétricas de séries temporais foram desenvolvidos por Box *et al.* (2015), cujos métodos projetam uma classe com quatro modelos, denominados como "Auto-Regressivos Integrados de Médias Móveis" (ARIMA). Os modelos representam relações de dependência entre as observações de uma série temporal por efeitos da influência de processos estocásticos, sendo possíveis representar ST estacionárias e não-estacionárias.

O modelo ARIMA (p,d,q) resulta da combinação de três procedimentos estatísticos (BOX *et al.*, 2015): i) autorregressão (AR(p)); ii) *integração*[1]; e, Médias Móveis (MA(q)). O modelo autorregressivo AR(p) de ordem p é indicado para séries temporais estacionárias. De acordo com esse modelo, o valor da observação atual s_t é definido por meio de um sistema linear

[1] Integração é o nome dado à operação de diferenciação, a qual consiste em tomar diferenças sucessivas da série original $S = (s_1, s_2, ..., s_m)$. A primeira diferença é denotada por $\Delta_{S_t} = s_t - s_{t-1}$

finito de observações prévias, conforme definido na Equação 2.5.

$$s_t = \phi_1 s_{t-1} + \phi_2 s_{t-2} + ... + \phi_p s_{t-p} + a_t \tag{2.5}$$

em que os valores $\phi_1, \phi_2, ...\phi_p$ correspondem aos pesos do modelo para observações prévias, a_t compreende o ruído branco em uma distribuição com média zero e variância constante, e o termo p corresponde o número de observações anteriores que devem ser levadas em consideração.

O modelo de médias móveis de ordem q, denotado por MA(q), admite que a observação atual é formada pela média ponderada das q observações anteriores. Dessa forma, o valor da variável atual s_t é definida por um sistema linear finito das observações prévias s_t, conforme definido na Equação 2.6.

$$s_t = s_t + \theta_1 s_{t-1} + \theta_2 s_{t-2} + ... + \theta_p s_{t-q} \tag{2.6}$$

no qual os valores $\theta_1, \theta_2, ...\theta_p$ correspondem aos coeficientes do modelo para os valores prévios, em que a soma é igual a 1, e o termo q determina o número de observações passadas que devem ser levadas em consideração.

O modelo autorregressivo de médias móveis de ordem p, q é denominado ARMA(p, q). Esse modelo é formado pela união do modelo Autorregressivo AR(p) e o de médias móveis MA(q), sendo utilizado para séries cujo valor de uma variável no instante t é definida em função de valores defasados da mesma variável em instantes anteriores (BOX *et al.*, 2015), definido na Equação 2.7.

$$s_t = \phi_1 s_{t-1} + .. + \phi_p s_{t-p} + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} ... + \theta_q a_{t-q} \tag{2.7}$$

o valor atual da observação s_t é definido a partir de um sistema linear finito de observações prévias AR(p), somado a outro sistema linear finito dos valores prévios de ruídos aleatórios a_t MA(q),em que ϕ e θ são os coeficientes do modelo autorregressivo e de médias móveis, respectivamente.

O modelo autorregressivo, integrado e de médias móveis **ARIMA** (p, d, q) é usado para **representar séries temporais não estacionárias com tendência linear** (positiva ou negativa) (BOX *et al.*, 2015). Uma série estacionária pode ser definida como w_t, sendo a diferença de ordem d da série não estacionária s_t, definida na Equação 2.8.

$$w_t = \Delta^d s_t \tag{2.8}$$

desse modo, s_t é definida como uma integral de w_t, e portanto, caracteriza-se que s_t segue um modelo autorregressivo, *integrado*, de médias móveis de acordo com a Equação 2.9.

$$s_t = \phi_1 s_{t-1} + .. + \phi_p w_{t-p} + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} ... + \phi_q a_{t-q} \tag{2.9}$$

em que $w_t = \Delta^d s_t = \Delta(\Delta^{d-1} s_t)$, o qual d indica o grau do operador de diferença, ϕ_p e θ_q são os coeficientes do modelo autorregressivo com comprimento de defasagem p, e de médias móveis com comprimento de defasagem q, respectivamente. Em resumo, o uso do ARIMA assume que a d-ésima diferença entre as observações da série pode ser representada por um processo estacionário capaz de ser estimado por um modelo ARMA. Dessa forma, a ST que apresenta tendência não-explosiva, ou seja, não-estacionariedade homogênea, assim como estacionárias podem ser modeladas pelo ARIMA.

2.2.2 Modelos não Paramétricos

Os modelos paramétricos têm demonstrado resultados de boa precisão, mas eles enfrentam dificuldades quando aplicados a situações de previsão de séries que não seguem padrões comuns, como é o caso das séries financeiras. Isso se deve, em grande parte, à presença de um alto nível de ruído, à natureza caótica e não linear inerente a esse tipo de dado. Recentemente, a literatura tem visto a emergência de modelos não paramétricos como alternativas para enfrentar as limitações dos modelos paramétricos. Dentre essas alternativas, destacam-se modelos preditivos como a Regressão de Vetores de Suporte (SVR - *Support Vector Regression*) (DRUCKER *et al.*, 1997), Redes Neurais Artificiais (RNA) (YEGNANARAYANA, 2009) e abordagens baseadas em *Transformers* aplicadas à séries temporais (LIM *et al.*, 2021). Estudo que utilizam de modelos não paramétricos têm demonstrado resultados promissores na tarefa de previsão em diversos domínios (VERMA; SAHU; SAHU, 2023).

2.2.2.1 Support Vector Regression

Os modelos de regressão têm como objetivo estimar a relação entre uma variável dependente (preditora) e variável(eis) independente(s) (características). A variável independente é um dado observado e frequentemente representado por um vetor X_i (onde i denota uma observação nos dados), enquanto a variável dependente é um dado observado e denotado como o escalar Y_i. Parâmetros desconhecidos são comumente representados por um vetor β, e os termos correspondentes aos erros observados nos dados são denotados pelo escalar e_i. O relacionamento linear da variável Y_i em relação a X_i pode ser estabelecido por meio de uma função $f(X)$. O modelo da equação $Y = f(X)$ é considerado simples (no caso de X_i ser univariado) quando representa uma relação causal entre duas variáveis (Equação 2.10), e é considerado multivariado quando envolve duas ou mais variáveis (Equação 2.11). Certos modelos de regressão propõem que Y_i é

uma função de X_i e β, enquanto e_i representa um termo de erro que pode incorporar valores não modelados de Y_i ou ruído estatístico aleatório (DRUCKER *et al.*, 1997).

$$Y_i = \beta_0 + \beta_1 X + e_i \tag{2.10}$$

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e_i \tag{2.11}$$

em que X_i são amostras de dados (ou s_t especificada na Seção 2.1.1 - Análises de Séries Temporais), Y_i o valor para ser predito. O desafio dos modelos de regressão em métodos preditivos é estabelecer uma função $f(X)$ que melhor ajusta os dados e o resultado no valor mais preciso de Y_i, mostrado na Equação 2.12.

$$Y_i = f(X_i, \beta) + e_i \tag{2.12}$$

A variável independente (X_i) da equação de regressão pode ser representada tanto por séries temporais quanto por dados de uma representação vetorial de textos (SMOLA; SCHÖLKOPF, 2004). Para situações em que o problema de regressão está inserido em um cenário caótico, o modelo de regressão não linear é mais apropriado (DRUCKER *et al.*, 1997; SMOLA; SCHÖLKOPF, 2004). O *Support Vector Regression* (SVR) é um modelo não linear semelhante com a Máquina de Suporte de Vetores (do inglês, Support Vector Machine (SVM)) usado para tarefas de classificação. A função ($f(BR_i)$) não linear usada para estimar um valor da série temporal Y_i é mostrada na Equação 2.13.

$$Y_i = f(BR_i) = \sum_{j=1}^{N} (\alpha_j - \alpha_j^*) K(X_j, X_i) + b \tag{2.13}$$

em que (b) representa o período, $K(,)$ representa a função kernel que transforma os dados em um espaço de alta dimensionalidade para permitir uma separação linear, α_j e α_j^* são os multiplicadores não negativos para cada observação X_j (também chamada variável dupla). A função Kernel $K(,)$ pode ser escolhida de acordo com as características do conjunto de dados (*datasets*). Os Kernels mais comuns são o Polinomial, Função de Base Radial (do inglês, Radial Base Function (RBF)) e a Função Sigmoid, conforme apresentado na Tabela 1.

Tabela 1 – Funções do Kernel.

Kernel Name	**Kernel Function**
Polynomial	$K(x_j, x_k) = (1 + x_j x_k)^q$
RBF	$K(x_j, x_k) = exp(-\gamma \lvert\lvert x_j - x_k \rvert\rvert^2)$
Sigmoid	$K(x_j, x_k) = tanh(\gamma(x_j, x_k) + r)$

Segundo Drucker *et al.* (1997), o método SVR precisa ser ajustado para que os multiplicadores α_j e α_j^* sejam estimados no processo de otimização do SVR, conforme a Equação 2.14 que representa a função objetivo para ser minimizada.

$$L(\alpha) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(X_i, X_j) + \varepsilon\sum_{i=1}^{N}(\alpha_i - \alpha_i^*) - \sum_{i=1}^{N} y_i(\alpha_i - \alpha_i^*) \tag{2.14}$$

em que

$$\begin{aligned} &\sum_{i=1}^{N}(\alpha_n - \alpha_n^*) = 0 \\ &\forall n : 0 \leq \alpha_n \leq C \\ &\forall n : 0 \leq \alpha_n^* \leq C \end{aligned} \tag{2.15}$$

no qual K é o Kernel, ε define uma margem de tolerância o qual nenhuma penalidade é dada a erros de previsão; e C é uma constante positiva previamente definida que controla a penalidade para observações que excedem a margem, o que também contribui para evitar o *overfitting* excessivo. Técnicas de otimização de programação quadrática podem resolver o problema de minimização do método SVR não linear. Uma alternativa é utilizar a SMO (do inglês, *Sequential Minimal Optimization*) apresentado por Drucker *et al.* (1997), frequentemente usada em problemas relacionados ao SVR.

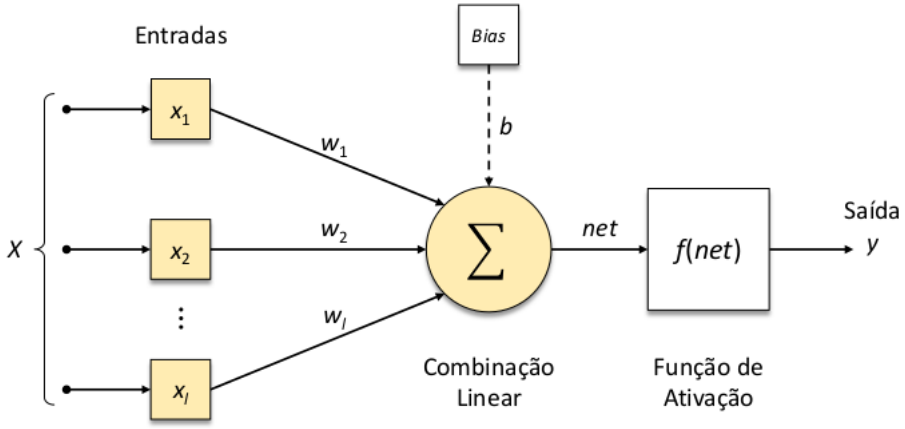
2.2.2.2 Redes Neurais Artificiais (RNA)

As Redes Neurais Artificiais (RNAs) têm ganhado destaque nas tarefas de previsão de séries temporais devido à sua capacidade de modelar e prever padrões não lineares. Proposto por Rosenblatt (1958), a estrutura base da RNA foi inspiradas no funcionamento do cérebro humano, consistindo em uma técnica de aprendizado de máquina que visa replicar a maneira como o cérebro realiza tarefas específicas ou funções de interesse. Essa abordagem permite que as RNAs contenham unidades de processamento interconectadas, semelhantes às unidades neurais no cérebro. Essas conexões geralmente possuem pesos associados, que desempenham o papel de armazenar informações aprendidas a partir dos padrões presentes nos dados. Em grande parte dos casos, esses pesos funcionam como uma forma de memória para a rede (HAYKIN, 2007).

Proposto para representar um neurônio biológico, um neurônio artificial (ou *Perceptron*) consiste fundamentalmente em um conjunto de entradas (X) e uma saída (y), juntamente com um núcleo de processamento que opera de acordo com uma regra ou função matemática específica, chamada de **Função de Ativação**. Cada entrada é multiplicada por um peso correspondente (w_I), resultando em entradas ponderadas. Essas entradas ponderadas são então somadas, gerando um valor denominado *NET*, que é posteriormente comparado com um limite de ativação determinado

para o neurônio. Se o valor de *NET* alcançar ou exceder o limite de ativação, o neurônio é ativado. O esquema do neurônio é ilustrado na Figura 6.

Figura 6 – Estrutura do Perceptron.



Fonte: (HAYKIN *et al.*, 2009 apud PARMEZAN; SOUZA; BATISTA, 2019).

Um neurônio singular recebe l entradas de dados, representadas como $x_i \in X$. Cada elemento i de X é associado a um peso sináptico w_i. Os pesos sinápticos podem ter valores positivos ou negativos, refletindo a importância relativa de cada entrada no processo de cálculo. A combinação linear das entradas ponderadas pelos seus respectivos pesos, somada a um limiar (*bias*) $b \in \mathbb{R}$, resulta no valor *net* conforme indicado na Equação 2.16. O valor calculado *net* é então fornecido à função de ativação f, que determina a saída y do neurônio.

$$net = \sum_{i=1}^{l} w_i x_i + b \tag{2.16}$$

O valor de *bias* (b) têm objetivo de corrigir, diminuir ou aumentar o valor de *net*. Para casos em que a classificação é considerada linearmente separável, f corresponde a uma função do tipo:

$$f(net) = \begin{cases} 1 & \text{se } net > 0 \\ 0 & \text{se } net \leqslant 0 \end{cases} \tag{2.17}$$

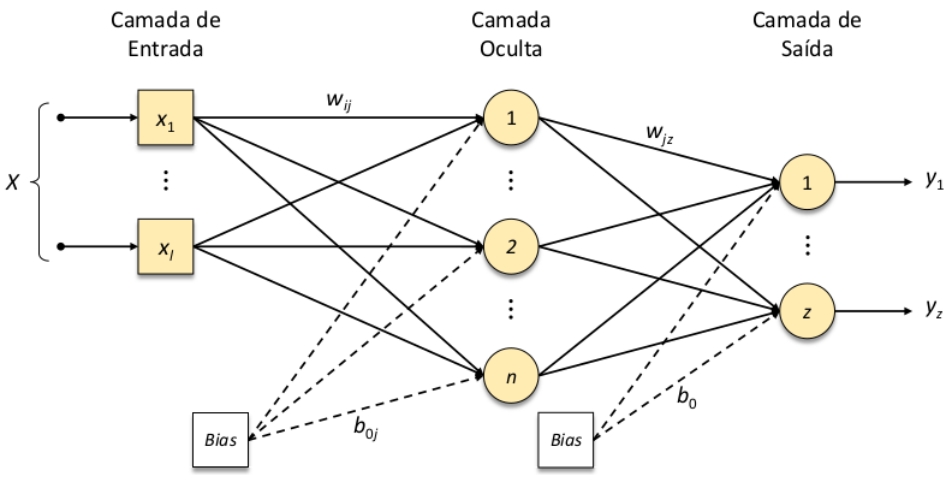
Em relação ao valor resultante de y, pode ser tanto binário $y \in \{0,1\}$ ou $y \in \{-1,1\}$, quanto contínuo, em que $y \in \mathbb{R}$. Além disso, outros tipos de função de ativação podem ser usadas. O aprendizado ocorre por meio da aplicação do algoritmo de convergência do *Perceptron*, uma regra de ajuste de erros. Os pesos sinápticos de cada *Perceptron* são adaptados usando um processo iterativo de aprendizado finito. O objetivo é encontrar um conjunto de pesos w que satisfaça as igualdades da função degrau (HAYKIN, 2007).

Devido a críticas sobre a limitação do *Perceptron* em resolver problemas não-linearmente separáveis (MINSKY; PAPERT, 1969), surgiram propostas para adicionar camadas intermediá-

rias nas estruturas das redes neurais. A adoção de modelos com mais de duas camadas ganhou impulso com a introdução do algoritmo de aprendizado *Backpropagation* (MCCLELLAND *et al.*, 1986). Esse algoritmo permitiu que redes neurais com camadas intermediárias fossem treinadas de maneira eficiente em duas etapas sequenciais. Primeiro, um padrão de entrada é apresentado à camada inicial de dados, e o processamento é propagado camada a camada pela rede até que a saída final seja produzida. Em seguida, o erro é calculado comparando a saída obtida com a saída desejada. Esse erro é retropropagado da camada de saída até a camada de entrada, e os pesos das conexões nas camadas intermediárias são ajustados usando a regra delta generalizada (HAYKIN *et al.*, 2009). Esse processo de retropropagação é repetido até que a rede convirja para um estado em que seja capaz de representar todos os padrões do conjunto de treinamento.

Um modelo de rede neural com múltiplas camadas treinado pelo algoritmo de *Backpropagation* é denominado de *Multilayer Perceptron* (MLP). A estrutura de uma rede *MLP* com três camadas é apresentada na Figura 7.

Figura 7 – Estrutura de uma rede MLP com uma camada oculta.



Fonte: (HAYKIN *et al.*, 2009 apud PARMEZAN; SOUZA; BATISTA, 2019).

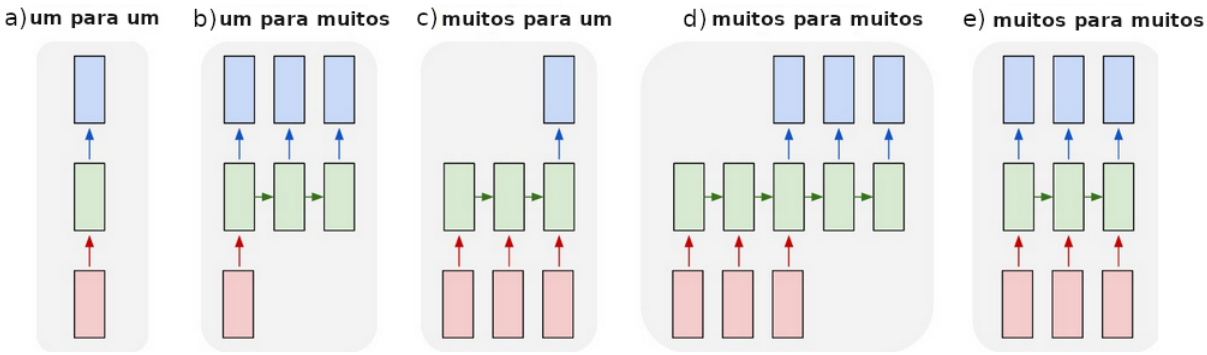
Considerado como um único neurônio na camada de saída, o z-ésimo elemento da rede *MLP* pode ser representada do seguinte modo:

$$y = f\left(\sum_{j=1}^{n} w_j f\left(\sum_{i=1}^{l} w_{ij}x_i + b_{0j}\right) + b_0\right) \tag{2.18}$$

No modelo esquemático do *MLP*, pode haver uma ou mais camadas ocultas entre as camadas de entrada e saída. A precisão do resultado da camada de saída está associada a três aspectos principais: *i*) determinação do número de camadas ocultas; *ii*) definição do número de neurônios em cada uma das camadas; e *iii*) especificação dos pesos sinápticos que interconectam os neurônios nas diferentes camadas da rede (HAYKIN *et al.*, 2009).

A MLP é um modelo simples destinado a realizar a classificação binária. No entanto, para tarefas de regressão e especialmente para a previsão de valores em séries temporais, as Redes Neurais Recorrentes (RNR) são mais indicadas (SALEHINEJAD *et al.*, 2017). As RNRs pertencem a uma classe de RNA em que os *perceptrons* ocultos recebem sinais tanto da camada de entrada quanto da camada oculta na iteração de tempo anterior. Isso permite que a camada oculta atue como uma memória, armazenando informações dos dados observados e recuperando informações do estado oculto anterior a cada período de tempo. Além disso, o estado oculto no mesmo período pode influenciar a camada de saída, caso seja o momento da previsão, e também fornece informações para o estado oculto do próximo período (GOODFELLOW; BENGIO; COURVILLE, 2016). As RNRs possuem a flexibilidade de processar diversas sequências de entrada e saída, com a escolha da arquitetura dependendo da aplicação específica do modelo. A Figura 8 ilustra as diferentes estruturas das RNRs utilizadas em tarefas de previsão.

Figura 8 – Cada retângulo é um vetor e as setas representam funções (por exemplo, multiplicação de matriz). Os vetores de entrada estão em vermelho, os vetores de saída estão em azul e os vetores verdes mantêm o estado oculto da RNN.



Fonte: Adaptado de Karpathy (2015).

Dentro das estruturas das Redes Neurais Recorrentes (RNRs), a abordagem “one to one” se refere a uma rede neural clássica (*feedforward*) com tamanho fixo de entrada e saída. Esse caso é comum em tarefas como classificação de imagens. Em aplicações “one to many” (Figura 8 (b)), os dados são usados para definir o estado oculto uma única vez, e esse estado é então utilizado para gerar saídas ao longo de vários períodos de tempo. Esse tipo de estrutura é observado, por exemplo, em modelos para descrever imagens (*image-captioning*). No contexto “many to one” na Figura 8 (c), os dados da sequência de entrada são processados para gerar uma previsão após a leitura completa da entrada. Isso é comum em tarefas como classificação de textos (análise de sentimento) e previsão de séries temporais.

Há também duas variações do caso “many to many”. Na primeira abordagem (Figura 8 (d)), os dados são lidos por um período antes de as previsões serem geradas. Nesse cenário, existe um atraso temporal entre a leitura da sequência de entrada e a produção das saídas. Isso é encontrado em tarefas como tradução de textos ou legendagem de vídeos, em que a RNR primeiro processa a entrada antes de começar a gerar a sequência de saída. O último tipo de “many to many” (Figura 8 (e)) ocorre quando há sequências na entrada e na saída da rede, e

cada entrada corresponde àuma saída no mesmo período de tempo. Essa estrutura é utilizada em séries temporais em que a previsão para o próximo período é baseada nos dados dos períodos anteriores, gerando uma saída para cada entrada (OLAH, 2021).

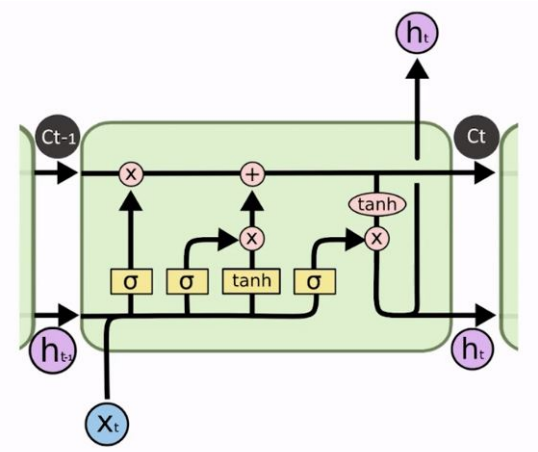
De modo resumido, o estado oculto da RNR recebe informações das variáveis independentes x_t e o próprio processamento do estado oculto em h_{t-1} e o peso w_h no período de tempo anterior. A Equação 2.19 representa uma RNR de estrutura "many to one".

$$\hat{y} = f\left(\sum_{j=1}^{n} h_j w_j f\left(\sum_{i=1}^{l} w_{ij} x_i + w_h h_{j-1} + b_h\right) + b_0\right) \tag{2.19}$$

Note que a diferença entre as Equações 2.18 e 2.19 é a informação do estado oculto do período anterior $w_h h_{t-1}$. Repare que os parâmetros que fazem a transição da informação entre os estados ocultos de diferentes períodos são sempre os mesmos. Isso significa que a RNR compartilham parâmetros através do tempo.

Consideradas como evoluções das RNRs, as redes do tipo *Long Short-Term Memory* (LSTM) foram desenvolvidas para resolver um problema comum, os gradientes que fazem a relação temporal "explodem" ou "desaparecem" ao longo do tempo (GREFF *et al.*, 2016). Para superar esse problema, os *perceptrons* das redes LSTM mantém uma célula exclusiva para o armazenamento e fluxo da memória, além de portões do tipo entrada (x_t), saída (h_t) e esquecimento que controlam esse fluxo (Figura 9). Uma variação da LSTM são as *Gated Recurrent Network* (GRU) (CHO *et al.*, 2014). Sua estrutura alteram os portões de entrada e saída por um portão de atualização, o qual é responsável por controlar o quanto de informação deve reter e quanto deve atualizar. No lugar do portão de esquecimento possui um portão de *reset*, cuja localização na estrutura é diferente da LSTM (DEY; SALEM, 2017).

Figura 9 – Estrutura básica da LSTM.



Fonte: (HOCHREITER; SCHMIDHUBER, 1997 apud GERS *et al.*, 2015)

As GRUs tem aplicação similar às redes LSTMs, porém mais rápidas e fáceis de treinar. Entretanto, o desempenho das GRUs podem ser menos expressivos do que as redes LSTMs (FU;

ZHANG; LI, 2016). Existem vários procedimentos e estruturas de previsão capaz de generalizar e mapear comportamentos não lineares pela geração dos dados de entrada (SEZER; GUDELEK; OZBAYOGLU, 2020a). Mesmo em situações aparentemente randômicas, tanto as RNA's como o SVR, quando utilizados para tarefas de previsão de séries temporais, assumem a premissa que os dados históricos podem ser uma base de conhecimento para previsão de dados futuros. No entanto, a configuração experimental deve ser cuidadosamente aplicada quando considerado diferentes tipos de previsão de séries temporais.

É importante destacar que existem outros modelos de previsão de séries temporais considerados estado da arte na literatura, como o DLinear (ZENG *et al.*, 2023), PatchMixer (GONG; TANG; LIANG, 2023), Temporal Fusion Transformers (TFT) (LIM *et al.*, 2021), e PatchTST (NIE *et al.*, 2022). No entanto, esses modelos não foram utilizados nas avaliações propostas na presente tese. Isso se deve à necessidade de uma exploração mais detalhada que está além do escopo desta tese. Pesquisas futuras podem integrar esses modelos para expandir e complementar os resultados apresentados na proposta da presente tese.

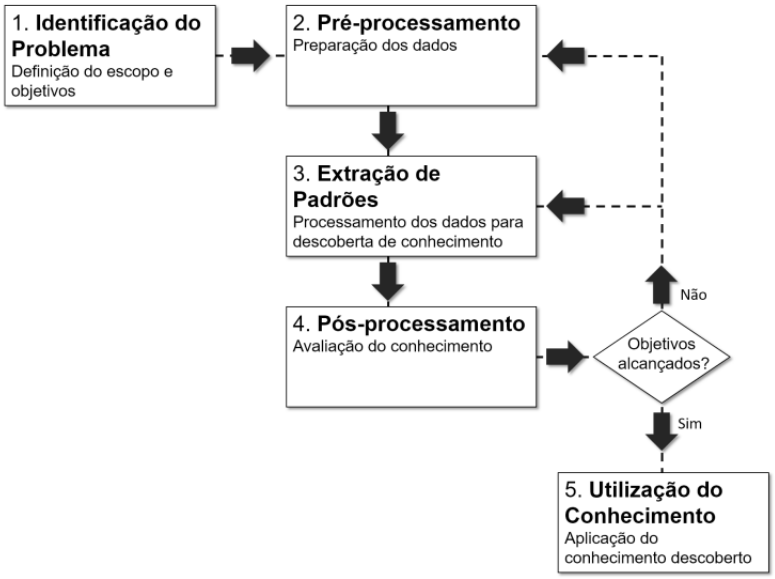
2.3 Mineração de Textos

A Mineração de Textos (MT), também conhecida como a mineração de dados textuais é considerada uma instância da Mineração de Dados, porém com objetivo de extrair informações importantes de textos de modo automatizado. A principal tarefa da MT é transformar textos em representações que sejam manipuláveis por modelos de aprendizado de máquina com finalidade de extrair padrões e tendências que podem ser úteis no processo de tomada de decisão (AGGARWAL, 2014).

De modo geral, as atividades da MT podem ser divididas em cinco etapas, que são: Identificação do Problema, Pré Processamento, Extração de Padrões, Pós-Processamento e Utilização do Conhecimento, conforme ilustradas na Figura 10. A etapa de **Identificação do problema** têm objetivo de guiar o processo de MT como um todo, e consiste da especificação dos objetivos e na delimitação do escopo da mineração. Além disso, serão identificadas as coleções de textos adequadas para o problema em questão, e como os resultados finais da mineração serão utilizados (SINOARA *et al.*, 2019). Essa etapa, em geral deve ser acompanhada por um especialista do domínio da aplicação, que irá direcionar os problemas recorrentes que devem ser solucionados.

A etapa de **Pré-processamento** preocupa-se na preparação dos dados para extração de padrões. Após definir o escopo e os objetivos do processo, nesta etapa busca-se colocar os dados em um formato adequado para a extração de conhecimento, normalmente realizadas atividades de tratamento, limpeza e redução do volume de dados disponível na base de dados. Na etapa de pré-processamento os documentos são representados de modo que podem ser processados por algoritmos de extração de padrões.

Figura 10 – Processos de Mineração de Textos.



Fonte: (SINOARA *et al.*, 2019) adaptado de (REZENDE, 2003).

Após a coleção de documentos estar devidamente formatada e tratada, inicia-se a etapa de **Extração de Padrões**. As tarefas a serem realizadas são definidas de acordo com o objetivo final do processo de extração de conhecimento. Na etapa de Extração de Padrões, aplica-se um algoritmo de aprendizado adequado para extrair padrões dos dados pré-processados. O algoritmo é escolhido com base nos dados disponíveis e no tipo de conhecimento que se deseja extrair. Nesse caso, quando o objetivo do processo de MT é o organizar a coleção de documentos, há duas tarefas que podem ser aplicadas: classificação e agrupamento. Por outro lado, quando o objetivo é auxiliar na predição de um valor quantitativo, utiliza-se tarefas de regressão.

Com os padrões extraídos da coleção de dados, a próxima tarefa é avaliar e interpretar os resultados na etapa de **Pós-processamento**. Assim como as outras etapas, a avaliação do conhecimento deve ser guiada pelos objetivos definidos no início do processo. Pode-se avaliar diversos aspectos do conhecimento extraído, como a representatividade, a interpretabilidade, a inovação, a aplicabilidade e validade. Recomenda-se que essa avaliação seja realizada em conjunto de com especialista do domínio, ou mesmo por meio da aplicação de medidas subjetivas.

Caso o conhecimento extraído atenda aos objetivos estabelecidos inicialmente, ou caso cumpra algum dos aspectos definidos anteriormente, considera-se que a MT foi bem sucedida, e que o conhecimento obtido está pronto para ser utilizado pelos usuários (etapa de **Utilização do Conhecimento**). Caso contrário, outro ciclo deve ser executado, realizando adaptações nas atividades de preparação dos dados ou nos parâmetros da extração de padrões. Mudanças nos objetivos estabelecidos na etapa de pré-processamento ou na identificação do problema podem ser realizadas, agregando mais informações em relação as etapas realizadas anteriormente.

O objetivo central do presente trabalho de Doutorado é a aplicação de técnicas avançadas de pré-processamento de textos e modelos de linguagem para enriquecer séries temporais com

informações textuais. Diante disso, nas seções subsequentes, discussões serão aprofundadas sobre estratégias particulares de pré-processamento textual e investigar modelos de **Representação de espaço vetorial**, englobando desde os **Modelos não baseados em contexto** (não semântico), até aquelas **com base em contexto** (semântico). Essas etapas são fundamentais para estabelecer a base sólida necessária para a fusão bem-sucedida de informações provenientes de fontes textuais com as séries temporais, visando aprimorar significativamente as capacidades de análise e previsão.

2.3.1 Representação de espaço vetorial

No intuito de transformar conjuntos de documentos em uma estrutura matricial atributo-valor, é adotada a abordagem de representação de documentos por meio de um modelo espaço-vetorial, assemelhando-se à maneira pela qual dados estruturados são representados. A premissa é retratar um documento como um ponto no espaço, definido por suas coordenadas vetoriais. Nesse contexto, a proximidade entre pontos no espaço vetorial denota a similaridade entre os documentos (AGGARWAL, 2014). Uma abordagem amplamente reconhecida é a *Bag-of-Words* (BoW), que representa a representação textual mais convencional, em que cada termo corresponde a uma palavra presente na coleção de documentos.

Representações tradicionais de textos, como a abordagem *Bag-of-Words* (BoW), são frequentemente representadas em forma de matriz documento-termo. Nesse contexto, considere uma coleção de documentos $D = [d_1, d_2, ..., d_{k-1}, d_k]$ composta por k documentos, e um conjunto de b termos presentes nessa coleção $T = [w_1, w_2, ..., w_{b-1}, w_b]$. A representação da coleção de documentos pode ser concebida como uma matriz documento-termo, composta pela combinação de k vetores, cada um com b dimensões. Essa estrutura é ilustrada de maneira esquemática na Figura 11.

Figura 11 – Ilustração da representação de espaço vetorial de k documentos e b termos como uma matriz documento-termo.

	w_1	w_2	...	w_{b-1}	w_b
d_1	pd_{1,w_1}	pd_{1,w_2}	...	$pd_{1,w_{b-1}}$	pd_{1,w_b}
d_2	pd_{2,w_1}	pd_{2,w_2}	...	$pd_{2,w_{b-1}}$	pd_{2,w_b}
⋮	⋮	⋮	⋱	⋮	⋮
d_{k-1}	pd_{k-1,w_1}	pd_{k-1,w_2}	...	$pd_{k-1,w_{b-1}}$	pd_{k-1,w_b}
d_k	pd_{k,w_1}	pd_{k,w_2}	...	$pd_{k,w_{b-1}}$	pd_{k,w_b}

Elaborado pelo Autor.

Os valores da matriz documento-termo correspondem ao peso de cada termo em cada documento, ou seja, p_{d_k,w_b} corresponde ao peso do termo w_j no documento d_i. Geralmente, os valores dos pesos são calculados com base na frequência dos termos nos documentos. As medidas mais comuns são: i) Fequência do Termo (do inglês, *Term Frequency* (TF)), que corresponde ao

número de vezes que o termo w_j ocorre no documento d_i; ii) frequência do termo ponderada pelo inverso da frequência de documento (do inglês, *Term Frequency-Inverse Document Frequency* (TF-IDF)); e, iii) binária, correspondente ao valor 1 ou 0, presença e ausência do termo do documento, respectivamente. No presente trabalho de Doutorado, é proposto um novo método de atribuição de peso do termo w_j para o documento d_i, apresentado na Seção 5.1.2. Essa abordagem, denominada TD-BERT (*Terms and Documents from pre-trained BERT models*) utiliza a estrutura tradicional da *BoW* e os recursos semânticos do modelo *Transformers* (FILHO *et al.*, 2022).

Além do cálculo dos pesos, existem diferentes possibilidades para definição dos termos. A *BoW* adota que os termos são independentes e a ordem com que eles ocorrem nos documentos não é levado em consideração. Uma alternativa, é manter a relação de ordem entre as palavras usando sequências de palavras (n-gramas). Nessa representação, cada termo é formado por uma sequência de n palavras que ocorrem pelo menos em um documento na coleção de textos. No entanto, o uso de n-gramas aumenta substancialmente o número de termos na representação matriz atributo valor (SINOARA *et al.*, 2019). Além disso, termos formados por mais de uma palavra são mais específicos do que termos formados por apenas uma palavra, oque ocorre na diminuição da frequência dos termos na coleção de documentos. Por exemplo, o termo "Agronegócio Brasileiro" será menos frequente que os termos independentes "Agronegócio" e "Brasileiro", uma vez que estes podem ocorrer em outros contextos. Esses fatores acentuam o problema da alta dimensionalidade e alta esparsidade, situações que já ocorrem na representação BoW.

As características apresentadas anteriormente fazem com que muitos dos métodos de extração de padrões sejam ineficientes ao lidar com representações de alta dimensionalidade e esparsas. Dessa forma, uma alternativa é reduzir o número de termos, usando algumas técnicas de pré-processamento, listadas a seguir:

- **Remoção de Stopwords**: visa eliminar palavras que não trazem informação relevante para o processo de Mineração de Textos. Geralmente, as funções de artigo, preposições, pronomes e conjunções são eliminadas na etapa de pré processamento. Outra possibilidade é identificar *stopwords* específicas do domínio de aplicação, os quais são frequentes na coleção de documentos e não oferece ganho de informação para uma determinada aplicação.

- **Normalização**: objetiva eliminar diversas variações que as palavras podem assumir, como por exemplo variações de gênero, número de substantivos e conjugações dos verbos. A normalização pode ser realizada por três técnicas: i) radicalização (*stemming*) que reduz cada palavra para o seu radical; ii) lematização, reduz cada palavra a seu lema (ou forma canônica), ou seja, substantivos e adjetivos são reduzidos a forma masculina singular, e verbos são reduzidos ao infinitivo; e, iii) substantivação, transforma a palavra para que ela

tenha o comportamento sintático/semântico semelhante de um substantivo.

- **Seleção de atributos**: visa selecionar os atributos mais relevantes da coleção de documentos, tornando o conjunto de atributos mais conciso, mas não menos representativo em relação ao conjunto original. Pretende-se identificar os atributos que são importantes para o problema, selecionando os que tem maior influência na definição da classe e/ou eliminando os atributos redundantes e com alto grau de ruído. A seleção dos atributos podem ser realizada por uma medida de avaliação com base em *ranking* ou usando um corte limiar.

As técnicas de pré-processamento de textos reduzem a dimensionalidade e mantém os termos como atributos da representação. Outras técnicas podem ser usadas para reduzir a dimensionalidade da representação, mas essas abordagens podem gerar novos atributos e não garanten uma relação explícita com termos da representação inicial. Rossi (2016) classifica essas técnicas em duas categorias: extração de atributos e extração de tópicos. Uma das principais técnicas de extração de atributos refere-se o *Latent Semantic Indexing* (LSI) e as técnicas de extração de tópicos é o *Latent Dirichlet Allocation* (LDA) (AGGARWAL, 2014). Por meio das técnicas LSI e LDA, formas alternativas de expressar o mesmo conteúdo são reduzidas a uma representação comum. Com isso, além de reduzir a dimensionalidade, esforços também são realizados para considerar a semântica dos textos. Apesar de remover ruídos causados por sinônimos e termos polissêmicos, os textos ainda são tratados como um conjunto de palavras independentes e desordenadas (SINOARA *et al.*, 2019). Dessa forma, os relacionamentos semânticos entre as palavras contidas nos textos não são representados. Para superar essas limitações, representações semânticas com recursos linguísticos, os *text embeddings*, ganhou muita atenção em semântica distribucional na última década. Essas representações são apresentadas a seguir.

2.3.2 Representações independentes de contexto

Em aplicações que demandam um nível considerável de compreensão linguística, uma abordagem frequentemente adotada é a representação de textos por meio de modelos de linguagem estatísticos. Esses modelos atribuem uma probabilidade P a uma sequência de palavras $w_1, w_2, ..., w_k$. A maneira mais simples de calcular $P(w_1, w_2, ..., w_k)$ é empregar a regra da cadeia à sequência, como demonstrado na Equação 2.20 (AGGARWAL; AGGARWAL, 2018):

$$\begin{aligned} P(w_1, w_2, ..., w_b) &= P(w_1) \cdot P(w_2|w_1) \cdot P(w3|w_1, w_2) \cdot ... \cdot P(w_b|w_1, w_2, ...w_{b-1}) \\ &= \prod_{i=1}^{b} P(w_i|w_1, ..., w_{i-1}) \end{aligned} \tag{2.20}$$

A distribuição de probabilidade P é calculada usando a contagem (*cont*) da ocorrência de termos de frase no conjunto de treinamento (Equação 2.21)

$$P(w_i|w_1,...,w_{i-1}) = \frac{P(w_1,...w_i)}{P(w_1,...w_{i-1})} = \frac{cont(w_1,...w_i)}{cont(w_1,...w_{i-1})} \tag{2.21}$$

Um problema do modelo estatístico é a contagem do grupo de termos é difícil de ser estimada com precisão para grupos grandes de palavras, isso pode resultar em um numerador e denominador próximos de 0. Para minimizar este problema, uma estratégia é considerar o pressuposto de *Markov*, que considera apenas os últimos $n-1$ *tokens* para estimar a probabilidade condicional de um *token*, resultando em um modelo *n-gram*. Dessa forma, o pressuposto *Markoviano* pode ser estimado conforme descreve a Equação 2.22

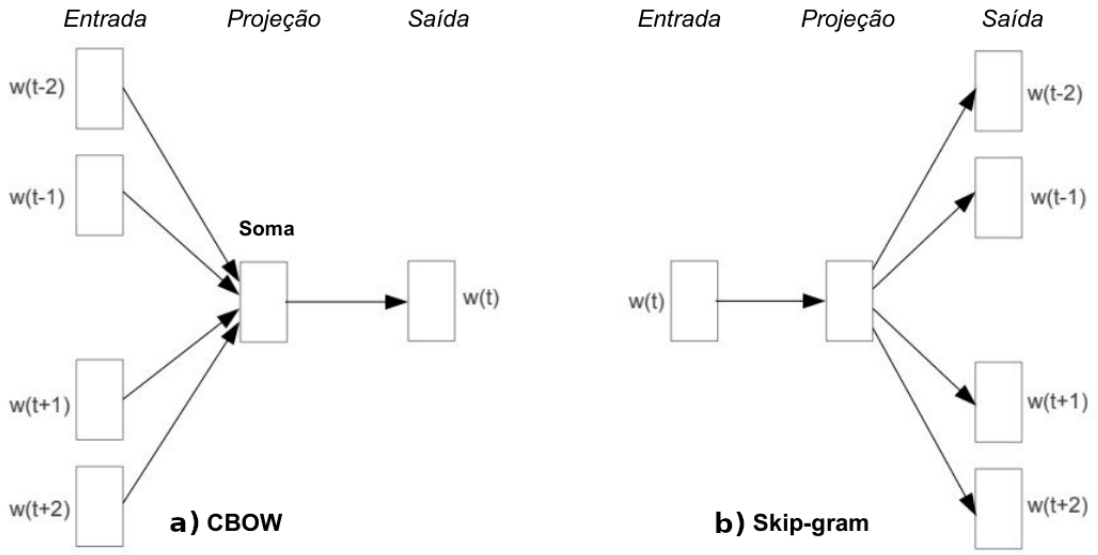
$$P(w_i|w_1,...,w_{i-1}) \approx P(w_i|w_{i-n+1},...,w_{i-1}) = \frac{cont(w_{i-n+1},...w_i)}{cont(w_{i-n+1},...w_{i-1})} \tag{2.22}$$

Tendo em vista que a grande quantidade de valores de n proporcionam melhores discriminações, a quantidade de dados disponíveis geralmente não é suficiente para que estimativas confiáveis sejam feitas. Além disso, em modelos em *n-gramas*, pequenas variações de uma sentença podem causar grandes efeitos na estimativa do modelo.

A fim de superar as limitações expostas anteriormente, na literatura encontram-se abordagens que permitem gerar *embeddings* para todas as palavras contidas em um documento. Os autores Bengio, Ducharme e Vincent (2000) propuseram pela primeira vez um modelo *word embeddings* com base em redes neurais. Este estudo enfrentou a complexa tarefa de modelagem estatística de linguagem, lidando com desafios inerentes à maldição da dimensionalidade ao aprender representações distribuídas para palavras. Essas representações permitiram que informações de sentenças de treinamento influenciassem um amplo espectro de sentenças semanticamente relacionadas, resultando em melhorias substanciais sobre modelos de n-gramas convencionais. Os resultados destacaram a eficácia das redes neurais na função de probabilidade e na exploração de contextos mais amplos.

Um modelo neural de linguagem bem consolidado é o *Word2Vec* (MIKOLOV *et al.*, 2013), treinado em duas etapas. Na primeira etapa, vetores de palavras contínuas são apreendidos por meio de um modelo simples. Na segunda etapa, um modelo de linguagem neural *n-gram* é treinado a partir das representações distribuídas de palavras (BENGIO; DUCHARME; VINCENT, 2000). O *Word2Vec* realiza o treinamento de *embeddings* considerando duas variantes: *Skip-gram* e *Continuous Bag-of-Words* (CBOW). CBOW computa a probabilidade condicional de uma palavra alvo ocorrer dado as palavras que estão próximas, isto é, observa o contexto em que a palavra está inserida. O *Skip-Gram* procura prever as palavras que estão ao redor dessa palavra, ou seja, tenta prever o contexto de ocorrência da palavra alvo (Figura 12). Segundo os autores, a principal contribuição do trabalho é a possibilidade de treinar *word embeddings* usando grande conjunto de dados com bilhões de palavras e baixo custo computacional.

Figura 12 – Variantes do modelo Word2Vec. Na esquerda é apresentada a arquitetura do modelo CBOW e na direita é apresentada a arquitetura do algoritmo Skip-gram.



Fonte: Figura adaptada de (MIKOLOV *et al.*, 2013).

O objetivo do modelo CBOW é predizer a i-ésima palavra w_i de uma sentença utilizando uma janela de tamanho t ao redor da palavra. Dessa forma, as palavras

$$w_{i-t}w_{i-t+1},\ldots,w_{i-1}w_{i+1},\ldots,w_{i+t-1}w_{i+t} \tag{2.23}$$

são usadas para predizer a palavra alvo w_i. A arquitetura CBOW é similar a uma rede neural *feed forward* (BENGIO; DUCHARME; VINCENT, 2000). A rede é composta por uma camada de entrada, uma camada de projeção, uma camada oculta e uma camada de saída (LE; MIKOLOV, 2014). Cada palavra é mapeada para um vetor único, representado como uma coluna na matriz W e cada coluna é indexada pela posição da palavra do vocabulário. A concatenação ou soma dos vetores é usado como características para a predição da próxima palavra na sentença. Ou seja, dado a sequência de palavras de treinamento $w_1, w_2, w_3, \ldots, w_t$, o objetivo do modelo é maximizar a média do log de probabilidade, conforme Equação 2.24.

$$\frac{1}{T}\sum_{t=k}^{T-k}\log p(w_t|w_{t-k},\ldots,w_{t+k}) \tag{2.24}$$

A tarefa de predição é feita geralmente por um classificador multiclasse, como o *Softmax*, fundamentada na equação 2.25:

$$p(w_t|w_{t-k},\ldots,w_{t+k}) = \frac{e^{y_{w_t}}}{\sum_i e^{y_i}} \tag{2.25}$$

Para cada y_i é o log de probabilidade não normalizado de cada palavra de saída i, expressado na Equação 2.26, em que U e b são parâmetros do classificador *Softmax* e h é

construído pela concatenação ou média dos vetores de palavras, extraídos de W.

$$y = b + Uh(w_{t-k}, ..., w_{t_k}; W) \tag{2.26}$$

A rede neural é geralmente treinada usando gradiente descendente estocástico, na qual o gradiente é obtido via retro propagação (RUMELHART; HINTON; WILLIAMS, 1986).

O modelo neural *skip-gram* possui arquitetura similar ao CBOW, entretanto, em vez de predizer uma palavra dado um contexto, o objetivo do modelo é predizer o contexto $w_{i-t}w_{i-t+1}, ..., w_{i-1}w_{i+1}$ ao entorno da *i*ésima palavra da sentença w_i. Em outras palavras, considera-se uma palavra como entrada do classificador e são realizadas predições de palavras que ocorram antes e depois da palavra atual. As palavras que estão geralmente mais próximas a palavra de entrada recebem pesos maiores, enquanto que as palavras mais distantes recebem pesos menores, devido àmenor amostragem nos exemplos de treinamento (MIKOLOV *et al.*, 2013).

Algumas variações surgiram para a criação de *embeddings* ao nível de sentenças, parágrafos e documentos surgiram, considerando os modelos CBOW e *Skip-Gram*. O modelo Doc2Vec (LE; MIKOLOV, 2014), também conhecido como *Paragraph Vector*, permite a representação de documentos inteiros em vetores densos e tamanho fixo. O modelo captura contextos de palavras em relação a um documento, possibilitando o aprendizado de representações contínuas para textos completos. O *Global Vectors for Word Representations* (GloVe) (PENNINGTON; SOCHER; MANNING, 2014) é um modelo de linguagem que mapeia palavras em um espaço vetorial de maneira a capturar relações semânticas e sintáticas entre as palavras. A principal contribuição do GloVe foi de capturar informações distribucionais de palavras, usando uma matriz de coocorrência ponderada. O *FastText* é uma extensão do *Word2Vec* que não apenas representações de palavas, mas também de subpalavras (*n-gramas*). Isso possibilita a captura de informações morfológicas e subpalavras em representações, o que torna eficaz para idiomas com morfologias complexas, e por isso, lida com palavras raras ou fora do vocabulário do contexto (BOJANOWSKI *et al.*, 2017).

Na época em que foram propostos, algumas limitações foram apontadas na literatura para os modelos independentes de contexto. Segundo Lucy e Gauthier (2017) algumas classes de características acabam sendo mal representadas pelos métodos avaliados, faltando elementos fundamentais de semântica e alguns domínios semânticos são particularmente afetados por esses problemas. Outra limitação é que as *word embeddings* geradas por os modelos apresentados anteriormente são representações livres de contextos (MCCANN *et al.*, 2017), isso significa que após o treinamento uma palavra terá sempre o mesmo *word vector* independente do contexto que a palavra está inserida. Representações dependentes de contexto têm sido propostos com finalidade de superar as limitações apresentada, como os modelos *transformers*.

2.3.3 Representações dependentes de contexto

Apesar das abordagens anteriores sanarem parte das desvantagens do modelo *Bag-of-Words*, as *embeddings* geradas pelos modelos são livres de contexto. Por exemplo, considere o exemplo, D1 = "Acendi a vela quando a luz acabou" e D2 = "Içamos a vela, e o navio partiu". A palavra "vela" é utilizada em contextos diferentes, no entanto, recebe a mesma representação vetorial (*embeddings*), independente do contexto em que a palavra está inserida. Para lidar com casos polissêmicos, novas abordagens foram propostas na literatura para considerar estrutura semântica e sintática dos documentos de textos. O *Embeddings from Language Models* (ELMo) é um modelo contextual de representação de palavras que modela características complexas de palavras e como seus usos variam entre diferentes contextos linguísticos (PETERS *et al.*, 2018). As *embeddings* geradas por ELMo são apreendidas por meio de um modelo de linguagem bidirecional profundo, treinado por uma *Long Short Term Memory* (LSTM) usando um grande corpus textual.

As *embeddings* apreendidas por ELMo são funções aprendidas de estados internos de uma LSTM bidirecional com L camadas de propagação (*forward*), que aprendem informações contextuais considerando a sequência de palavras da direita para a esquerda, predizendo o próximo token t_k da sequência com base em informações históricas, usando uma camada *Softmax* (Equação 2.27):

$$p(t_1, t_2, ..., t_n) = \prod_{k=1}^{N} p(t_k | t_1, t_2, ..., t_{k-1}) \tag{2.27}$$

Os resultados obtidos são combinados com a saída de L camadas de LSTM de retro propagação (*backward*), que executam sobre a sequência inversa de palavras, predizendo o token anterior dado o contexto do futuro na Equação 2.28:

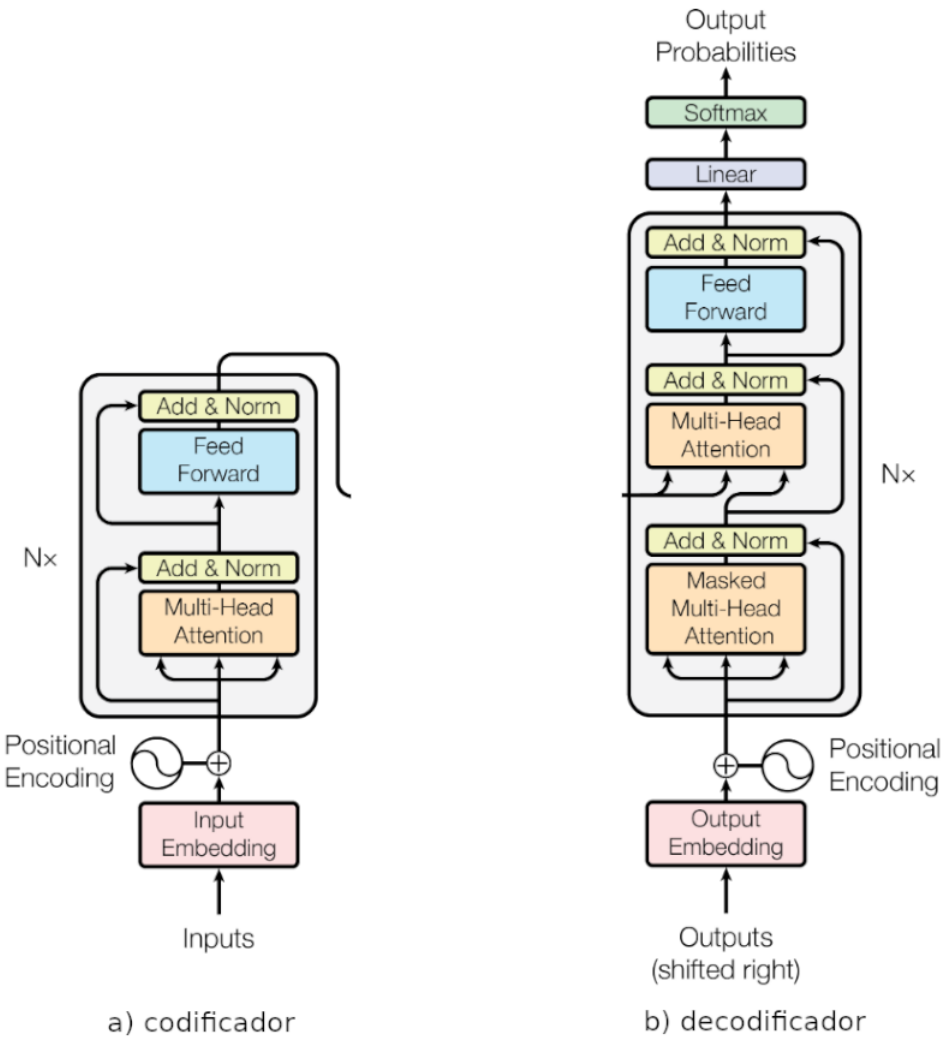
$$p(t_1, t_2, ..., t_n) = \prod_{k=1}^{N} p(t_k | t_{k+1}, t_{k+2} ..., t_N) \tag{2.28}$$

A combinação das camadas de propagação e retro propagação é feita pela rede neural bidirecional, que maximiza o *log* da probabilidade de ambas as direções, visando predizer a próxima palavra da sentença. Embora o modelo de linguagem tenha muitas vantagens, também apresenta algumas desvantagens, como: custo computacional elevado, tempo de inferência mais longo, necessidade de textos anotados e interpretabilidade limitada.

Para superar as desvantagens do modelo ELMo, modelos de *Transformers* foram projetados com base em um mecanismo de atenção, capaz de modelar dependências globais entre a entrada e saída (VASWANI *et al.*, 2017). Ao contrário das Redes Neurais Recorrentes, o modelo *tranformers* não exige que os dados sequenciais sejam processados na ordem. Por exemplo, se os dados de entrada é uma frase em linguagem natural, o *tranformers* não precisa processar o início, antes do final da frase. Devido a esse recurso, o tempo de treinamento é reduzido por

permitir a paralelização no processamento do modelo (VASWANI *et al.*, 2017). O modelo foi projetado com base na arquitetura de codificar e decodificar em duas camadas (Figura 13): a) pré-processar interativamente a entrada em uma camada de cada vez; e, b) extrair-padrões por um conjunto de camadas para decodificar interativamente a saída do codificador.

Figura 13 – Arquitetura do modelo *Transformer*.



Fonte: (VASWANI *et al.*, 2017).

Tanto o codificador quanto o decodificador apresentam componentes semelhantes. Segundo Vaswani *et al.* (2017), cada componente tem a função:

- *Input embedding*: recebe simultaneamente as palavras da sentença e as mapeia para suas respectivas *embeddings*;
- *Positional encoding*: armazena informações sobre a posição relativa ou absoluta de um token dentro da sequência de palavras, fazendo com que o modelo aprenda informações sobre as posições dos tokens para sequências de diferentes frequências;

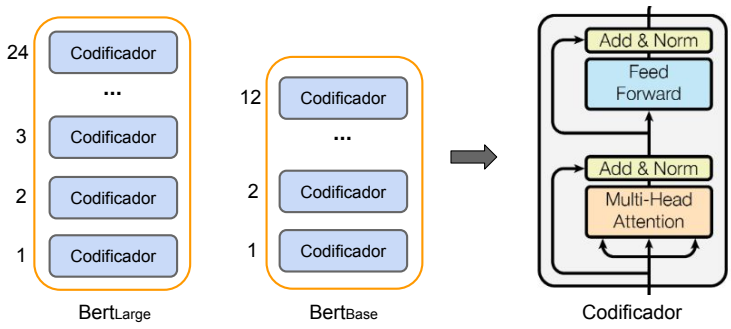
- *Self-attention e Multi-head*: permite que o modelo atenda um conjunto de informações de diferentes subespaços de representação em diferentes posições;
- *Add and Normalization*: recebe valores de entrada e os vetores de saída da camada anterior, realiza a soma e normaliza para facilitar o processo de otimização e garantir que o *positional encoding* se mantenha estável durante o processo;
- *Feed Forward*: composta por várias redes neurais de propagação que processam palavras e sentenças de forma paralela. Cada rede neural é composta por duas camadas totalmente conectadas;
- *Output Embeddings*: recebe as palavras da sentença que foram preditas na etapa anterior;
- *Self-attention* e *Masked multi-head*: funcionamento similar ao *self-attention e multi-head*, porém utiliza apenas palavras já preditas pelo mecanismo de atenção;
- *Linear*: a saída do decodificador é um vetor numérico. Nesse caso, para que ele seja convertido em um palavra, são necessários dois processos tradicionais (Linear e Softmax). O primeiro é o processamento da camada linear, em que uma rede neural totalmente conectada recebe o vetor de saída e o projeta em um vetor chamado *logits vector*. Este vetor possui o mesmo tamanho do vocabulário da coleção de documentos, de modo que cada posição deste vetor contenha uma pontuação atribuída a cada palavra;
- *Softmax*: esta camada recebe o *logits vector* e transforma a pontuação de cada palavra em probabilidade. A palavra que recebe maior probabilidade é selecionada.

Proposto por Devlin *et al.* (2018), *Bidirectional Encoder Representations from Transformers* (BERT) é uma arquitetura projetada para pré-treinar representações bidirecionais de textos não rotulados, baseado na implementação original (VASWANI *et al.*, 2017). O modelo pré-treinado do BERT possui o número de camadas (L) no codificador, a dimensão do *embeddings* de entrada (H) e o número de entradas do *Self-Attention Head* (A). Inicialmente, os modelos pré-treinados do BERT foram propostos em dois tamanhos: $BERT_{BASE}$ = (L=12, H=768, A=12, Parâmetros=110M) e $BERT_{BASE}$ = (L=24, H=1024, A=16, Parâmetros=340M), conforme ilustrado na Figura 14.

Ao contrário dos modelos direcionais, que consideram a entrada de textos sequencialmente, o codificador *transformer* lê toda a sequência de palavras de uma vez. Esse modelo é um contraste com as representações tradicionais que consideravam uma sequência de texto da esquerda para a direita, ou treinamento combinado nos dois sentidos (direita para esquerda e esquerda para direita). De modo diferente, o pré-treinamento do BERT usa duas tarefas não supervisionadas: i) *Masked Language Model* (MLM); e, ii) *Next Sentence Prediction* (NSP).

Ao contrário dos modelos direcionais, que consideram a entrada de textos sequencialmente, o codificador *transformer* lê toda a sequência de palavras de uma vez. Esse modelo é

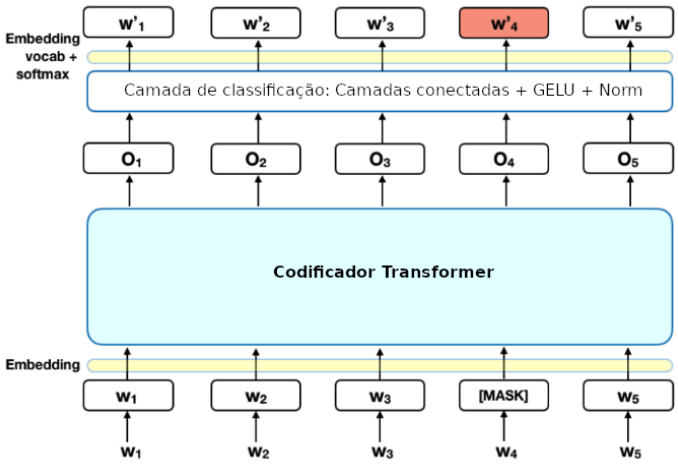
Figura 14 – Modelo de pré-treinamento do BERT.



Fonte: Adaptado de (VASWANI *et al.*, 2017).

um contraste com as representações tradicionais que consideravam uma sequência de texto da esquerda para a direita, ou treinamento combinado nos dois sentidos (direita para esquerda e esquerda para direita). De modo diferente, o pré-treinamento do BERT usa duas tarefas não supervisionadas: i) *Masked Language Model* (MLM); e, ii) *Next Sentence Prediction* (NSP). Na **primeira etapa** de MLM, uma porcentagem dos *tokens* de entrada são selecionadas aleatoriamente e substituídos pelo *token* (MASK). O desafio do BERT é predizer o valor original das palavras mascaradas, baseado no contexto de outras palavras não mascaradas na sentença. Nesse cenário, a predição das palavras na saída do modelo requer: adicionar uma camada de classificação no topo do decodificador; multiplicar os vetores de saída por uma matriz *embedding*, transformando em uma dimensão no vocabulário; e, calcular a probabilidade de cada palavra no vocabulário com $softmax$. A Figura 15 ilustra os procedimentos de MLM.

Figura 15 – Implementação da etapa de MLM no BERT.

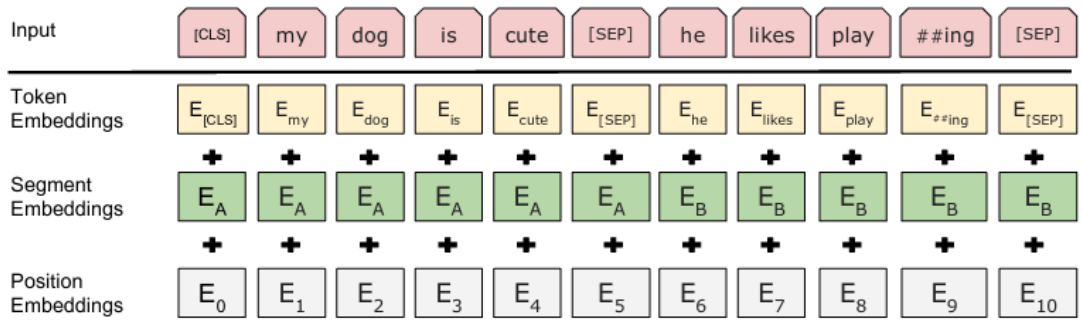


Fonte: Adaptado de Academy (2021).

A saída final do codificador não é usada diretamente para previsões, em vez disso, uma camada totalmente conectada com ativação de *Gaussian Error Linear Unit* (GELU) adicionada no meio. A função de perda é apenas considerada nos valores preditos para as palavras mascaradas, tornando o aprendizado baseado no contexto.

Na **segunda etapa** de NSP, o modelo recebe pares de sentenças como entrada e aprende a predizer se a segunda sentença é subsequente a sentença original do documento. Durante o treinamento, 50% das entradas são um par em que a segunda sentença é a sentença subsequente do documento original, enquanto que nos outros 50%, uma frase aleatória do corpus é escolhida como a segunda sentença. A suposição é que a sentença será desconectada da primeira. Para ajudar o modelo a distinguir entre as duas sentenças no treinamento, as entradas é processadas conforme ilustra a Figura 16, antes de entrar no modelo de codificação.

Figura 16 – Representação de entrada de BERT. Os embeddings de entrada são a soma dos embeddings de token, os embeddings de segmentação e a incorporação de posição.



Fonte: (DEVLIN *et al.*, 2018).

As entradas são processadas da seguinte maneira: um token CLS é inserido no começo da sentença; o token SEP é usado como um separador entre diferentes sentenças; *Token Embeddings* são representações numéricas das palavras na entrada da sentença; o *Segment Embeddings* são usados para ajudar o BERT a distinguir entre diferentes sentenças em cada entrada; o *Position Embeddings* é adicionado para cada token indicando a sua posição na sequência das sentenças. Para predizer se a segunda sentença está realmente conectada à primeira, as seguintes etapas são executadas: i) toda a sequência de entrada passa pelo modelo *transformer*; ii) a saída do token CLS é transformada em um vetor ($2x1$), usando uma camada de classificação simples (matrizes aprendidas com pesos e *bias*); iii) calcular a probabilidade da próxima sentença (*IsNextSequence*) com $Softmax$.

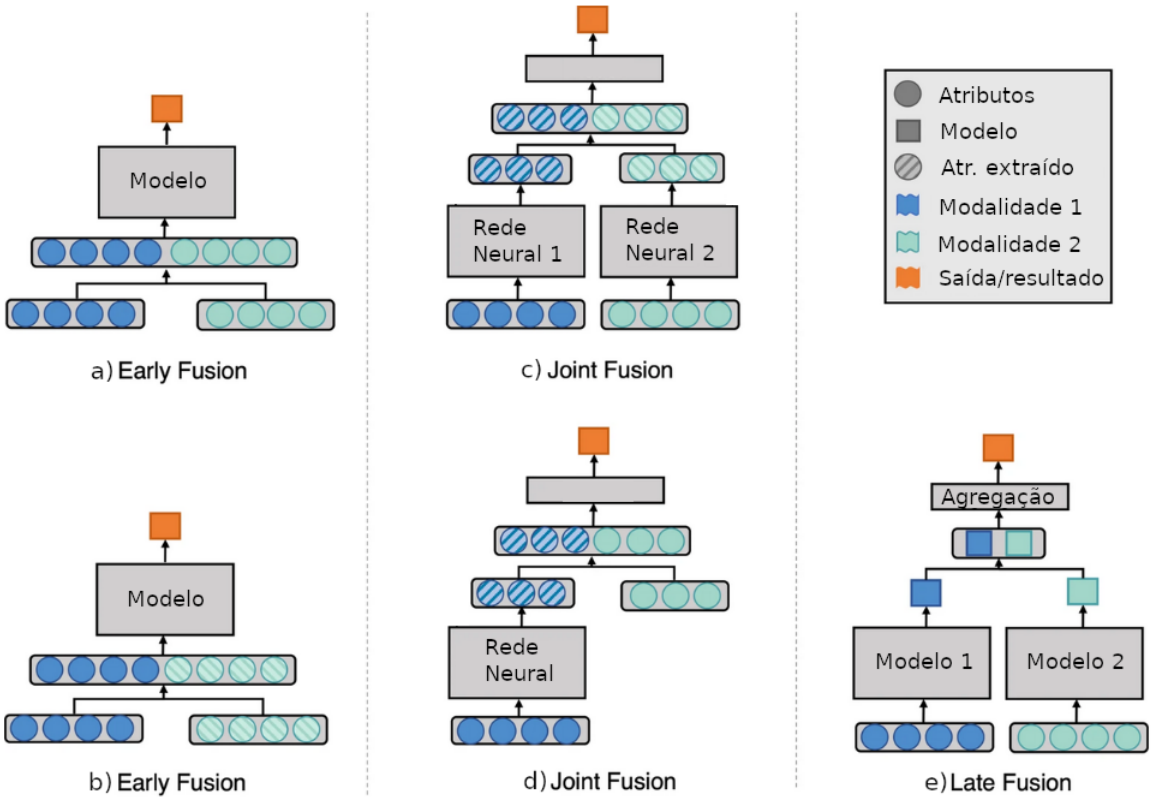
Ao treinar o BERT, o *Masked* LM e *Next Sentence Prediction* são treinados em conjunto, objetivando minimizar a função de perda usando as duas estratégias. Nesse sentido, o BERT pode ser ajustado com apenas uma camada de saída adicional para criar modelos robustos para uma variedade de tarefas. Esses ajustes não precisam ser modificados substancialmente na arquitetura do modelo, podendo ser refinado para uma tarefa específica do domínio.

2.4 Fusão de Informação

A fusão de informações é uma abordagem crucial na análise de dados provenientes de múltiplas fontes ou modalidades diferentes, com o objetivo de melhorar a qualidade da

informação e a robustez dos resultados obtidos. Essa técnica desempenha um papel fundamental em áreas como processamento de imagem (HUANG *et al.*, 2020), sistemas de *blockchain* (LIANG *et al.*, 2021) e domínios financeiros (ZHONG; HITCHCOCK, 2021). A ideia subjacente à fusão de informações é combinar os pontos fortes de diversas fontes de dados para obter uma visão mais completa e precisa do problema em questão. Existem várias estratégias de fusão de informações, cada uma com suas próprias vantagens e aplicações específicas. Três abordagens amplamente utilizadas são a Fusão Inicial (*Early Fusion*), a Fusão Conjunta (*Joint Fusion*) e a Fusão Tardia (*Late Fusion*) (HUANG *et al.*, 2020).

Figura 17 – Modelos de arquiteturas para fusão de diferentes estratégias. *Early Fusion* (Figura da esquerda); *Joint Fusion* (Figura do meio); e, *Late Fusion* (Figura da direita).



Fonte: Adaptado de Huang *et al.* (2020).

O **Early Fusion**, também conhecida como fusão em nível de recurso, envolve o processo de combinar diversas modalidades de entrada em um único vetor de características, antes de alimentar um único modelo de aprendizado de máquina para treinamento (Figura 17 (a) e (b)). Essas modalidades de entrada podem ser combinadas de várias maneiras distintas, tais como a concatenação, a adição ou a média dos dados de entrada no modelo de predição. A fusão dos dados originais representa o que denomina-se de fusão inicial do tipo I, enquanto a fusão dos dados extraídos, que podem ser obtidos por meio de extração manual ou representações aprendidas de outras redes neurais, caracteriza a fusão inicial do tipo II (HUANG *et al.*, 2020).

O **Joint Fusion** (Figura 17 (c) e (d)) refere a combinação de representações de recursos aprendidos em camadas intermediárias de redes neurais com recursos de outras modalidades, que

são usados como entrada para um modelo final. Ao contrário da *early fusion*, na *Joint Fusion*, a perda resultante da previsão é retroalimentada para as redes neurais com a extração de recursos durante o treinamento. Isso tem o efeito de melhorar as representações de recursos à medida que o modelo é refinado ao longo das iterações de treinamento. A implementação da fusão conjunta é realizada com redes neurais devido à capacidade delas de propagar a perda (*loss*) do modelo de previsão de volta para o(s) modelo(s) de extração de recursos. Quando todas as modalidades contribuem para as representações de recursos, é considerada uma fusão conjunta do tipo I (Figura 17 (c)). No entanto, nem todas as situações de entrada exigem que a etapa de extração de recursos seja configurada como uma fusão conjunta, dando origem à fusão conjunta do tipo II (Figura 17 (d)) (HUANG *et al.*, 2020).

O **Late Fusion** (Figura 17 (e)) envolve a combinação de previsões provenientes de diversos modelos para tomar uma decisão final, o que a torna frequentemente referida como fusão de nível de decisão. Nesse cenário, diferentes modelos são treinados utilizando modalidades distintas, e a decisão final é tomada utilizando uma função de agregação para combinar as previsões dos vários modelos. Algumas exemplos de funções de agregação incluem média, votação por maioria, votação ponderada ou até mesmo um metaclassificador que toma como entrada as previsões de cada modelo individual. A escolha da função de agregação é tipicamente determinada de forma empírica e pode variar conforme a aplicação e as modalidades de entrada envolvidas (HUANG *et al.*, 2020).

2.5 Métricas de Avaliação

A seleção das métricas de avaliação desempenha um papel de extrema importância na verificação da precisão das previsões em uma série temporal, uma vez que os modelos preditivos devem ser ajustados para aprimorar continuamente essas métricas em cada etapa de previsão. Especificamente no contexto da previsão de séries temporais, medidas de erro desempenham um papel fundamental na quantificação da discrepância entre os valores reais observados $(y_1, y_2, ..., y_h)$ e os valores previstos pelo modelo $(\hat{y}_1, \hat{y}_2, ...\hat{y}_h)$. Em situações em que o horizonte de previsão (h) é extenso, índices de desempenho que possibilitam uma avaliação abrangente dos dados previstos são extraídos, a fim de obter uma estimativa precisa da configuração experimental.

Uma das medidas mais amplamente empregadas para calcular erros de previsão é o Erro Médio Percentual Absoluto (MAPE), cuja formulação é expressa pela Equação 2.29.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| * 100 \tag{2.29}$$

em que n é o número de amostras de teste, y_i é o valor atual e $\hat{y}_i$ é o valor predito. O resultado do MAPE é expresso como um valor percentual que estabelece a relação entre o valor predito e o

valor real da série. No entanto, é importante ressaltar uma limitação prática do MAPE: quando uma série temporal possui valores zero, ocorre uma divisão inadequada por zero, o que prejudica sua aplicação.

O Erro Quadrático Médio (MSE) e Raiz do Erro Quadrático Médio (RMSE), conforme definidos nas Equações 2.30 e 2.31, respectivamente, constituem alternativas ao problema identificado com o MAPE.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.30}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.31}$$

O MSE é semelhante ao Média do Erro Absoluto (MAE), porém é mais sensível a erros significativamente grandes, uma vez que eleva as diferenças individuais ao quadrado. . O resultado do MSE é sempre positivo, e um valor de MSE igual a zero indica uma correspondência perfeita entre as previsões do modelo e os valores reais, o que significa que o modelo é capaz de reproduzir os dados de forma precisa. Por outro lado, o RMSE é frequentemente empregado para expressar os valores dos erros na mesma escala da variável em análise (HYNDMAN; ATHANASOPOULOS, 2018). É importante notar que tanto o MSE quanto o RMSE são sensíveis a escalas, ou seja, não são adequados para comparações entre séries que possuam variáveis com escalas distintas.

Introduzido por Hyndman e Koehler (2006), o Erro Médio Absoluto Escalado (MASE) é uma alternativa ao emprego de erros percentuais quando se busca comparar a precisão de previsões em séries com escalas distintas. A Equação 2.32 descreve a métrica MASE.

$$MASE = mean(|q_j|) \tag{2.32}$$

em que

$$q_j = \frac{e_j}{\frac{1}{T-1}\sum_{t=2}^{T}|y_t - y_{t-1}|} \tag{2.33}$$

y_t é a observação da série temporal, e_t é o erro da previsão para o dado período.

A avaliação dos resultados do modelo preditivo por meio das métricas apresentadas acima é tão importante quanto a análise qualitativa da distribuição gráfica dos erros de predição. Para a robustez do modelo, o retrato dos erros ao longo da série precisa compor um padrão aleatório, e não uniforme. Dessa forma, o uso de mais de uma métrica de avaliação é recomendado para analisar as flutuações erráticas das previsões.

2.6 Considerações Finais

No decorrer deste capítulo, fundamentos que constituem a base deste trabalho de doutorado foram explorados e delineados. Abordou-se conceitos fundamentais sobre séries temporais, compreendendo suas classificações, tipologias e elementos essenciais para a compreensão do comportamento e estrutura das séries temporais. De modo específico, as séries financeiras, consideradas neste trabalho, são caracterizadas como univariadas/multivariadas, discretas, estocásticas e não estacionárias, delineando a complexidade e a dinâmica desses dados. Por outro lado, os dados textuais são oriundos de fontes em línguas portuguesa e inglesa, representando um complemento fundamental para o enriquecimento das representações, tanto das séries quanto dos textos. O uso de estratégias de fusão entre diferentes tipos de dados emerge como um dos pilares deste estudo, seja para tarefas de regressão ou de classificação.

A previsão de séries e a classificação de textos financeiros é especialmente desafiadora, devido à influência de fatores não previsíveis. Recentemente, técnicas de mineração de texto têm sido empregadas para complementar séries temporais, a fim de considerar fatores externos e melhorar a precisão dos modelos. A integração de diferentes fontes de dados em modelos multimodais apresenta desafios relacionados às estratégias de fusão e previsão. Neste estudo, as estratégias de previsão de uma única etapa à frente são prioritárias, pois os dados textuais possuem limitações para a aplicação em previsões de múltiplas etapas à frente em cenários do mundo real. Além disso, as estratégias de fusão, como a *early fusion*, desempenham um papel fundamental ao considerar dados de séries temporais e textos no estágio de pré-processamento para os modelos preditivos. Em contraste, abordagens como *joint fusion* e *Late Fusion* são direcionadas à melhoria do desempenho dos modelos preditivos, focando na combinação dos resultados para aprimorar a precisão da previsão, sem enriquecer as representações para os modelos.

Considerando as especificações apresentadas anteriormente, uma grande variedade de estudos na literatura têm abordado a multimodalidade de dados para tarefas preditivas, seja de classificação ou regressão. Na sequência, um mapeamento sistemático que contempla as técnicas de fusão apresentadas no presente capítulo será apresentado.

CAPÍTULO

3

MAPEAMENTO SISTEMÁTICO

No presente capítulo é apresentado um mapeamento sistemático dos estudos relacionados às tarefas preditivas do mercado financeiro em que consideram dados de séries temporais e de textos. Inicialmente é apresentado o protocolo de pesquisa (Seção 3.1) com definições das questões de pesquisa, estratégias e definições das *strings* de buscas, critérios de inclusão e exclusão dos estudos selecionados, e os procedimentos de extração dos dados. Os resultados (Seção 3.2) dos estudos selecionados são apresentados e discutidos considerando três abordagens de fusão, e por fim, as questões de pesquisa são respondidas e o relacionamento entres os estudos é apresentada.

3.1 Protocolo de Pesquisa

Uma metodologia de pesquisa foi adotada com base nas diretrizes apresentadas por (KITCHENHAM; CHARTERS, 2007), que recomenda o seguinte conjunto de etapas: i) definição de um conjunto de questões a serem respondidas para agregar valor à investigação (conforme a subseção 3.1.1); ii) apresentação, avaliação e definição da estratégia de busca (conforme subseção 3.1.2); iii) estabelecimento e aplicação de critérios de inclusão e exclusão (abordados na subseção 3.1.2); iv) análise minuciosa dos estudos selecionados e descrição de todas as informações extraídas (conforme subseção 3.1.3). Essas etapas são minuciosamente detalhadas para atender às questões de pesquisa definidas neste mapeamento sistemático.

3.1.1 Definição das questões de pesquisa

O objetivo principal deste mapeamento sistemático consiste em mapear estudos que combinam séries temporais e informações textuais para tarefas de previsão. Além disso, explorou-se o estado da arte que envolve estudos qualitativos e avaliações do mercado financeiro. Com base nesse objetivo, as seguintes Questões de Pesquisa (QP) são identificadas:

- **QP1**: Quais são as estratégias de fusão mais usadas para combinar dados de séries temporais e de textos para tarefas preditivas?
- **QP2**: Quais são os modelos de previsão mais usados para tarefas de preditivas que usam dados multi-modais no mercado financeiro?
- **QP3**: Quais são as representações vetoriais de texto mais utilizadas para combinar com séries temporais?
- **QP4**: Quantos estudos realizaram a fusão de informações entre séries temporais e informações textuais para tarefas preditivas nos últimos anos?

QP1 consiste em uma pesquisa das técnicas amplamente utilizadas para enriquecer séries temporais com dados textuais, seja por meio de jornais, sites ou redes sociais. QP2 concentra-se em modelos capazes de realizar tarefas de previsão, aplicando uma combinação de informações textuais e séries temporais. QP3 lida com as representações textuais que podem ser utilizadas com séries temporais para tarefas de previsão. Por fim, QP4 investiga o número de trabalhos publicados nos últimos anos que utilizaram a fusão de informações e tarefas preditivas para o mercado financeiro.

3.1.2 Estratégia de Pesquisa

Considerando as questões de pesquisa, identificou-se as palavras-chave usadas para mapear os estudos relacionados a esta Tese. A Tabela 2 mostra os termos usados no levantamento bibliográfico.

Tabela 2 – Termos usados para criar as *strings* de busca.

Termos	**Combinação**
Times series	"*texts embeddings*", "BERT", "*Bag-of-Words*"
Texts	"*time series*", "*sentiment*"
Financial Market	"*stock market*", "*commodity*"
Information Fusion	"*multimodal*", "*multimodality*"

Fonte: elaborado pelo Autor.

A estratégia adotada foi de incluir termos mais abrangentes para que fosse possível alcançar uma ampla identificação de estudos primários. Nesse sentido, a seguinte *string* de busca foi definida: (("time series" AND "text embedding" OR "BERT" OR "Bag-of-Word") OR ("texts" AND "time series" AND "sentiment")) AND ("financial market", OR "stock market" OR "commodity") AND ("information fusion" OR "multimodal" OR "multimodality").

A busca por estudos foi realizada na plataforma do Google Scholar, onde foi possível ter acesso às principais bibliotecas digitais, como: ACM Digital Library, IEEE, Springer e Elsevier.

3.1.3 Critérios de Inclusão e Exclusão

Após a pesquisa bibliográfica inicial e a leitura do resumo de todos os resultados, os seguintes critérios de inclusão e exclusão são aplicados aos estudos para o processo de seleção para análise detalhada. Em relação aos critérios de inclusão, vale ressaltar que são selecionados os estudos que atenderam pelo menos a um dos seguintes itens:

- **I1**: Os estudos devem estar diretamente relacionados à aplicação de tarefa preditiva multimodal que envolve séries temporais e textos.
- **I2**: Os estudos devem fornecer resultados que envolvam modelos preditivos multimodal com dados quantitativos e/ou qualitativos.
- **I3**: Os estudos devem abordar aplicações sensíveis ao contexto de modelos de predição usando séries temporais e dados textuais.
- **I4**: O contexto a ser explorado envolve a predição de séries temporais ou classificação de textos relacionadas ao mercado financeiro e/ou mercado de commodities.

O estudo a ser incluído deve atender a pelo menos um critério de inclusão. Assim, os estudos devem estar relacionados à análise preditiva combinada com dados textuais e séries temporais. Por outro lado, os critérios de exclusão são aplicados aos resultados para concluir o processo de seleção final na etapa seguinte. Os critérios definidos são:

- **E1**: Estudos que usam apenas séries temporais ou informações textuais/mineração de textos.
- **E2**: Estudos que apresentam apenas opiniões sem evidências empíricas.
- **E3**: Estudos que analisam a correlação entre dados de séries temporais e informações de textos.
- **E4**: Trabalhos do mesmo autor com títulos semelhantes em periódicos diferentes ou duplicados.
- **E5**: Estudos ou dissertações com trabalhos não publicados em periódicos/conferências/-congressos.
- **E6**: Trabalhos de revisão ou de mapeamento sistemático.

Os estudos que consideram apenas dados de séries temporais ou apenas informações textuais serão excluídos. Além disso, serão excluídos estudos que apenas apresentam opiniões sem evidências empíricas, ou publicados há mais de dez anos, ou teses/dissertações que não foram publicadas em periódicos, conferências ou congressos. Consideramos os últimos dez anos

devido ao aumento substancial de trabalhos publicados neste período, como apresentado nos resultados.

3.1.4 Procedimento de extração de dados

O procedimento de extração de dados está dividido em três etapas de seleção e a etapa final de levantamento dos dados e dos métodos empregados nos estudos. As três primeiras etapas (ETP) são apresentadas a seguir. Detalhes do quantitativo dos trabalhos incluídos e excluídos por etapa são apresentados na Tabela 3.

- **ETP 1**: A etapa inicial utilizou as *strings* de busca previamente definida e totalizou em 331 documentos. Nesta fase, foi possível identificar 30 estudos duplicados, que foram excluídos, resultando em 301 estudos.
- **ETP 2**: Na segunda etapa, uma seleção foi feita lendo o título e o resumo e observando os critérios de inclusão e exclusão pré definidos na seção anterior. Como resultado, 60 estudos foram incluídos para a última etapa.
- **ETP 3**: Na terceira etapa, foi realizada a leitura completa dos documentos, na qual os dados referentes às questões de pesquisa foram extraídos e apresentados na Seção 3.2. Após a leitura de 60 documentos, foram excluídos 18 e incluídos 42 estudos para etapa final.

Tabela 3 – Detalhes das etapas de seleção da extração de dados.

Etapa	Descrição	Inicial	Excluídos	Incluídos
1	Buscas usando *strings* de pesquisa	331	30	301
2	Seleção lendo o título e resumo	301	241	60
3	Leitura completa do documento	60	18	42

Fonte: Elaborado pelo Autor.

Após o processo final de seleção, uma tarefa de agregação de dados foi realizada para extrair os atributos de cada documento para uma análise minuciosa dos métodos utilizados em cada trabalho. Com esta análise, algumas questões de pesquisa são respondidas na sequência do capítulo.

3.2 Análises dos Resultados

Esta seção apresenta uma análise dos resultados obtidos por meio do protocolo de pesquisa demonstrado na Seção 3.1. Serão analisados integralmente 42 estudos com o objetivo de identificar os tipos de dados, as estratégias de fusão e modelos utilizados para as tarefas

preditivas. Os critérios de exclusão são rigorosamente aplicados para mapear o estado da arte no contexto da pesquisa.

Quanto ao processo de seleção, as porcentagens relativas a cada critério de exclusão estão detalhadas na Tabela 4. Destaca-se que apenas um critério de exclusão foi considerado para cada estudo, seguindo a ordem de prioridade dos critérios. Por exemplo, se um estudo possui os critérios E1 e E5, considerou-se apenas o critério E1 para cálculo final do número e porcentagem dos estudos excluídos. O total de estudos excluídos (289) nas três etapas é apresentado por critérios na Tabela 3.

Tabela 4 – Número e porcentagem de estudos excluídos.

Critérios de Exclusão	E1	E2	E3	E4	E5	E6
Número	77	20	11	87	57	37
Percentual(%)	26.64	6.92	3.81	30.10	19.72	12.80

Fonte: Elaborado pelo Autor.

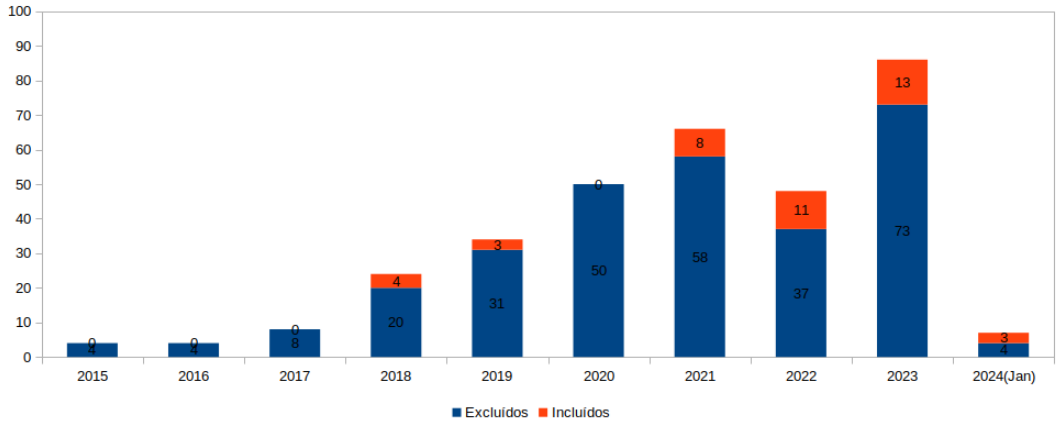
A análise detalhada na Tabela 4 proporciona uma visão aprofundada das porcentagens referentes aos critérios de exclusão adotados durante as etapas de seleção. Notavelmente, o critério E1 desempenhou um papel significativo, resultando na exclusão de 77 estudos, representando 26.64% do total. Em contrapartida, o critério E2 foi aplicado em 20 estudos, correspondendo a 6.42%. O critério E3 registrou a menor utilização, excluindo 3.81% dos estudos, enquanto o critério E4 mostrou-se o mais abrangente, resultando na exclusão de 30.10% dos trabalhos (87). Os critérios E5 e E6 contribuíram para as exclusões de forma expressiva, atingindo percentuais de 19.72% e 12.80%, respectivamente. Essa análise percentual demonstra a influência de cada critério no processo seletivo, permitindo uma compreensão precisa do perfil dos estudos excluídos.

3.2.1 Trabalhos selecionados

Tarefas preditivas no mercado financeiro têm atraído cada vez mais a atenção de pesquisadores (CHU *et al.*, 2023). Como resultado, o número de publicações cresceu substancialmente nos últimos anos, referentes a trabalhos que consideram estratégias de fusão de informação para previsão do mercado financeiro. De acordo com os resultados obtidos neste mapeamento sistemático, a Figura 18 demonstra o aumento de estudos que utilizam séries temporais e/ou informações textuais para tarefas de previsão do mercado financeiro.

Conforme ilustrado na Figura 18, observa-se que trinta e dois estudos foram propostos nos últimos três anos (2021-2023), evidenciando um aumento significativo na quantidade de pesquisas que propuseram estratégias de fusão de dados de séries temporais e informações textuais (representadas pela etiqueta laranja). Embora tenha havido uma diminuição no total de estudos em 2022, a tendência de crescimento linear dos estudos incluídos permaneceu consistente.

Figura 18 – Número de estudos incluídos/excluídos por ano.



Fonte: Elaborado pelo Autor.

Diante desse cenário, prosseguiu-se com a análise dos trabalhos incluídos nos mapeamentos para abordar as questões de pesquisa levantadas neste estudo de mapeamento.

As técnicas mais utilizadas para a fusão de dados multimodais são o Early Fusion (EF), Joint Fusion (JF) e Late Fusion (LF). Tendo em vista que os modelos de fusão assumem estratégias distintas, análises de cada estudo são apresentadas em três Tabelas (5, 6 e 7). As Tabelas oferecem uma descrição dos dados das séries temporais, as representações textuais, configuração experimental e modelos preditivos utilizados em cada estudo. Em termos gerais, dois tipos de avaliação são considerados nos estudos: regressão e classificação. Tarefas de regressão são empregadas em séries temporais para previsões numéricas, enquanto a classificação pode ser aplicada para prever tendências, polaridade e sentimentos do mercado financeiro. A Tabela 5 demonstra dois estudos incluídos com a estratégia de fusão *late fusion*.

Tabela 5 – Estudos com estratégias de *Late Fusion*.

Referência	**Série Temp.**	**Rep. Textual**	**Configuração**	**Modelos**
(AWAD; ELKAFFAS; FAKHR, 2023)	S&P, Yahoo, and NASDAQ	BERT and TF-IDF	Treino / Teste	DRL
(WANG *et al.*, 2023)	DJIA	SenticNet	Traino / Teste	MLP, NB, RF, LSTM.

Fonte: Elaborado pelo Autor.

Entre os quarenta e dois estudos incluídos para etapa final do mapeamento sistemático, apenas duas abordagens empregaram a estratégia *late fusion*. O estudo proposto por (AWAD; ELKAFFAS; FAKHR, 2023) utiliza o BERT e representação TF-IDF para prever sentimentos de notícias de mídia social. Em paralelo, uma técnica de *Variaton Mode Decomposition* (VMD) e LSTM são utilizados para predição de série de preço. Essas duas abordagens são integradas por um *Deep Reinforcement Learning* (DRL) para combinar sentimentos e predição de mercado do valor de abertura do próximo dia. No próximo estudo (WANG *et al.*, 2023), uma modelo de fusão é proposto em três fases : A primeira consiste em previsões intermediarias (*Multi Layer Perceptron* (MLP), *Naive Bayes* (NB), *Random Forest* (RF), *Long Short-Term Memory* (LSTM))

por meio de indicadores técnicos da séries de preços, a segunda considera apenas uma pontuação de análise de sentimento léxico (SenticNet). Por fim, as duas previsões são combinadas para determinar a previsão de tendência para o dia seguinte de cinco empresas do Dow Jones Industrial Average (DJIA).

A estratégia de fusão *Early Fusion* consiste no processo de combinar duas ou mais modalidades de entrada em um único vetor de características, antes de alimentar um único modelo de aprendizado de máquina para treinamento e teste. A Tabela 6 apresenta dezenove estudos que utilizaram da estratégia EF.

Tabela 6 – Estudos com estratégias de *Early Fusion*.

Referência	**Série Temp.**	**Rep. Textual**	**Configuração**	**Modelos**
(ZHANG *et al.*, 2018)	China A-Share	Word2Vec	Train / Test (75% / 25%)	SVM
(PICASSO *et al.*, 2019)	NASDAQ100	SenticNet Emb.	Increase Window	RF, SVM, RNN
(LI; SHANG; WANG, 2019)	Pretróleo bruto	TF-IDF (LDA)	Train / Test	RF, SR, LR
(ZHANG *et al.*, 2019)	China A-Share	Word2Vec	-	Hidden Markov
(LI; WU; WANG, 2020)	Hong Kong	SenticNet 5	Train, Val, Test (80%, 10%, 10%)	LSTM
(ZHOU *et al.*, 2020)	China A-Share	HowNet	Train / Test (75% / 25%)	SVM
(NTI; ADEKOYA; WEYORI, 2021)	Ações (Ghana)	-	Train / Test (75% / 25%)	LSTM
(ZHANG; YANG; ZHOU, 2021)	BGI Genomics	SnowNLP and Senta	Train / Test	LSTM-Attention
(HAO *et al.*, 2021)	Ações (Taiwan)	Skip-gram	Train / Test	SVM Fuzzy
(FARIMANI *et al.*, 2022)	Forex and Criptomoeda	BERT	Train / Val / Test	LSTM
(LI *et al.*, 2022)	Ações da China	Knowledge Graphs	Train / Val / Test	LSTM
(SAWHNEY *et al.*, 2021)	S&P 500	BERT	Train / Val / Test	LSTM
(KHALIL; PIPA, 2022)	Mercado de ações	Score Pos Tagging	Train / Test	LSTM
(YE *et al.*, 2022)	Bitcoin	VADER	Train / Val / Test	Ensemble (LSTM / GRU)
(FILHO; MARCACINI; REZENDE, 2022)	Commoditie (soja)	TF-IDF	Train / Test	SVR and Ensemble models
(TADPHALE *et al.*, 2023)	Forex (USD-INR)	BERT	Train / Val / Test	LSTM
(GU *et al.*, 2023)	DJIA and S&P 500	VADER	Train / Val / Test	GRU
(WANG; HSIAO; LIOU, 2023)	Ações de Taiwan	TF-IDF	Train / Test	SVR
(AVRAMELOU *et al.*, 2024)	Criptomoeda (USDT)	BERT	Train / Test	LSTM

Fonte: Elaborado pelo Autor.

Entre os dezenove estudos incluídos na estratégia *Early Fusion*, nota-se que a maioria deles utilizou dados do mercado de ações. Alguns estudos não forneceram informações deta-

lhadas sobre os dados, as representações textuais e as configurações experimentais utilizadas nas avaliações. Grande parte das propostas apresenta o modelo LSTM como uma abordagem inovadora, variando a combinação entre a representação de série temporal e os dados textuais como entrada no modelo. Dada a quantidade considerável de estudos analisados na terceira etapa do mapeamento sistemático, são especificados apenas os modelos de fusão empregadas em cada estudo, discutidas sequencialmente conforme apresentado na Tabela 6.

Os dados do mercado de ações da China (*A-Share*) são combinados com *embeddings* (*word2vec*) de notícias da web e sentimentos dos usuários das mídias sociais (ZHANG *et al.*, 2018). Nesse contexto, um tensor de decomposição é construído, e, em seguida, o modelo SVM é aplicado para a classificação das tendências de mercado. Em um estudo subsequente (PICASSO *et al.*, 2019), são combinados dados de análise técnica (séries de preços) e análise fundamentalista (*embeddings SenticNet*) para prever as tendências do mercado NASDAQ100. Outra abordagem interessante proposta por Li, Shang e Wang (2019) utiliza manchetes de notícias sobre o petróleo bruto para extrair contextos e tópicos, utilizando o método de *Latent Dirichlet Allocation* (LDA). Esse enfoque envolve um processo de agrupamento baseado no LDA, cujos resultados são, então, combinados com dados de séries temporais para prever os preços do petróleo, utilizando três modelos preditivos (RF, SR, LR).

Um estudo propõe um modelo de Markov para integrar eventos de notícias com dados quantitativos no mercado de ações A-Share da China (ZHANG *et al.*, 2019). A abordagem visa prever os estados futuros dos preços de várias ações correlacionadas simultaneamente. Em outro trabalho que utiliza o modelo LSTM (LI; WU; WANG, 2020), os autores representam dados numéricos por meio de indicadores técnicos e os combinam com quatro vetores de sentimentos obtidos através de técnicas de análise de sentimento do *SenticNet* 5. Além disso, em um estudo adicional (ZHOU *et al.*, 2020), os autores propõem quatorze variações com o *HowNet*[1] chinês. Abordagens de *Part-of-speech* (POS) são empregadas para determinar a polaridade da mensagem e de cada palavra, o grau de advérbios e as palavras de negação. As pontuações emocionais dos textos são combinadas com dados de séries temporais para prever os preços das ações da A-Share na China

Uma abordagem integra seis fontes heterogêneas (ações, websites, fóruns, entre outros) e utiliza um modelo CNN para a integrar os dados e selecionar as *features* (NTI; ADEKOYA; WEYORI, 2021). Em seguida, uma rede LSTM é empregada para a previsão de preços no mercado de Ghana. Uma variação do modelo LSTM, utilizando um mecanismo de atenção, é proposta por Zhang, Yang e Zhou (2021). O estudo incorpora análise de sentimentos obtidos por SnowNLP[2] e Senta[3], gerando quatro características numéricas. Essa representação, juntamente

[1] HowNet é uma base de conhecimento baseada em sememas. A base predefine um conjunto de 2.000 sememas e os utiliza para anotar mais de 100.000 palavras em chinês e inglês.
[2] GitHub do SnowNLP: https://github.com/isnowfy/snownlp
[3] Github do Senta: https://github.com/baidu/Senta

com dados das ações da BGI Genomics[4], é utilizada como entrada no modelo LSTM com mecanismo de atenção para extrair recursos-chave de múltiplas fontes de dados e avaliar o desempenho na previsão de preços das ações. Os autores Hao *et al.* (2021) propõem um SVM com hiperplano *Fuzzy* para integrar informações textuais na predição de tendências das ações de Taiwan. Recursos obtidos por meio de LDA e agrupamentos de *word embeddings* (Skip-gram) são utilizados. Além disso, uma técnica de *Particle Swarm Optimization* (PSO) é empregado para filtrar características com desempenho preditivo abaixo da média e integrá-las com dados de mercado.

Os autores Farimani *et al.* (2022) calcularam a correlação da série temporal com a distribuição de probabilidade das *news embeddings* (BERT), usando *fine-tuning* para análise de sentimento no domínio financeiro. A estratégia de fusão combina o humor da série temporal e dados de mercado, e posteriormente, uma RNN é empregada para extração de características, seguida de um modelo LSTM para regressão de preços. Outro estudo, apresenta um *framework* que integra técnicas denominadas modelos de pontuação e *screening*, usando quatro tipos de dados (LI *et al.*, 2022). Um modelo utiliza o Seq2Seq[5] e dados históricos de ações, enquanto outro modelo emprega uma estratégia de discretização *bottom-up*. O *screening* é composto por um modelo discriminativo e um modelo de mídia com base na relação de um grafo ponderado. Esses dois tipos de modelos são fundidos, e o LSTM é utilizado para prever ações da China.

Para prever o índice S&P500, um estudo define um modelo com base na entropia e quantifica variações temporais (preços) em dados de mercados afetados (texto e preços) de modo hierárquico (SAWHNEY *et al.*, 2021). O modelo utiliza um BERT financeiro *fine-tuned* de notícias/tweets e dados da entropia de séries temporais como entrada para o modelo LSTM. Em abordagem usando o LSTM (KHALIL; PIPA, 2022), propõem um modelo de representação com índice de sentimento usando pontuação de *Pos Tagging* e dados de séries de preços do mercado de ações. A proposta utiliza dez fontes de dados textuais, tais como: notícias de mídia convencional, mídia impressa, mídia social, feeds de notícias, blogs, portais de consultoria a investidores, opiniões de especialistas, atualizações de corretores, informações baseadas na web, notícias internas da empresa e anúncios públicos sobre políticas e reformas. Os autores Ye *et al.* (2022) empregam uma abordagem semelhante para prever os preços do Bitcoin nos próximos trinta minutos. O modelo integra dados de preços, indicadores técnicos e índices de sentimento, utilizando o LSTM e GRU com a técnica *stacking ensemble*[6]. O *Valence Aware Dictionary and sEntiment Reasoner* (VADER)[7] é utilizado para gerar índices de sentimentos e combinado com indicadores técnicos, sendo utilizado como entrada para *stacking ensemble*.

Um dos resultados deste trabalho de doutorado é um estudo que explorou o uso limitado

[4] A BGI Genomics é líder mundial no fornecimento de soluções integradas de medicina de precisão.
[5] Github do Seq2Seq: https://github.com/google/seq2seq
[6] Stacking ensemble é um modelo que aprende como combinar melhor as previsões de vários modelos de aprendizado de máquina com bom desempenho.
[7] Github do VADER: https://github.com/cjhutto/vaderSentiment

de termos específicos de domínio para gerar representações textuais de baixa dimensão e combiná-las com séries temporais Filho *et al.* (2020). Essa representação enriquecida é empregada para prever os preços diários futuros dos mercados de milho e soja em um cenário real, utilizando a estratégia de janela deslizante diária para prever o preço do próximo dia. Os autores Tadphale *et al.* (2023) assumem que as taxas de câmbio no mercado FOREX, especificamente USD/INR (valor cambial do dólar americano em relação à Rúpia Indiana), são influenciadas por vários fatores, como sentimento de mercado, outros mercados ativos, ouro, petróleo bruto e outros. O modelo de representação híbrida proposto é uma rede neural LSTM que recebe uma matriz de cinco dimensões em duas camadas, onde cada dimensão representa um fator de mercado. No próximo estudo (GU *et al.*, 2023), é proposto um modelo em três etapas: mapeamento de sentimento usando o VADER, modelagem temporal de séries temporais usando GRU e outra camada densa para previsão de preços futuros. Especificamente, o modelo utiliza o VADER e uma medida de sentimento "auto-definida" para extrair características léxicas, combinando-as com séries de preços. A GRU é empregada para capturar possíveis dependências de sentimento e prever os preços da DJIA e S&P500.

Uma abordagem recente integra diversas fontes de informação sobre o mercado de ações de Taiwan, incorporando notícias financeiras e volumes de negociação de empresas *blue-chip*[8] (WANG; HSIAO; LIOU, 2023). O modelo apresenta um novo classificador baseado na opinião de notícias (NOC) sem um dicionário predefinido. O foco do estudo é empregar o método *chi-square*[9] para extrair palavras e frases de artigos de notícias. Os rótulos das notícias são gerados e combinados com indicadores técnicos das ações, e um modelo *Multiple-Kernel Support Vector Regression* (MKSVR) é usado para prever o mercado de ações. Por fim, um estudo é proposto com base em um modelo pré-treinado com dados financeiros (AVRAMELOU *et al.*, 2024). Dados de sentimentos são extraídos usando o analisador de sentimento FinBERT, e as *features* relacionadas aos preços são capturadas em vários aspectos, oriundos dos valores de Abertura, Maior, Menor e Fechamento. Especificamente, duas camadas lineares, seguidas por uma função de ativação ELU, são empregadas como camadas de *embeddings* para lidar com dados de diferentes modalidades. A avaliação envolve treinar um agente *Deep Reinforcement Learning* (DRL) com *Proximal Policy Optimization* (PPO), e na etapa seguinte, utiliza-se um modelo LSTM para a predição do valor da criptomoeda.

A seguir, são apresentados os estudos disponíveis na literatura que empregaram a estratégia de *Joint Fusion*. Nessa abordagem, a combinação das representações ocorre por meio de recursos aprendidos em camadas intermediárias de redes neurais ou modelos de mecanismo de atenção. Em outras palavras, as representações vetoriais utilizadas como entrada nos modelos

[8] O conceito é usado para se referir a ações que são responsáveis por movimentar grandes volumes na Bolsa

[9] Um teste qui-quadrado é um teste de hipótese estatística usado na análise de tabelas de contingência quando os tamanhos das amostras são grandes. Em termos mais simples, este teste é usado principalmente para examinar se duas variáveis categóricas são independentes e influenciam a estatística do teste.

de predição foram processadas conjunta ou simultaneamente, gerando uma nova representação vetorial que representa dados multi-modais. Nesse contexto, é comum o uso de *embeddings* de textos e modelos de *multi-head attention* para estabelecer uma relação entre os dados de outras modalidades. Além disso, novas abordagens têm sido apresentadas para representar dados multi-modais em modelagens de grafos, por meio de *Graph Neural Network*. A Tabela 7 apresenta os estudos incluídos que utilizaram a estratégia de *Joint Fusion*.

Tabela 7 – Estudos com estratégias de *Join Fusion*

Referência	Série Temp.	Rep. Textual	Configuração	Modelos
(XU *et al.*, 2020)	Stock market	Glove	Train / Test	SMPN
(CHENG *et al.*, 2022)	China A-shares	BERT	Train / Test (rolling window)	GNN
(WINDSOR; CAO, 2022)	USD/CNY exchange rate	BERT	-	LSTM
(WANG *et al.*, 2022)	CSI 300	BERT	Train / Val / Test	GNN
(CHEN; HUANG, 2021)	S&P 500 index	VADER	Train / Test (rolling window)	LSTM
(FARIMANI *et al.*, 2021)	Forex index	BERT	Train / Val / Test	LSTM
(ZHANG *et al.*, 2022)	stock market	Embeddings	Train / Val / Test	Transformers
(DARADKEH, 2022)	stock market	Score BiLSTM	Train / Val / Test	LSTM
(HUANG *et al.*, 2022)	S&P 500 index and CSI 300	BERT	Train / Test	Graph (Attention)
(LI *et al.*, 2022)	CSI 300	Embeddings	Train / Test	Graph
(LIN *et al.*, 2022)	DJIA	Spacy embedding	Train / Test	CNN and LSTM
(YI *et al.*, 2023)	Chine stock market	Word2Vec	Train / Val / Test	LSTM
(JI *et al.*, 2023)	pharmaceutical stocks	BERT	Train / Test (cross validation)	LSTM
(ESLAMIEH; SHAJARI; NICKABADI, 2023)	DJIA	Word2Vec	Train / Test	BiLSTM
(MA *et al.*, 2023)	Chine stock market	BERT	Train / Val / Test	BiLSTM
(YANG *et al.*, 2023)	Stock big tech	BERT	Train / Val / Test	RNN, RF, ARNN, ACNN, GAT
(USMANI; SHAMSI, 2023)	PSX	VADER/HIV4 and LM	Train / Test	LSTM
(LIU *et al.*, 2023)	China A-Share	sentiment vectors (Sent)	Train / Test (sliding window)	Multi-head attention
(CHANG; ZHANG, 2023)	stock market	TF-IDF and Sentiment Score	Train / Test	CNN
(TAN *et al.*, 2024)	CSI 300	BERT	Train / Val / Test	CHARM
(ZHANG *et al.*, 2024)	stock market	Glove	Train / Val / Test	CoATSMP

Fonte: Elaborado pelo Autor.

Dentre os vinte e um estudos apresentados na Tabela 7, dez utilizaram modelos preditivos LSTM, quatro consideram modelos GNN, três empregaram *multi-head attention*, e cinco optaram por modelos variados ou integrados. Um ponto importante a destacar é a utilização de *multi-head attention* integradas com diferentes modelos de Grafos, LSTM e transformers. Um estudo pioneiro incorporou o mecanismo de atenção para prever movimentos de ações (XU *et al.*,

2020). A proposta, denominada *Stock Movement Predictive Network* (SMPN), obteve como principal resultado a redução de ruídos nas representações textuais (Glove) e uma melhoria no refinamento de *embeddings* contextuais. Essa abordagem utilizou *tweets* e séries de preços relacionados ao mercado de ações. Outra abordagem pioneira considerou uma GNN com dados multi-modais para prever séries temporais financeiras (CHENG *et al.*, 2022). O modelo proposto prevê movimentos de preços integrando fontes de relacionamentos *lead-lag*[10], incluindo preços históricos, fontes informativas e estratégias de grafos de conhecimento. A GNN heterogênea é construída com as fontes como nós e as relações do grafo de conhecimento como arestas. Além disso, um mecanismo de atenção de duas fases é empregado para garantir a interpretabilidade do modelo.

O próximo estudo apresenta um modelo que utiliza dois módulos LSTM paralelos para extrair *features* abstratas de cada modalidade e outra camada de representação compartilhada que funde essas *features* (WINDSOR; CAO, 2022). Para a modalidade de texto, o BERT é utilizado para analisar sentimentos em microblogs de redes sociais. O modelo incorpora indicadores de mercado e sentimentos dos investidores, tratando os dois tipos de dados de forma diferente para prever a taxa de câmbio USD/CNY. Em uma abordagem alternativa, utilizando GNN, um modelo é proposto considerando dados de negociação do mercado de ações e notícias como tipos distintos de nós (WANG *et al.*, 2022). A modelagem combina características de nós com base em arestas, pesos e direção de transferência de informações usando GRU para agregação de nós dos subgrafos. Os vértices agregados criam um grafo heterogêneo, com mecanismos de atenção construídos para nós homogêneos e heterogêneos. O modelo classifica o gráfico heterogêneo para prever a tendência do mercado de ações. Em outra abordagem proposta por (CHEN; HUANG, 2021), um modelo desenvolve um sistema de negociação que utiliza três técnicas: aprendizagem por reforço, análise de sentimento e aprendizado multimodal. Na primeira etapa, dados de preços e notícias são processados por *Deep Recurrent Neural Networks* (DRNN) de modo separado. Na segunda, o VADER é utilizado para extrair pontuações referentes à polaridade das notícias. Por fim, as *features* extraídas do DRNN são mescladas usando uma fusão inicial e as *features* são combinadas em uma representação única antes da classificação.

Em outro estudo que considera o modelo LSTM (FARIMANI *et al.*, 2021), uma abordagem é proposta para combinar a representação de documentos com base na distribuição de palavras sobre conceitos econômicos e no sentimento da notícia. O método, denominado Bag-of-Economic-Concepts baseado em BERT (BERT-BoEC), captura características temporais na distribuição de conceitos latentes no fluxo de notícias e o escore do sentimento dos títulos de notícias, empregando uma rede neural convolucional. Também extrai recursos informativos de indicadores técnicos. Essas representações e *embeddings* de indicadores técnicos são usadas para a previsão de séries temporais do índice do Forex. Os autores Zhang *et al.* (2022) propõem um modelo denominado *Transformer Encoder-based Attention Network* (TEANet) que emprega

[10] Um efeito lead-lag, especialmente em economia, descreve a situação em que uma variável (principal) é correlacionada de forma cruzada com os valores de outra variável (atrasada) em momentos posteriores.

mecanismos de atenção para capturar dependências entre os dados multi-modais. O modelo *transformers* possui dois codificadores e decodificadores empilhados com composições semelhantes. A primeira entrada é um pequeno texto de amostra, e a segunda é a série de preços. O modelo utiliza um extrator de *features*, incorporando um encoder *transformers*. A concatenação das *features* processadas é empregada para analisar os tweets e os preços das ações, integrando a influência de diversos fatores para a previsão do movimento das ações.

Um estudo embasou um modelo híbrido CNN-BiLSTM para prever tendências no mercado de ações, integrando eventos de notícias, padrões de sentimento e dados financeiros quantitativos (DARADKEH, 2022). Uma estratégia fundamental envolve o cálculo de pontuações de sentimento para itens de notícias utilizando BiLSTM, seguido pela fusão dessas pontuações com dados de ações. A estrutura proposta é avaliada por meio de dois estudos de caso nos setores imobiliário e de comunicações, utilizando dados coletados do Mercado Financeiro de Dubai (do inglês *Dubai Financial Market* - DFM). Em um estudo subsequente, uma abordagem baseada em grafo considera dados do mercado de ações, dados textuais, numéricos e relacionais (HUANG *et al.*, 2022). Um modelo de codificação baseado em LSTM é introduzido para dados de preços de ações, visando capturar dependências de longo prazo, enquanto os dados textuais são representados usando embeddings (BERT). As representações de nós de cada módulo de extração de *features* são atualizadas usando uma rede de atenção de grafo multinível (ML-GAT) em um grafo isomórfico convertido do Wikidata. Esse processo agrega seletivamente informações de vários tipos de relacionamento, combina-as com a representação do nó e, em seguida, alimenta o resultado em uma camada de previsão específica para a previsão. Outro estudo que utiliza GNN (LI *et al.*, 2022) emprega pesos nas arestas e mecanismos de transmissão de informações adaptados para dados de subgrafos, a fim de completar uma técnica de *screening* dos nós. A modelagem do grafo é realizada utilizando GRU e LSTM para agregar nós aos subgrafos. O mecanismo de atenção do metacaminho é integrado na GNN para classificar a volatilidade do mercado de ações.

O próximo estudo apresenta uma Rede Convolucional com Atuação Espaço-Temporal (Spatial-Temporal Attention-based Convolutional Network (STACN)), que capitaliza as vantagens de um mecanismo de atenção, uma CNN e LSTM para extrair informações textuais e numéricas visando a previsão do preço das ações (LIN *et al.*, 2022). O estudo introduz um *autoencoder* sequencial para processar textos, capturando relações entre cada par de textos e uma frase, gerando informações representativas para cada frase. Essa representação, juntamente com os dados de mercado, é conectada e utilizada como entrada nos modelos de previsão. Em outro estudo, os autores Yi *et al.* (2023) evitam os métodos convencionais léxicos e empregam *embeddings* de caracteres para a classificação de textos. Os dados de texto são rotulados manualmente com base no conhecimento financeiro. Em seguida, uma Rede Neural Convolucional Unidimensional (one-Dimensional Convolutional Neural Network (1DCNN)) é projetada para a classificação de sentimentos dos textos em nível de caractere, visando avaliar a confiabilidade das anotações de texto. Por fim, considerando a sequência temporal dos preços e a continuidade

da pós influência, combinam os dados de indicadores técnicos e sentimentos, e empregam o modelo LSTM para estimar as flutuações do mercado de ações em diferentes setores. Em uma abordagem adicional apresentada por Ji *et al.* (2023), os autores propõem um modelo de previsão de preços de ações utilizando a Memória de Longo Prazo com Atenção (Attention-based Long Short-Term Memory (ALSTM)), incorporando dados de preços, indicadores técnicos e informações léxicas de sentimento das mídias sociais. A principal contribuição inclui a integração de um modelo de classificação *fine-tuned* BERT e sentimento léxico combinados em uma única representação. Essa representação combinada, juntamente com preços e indicadores técnicos, serve como entrada para o modelo ALSTM

Um estudo propôs o *User2Vec*[11], que utiliza vetores representativos de mensagens de investidores de uma rede social no formato de *embeddings*, considerando o sentimento da mensagem, uma pontuação do investidor e dados do mercado (ESLAMIEH; SHAJARI; NICKABADI, 2023). A contribuição do estudo inclui a apresentação de um modelo LSTM bidirecional empilhado e uma CNN unidimensional (1D-CNN) para agregar todos os dados e prever o futuro do DJIA. O modelo *Multi-source Aggregated Classification* (MAC), proposto por Ma *et al.* (2023), integra três fontes de dados: dados de transações e indicadores técnicos, notícias sobre as ações-alvo e notícias sobre suas ações relacionadas. Para representar informações de notícias, um modelo pré-treinado chinês RoBERTa é utilizado para alinhar com notícias da empresa e preços das ações, visando prever o movimento do próximo dia. Em seguida, o *Graph Convolutional Networks* (GCN) é empregado para modelar as conexões entre a empresa-alvo e empresas relacionadas. Por fim, as fontes de dados são concatenadas e utilizadas como entrada no *Bidirectional* LSTM (BiLSTM) para predizer o preço do movimento da ação de seis empresas chinesas. No próximo estudo, uma abordagem introduz *Sequential Three-Way Decisions* (S3WE) para prever o preço de ações de grandes empresas de tecnologia (YANG *et al.*, 2023). O modelo constroem três tipos de modelos de multigranularidade para combinar preços e dados textuais: Análise de sentimento (senthigh); dados de alta frequência com *features* extraídas por BERT (berthigh); e dados de baixa frequência com *features* BERT (bertlow). A principal contribuição do estudo foi adicionar um mecanismo de atenção dentro da CNN e RNN para lidar com grandes *embeddings* e séries temporais.

Para prever a bolsa de valores do Paquistão, os autores Usmani e Shamsi (2023) propuseram o *LSTM-based Weighted Categorized News* (WCN-LSTM), utilizando três dicionários de sentimento (VADER, HIV4 e LM). O modelo integra categorias de notícias com seus pesos léxicos aprendidos simultaneamente, abrangendo pontuações de sentimento de notícias relacionadas ao mercado, ao setor e às ações. Cada categoria de notícias é processada por uma camada LSTM e concatenada em uma camada densa, que é então mesclada com o preço das ações e indicadores técnicos em outra camada densa. Outro modelo é proposto com base na estrutura *multi-head attention* em dois módulos (LIU *et al.*, 2023). O primeiro realiza o pré-processamento

[11] Github do User2Vec: https://github.com/samiroid/usr2vec

específico em indicadores técnicos do mercado de ações, indicadores financeiros e comentários de investidores individualmente. O segundo estrutura uma *multi-head attention* com dados que incorporam três *head attentions*: tecnologia, financeiro e opinião pública. O modelo utiliza a representação com dados heterogêneos gerados no estágio de pré-processamento e utiliza um mecanismo *multi-head attention* estruturado como modelo de previsão de ações da China A-share. Uma abordagem mais simples aplica um método que "julga"a tendência das ações com base nos sentimentos expressos nos comentários por especialistas em fóruns (CHANG; ZHANG, 2023). O modelo, chamado *Convolutional Neural Network with Sentiment Check* (CNN-SC), utiliza pontuações léxicas e TF-IDF das notícias combinadas com séries temporais para a previsão do preço das ações.

O penúltimo estudo, denominado *Context-Aware Hierarchical Attention Mechanism* (CHARM), é proposto por Tan *et al.* (2024). Inicialmente, sinais de mercado heterogêneos são fundidos e visualizados por meio de uma abordagem de aprendizagem baseada em tensores para compreender seus efeitos interativos nos movimentos das ações. Em seguida, o CHARM é introduzido para processar mídia textual, e, por fim, é apresentada uma representação parametrizada baseada em tensores para fundir múltiplos sinais de mercado. O último estudo da Tabela 7 apresenta uma avaliação dos métodos de fusão: concatenação, soma e *soft fusion*. Os autores Zhang *et al.* (2024) propõem o *Collaborative Attention Transformer fusion model for Stock Movement Prediction* (CoATSMP), que incorpora extração paralela de *features* de textos e preços, uma função em nível de parâmetro e um módulo de processamento em conjunto. O modelo emprega um *Transformer* baseado na função *Gaussian Error Linear Units* (GELU) para interagir e otimizar recursos de texto e preço de ações. Por fim, as *features* são processadas em conjunto e fundidas para completar a tarefa de previsão.

3.2.2 Respondendo as questões de pesquisa

Na abordagem da Questão de Pesquisa **QP1**, considerando os quarenta e dois estudos incluídos na fase final do mapeamento sistemático, observou-se que dois deles utilizaram técnicas de *Late Fusion*, enquanto dezenove optaram por *Early Fusion*, e vinte e um propuseram estratégias de *Joint Fusion*. É relevante destacar que os estudos que empregam estratégias de *Early Fusion* são mais antigos, representando uma abordagem pioneira em comparação com os estudos que adotaram *Joint Fusion*. Com o advento dos modelos *Transformers* e *Multi-head attention* nos últimos anos, observou-se um aumento significativo em estudos recentes que buscam estabelecer relações de colinearidade entre dados multi-modais. Esse fenômeno sugere uma tendência em direção a abordagens mais avançadas e eficazes na integração de informações provenientes de diferentes modalidades.

Para abordar a Questão de Pesquisa **QP2**, examinamos os quarenta e dois estudos detalhados nas Tabelas 5, 6 e 7. Dentre eles, cerca de vinte e dois optaram por empregar Redes Neurais Recorrentes, notadamente o LSTM. De maneira geral, observa-se uma predominância

de modelos baseados em redes neurais entre os estudos incluídos neste mapeamento sistemático. No entanto, é notável o crescimento do uso de estratégias de *Multi-head attention*, especialmente quando combinadas com uma variedade de modelos, indicando uma tendência recente na pesquisa. Essa evolução sugere uma busca por abordagens mais avançadas e flexíveis na integração de informações multimodais.

A Questão de Pesquisa **QP3** aborda as representações vetoriais de textos utilizadas em conjunto com séries temporais. Dos estudos selecionados, três empregaram representações baseadas na frequência de termos, quinze adotaram representações dependentes de contexto (como BERT), vinte e três optaram por representações independentes de contexto, como Word2Vec, enquanto um estudo não especificou essa informação. Em linhas gerais, as representações independentes de contexto são as mais prevalentes para tarefas preditivas, considerando todos os estudos.

Ao analisar as representações conforme a estratégia de fusão (EF, LF), dos estudos apresentados nas Tabelas 5 e 6, quatro utilizaram TF-IDF, doze optaram por representações dependentes de contexto, e cinco escolheram representações independentes de contexto. Observa-se que as representações que incorporam dados léxicos, por meio de pontuações obtidas por *Pos-tagging*, são amplamente adotadas nos estudos que consideram estratégias EL e LF. Esse resultado decorre do benefício de utilizar representações de baixa dimensão, superando desafios relacionados à maldição da dimensionalidade e à esparsidade.

Em relação à **QP3**, na estratégia *Joint Fusion* (Tabela 7), onze estudos preferiram representações independentes de contexto, enquanto dez optaram por representações dependentes de contexto. Uma vantagem evidente da estratégia JF é a capacidade de considerar representações multimodais de alta dimensão, processando-as em camadas intermediárias de modelos neurais, seja de maneira paralela ou integrada, e gerando uma nova representação vetorial como entrada para modelos de previsão. Dessa forma, o uso de representações dependentes de contexto para a estratégia JF pode ser considerado mais indicado do que para modelos de fusão EF e LF.

Para responder à **QP4**, a Tabela 3 apresenta detalhes de cada etapa de seleção dos estudos. Inicialmente, trezentos e trinta e um estudos foram selecionados com as *strings* de busca, dos quais 30 foram excluídos por serem estudos duplicados. Dos trezentos e um estudos restantes, sessenta foram considerados para leitura completa do documento. Na terceira etapa, apenas quarenta e dois foram selecionados para análise e agregação das informações. Vale ressaltar que o primeiro estudo incluído foi publicado em 2018, e observa-se uma quantidade crescente nos últimos anos de trabalhos que consideram dados multi-modais para a previsão do mercado financeiro.

3.2.3 Relacionamento entre os estudos

Esta seção analisa as citações entre os trabalhos selecionados na última etapa do mapeamento sistemático, conforme apresentado nas Tabelas 5, 6 e 7. A Figura 19 apresenta um sociograma que ilustra a relação de citações entre todos os trabalhos incluídos no mapeamento sistemático. As setas no sociograma representam as citações direcionadas, as cores verde, laranja e roxo representam trabalhos que utilizam de estratégias de *Early Fusion*, *Joint Fusion* e *Late Fusion*, respectivamente. Para visualizar a sequência das citações, o sociograma está temporalmente ordenados de cima para baixo.

Conforme pode ser observado na Figura 19, os estudos selecionados para análise final do mapeamento sistemático são relacionados entre eles. Considerando todos estudos, quarenta e oito citações foram levantadas, sendo que: i) treze estudos são citados; ii) vinte e nove não são citados; e, iii) seis não citam e não são citados. Os estudos publicados por Zhang *et al.* (2018), Picasso *et al.* (2019) e Li, Wu e Wang (2020) são os que tiveram mais citações, sendo que cada um recebeu nove, totalizando em vinte e sete considerando apenas os três trabalhos. Além de serem estudos pioneiros com estratégia *Early Fusion*, esses apresentam semelhanças quanto ao uso de representações textuais independentes de contextos (Word2Vec, SenticNet).

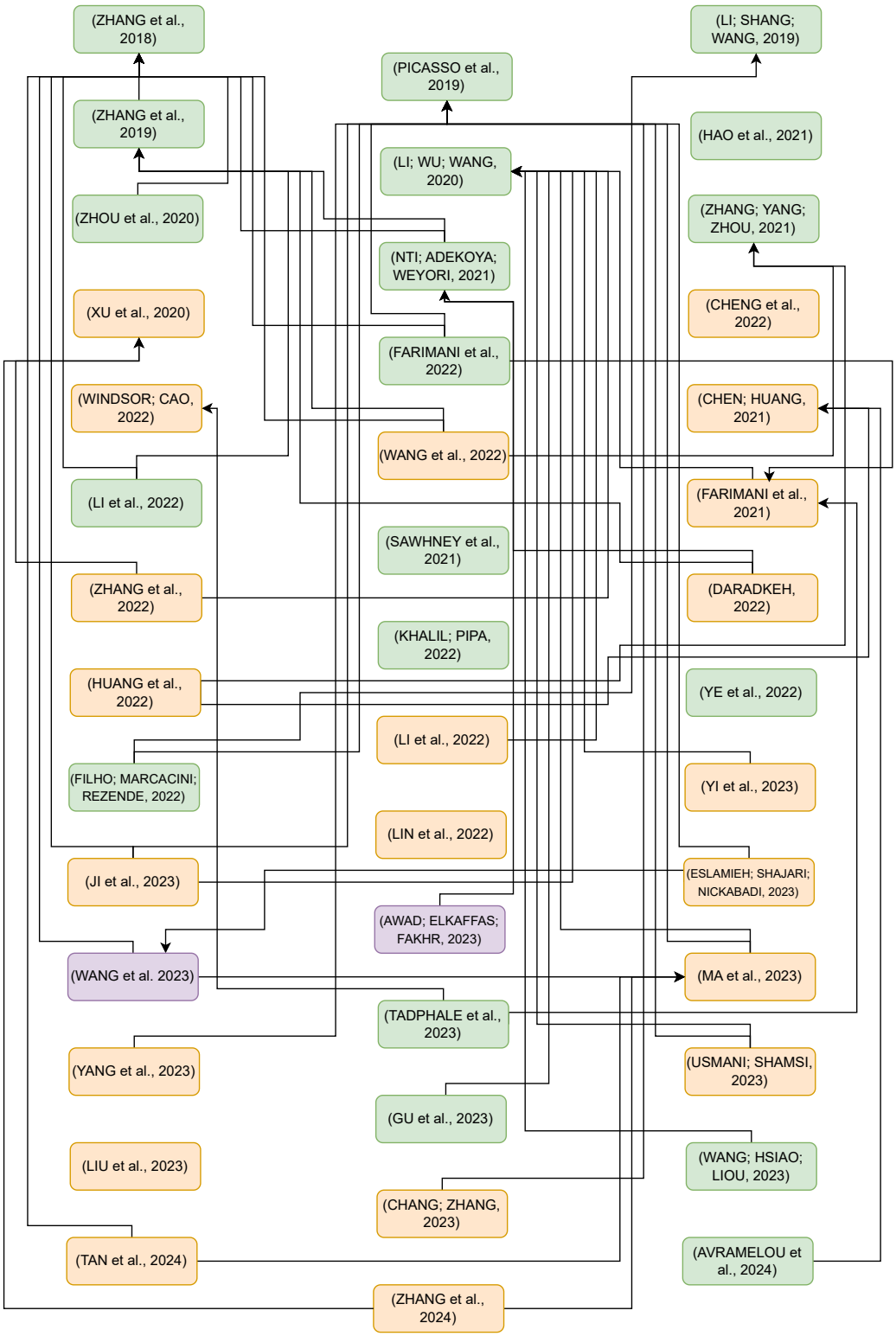
Abordagens publicadas recentemente têm recebido atenção de outros trabalhos, como é o caso de Ma *et al.* (2023). Este estudo apresenta relevância por integrar três tipos de dados, utilizar de representação textuais atuais, e técnicas de GCN e BiLSTM para modelar conexões entre os dados multi-modais. Naturalmente, os estudos publicados recentemente apenas citam trabalhos anteriores, e uma semelhança observada são que esses esses utilizam de técnicas de *Joint Fusion* tendem a citar trabalhos que consideram a mesma estratégia.

3.3 Considerações Finais

Nos últimos anos, tem havido um aumento significativo no interesse por integrar diferentes tipos de dados para aprimorar o desempenho dos modelos ou considerar fatores externos aos dados. As tarefas de mineração de textos desempenharam um papel crucial nesse cenário, proporcionando uma estruturação eficiente para as representações vetoriais. Uma variedade de técnicas de fusão de informações pode ser aplicada para enriquecer essas representações, resultando em uma compreensão mais abrangente e precisa dos dados. Nesse contexto, este capítulo apresentou um mapeamento sistemático com o objetivo de explorar os modelos de representação disponíveis na literatura, bem como os modelos empregados em suas respectivas tarefas preditivas. Essa abordagem permite uma visão abrangente das práticas atuais, destacando as diversas estratégias utilizadas para integrar informações textuais em séries temporais.

Os estudos analisados neste mapeamento destacam vários desafios de pesquisa. As abordagens pioneiras apresentadas na literatura, ao explorarem estratégias multi-modais, optaram predominantemente por utilizar técnicas de *Early Fusion*. De maneira geral, essas abordagens

Figura 19 – Sociograma de citações dos estudos selecionados.



Fonte: Elaborado pelo autor.

consideraram representações vetoriais de textos e séries temporais, buscando incorporar esses dados em uma representação única. Isso foi feito por meio de estratégias como concatenação

linear, soma das *features* ou processamento paralelo. Essas estratégias envolvem dados pré-processados separadamente, os quais são posteriormente utilizados como entrada nos modelos de previsão. Uma vantagem notável desses modelos é a consideração dos dados nativos de cada modalidade. No entanto, é importante salientar que essas abordagens também apresentam limitações, como a falta de exploração de interações mais complexas entre as modalidades e a potencial perda de informações valiosas na etapa de fusão.

Com os avanços dos modelos *Transformers*, GNN e a combinação de técnicas de *multi-head attention* com modelos variados, abordagens recentes têm emergido para pré-processar dados de maneira conjunta. O objetivo é aprender correlações significativas entre dados multimodais e gerar representações que capturem interações mais complexas entre eles. Entretanto, uma desvantagem potencial dessas estratégias é a possível perda de interpretabilidade nos modelos de previsão. Em contrapartida, essas representações podem ser geradas em dimensões mais compactas e, dessa forma, facilitar a compreensão dos padrões subjacentes, ao mesmo tempo que podem oferecer um desempenho aprimorado nas tarefas preditivas. Essa tensão entre complexidade e interpretabilidade destaca um ponto importante e que merece atenção de investigação aprofundada em futuras pesquisas na área.

CAPÍTULO

4

ENRIQUECENDO SÉRIES TEMPORAIS COM INFORMAÇÃO DE TEXTOS PARA TAREFAS DE REGRESSÃO

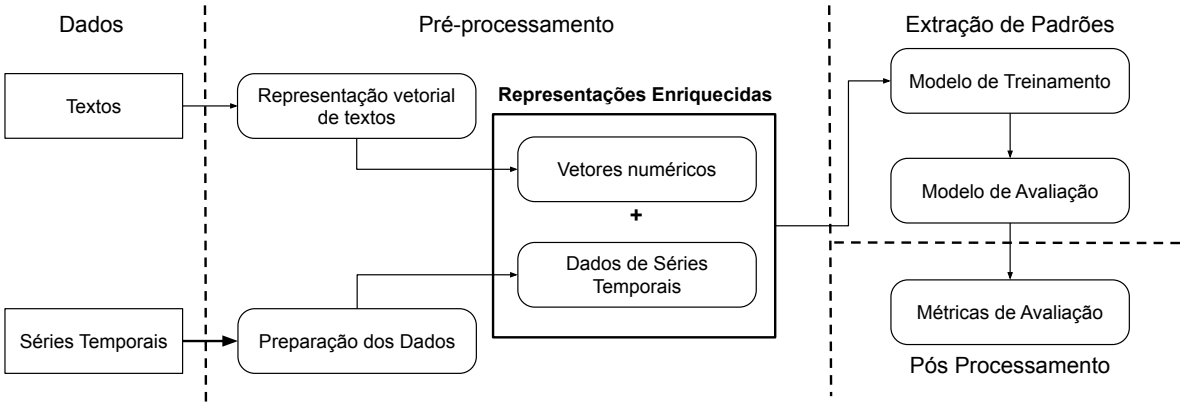
Após conduzir estudos preliminares e realizar um mapeamento sistemático, o presente capítulo apresenta uma proposta com diferentes abordagens de enriquecer séries temporais com informações textuais, com ênfase na tarefa de regressão. Durante a investigação, dados do mercado da soja/milho são utilizados em português e inglês. Inicialmente, foi enfrentado um desafio em relação à disponibilidade de conjuntos de textos alinhados temporalmente com os dados de séries temporais. Dessa forma, as duas primeiras abordagens realizaram a previsão de preços dos contratos futuros mensais de commodities foram abordadas por meio do agrupamento de notícias intra-mês. Este enfoque resultaram na publicação de dois trabalhos (FILHO *et al.*, 2020; FILHO; MARCACINI; REZENDE, 2021). Os resultados detalhados de cada abordagem podem ser encontrados nas Seções 4.1 e 4.2, respectivamente.

Posteriormente, para realizar previsões intra-dia no contexto do mercado da soja, foi necessário implementar técnicas de *web scraping* para coletar, no mínimo, uma notícia por dia e representativa do mercado da soja. Com base nos conhecimentos adquiridos nas avaliações anteriores, a terceira abordagem apresenta uma representação que incorpora um conjunto finito de termos essenciais do domínio para enriquecer a série de preços. A estratégia de previsão de uma única etapa à frente foi empregada para antecipar o preço da soja no próximo dia. Essa abordagem resultou na publicação do trabalho (FILHO; MARCACINI; REZENDE, 2022), que é detalhado na Seção 4.3.

A estratégia de *early fusion* é adotada em ambas configurações experimentais, uma vez que o objetivo é de propor representações enriquecidas para modelos de previsão em tempo real. Em cada instância, diferentes tipos de séries temporais e coleção de documentos são empregados em tarefas de previsão. Apesar das variações nas estratégias de previsão e nos dados utilizados

em cada avaliação, a abordagem de enriquecer séries temporais permaneceu-se consistente. A Figura 20 ilustra as etapas executadas nas três abordagens propostas nesse capítulo, apresentadas nas Seções 4.1, 4.2 e 4.3.

Figura 20 – Etapas realizadas nas três abordagens da presente proposta.



Fonte: Elaborado pelo Autor.

A primeira etapa de **Dados**, envolve a aquisição da coleção de documentos de texto alinhados temporalmente com os dados de séries temporais. Na etapa de **Pré-processamento**, os dados são transformados e as representações são enriquecidas. As três abordagens seguem a mesma estratégia de *Early Fusion*, em que as representações de textos são incorporadas com os dados de séries temporais. A última etapa, de **Extração de padrões**, engloba a utilização de estratégias de previsão e modelos preditivos. Nesta fase, a ênfase é propor diferentes estratégias de previsão usando as representações enriquecidas.

Na etapa de pré-processamento, uma série temporal S de tamanho m é definida como uma sequência de observações $S = (s_1, s_2, ..., s_m)$, em que $s_t \in \mathbb{R}^d$ representa uma observação s no tempo s_t com $d > 1$ (ST multivariada). Na etapa de treinamento do modelo de previsão, uma subsequência de tamanho r é extraída da série temporal S. Dessa forma, uma subsequência $S_u = (s_u, s_{u+1}, ..., s_{u+r})$ é definida em que u indica o período de tempo da primeira observação da subsequência, com $1 \leq u \leq m - r$. Várias subsequências de S com tamanho r são extraídas com horizonte de previsão h. Dessa forma, cada subsequência S_u é associada com uma variável dependente y_{u+h} (etapas à frente), assim gerando um conjunto de treinamento.

$$X = \{(S_{u_1}, y_{(u+h)_1}), (S_{u_2}, y_{(u+h)_2}), ..., (S_{u_n}, y_{(u+h)_n})\} \tag{4.1}$$

em que o X representa todo o conjunto de treinamento.

As três abordagens subsequentes são realizadas com intuito de obter modelos de representação de séries temporais enriquecida com textos, sejam enriquecidas por uma BoW ou recursos semânticos. Nas avaliações que possuem a BoW, tarefas para diminuir o problema da dimensionalidade e esparsidade são realizadas, como: remoção de *stopwords*, uso de n-gramas e a

radicalização dos termos. Em seguida, alinhamentos de todos documentos de textos relacionados para cada subsequência da série temporal são realizados, isto é, o conjunto de documentos que estão no período de tempo S_{u+r} e suas respectivas representações no espaço vetorial, definida na Equação 4.2 (*Features* de Textos).

$$\begin{aligned} FT(u,r) &= Q(T,u,r) \\ &= \{B(d_1), B(d_2), ..., B(d_k)\} \\ &= \{\vec{v}_{d_1}, \vec{v}_{d_2}, ..., \vec{v}_{d_k}\} \end{aligned} \tag{4.2}$$

em que FT é um subconjunto (Q) de textos no tempo (T), u indica o período de tempo para primeira coleção de documentos (textos), e r o tamanho da subsequência (exemplo, intervalos de dias ou meses). A representação vetorial B de cada documento (d_k) é expressada como um vetor $\vec{v}_{d_k}$. Assim, cada vetor numérico associado com a subsequência é expressado como um vetor médio dos vetores de todos documentos de textos, conforme definido na Equação 4.3 (*Features* Enriquecidos).

$$FE(u,r) = \sum_{\vec{v}_d \in FT(u,r)} \frac{\vec{v}_d}{|FT(u,r)|} \tag{4.3}$$

A Representação Enriquecida (RE) da subsequência é formada por um vetor concatenado entre os atributos das séries temporais e os *Features* Enriquecidos, $RE(u,r) = S(u,r) \oplus FE(u,r)$, ilustrada na Figura 21.

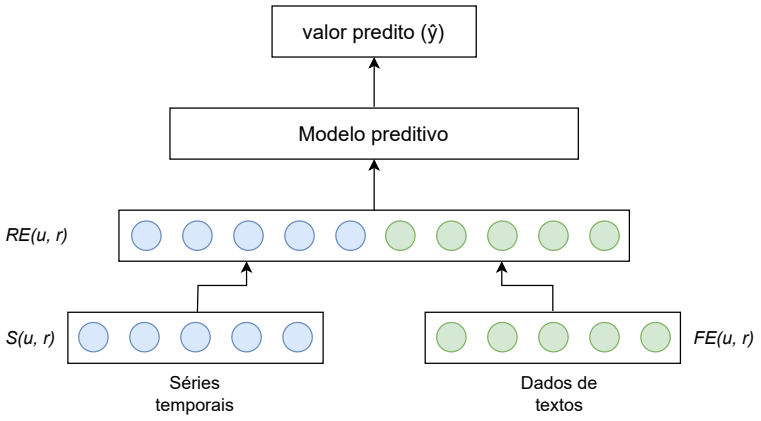


Figura 21 – Modelo *early fusion* entre dados de séries temporais e dados de textos.

Em cada avaliação, diferentes tamanhos de r é empregado para previsão de h etapas a frente da variável y, expressado na Equação 4.4.

$$X = \{(RE_{u_1}, y_{(u+h)_1}), (RE_{u_2}, y_{(u+h)_2}), ..., (RE_{u_n}, y_{(u+h)_n})\}. \tag{4.4}$$

Na sequência, diferentes *vetores numéricos* de textos e séries temporais, ilustrados na Figura 20, são apresentadas na etapa de pré-processamento. Representações textuais com base na matriz-atributo valor (como a BoW) e vetores *embeddings* são considerados como variável $FE(u,r)$ para enriquecer as séries temporais $S(u,r)$. Nas próximas Seções, três diferentes abordagens são avaliadas, apresentando os **dados**, o **pré-processamento** para obter as representações enriquecidas, **configuração experimental** e o **resultados/discussão** de cada avaliação.

4.1 Representações de Séries Temporais enriquecidas com BoW

Na primeira abordagem, um modelo de representação com dados textuais e séries temporais é apresentado para previsão do preço médio (mensal) do mercado da soja e do milho (FILHO *et al.*, 2020). O método considera a representação de BoW e séries temporais referente aos dados apresentados na Tabela 8. Tendo em vista que os dados das séries temporais possuem muitos atributos, uma estratégia oriundo do modelo da árvore de decisão é empregado para obter os atributos com maior ganho de informação. Dessa forma, na avaliação considero-se quatro representações: Séries Temporais (ST), Séries temporais combinadas com BoW (ST/Textos), ST com atributos extraídos da árvore de decisão (ST(DT)), e ST com atributos extraídos da árvore de decisão e enriquecidas com textos (ST(DT)/Texto).

Tabela 8 – Visão Geral das séries temporais e a quantidade de textos.

ST	Período	Meses	Atributos (ST)	Atributos (ST/DT)	Textos
Milho	Jan 2014 to Fev 2020	73	112	44	3671
Soja	Jan 2014 to Fev 2020	73	70	22	11254

Fonte: elaborado pelo Autor.

A fonte de dados de séries temporais utilizada na avaliação é proveniente do *World Agricultural Supply and Demand Estimates* (WAOB) do Departamento de Agricultura dos Estados Unidos (USDA), disponível no site do Kaggle[1]. A Tabela 8 descreve a Série Temporal (ST), o período, a quantidade de meses e dos atributos dos conjuntos de dados. Os atributos representam várias características de séries temporais, como área plantada, área colhida, rendimento, importações, oferta, demanda e outras estimativas dos países com maior produção de milho e soja. Os dados originais de preço são obtidos da *Chicago Board of Trade* (CBOT), disponíveis no site do CME[2]. A CBOT é uma corretora designada para o CME Group para contratos futuros de negociação de contratos de commodities agrícolas, e os preços praticados na CBOT são uma referência nos preços mundiais.

Com o objetivo de extrair de modo automático os atributos mais representativos das séries temporais, foi necessário considerar os dados da série temporal como uma tarefa de

[1] https://www.kaggle.com/ainslie/usda-wasde-monthly-corn-soybean-projections
[2] http://www.cmegroup.com/

classificação. Dessa forma, a cada etapa de tempo da série (mensal), dois rótulos numéricos são atribuídos para representar a variação no preço médio mensal do milho e da soja. O rótulo 0 representa que o preço estava neutro ou abaixo, enquanto o rótulo 1 representa que o preço subiu consideravelmente em relação ao mês anterior. A árvore de decisão é processada dez vezes (10 *folds*), e os atributos que não são considerados em nenhuma iteração (fold), isto é, descartados na construção de todos os modelos da árvore de decisão para a tarefa de classificação, também são descartados no modelo de regressão. É importante ressaltar que essa a estratégia é utilizada apenas para obter os atributos mais representativos nos conjunto de dados da série temporal.

Em relação a coleção de textos, este trabalho utilizou dados textuais extraídos do site Soybean & Corn Advisor[3]. Desde 2009, o site fornece notícias e informações diárias em inglês sobre a produção de soja e milho, relacionadas aos ciclos de crescimento sul-americanos, clima, infraestrutura, uso da terra, produção de etanol e combustíveis alternativos. A Tabela 8 (Textos) apresenta a quantidade de textos utilizados em cada avaliação da presente abordagem.

4.1.1 Pré-processamento

Na presente abordagem, o valor u das subsequências indica uma data. A data é usada para delimitar as etapas de tempo. Por exemplo, dada uma granularidade mensal de uma série temporal, se u representa o mês "Jan. 2020" de uma subsequência de tamanho $r = 3$, então os períodos de tempo envolvidos na subsequência são $(u,r) = \{Jan.2020, Fev.2020, Mar.2020\}$. Este período de tempo é importante para compor uma função de alinhamento temporal entre a série temporal e a base de conhecimento textual T. Seja $Q(T,u,r)$ uma função de alinhamento que retorna o conjunto de k documentos $T_{(u,r)}$ dado o intervalo de tempo em (u,r), usando a data de publicação de cada documento d.

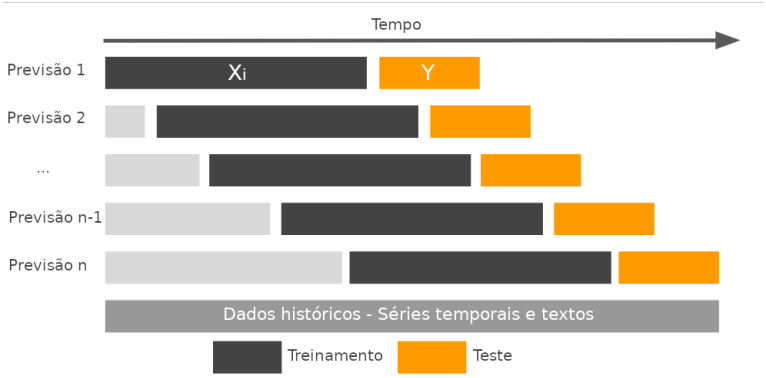
Os documentos $d \in T$ são representados em um modelo de espaço vetorial. Nesta abordagem, são extraídos atributos da BoW com n-grama = 2, excluindo termos (palavras) com ocorrência abaixo de 5% e acima de 95% nos textos. Métodos de mineração de texto disponíveis em (AGGARWAL, 2014) são utilizados, onde a função $B(d) = \vec{v}_d = \{w_1, w_2, ..., w_b\}$, com b definindo o tamanho do espaço BoW. A função $B(d)$ mapeia o texto em linguagem natural (contido em d) para uma representação vetorial de b dimensões.

4.1.2 Configuração experimental

Na avaliação da abordagem é utilizado o método de janela deslizante para as fases de aprendizado (treinamento) e validação (teste). A Figura 22 ilustra como o método é aplicado nesta abordagem. Conforme apresentado na Equação 2.11 (pág. 40), considere X a janela de treinamento, Y a variável dependente (valor previsto) e (n) o número de testes.

[3] http://www.soybeansandcorn.com

Figura 22 – Estratégia de janela deslizante



Fonte: Elaborado pelo Autor.

Na avaliação da presente abordagem, seis "janelas" são utilizadas para avaliar a precisão da previsão, em que $Y = 1$, $X_i = (2, 6, 12, 24, 36, 48)$ representa o conjunto de janelas definidas para treinamento, e $n = (71, 67, 61, 49, 37, 25)$ o número de testes (previsões), respectivamente. À medida que o tamanho das janelas aumenta no treinamento, o número de testes diminui. A variável y_i mostrada na Equação 2.29 (pág. 60) são os resultados Y em cada teste ilustrado na Figura 22.

4.1.3 Resultados e Discussão

O valor de MAPE é usado para avaliar o desempenho dos experimentos da presente abordagem. As representações ST, ST/Textos, ST(DT) e ST(DT)/Textos são aplicadas ao modelo de regressão SVR com três kernels: Polinomial(P), RBF(R) e Sigmoid(S) apresentados na Tabela 1 (pág. 40). A Tabela 9 apresenta os resultados da previsão do preço do milho. O valor em negrito representa o menor valor MAPE em cada janela (linhas da tabela), e o sublinhado o menor valor obtido na representação em cada modelo de regressão (coluna).

O modelo SVR(P) com a ST obteve dois resultados com o menor MAPE (3.82% e 3.74%), o ST(DT) obteve quatro resultados (4.58%, 5.59%, 4.11% e 3.69%), e o ST(DT)/Textos obteve um resultado (3.82%). Já modelo SVR(R) teve dois menores valores de MAPE para a representação ST/Textos (3.24% e 3.23%) e quatro para ST(DT)/Textos (4.48%, 4.82%, 3.67% e 3.52%). Enquanto para o SVR (S), os melhores resultados ficaram dispersos entre as representações. A Tabela 10 compara os melhores resultados (sublinhados) entre os modelos da Tabela 9.

Observa-se que a representação ST/Textos e o modelo SVR(R), com valor de 3.23%, obteve o melhor desempenho em relação à todos resultados (Tabela 9). Em relação aos resultados obtidos a partir da série temporal da soja, verifica-se que os valores são semelhantes aos resultados obtidos a partir da série temporal do milho. As representações ST(DT) e ST(DT)/Textos tiveram os menores valores de MAPE na configuração de parâmetros do modelo de regressão, conforme mostrado na Tabela 11. A Tabela 12 apresenta os resultados que obtiveram o menor MAPE de

Tabela 9 – Milho - Resultdos (MAPE)

Modelo	Treino	ST	ST/Textos	ST(DT)	ST(DT)/Textos
SVR(P)	2	4.59%	4.60%	**4.58%**	4.59%
	6	5.65%	5.64%	**5.59%**	5.62%
	12	**3.82%**	3.83%	3.83%	**3.82%**
	24	4.19%	4.15%	**4.11%**	4.16%
	36	**3.74%**	3.79%	3.79%	3.78%
	48	3.80%	3.83%	**3.69%**	3.72%
SVR(R)	2	4.57%	4.49%	4.58%	**4.48%**
	6	5.22%	4.92%	5.13%	**4.82%**
	12	3.90%	3.70%	3.86%	**3.67%**
	24	3.81%	3.61%	3.73%	**3.52%**
	36	3.49%	**3.24%**	3.53%	3.36%
	48	3.48%	**3.23%**	3.45%	3.27%
SVR(S)	2	**4.61%**	4.62%	4.63%	4.64%
	6	5.68%	5.69%	**5.67%**	5.70%
	12	3.99%	**3.98%**	3.99%	3.99%
	24	4.42%	4.42%	4.42%	**4.41%**
	36	4.10%	**4.09%**	4.10%	4.10%
	48	4.14%	4.14%	**4.13%**	4.14%

Fonte: elaborado pelo Autor.

Tabela 10 – Milho - Melhores resultados

Modelo	ST	ST/Textos	ST(DT)	ST(DT)/Textos
SVR(P)	3.74%	3.79%	**3.69%**	3.72%
SVR(R)	3.48%	**3.23%**	3.45%	3.27%
SVR(S)	3.99%	**3.98%**	3.99%	3.99%

Fonte: elaborado pelo Autor.

cada modelo de regressão (valores sublinhados da Tabela 11).

Tabela 11 – Soja - Resultados (MAPE)

Modelo	Treino	ST	ST/Textos	ST(DT)	ST(DT)/Textos
SVR(P)	2	4.44%	4.45%	**4.41%**	4.46%
	6	6.24%	6.37%	**6.12%**	6.36%
	12	4.79%	5.35%	**4.52%**	5.41%
	24	**5.57%**	5.61%	5.58%	5.60%
	36	5.08%	5.52%	**4.45%**	4.88%
	48	3.90%	4.59%	**3.87%**	4.48%
SVR(R)	2	4.48%	4.31%	4.49%	**4.24%**
	6	6.27%	5.20%	6.27%	**4.96%**
	12	6.43%	4.50%	6.45%	**4.37%**
	24	5.67%	4.04%	5.60%	**3.79%**
	36	5.52%	4.77%	5.64%	**4.24%**
	48	6.61%	6.09%	6.50%	**4.88%**
SVR(S)	2	4.50%	4.49%	4.49%	**4.48%**
	6	6.60%	**6.58%**	6.59%	6.59%
	12	6.86%	6.86%	6.86%	**6.85%**
	24	5.77%	5.77%	**5.75%**	5.76%
	36	6.08%	**6.07%**	6.10%	6.09%
	48	7.47%	7.45%	7.45%	**7.43%**

Fonte: elaborado pelo Autor.

Observa-se que a representação ST(DT)/Textos e o modelo SVR(R), com valor de 3.79%,

Tabela 12 – Soja - Melhores resultados

Modelo	ST	ST/TexST	ST(DT)	ST(DT)/Textos
SVR(P)	3.90%	4.45%	**3.87%**	4.46%
SVR(R)	4.48%	4.04%	4.49%	**3.79%**
SVR(S)	4.50%	4.49%	4.49%	**4.48%**

Fonte: elaborado pelo Autor.

obteve o melhor desempenho em relação a todos resultados (Tabela 12). Analisando os resultados das Tabelas 9 e 11 relacionados ao modelo SVR (P), as representações ST(DT) alcançaram os menores valores de MAPE (3.69% e 3.87%, respectivamente), sendo obtido com o período de treinamento de 48 meses. Os resultados indicam que o tamanho de janela maior de treinamento obtém um grande volume de textos, e o conjunto de características extraídas podem ser mais discriminantes do que um tamanho de janela pequena.

Em comparação com os resultados do modelo SVR (R) nas Tabelas 9 e 11, as representações ST/Textos e ST(DT)/Textos obtiveram os menores valores de MAPE. Os resultados do milho obtiveram melhores resultados com janelas maiores para ST/Textos, e ST(DT)/Textos alcançou melhor precisão de previsão para um período de treinamento de 24 meses. ST/Textos para o milho, com 48 meses de treinamento, obteve a melhor precisão de previsão com MAPE de 3,23%, e 3,79% para a soja com 24 meses de treinamento.

Analisando os resultados do modelo SVR (S) nas Tabelas 9 e 11, as quatro representações obtiveram o menor valor de MAPE em diferentes períodos de treinamento. As representações ST/Textos e ST(DT)/Textos alcançaram menor MAPE em alguns cenários. Neste experimento, o melhor desempenho para os períodos de milho e soja não coincidiu. Para o milho, o período de treinamento de 12 meses obteve o menor valor de MAPE com 3,98%, enquanto para a soja, o período de 2 meses obteve o melhor desempenho com 4,48%.

Os resultados experimentais evidenciaram que as representações combinadas com textos aprimoram moderadamente o desempenho das previsões em comparação com modelos que consideram exclusivamente séries temporais. Esta análise ressaltou que as representações integradas com informações textuais apresentam uma alternativa valiosa para aprimorar a precisão das previsões de preços em tarefas de regressão. Não foram realizados testes de significância, pois não houve variação significativa entre os modelos preditivos. Essa avaliação está disponível no repositório do GitHub[4].

Apesar da simplicidade do modelo *early fusion* apresentado neste trabalho, até o momento da publicação, são identificados poucos trabalhos semelhantes a presente abordagem. Desde a sua divulgação, o método tem recebido citações em outros estudos relevantes, evidenciando sua contribuição para a área de pesquisa. No entanto, é importante ressaltar que a representação BoW empregada para enriquecer as séries temporais demonstra limitações, especialmente pela ausência de recursos semânticos. Essa limitação restringe a capacidade do modelo em capturar

[4] https://github.com/ivanfilhoreis/reisfilho_kdmile2020

eventos significativos que possam influenciar o cenário financeiro.

Diante desse contexto, avaliações subsequentes são conduzidos com o propósito de explorar representações mais avançadas, que incorporam aspectos semânticos. Adicionalmente, para análises complementares, outros modelos baseados em redes neurais recorrentes são empregados, ampliando assim a compreensão e avaliação do desempenho do método apresentado.

4.2 Representação de ST enriquecida com representação dependente de contexto.

Na segunda abordagem, um modelo de representação é apresentado usando séries temporais com recursos semânticos de textos para previsão do preço médio do mercado do milho e da soja (FILHO; MARCACINI; REZENDE, 2021). Além das representações de BoW consideradas em avaliações anteriores, representações vetoriais com base no modelo pré-treinado do BERT são utilizadas na variável FE da Equação 4.3 (85). Nessa avaliação, dados de séries temporais do milho e da soja mencionados na Tabela 8 foram atualizados em quantidade de meses (de 73 para 84 meses) e considerados para comparar o desempenho dos modelos de representações enriquecidas.

4.2.1 Configuração Experimental

A presente abordagem apresenta duas representações de séries temporais enriquecidas (*RE*): i) Séries Temporais concatenadas com BoW (TS/BoW); e, ii) Séries Temporais concatenadas com *texts embeddings* do BERT (TS/BERT). As duas representações são comparadas com os resultados das Séries Temporais (TS). A estratégia de janela deslizante com única-etapa à frente e horizonte de previsão $h = 1$ (previsão de uma etapa de tempo à frente) é realizada para avaliação. Diferentes tamanhos de janelas (X_i) são utilizados para o conjunto de Treino (12, 24, 36, 48 e 60), enquanto que o teste é uma única etapa a frente ($y = 1$), ou seja, a previsão do preço no próximo mês 22. Em relação aos dados textuais, manchetes são utilizadas em vez de notícias. Em relação à BoW, a representação vetorial de textos é considerada com termos unitários, excluindo os termos com ocorrência abaixo de 20% e acima 80% nos textos. Outras configurações foram testadas, porém, não tiveram ganho de desempenho significativos.

Os modelos de regressão SVR e LSTM são utilizados para avaliar o desempenho de previsão das representações. Os parâmetros utilizados no modelo SVR são os apresentados na Tabela 1 (pág. 40). Após uma avaliação de hiperparâmetros a ser utilizado, o modelo LSTM consistiu na seguinte estrutura: uma camada LSTM com 100 unidades ocultas e função de ativação tangente hiperbólica (*tanh*). Essa camada não retorna sequências (*return_sequences=False*); uma camada densa (*fully connected*) com 60 unidades e função de ativação ReLU; Uma camada densa com um unidade e função de ativação linear; O otimizador é o Adam com uma taxa de

aprendizado (*lr*) de 0.001; e, a função de perda (*loss*) usada é o erro médio quadrático (*mse*).

4.2.2 Resultados e Discussão

As Tabelas 13 e 14 apresentam os resultados da previsão do preço do milho e da soja, respectivamente. Os valores em negrito representam a média do menor valor de MAPE em cada janela de treinamento, os resultados sublinhados representam o melhor desempenho em cada representação e valores entre parênteses o menor MAPE entre todos os resultados de cada representação.

Tabela 13 – Milho: Resultados da previsão da Série Temporal enriquecidas com Inf. textuais.

Modelo	Treino	TS	TS/BoW	TS/BERT
SVR(P)	12	4.89%	4.93%	**4.88%**
	24	**5.17%**	5.20%	5.49%
	36	5.06%	**5.04%**	5.05%
	48	6.32%	6.25%	**5.99%**
	60	**6.83%**	6.84%	7.37%
SVR(R)	12	4.40%	(**4.27%**)	5.23%
	24	(**4.21%**)	4.28%	5.46%
	36	**4.27%**	4.29%	5.29%
	48	**4.76%**	4.80%	5.54%
	60	6.18%	**6.09%**	7.07%
SVR(S)	12	5.04%	**5.03%**	5.04%
	24	**5.35%**	5.36%	5.36%
	36	**5.24%**	5.25%	5.25%
	48	**5.61%**	5.62%	5.62%
	60	**7.00%**	7.01%	7.02%
LSTM	12	4.35%	**4.28%**	(4.29%)
	24	**4.38%**	5.15%	5.25%
	36	**4.54%**	5.02%	4.92%
	48	5.68%	5.55%	**5.33%**
	60	**6.56%**	7.11%	6.84%

Fonte: elaborado pelo Autor.

Analisando os resultados da previsão do preço do milho por modelos de previsão (valores em negrito), o modelo SVR (P) com séries temporais (TS) obteve dois resultados com menor MAPE (5.17% e 6.83%), o TS/BoW obteve um (5.04) e a representação TS/BERT atingiu dois melhores resultados (4.88% e 5.99%). O melhor desempenho do kernel polinomial foi o TS/BERT no período de 12 meses de treinamento com MAPE de 4.88%. O modelo SVR (R) usando os dados das TS obteve três resultados com melhores resultados (4.21%, 4.27% e 4.76%), a representação TS/BoW obteve dois (4.27% e 6.09%) e o TS/BERT não teve nenhum. O período de 24 meses de treinamento usando a TS atingiu o menor valor de MAPE com 4.21%. Em relação ao modelo SVR (S), os resultados foram semelhantes em cada janela de treinamento. O período de 12 meses alcançou o melhor desempenho com 5.03% usando o TS/BoW. Os resultados do LSTM usando a TS obteve três menores valores de MAPE (4,38%, 4.54% e 6.56%), TS/BoW obteve um (4.28%), e o TS/BERT também apenas um (5.33%). O menor valor de MAPE do modelo LSTM foi no período de 12 meses usando a TS/BoW com valor 4.28%. Observando o

Tabela 14 – Soja: Resultados da previsão da Série Temporal enriquecidas com Inf. Textuais

Modelo	Treino	TS	TS/BoW	TS/BERT
SVR(P)	12	**5.29%**	6.15%	6.58%
	24	6.06%	**5.78%**	6.39%
	36	**5.49%**	5.57%	6.04%
	48	**5.60%**	6.34%	6.56%
	60	**4.52%**	4.88%	6.52%
SVR(R)	12	4.76%	**4.36%**	6.57%
	24	4.61%	**(4.21%)**	6.19%
	36	4.93%	**4.61%**	6.52%
	48	5.39%	**4.82%**	7.76%
	60	5.65%	**5.34%**	8.26%
SVR(S)	12	**6.65%**	6.66%	6.66%
	24	**5.95%**	5.96%	(5.96%)
	36	**6.87%**	6.88%	6.88%
	48	**8.43%**	8.44%	8.45%
	60	**8.93%**	8.94%	8.95%
LSTM	12	**(4.31%)**	5.43%	6.15%
	24	**4.52%**	5.44%	6.43%
	36	**4.70%**	5.38%	6.11%
	48	**5.55%**	6.04%	6.83%
	60	**6.43%**	6.68%	7.77%

Fonte: elaborado pelo Autor.

melhor resultado de cada representação (valores entre parênteses), o modelo RBF no período de 24 meses com 4.21% para TS e 4.27% para TS/BoW no período de 12 meses, alcançaram melhores desempenhos. Em relação ao TS/BERT, o valor de 4.29% foi o melhor desempenho obtido da representação.

Em relação à previsão da soja, a representação TS aplicados no modelo SVR(P) obtiveram melhores resultados em relação ao TS/BoW e TS/BERT, sendo que o período de 60 meses com valor de 4.52% obteve o menor valor de MAPE. A representação TS/BoW para o modelo SVR(R) obteve os menores valores de MAPE, em que no período de 24 meses com valor de 4.21% alcançou o melhor desempenho. O modelo SVR(S) obteve os melhores MAPE usando a representação TS, em que o período de 24 meses com 5.95% teve o menor valor de MAPE. Em relação ao LSTM, a representação TS obteve os menores valores de MAPE em todas as janelas de treinamento. Analisando o melhor resultado de cada representação, observa-se que o período de 12 meses da TS, o período de 24 meses da TS/BoW e o período de 24 meses da TS/BERT atingiram os melhores resultados.

Analisando os resultados das Tabelas 13 e 14 relacionados ao modelo SVR(P), A TS obteve a maioria dos menores valores de MAPE, enquanto o TS/BERT obteve melhor desempenho com 4.88% somente no período de 12 meses de treinamento na previsão do milho. Observa-se que os melhores resultados do milho são os períodos de treinamento com tamanho de janela de 12, 24 e 36 meses. Enquanto nos períodos mais longos de treinamento na soja alcançaram os menores valores de MAPE.

Em relação aos resultados do modelo RBF das Tabelas 13 e 14, as representações TS

e TS/BoW alcançaram os menores MAPE. Os resultados do milho atingiram os melhores resultados nos períodos de 24, 36 e 48 meses. Enquanto na previsão da soja os menores valores de MAPE foram obtidos pela representação TS/BoW. Um ponto que deve ser destacado é que em ambos experimentos a janela de 24 meses de treinamento atingiram os melhores resultados.

Comparando os resultados do modelo Sigmoid das Tabelas 13 e 14, analisa-se que os resultados foram todos semelhantes entre as janelas de treinamento. Entretanto, seguiu o mesmo padrão do modelo RBF de obter os menores valores de MAPE para a janela de treinamento de 24 meses.

Observando os resultados do modelo LSTM das Tabelas 13 e 14, a representação ST alcançou o melhor desempenho em quase todas as janelas de treinamento. Nota-se que os menores valores de MAPE do milho foram obtidos com a janela de 12 meses, padrão que se repete para o menor MAPE da soja. O TS/BERT obteve apenas um resultado com melhor desempenho de previsão na janela de 48 meses.

Analisando os resultados de modo geral (Tabelas 13 e 14), observa-se que a representação TS/BERT obteve apenas três resultados com menor valor de MAPE, TS/BoW alcançou onze e TS obteve vinte e seis melhores resultados. Observando os resultados da previsão da soja, o valor de 4.21% da representação TS/BoW no período de 24 meses de treinamento, obteve o melhor desempenho entre todos os resultados. Em relação ao milho, o mesmo período de 24 meses com 4.21% para a ST, alcançou o menor valor de MAPE entre todos os resultados.

Conforme apresentado nos resultados, o enriquecimento de séries temporais com informações textuais não reduziram os porcentual dos erros na maioria das configurações experimentais. No entanto, em muitas situações ocorrem uma correlação entre as variações das séries temporais e a polaridade das notícias relacionadas ao domínio. Por exemplo, a Tabela 15 apresenta algumas manchetes do site *Soybean & Corn Advisor*[5] em períodos que as cotações do milho e da soja na CBOT apresentaram tendências de queda ou alta. Os rótulos na Figura 23 representam a data das manchetes da Tabela 15.

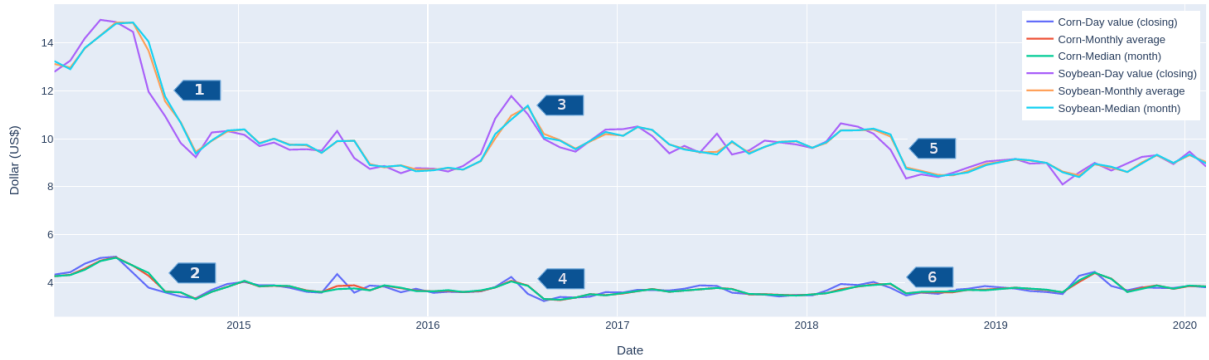
Tabela 15 – Amostras de notícias em momentos (rótulos) das séries de preços da Figura 23

Rótulo	**Data**	**Manchetes**
1	10-07-2014	*Os agricultores na Argentina terminaram a colheita da sua colheita de soja 2013/14, mas têm vendido muito lentamente a colheita, com apenas cerca de 45% da colheita vendida.*
1, 2	10-07-2014	*À medida que os preços do milho continuam a cair, os agricultores de Mato Grosso procuram culturas alternativas para cultivar a segunda colheita após a colheita da soja.*
3	06-06-2016	*Os preços da soja deverão atingir uma máxima recorde nos últimos anos, a alta do preço de R$ 100 por saca (aproximadamente $ 13,00 por bushel) em breve.*
4	13-06-2016	*Queda do preço do milho no Brasil devido à pressão inicial da colheita*
5, 6	26-06-2018	*O comércio de grãos no Brasil está mínimo há mais de três semanas.*
6	05-06-2018	*Analistas reduzem suas estimativas de milho brasileiro para 2017/18.*

Fonte: Soybean & Corn Advisor.

[5] http://www.soybeansandcorn.com

Figura 23 – Preço histórico do milho e da soja de 2014 à 2020 cotados no CBOT.



Fonte: Elaborado pelo Autor.

Os conteúdos das Manchetes da Tabela 15 apresentam simultaneidade com as variações dos preços do milho e da soja. Como exemplo, na data de 13-06-2016 (rótulo 4), o milho oscila com preço maior em comparação aos meses anteriores, porém nos próximos períodos teve uma leve queda. Analisando a manchete no mesmo período: "Queda do preço do milho no Brasil devido à pressão inicial de colheita", a manchete anuncia um motivo pelo qual ocorreu a queda do preço, porém não indica quais foram os fatores que levaram a esse movimento de mercado.

Com intuito de investigar alternativas para considerar fatores complexos e externos a séries temporais, este trabalho propôs uma análise no modelo de previsão usando duas representações de séries temporais enriquecidas: TS/BoW e TS/BERT. Os resultados indicaram que as representações enriquecidas não alcançam ganho considerável de desempenho. Apesar do uso de modelos preditivos robustos e atributos semânticos, as representações apresentadas não apresentaram um desempenho satisfatório em comparação com o modelo de previsão de séries temporais (ST). Acredita-se que ao enriquecer séries temporais com dados semânticos de textos, a maldição da dimensionalidade pode ter impactado negativamente nos modelos de regressão. Nesse contexto, avaliações futuras são conduzidas para identificar atributos específicos e relevantes do domínio, a fim de serem considerados para aprimorar a qualidade do enriquecimento das séries temporais. Essa abordagem visa superar as limitações observadas e otimizar a integração de séries temporais e dados textuais.

Devido à escassez de dados textuais relacionados ao domínio, uma limitação notável desta abordagem foi a dependência de apenas uma fonte de dados textuais para enriquecer as séries temporais. Além disso, as séries temporais utilizadas nos experimentos apresentam uma dependência temporal mensal. Para manter o alinhamento temporal entre as séries temporais e o conjunto de notícias, a média mensal do preço da commoditie foi adotada como a variável dependente do modelo preditivo. As representações apresentadas com os dados disponíveis exibiram algumas características que necessitavam de ajustes. Por exemplo, considere um conjunto de notícias de janeiro para prever o valor médio no mês de fevereiro, embora essa configuração possa ser válida em determinadas situações, a configuração não reflete a prática real na tomada de decisões. Nesse contexto, avaliações subsequentes são realizadas para avaliar

o desempenho da previsão de séries temporais em intervalos intra-dia, em vez de intra-mês.

4.3 Representação de ST enriquecida com termos específicos do domínio

Na terceira abordagem, um modelo de representação de séries temporais é enriquecido com características específicas do domínio, visando prever o preço da commoditie agrícola no dia seguinte (FILHO; MARCACINI; REZENDE, 2022). Considerando as limitações apresentadas nas duas primeiras abordagens, relacionadas à maldição da dimensionalidade e à relevância das previsões em um cenário real, um conjunto finito de termos extraídos de textos é apresentado para enriquecer as séries temporais. Essa abordagem envolveu trinta e três palavras-chave específicas e importantes do domínio (agrotermos) apresentadas na Tabela 16. Técnicas de mineração de textos são utilizadas para determinar a relação dessas palavras-chave com os documentos de texto. Em seguida, a série temporal é enriquecida com a representação vetorial dos textos em quatro modelos de regressão.

4.3.1 Pré-processamento

Uma das principais diferenças entre esta avaliação e as anteriores é a estratégia de previsão intra-dia e a incorporação de termos específicos do domínio (textos) na série temporal. Dessa forma, a cada etapa da previsão definimos uma sequência $S_u = (s_1, ..., s_u)$, onde u indica o período de tempo da última observação da série temporal. Cada sequência S_u está associada a um valor alvo de previsão y_{u+h}, onde h é o número de passos à frente, conhecido como previsão de um passo à frente com horizonte de previsão (h).

Uma estratégia é apresentada para obter uma representação da série temporal, que considera a ocorrência de palavras/termos específicos (trinta e três palavras) em textos que podem influenciar a série temporal. Dada uma sequência S_u, esta sequência é enriquecida com uma representação vetorial de textos (BoW) que calcula a ocorrência de palavras de domínio no período S_u. Primeiramente, é identificado via alinhamento temporal todos os documentos textuais relacionados à sequência (S_u) e suas respectivas representações no espaço vetorial, conforme definido na Equação 4.5 (Conjunto de palavras-chave).

$$\begin{aligned} KS(1,u) &= Q(T,u) \\ &= \{B(d_1), B(d_2), ..., B(d_k)\} \\ &= \{\vec{v}_{d_1}, \vec{v}_{d_2}, ..., \vec{v}_{d_k}\} \end{aligned} \tag{4.5}$$

em que KS é um subconjunto de textos (Q) com um texto por dia (T), e u indica o número de dias da sequência. A representação vetorial (B) de cada documento (d_k) é expressa como um vetor $\vec{v}_{d_k}$. O Term Frequency – Inverse Document Frequency (TF-IDF) é usado para refletir

a importância de uma palavra na coleção de documentos. Então, a representação de recurso associada à sequência é calculada como um vetor médio dos vetores do documento, conforme definido na Equação 4.6 (Recursos de palavras-chave):

$$KF(u,r) = \sum_{\vec{v}_d \in KS(u,r)} \frac{\vec{v}_d}{|KS(u,r)|} \tag{4.6}$$

A representação enriquecida é formada pela concatenação vetorial entre as observações da série temporal e os atributos da BoW (palavras-chave), $TK(u) = S(u) \oplus KF(u)$. Assim, podemos usar um conjunto de treinamento enriquecido

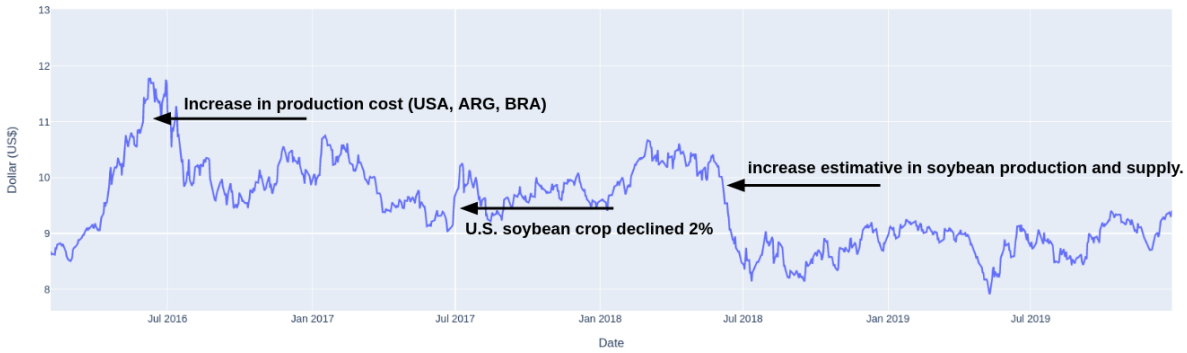
$$X = \{(TK_{u_1}, y_{(u+h)_1}), (TK_{u_2}, y_{(u+h)_2}), ..., (TK_{u_n}, y_{(u+h)_n})\} \tag{4.7}$$

nos modelos de regressão, conforme apresentado na próxima Seção.

4.3.2 Configuração experimental

Este trabalho apresenta uma avaliação de modelos de previsão para comparar o desempenho preditivo de três representações: séries temporais (TS), Séries Temporais enriquecidas com termos Específicos de Domínio (TSED) e somente termos específicos de domínio (DST). A série temporal considerada neste trabalho é da *Chicago Board of Trade* (CBOT), disponível no site do Grupo CME[6]. A Figura 24 apresenta a série de preços da soja. Os dados dados textuais extraídos do site *Soybean & Corn Advisor*[7] são utilizados para enriquecer a série temporal.

Figura 24 – Série de preço da soja - *Chicago of Board Trade* (CBOT)



Fonte: Elaborado pelo Autor.

A Figura 24 apresenta três exemplos de flutuações abruptas nas séries de preços. Ao analisar empiricamente os períodos das séries de preços, que alteram uma tendência (alta/baixa) ou flutuações abruptas em poucos dias, é constatado uma elevada ocorrência de palavras-chave nas notícias. A Tabela 16 descreve as palavras-chave específicas do domínio utilizadas para

[6] http://www.cmegroup.com/
[7] http://www.soybeansandcorn.com

enriquecer as séries temporais, o período do conjunto de dados, o tamanho dos conjuntos de dados de séries temporais e informações sobre dados textuais.

Tabela 16 – Visão geral de séries temporais e dados textuais usados na avaliação de experimentos.

Commodities	Milho e Soja
Período	2014-01-02 à 2020-12-30
Número de dias	1769
Atributos TS	Valores (Abertura, Fechamento, Máximo, Mínimo)
Número de notícias	1398
Palavras chaves específicas do domínio	colheita, safrinha, perdas, rendimento, estimativa, decepcionar, excelente, bom, chuvas, plantio, aumento, diminuição, preço, redução, vendas, adicional, completo, menor, baixo, mais, progresso, alto, doméstico, colheita, produção, declínio, custo, exportação, importação, sem notícias, registro, grande, crescente

Conforme mostrado na Tabela 16, o número de dias na série temporal é diferente do número de notícias. Portanto, o termo "sem notícias" é considerado para treinamento e teste nos dias em que não havia notícias no site para manter o alinhamento entre as séries temporais e os textos. As palavras chaves específicas do domínio foram definidas analisando os agrotermos disponíveis na Embrapa[8].

A estratégia de validação para séries temporais é utilizada para avaliar o modelo de representação apresentado na avaliação experimental (Figura 25). A primeira etapa de treinamento é realizada com 30% dos dados (F_1), e a cada iteração de validação cruzada, um dia é adicionado ao treinamento para prever o próximo passo a seguir. A variável y' na Equação 2.29 (pág. 60 representa a previsão dos preços das commodities h dias à frente, e n representa aproximadamente 1230 previsões (diariamente) realizadas na fase de teste.

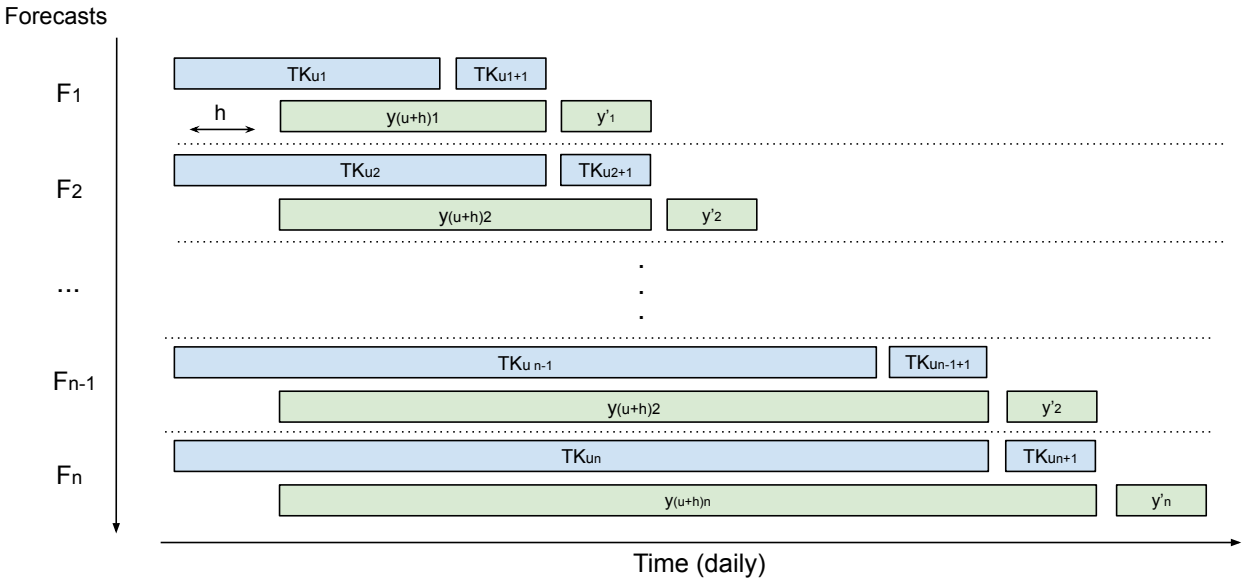
Quatro modelos de regressão não lineares são usados para prever dados de natureza caótica das séries temporais enriquecida com termos específicos do domínio, sendo: a Árvore de Regressão *Gradient Boosting* baseada em histograma (HGBR), Regressão de Vetores de Suporte (SVR), *Random Forest Regressor* (RF) e Regressor Bagging (BR). Para avaliar o desempenho e a validade do modelo é utilizado o indicador estatístico Mean Absolute Percentage Error (MAPE). Após a realização de diversos experimentos estruturados com diferentes configurações, os hiperparâmetros são definido na Tabela 17. Os hiperparâmetros são utilizados nas etapas de treinamento e teste do modelo.

4.3.3 Resultados e Discussão

A Tabela 18 apresenta os valores MAPE obtidos nas etapas de previsão. Na avaliação experimental foram considerados cinco tamanhos de h, ou seja, prevendo passos à frente de um a cinco dias à frente. Os valores em negrito são os menores valores MAPE de cada modelo

[8] https://sistemas.sede.embrapa.br/agrotermos/

Figura 25 – Modelo de Cross-validation para série temporal utilizado na avaliação.



Fonte: Elaborado pelo Autor.

Tabela 17 – Hyperparameters used in regression models.

Modelo	**Parâmetros**
HGBR	Default
SVR	Kernel RBF e gamma auto
RF	Depth = 4 e random state = 0
BR	Base SVR, Número do estimador = 10, random state = 0

Fonte: Elaborado pelo autor.

de regressão, e sublinhados representam os menores valores das representações TS, TSED e DST. A Figura 26 mostra o gráfico dos valores reais e previstos das commodities com horizonte de previsão $h = 1$. Os pontos vermelho e azul representam os dias em que a previsão atingiu o MAPE igual a zero para as representações TS e TSED, respectivamente. O nível de confiança das novas previsões pode ser medido pelo erro percentual médio obtido nos resultados da Tabela 18.

De acordo com os resultados apresentados na Tabela 18, a previsão do preço do milho considerando a representação TS, obteve os menores valores de MAPE (valores em negrito) em quase todas as configurações (h). Por exemplo, analisando os resultados de $h = 1$, o modelo SVR com a representação TS teve o menor valor MAPE com 1.145%, o RF teve o menor valor para a representação TSED com 1.168%, e o modelo SVR teve o MAPE mais baixo para representação DST com 6.056%. Este padrão de menor valor MAPE dos modelos de regressão para cada representação se repete para os demais horizontes de previsão (h).

Analisando os resultados da previsão do preço da soja na Tabela 18, o modelo HGBR obteve o menor valor MAPE para as representações TS e TSED para $h = 1$, com valores 0,982%

e 0,997%, respectivamente. Este padrão de menor valor MAPE dos modelos de previsão para cada representação não se repete para os demais horizontes de previsão h. Contudo, o modelo SVR obteve os menores valores de MAPE para a representação DST em todos os horizontes h, com valores 7.611%, 7.560%, 7.568%, 7.528 % e 7.506 %, respectivamente.

Conforme apresentado na Figura 26, as previsões da representação DST tiveram os resultados próximos a média da série de preços. Na sequência, será analisado apenas os resultados das representações TS e TSED que tiveram melhor desempenho (ou seja, resultados obtidos a partir dos valores sublinhados da Tabela 18). Além disso, a Tabela 19 compara o número de dias em que as representações menores valores de MAPE entre as representações (TS e TSED).

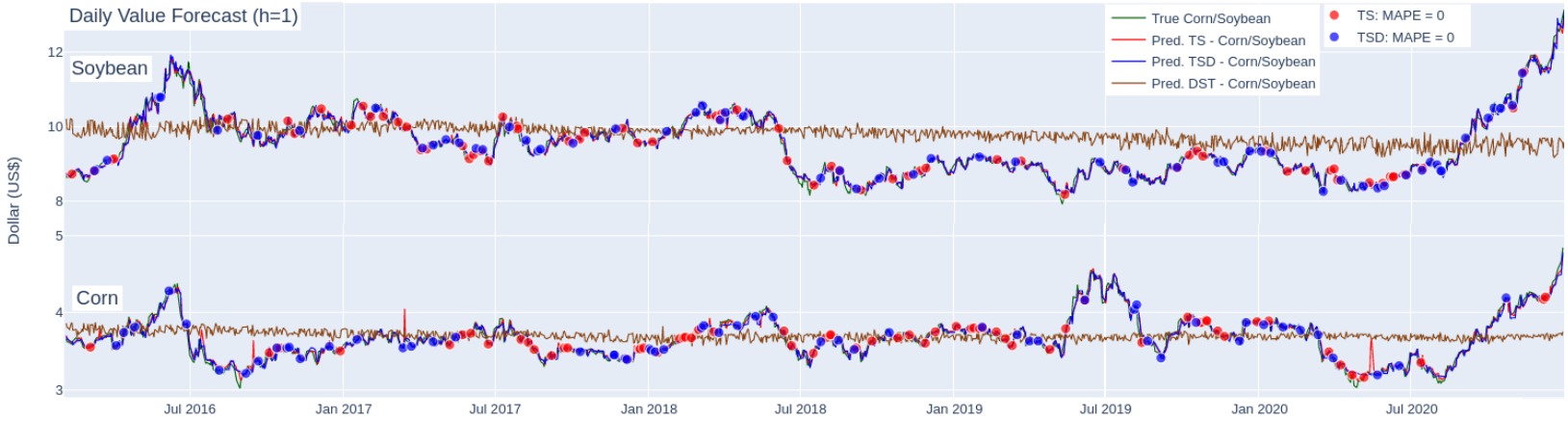
Analisando os resultados do milho na Tabela 19, a representação TS obteve 547 predições nas quais o valor do MAPE foi inferior ao TSED, 418 predições nas quais o TSED obteve melhor resultado em relação ao TS, e 272 nas quais ambas as representações obtiveram valores iguais para o horizonte ($h = 1$). Durante a fase de testes, algumas previsões obtiveram o valor MAPE igual a zero (0%), representado por pontos (vermelho e azul) na Figura 26. Neste caso, as representações TS e TSED obtiveram 69 e 57 previsões bem precisas, respectivamente. O melhor desempenho do TS sobre o TSED se repete com superioridade média de 16,7% em todos os horizontes de previsão (h).

Tabela 18 – Resultados do Milho e Soja com horizone de previsão (h)

	TS	TSED	DST	TS	TSED	DST	TS	TSED	DST	TS	TSED	DST	TS	TSED	DST
Milho															
Model		h = 1			h = 2			h = 3			h = 4			h = 5	
HGBR	**1,179**	1,186	7,554	**1,649**	1,687	7,578	**1,994**	2,021	7,579	**2,324**	2,341	7,522	**2,589**	2,607	7,48
SVR (RBF)	**1,145**	1,240	6,056	**1,566**	1,632	6,036	**1,888**	1,953	6,015	**2,168**	2,220	5,993	**2,407**	2,450	5,985
RF	**1,167**	1,168	7,133	1,594	**1,593**	7,098	**1,920**	1,929	7,076	2,218	**2,215**	7,076	2,455	**2,454**	7,061
BR	**1,173**	1,263	6,788	**1,572**	1,64	6,789	**1,907**	1,954	6,763	**2,189**	2,222	6,73	**2,418**	2,455	6,692
Soja															
HGBR	**0,982**	0,997	11,316	**1,375**	1,394	11,212	**1,714**	1,748	11,302	**1,987**	1,989	11,028	2,192	**2,157**	11,093
SVR (RBF)	1,022	**1,010**	7,611	1,382	**1,352**	7,560	1,696	**1,660**	7,568	1,947	**1,908**	7,528	2,147	**2,104**	7,506
RF	1,108	**1,107**	10,82	1,437	**1,434**	10,725	1,733	**1,728**	10,683	1,967	**1,964**	10,638	2,150	**2,142**	10,594
BR	**1,010**	1,027	7,807	1,369	**1,355**	7,791	1,659	**1,646**	7,772	1,906	**1,886**	7,727	2,104	**2,072**	7,672

Fonte: Elaborado pelo Autor.

Figura 26 – Valor diário previsto para milho e soja com horizonte (h=1).



Fonte: Elaborado pelo Autor.

Tabela 19 – Comparação do desempenho das representações considerando diferentes horizontes de previsão.

Milho					
Representações	**h = 1**	**h = 2**	**h = 3**	**h = 4**	**h = 5**
TS	547	570	545	570	489
TSED	418	455	466	480	441
TS = TSED	272	210	222	181	299
TS (MAPE 0%)	69	48	42	38	33
TSED (MAPE 0%)	57	48	50	30	28
Soja					
TS	586	584	586	582	587
TSED	526	507	536	554	578
TS = TSED	125	144	111	95	64
TS (MAPE 0%)	67	52	44	40	41
TSED (MAPE 0%)	60	48	43	43	46

Fonte: Elaborado pelo Autor.

Os resultados da previsão do preço da soja na Tabela 19 são semelhantes aos resultados do milho, em que a representação TS obteve um número mais significativo de previsões diárias em todos os horizontes de previsão h. Contudo, a superioridade do TS sobre o TSED é menor, com um valor médio de 7,6%. Por outro lado, o número de predições em que os valores do TSED MAPE são iguais ao TS, obteve um número inferior.

A frequência dos termos extraídos dos textos e incluídos na série temporal (TSED) foi investigado com mais detalhes, referente aos dias de previsão com erro MAPE igual a zero. A representação teve bom desempenho em dias com oscilações intra diárias abruptas na série de preços. A Tabela 20 apresenta exemplos para $h = 1$, em que a data representa o dia de publicação da notícia/manchete e previsão de dados; os valores em porcentagem representam a oscilação intra diária; e a frequência com que as palavras de domínio ocorrem nas notícias.

De acordo com os dados apresentados na Tabela 20, as palavras milho, exportação, aumento e produção possuem frequências de 1, 3, 1 e 4, respectivamente, nas notícias publicadas em 30/01/2020. Portanto, essas palavras são utilizadas como atributos na representação vetorial do TSED para a previsão do preço do milho em 31/01/2020. A medida Term Frequency - Inverse Document Frequency (TF-IDF) foi utilizada para medir a importância da palavra sobre documentos de texto. O valor TF-IDF é um fator de ponderação que aumenta proporcionalmente à medida que aumenta o número de ocorrências em um documento. Assim, palavras com alta frequência nos textos tiveram valores maiores, e palavras com pouca ocorrência tiveram valores menores na representação do TSED. No entanto, a representação TSED é baseada em palavras independentes e não expressa relações de palavras, sintaxe de texto ou semântica.

Também foi investigado o desempenho da previsão de preços para a representação TS nas datas mencionadas na Tabela 20. A representação do TS não teve um bom desempenho nos dias mencionados. Além disso, nos dias em que a representação do TS teve um desempenho superior ao do TSED, ocorreram frequentemente três situações: i) não havia notícias publicadas nas datas; ii) não tinham muita frequência de palavras-chave de domínio; iii) o conteúdo da notícia não

Tabela 20 – Notícias publicadas nos dias anteriores em que a série de preços apresentava oscilações anormais.

Milho				
Data	**Manchete**	**Predição**	**Intra dia**	**Ocorrência das palavras chaves**
2020/01/30	Brazil to be a Major Exporter of Food to India in the Coming Years.	2020/01/31	1,05%	corn(1), export(3), increase(1), production(4)
2018/07/19	Brazilians may be missing Selling Opportunity due to Freight Dispute.	2018/07/20	-1,40%	additional(2), corn(1), cost(5), crop(6), estimate(2), harvest(1), high(4), import(1), increase(4), large(3), planting(1), production(1), rains(3), record(2)
2018/05/23	Initial Impact of Truck Strike on Brazilian Agriculture Sector.	2018/05/24	-1,47%	corn(2), cost(1), crop(1), domestic(1), export(10), good(1), harvest(1), high(2), increase(2), large(3), price(2), production(4), rains(7), record(2), safrinha(1)
Soja				
2020/11/09	Brazil Importing U.S. Soybeans.	2020/11/10	3,24%	additional(3), domestic(3), export(2), harvest(2), high(1), import(7), large(2), planting(1), price(1), rains(1), record(2), sales(2), soybean(18)
2020/10/14	Full-Season Corn in Southern Brazil 39% Planted, About Average.	2020/10/15	-1,22	additional(1), crop(7), domestic(1), estimate(6), good(1), growing(3), harvest(2), high(3), increase(2), planting(13), price(4), production(3), rains(2), record(3), reduction(1), safrinha(11), soybean(3),
2017/02/07	Brazilian Government Announces Upgrade of Port of Santos.	2017/02/08	-0,84	complete(1), cost(1), export(4), good(1), import(4), increase(1), large(1), low(1), production(1), record(1), soybean(1)

Fonte: Elaborado pelo Autor.

representava com exatidão o domínio da aplicação. Em relação aos dois últimos, modelos de representação que considerem a semântica, a estrutura linguística e o contexto dos textos podem ser considerados para mitigar essa limitação, como os modelos de linguagem neural.

Em geral, modelos de representação de séries temporais que incorporam informações textuais raramente terão melhor desempenho em todas as etapas de previsão. Entretanto, o modelo apresentado surge como uma alternativa para antecipar oscilações abruptas em séries temporais. Além disso, em trabalhos futuros, representações enriquecidas podem contribuir para a explicabilidade dos modelos preditivos, muitas vezes considerados “caixas pretas". Trabalhos futuros podem explorar a extração mais detalhada de informações textuais, como entidades nomeadas, relações causais e técnicas que incorporem aspectos semânticos para aprimorar as séries temporais. Contudo, é essencial notar que, devido às experiências das avaliações anteriores, essas representações não devem conter dimensões excessivamente altas, pois isso pode prejudicar os modelos de regressão.

4.4 Considerações finais

Neste capítulo, uma proposta de enriquecer séries temporais com informações textuais foi apresentada, utilizando a estratégia de *early fusion*. Três abordagens foram apresentadas: i) enriquecimento de séries temporais com Bag-of-Words (BoW); ii) enriquecimento com dados dependentes de contexto; e, iii) enriquecimento com características específicas do domínio. A primeira abordagem, que consiste em enriquecer com dados BoW, demonstrou melhorias de desempenho em alguns cenários de avaliação. No entanto, observou-se que as representações baseadas em BoW trouxeram consigo a maldição da dimensionalidade e esparsidade nos dados. Como principal aprendizado dessa abordagem, evidenciou-se que modelos de regressão que lidam com representações extensas e esparsas podem não resultar em ganhos significativos para a previsão.

Na segunda abordagem, explorou-se a possibilidade de enriquecer séries temporais com embeddings de textos. A suposição inicial era que ao incorporar dados semânticos às séries temporais, seria possível considerar fatores qualitativos de mercado e aprimorar as tarefas de regressão. Entretanto, as dimensões das representações enriquecidas resultaram em tamanhos consideráveis, o que levou ao mesmo problema observado na avaliação anterior com representações extensas. Outra observação foi de evidenciar que dados textuais podem contribuir na previsão de preços em etapas de tempo menores de previsão. Diante desse desafio, pesquisas subsequentes foram conduzidas para desenvolver representações enriquecidas com informações específicas do domínio.

Na abordagem subsequente, a previsão foi conduzida considerando um conjunto finito de características em etapas intra-dia. Dessa forma, análises foram realizadas para enriquecer séries temporais com palavras-chave do domínio, indicativas de potenciais mudanças na série de preços. Os resultados evidenciaram que, em alguns cenários, a representação enriquecida obteve melhor desempenho do que as representações brutas de séries temporais. Contudo, os resultados não apresentaram um padrão consistente de previsão e, em alguns cenários, as previsões assertivas dos preços pareciam ocorrer de forma aleatória. Diante desse cenário, formulou-se uma hipótese de pesquisa de que dados textuais rotulados poderiam contribuir de forma mais eficaz em tarefas de previsão, especialmente com abordagens e arquiteturas multimodais. O próximo capítulo apresenta uma proposta com diferentes abordagens para integrar dados de séries temporais no processo de classificação de notícias do agronegócio.

CAPÍTULO

5

CLASSIFICAÇÃO DE TEXTOS USANDO DADOS DE SÉRIES TEMPORAIS

Neste capítulo é apresentado uma proposta, separada em duas abordagens diferentes, com o foco em integrar dados de séries temporais em tarefas de classificação de textos. Assim como na proposta anterior, desafios foram enfrentados quanto a escassez de notícias rotuladas temporalmente alinhadas com as séries temporais. Diante dessa lacuna, a primeira abordagem concentrou-se em utilizar dados de séries temporais para rotulagem automática de notícias. Os resultados e metodologias desenvolvidos nesse contexto são detalhados nas publicações (FILHO *et al.*, 2022; TRINDADE *et al.*, 2022), e apresentados na Seção 5.1.

Como delineado no capítulo de mapeamento sistemático, os modelos de aprendizado utilizados na literatura para tarefas preditivas no mercado financeiro, relacionados aos modelos multimodais, predominantemente adotam modelos de redes neurais (RNN, LSTM e outros) ou modelos com base na arquitetura *Transformers*. Dessa forma, a segunda abordagem apresenta estratégias de incorporar os benefícios do Graph Neural Networks (GNN) em modelagens com dados heterogêneos. Dessa forma, uma abordagem inovadora de agrupamento de subsequências de séries temporais, conectadas com *text embeddings* por meio de modelagens de GNN, é apresentada na Seção 5.2 e está em revisão (FILHO *et al.*, 2024).

Para o desenvolvimento da segunda abordagem, foi necessário a obtenção e rotulação manual de um conjunto de notícias do mercado financeiro. Um processo de rotulação de notícias do mercado da soja foi estruturado e realizado em colaboração com profissionais do setor e estudantes de graduação. Uma especialista no domínio econômico do agronegócio estabeleceu critérios para classificar as notícias como relevantes ou não relevantes. Uma equipe composta por estudantes foi treinada e supervisionada no processo assistido de rotulação de aproximadamente onze mil notícias. A metodologia utilizada no processo de rotulação está detalhada no trabalho (FILHO *et al.*, 2024) e disponível no repositório digital[1].

[1] https://data.mendeley.com/datasets/f8fdmpp6yh/2

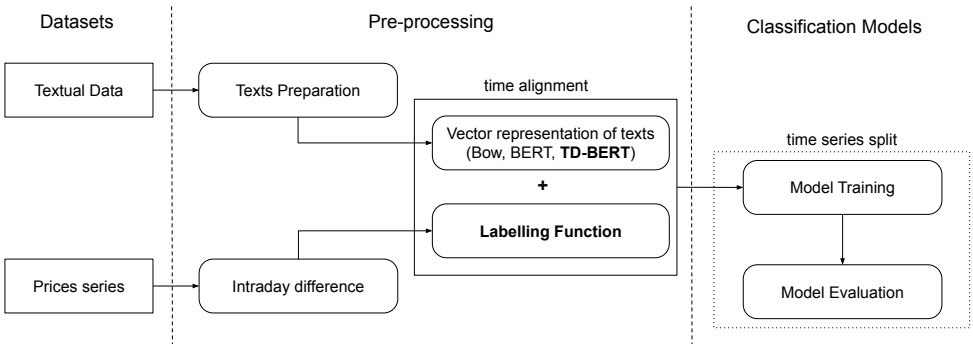
5.1 Classificação de textos por meio de rotulação fraca

Os modelos preditivos geralmente são criados a partir de um conjunto de dados contendo muitas amostras de treinamento, cada uma correspondendo a um objeto ou evento. Neste contexto, o desempenho dos modelos de aprendizado de máquina depende da disponibilidade de dados rotulados em grande quantidade e qualidade suficientes (BOECKING *et al.*, 2020). No entanto, os dados anotados para alguns domínios podem ser escassos, e o processo típico de obtenção de rótulos com especialistas que inspecionam amostras individuais é geralmente caro e demorado. Nessa etapa do trabalho, deparou-se com o desafio da escassez de notícias rotuladas temporalmente alinhadas com as séries temporais. Diante desse problema, a estratégia inicial foi de apresentar técnicas de aprendizagem por meio de rotulação automática de notícias, utilizando dados de séries temporais. Essa estratégia é conhecida na literatura como rotulagem por supervisão fraca (ZHOU, 2018).

A supervisão fraca fornece uma alternativa significativamente prática e barata em relação à anotação tradicional, reduzindo a necessidade de humanos rotularem manualmente grandes conjuntos de dados para treinar modelos de aprendizado de máquina (CHEN; XIU; DING, 2022; BOECKING *et al.*, 2020; WANG *et al.*, 2020). Os pesquisadores empregaram essa técnica para oferecer suporte a muitas aplicações, incluindo anotação e detecção de notícias falsas (HELMSTETTER; PAULHEIM, 2021; SHU *et al.*, 2020; WANG *et al.*, 2020), rotulagem de imagens de postagens de mídia social (DAI *et al.*, 2021), reconhecimento de entidades nomeadas (LISON *et al.*, 2020), e classificação de textos usando fontes externas (CHEN; XIU; DING, 2022; RATNER *et al.*, 2017).

Habitualmente, os eventos que alteram o comportamento do mercado financeiro são frequentemente relatados em notícias de texto. Assim, nessa etapa do trabalho é proposto técnicas que utilizam séries de preços de commodities do agronegócio para rotular textos curtos que correspondem a notícias agrícolas. A abordagem rotula fracamente as manchetes de notícias de acordo com o nível e os padrões de tendência da série temporal. A Figura 27 ilustra as etapas realizadas nesta abordagem.

Figura 27 – Modelo conceitual do método proposto.



Fonte: Elaborado pelo Autor.

A primeira etapa apresenta uma **Função de Rotulagem** de texto curto do mercado de commodities, usando dados de séries temporais (Seção 5.1.1). Além disso, apresenta um modelo de representação vetorial de texto baseado em BoW que adota uma medida de distância entre Termos e Documentos a partir de modelos BERT pré-treinados, denominado **TD-BERT** (Seção 5.1.2). A **Configuração Experimental** é apresentado na Seção 5.1.3 e os **Resultados e Discussão** das abordagens na Seção 5.1.4.

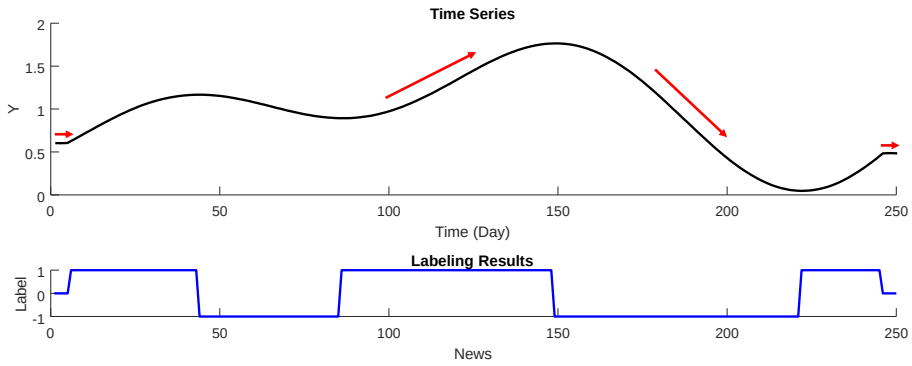
5.1.1 Função de rotulagem

Uma série de preços S de tamanho m é definida como uma sequência ordenada de observações, *ou seja*, $S = (s_1, s_2, ..., s_m)$, onde s_t representa uma observação s no momento t. Os documentos textuais D também são uma sequência ordenada $D = (d_1, d_2, ..., d_k)$, em que d_t é um texto d no tempo t e tamanho n. Portanto, é atribuído via alinhamento temporal um rótulo (-1, 0 ou 1) aos textos usando a seguinte equação:

$$d_t = \begin{cases} -1 & if \quad s_{t+lag} < (s_t + s_{t-lag})/2 \\ 1 & if \quad s_{t+lag} > (s_t + s_{t-lag})/2 \\ 0 & \text{caso contrário} \end{cases} \tag{5.1}$$

o texto d_t recebe um rótulo de acordo com os padrões de nível e tendência da série temporal S. A constante lag corresponde ao período sazonal da série temporal em número de observações. Para exemplificar, a Figura 28 ilustra o resultado da Equação 5.1 aplicada a uma série temporal sintética com $lag = 5$.

Figura 28 – Ilustração de como a função de rotulagem funciona.



Fonte: Elaborado pelo Autor.

Esta função visa capturar os comportamentos estáveis, crescentes e decrescentes da série temporal para atribuir rótulos a textos curtos organizados cronologicamente no tempo.

5.1.2 TD-BERT: Uma representação híbrida

Na presente abordagem, um novo modelo de representação matricial de textos é apresentada, em que incorpora a estrutura da BoW e considera pesos provenientes de modelos

pré-treinados do BERT. Primeiramente, uma coleção de documentos $D = [d_1, d_2, ..., d_k]$ é extraída contendo k documentos e um conjunto $T = [w_1, w_2, ..., w_b]$ com b termos. A estrutura é similar aos procedimentos apresentados na Seção 2.3.1 (Capítulo 2 - Pág. 48). No entanto, o *sentence transformers* dos modelos pré-treinados do BERT são usados para obter vetores que representam documentos e termos. A representação textual D com *sentence transformers* é definida como $DS = ([B_1], [B_2], ..., [B_k])$, em que B_i é um vetor BERT de h posições, representando um documento d_i no tempo t. A representação de termos com o sentence transformers é definida como $TS = ([W_1], [W_2], ...[W_b])$, em que W_j é um vetor BERT de h posições que representa um termo w_j. O conjunto de documentos é representado como uma matriz documento-termo constituída pela distância de cosseno c de cada k composta por b dimensões, conforme ilustrado na Figura 29.

Figura 29 – Ilustração da representação de documentos k como uma matriz de documento-termo.

	W_1	W_2	...	W_{b-1}	W_b
B_1	$c(B_1,W_1)$	$c(B_1,W_2)$	...	$c(B_1,W_{b-1})$	$c(B_1,W_b)$
B_2	$c(B_2,W_1)$	$c(B_2,W_2)$	...	$c(B_2,W_{b-1})$	$c(B_2,W_b)$
⋮	⋮	⋮	⋱	⋮	⋮
B_{k-1}	$c(B_{k-1},W_1)$	$c(B_{k-1},W_2)$	...	$c(B_{k-1},W_{b-1})$	$c(B_{k-1},W_b)$
B_k	$c(B_k,W_1)$	$c(B_k,W_2)$	...	$c(B_k,W_{b-1})$	$c(B_k,W_b)$

Fonte: (FILHO *et al.*, 2022).

Os valores da matriz corresponde a distância de cosseno de cada termo em relação à cada documento, isto é, $c(B_k, W_b)$ refere-se à distância entre os vetores W_j and B_i. Os valores vetoriais TS (Termos) e DS (Documentos) são atribuídos de acordo com um modelo BERT pré-treinado, denominado TD-BERT. O desempenho da classificação da representação TD-BERT é avaliado considerando três modelos pré-treinados: BERT base multilíngue (TD-BERT), DistilBERT base multilíngue (TD-DistilBERT) e BERT base português (TD-BERTimbau) (FILHO *et al.*, 2022). Na avaliação, considerou-se o total de nove representações textuais, incluindo BoW, BERT e TD-BERT, e diferentes paradigmas de aprendizagem.

5.1.3 Configuração Experimental

Avaliações de supervisão fraca são apresentadas utilizando dois conjuntos de dados de commodities agrícolas: milho e soja. Textos e séries históricas do milho e soja são utilizados na avaliação. Os dados textuais em português foram extraídos de um site de notícias agrícolas[2]. Fundada em 1997, a *Notícias Agrícolas* é uma das mídias mais influentes do agronegócio brasileiro. Tabela 27 descreve o período do conjunto de dados, o número de dias e informações sobre os dados de texto.

[2] https://www.noticiasagricolas.com.br/

Tabela 21 – Visão geral das séries temporais e dados textuais utilizados na avaliação experimental.

Commodities	Milho e Soja
Período	05-01-2015 à 10-12-2021
Número de dias	1753
Atributos da ST	Valores (Abertura, **Fechamento**, Máx., Mín.)
Número de Manchetes/Notícias	7172 (Milho) - 8394 (Soja)

Fonte: Elaborado pelo Autor.

Os dados de séries temporais foram extraídos do Centro de Estudos Avançados em Economia Aplicada (CEPEA) da Universidade de São Paulo (USP). O valor de fechamento da série de preços foi utilizado para calcular a diferença de preços intradia.

O desempenho dos modelos preditivos são avaliados considerando rótulos fracos em cenários binários positivos e negativos. O cenário Binário Positivo (PB) possui os rótulos [0, 1] e o Binário Negativo (NB) possui os rótulos [-1, 0]. A Tabela 22 apresenta exemplos de manchetes rotuladas de acordo com a função formalizada na Seção 5.1.1.

Tabela 22 – Amostras de notícias rotuladas usando a função de rotulagem.

Com.	**Data**	**Manchete**	**lab.**
Milho	12-01-2016	Dólar sobe nesta 4ª com atenção à política interna; milho acompanha	1
	21-06-2016	Preços do milho recuam até 15% no Brasil com colheita da 2ª safra	-1
	27-03-2017	Incerteza sobre a demanda por milho resulta em nova queda de preço	-1
	27-02-2018	USDA reporta a venda de 130 mil toneladas para destinos desconhecidos	1
Soja	10-05-2016	Chuva do início de outubro ainda não foi suficiente para as lavouras no Sul do MS	1
	30-01-2017	Com queda do dólar e perspectiva de safra elevada, preço da soja cai no Brasil	-1
	30-11-2018	Soja opera estável na Bolsa de Chicago observando início da reunião do G20	-1
	21-09-2020	USDA informa nova venda de 435 mil t de soja para China e demais destinos	1

Fonte: Elaborado pelo Autor.

Entre as 7.172 manchetes de milho, 3.209 são rotuladas como negativas (-1), 66 como neutras (0) e 3.897 como positivas (1). Em relação às manchetes sobre soja, 3.681 são rotuladas como negativas, 82 como neutras e 4.631 como positivas. Para fazer avaliação binária, os rótulos negativos são atribuídos como neutros no cenário PB e, no cenário NB, os rótulos positivos também são alterados para neutros.

Esta abordagem aplica representações com base na BoW, modelos NLM pré-treinados e o modelo TD-BERT. Na modelagem BoW, considerou-se técnicas de ponderação de três termos: Binário, TF e TF-IDF. Apenas versões unigrama é utilizada de cada dos termos de

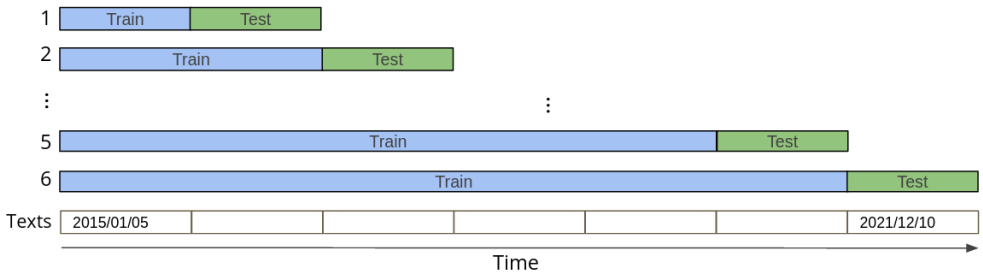
ponderação. Nesses modelos, aplicou-se a limpeza de texto para diminuir a dimensionalidade dos dados e aumentar a qualidade da representação. Segundo (AGGARWAL; AGGARWAL, 2018), esse processo melhora a qualidade dos resultados das classificações. As etapas de limpeza são: (1) conversão das palavras para minúsculas e remoção de acentos; (2) remoção de sinais de pontuação e caracteres alfanuméricos; (3) remoção de stopwords; e (4) radicalização de palavras.

Três modelos de linguagem neural pré-treinados são utilizados para avaliar técnicas de supervisão fraca: versões multilíngues (M) do BERT, M. DistilBERT e a versão em português BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020). Nos modelos pré-treinados, não utilizou-se técnicas de limpeza de texto para manter a estrutura do texto original. Assim, os *text embeddings* de modelos treinados são empregados como entrada para os modelos preditivos. Além disso, utilizou-se os modelos pré-treinados para construir os modelos propostos: TD-BERT (TD-Be), TD-DilstilBERT (TD-Di) e TD-BERTimbau (TD-Ba).

Cinco modelos de classificação tradicionais são utilizados: MLP, SVM, KNN, GNB e MNB. Os parâmetros dos modelos de aprendizado são: MLP arquitetura = {1, 3, 6} , neurônios = {50. 150} , momento = 0.9, and learning rate = 0.001; SVM kernels = {RBF, sigmoid, polinomial, linear} e C = 1.0 (penalty parameter); kNN: k $\in$ [3, 7, 11, 15] and métrica cosseno; GNB: Variância smoothing = [1e–01 : 1e–23]; MNB alpha = {1.0. 0.9, 0.5, 0.4, 0.1, 0.0} e probabilidade classs priore = {True, False}.

A estratégia de avaliação de divisão de série temporal foi empregada para considerar a dependência temporal dos dados textuais, *ou seja*, notícias "do passado" são treinadas para avaliar um cenário futuro. Assim, sete *splits* são utilizados para oito avaliações. Nesta configuração, cada divisão representa um ano do conjunto de dados textuais. Não se considerou interessante inserir um conjunto de validação dentro do treinamento, pois o foco foi em avaliar a capacidade do modelo de generalizar para períodos futuros com base em dados passados. A Figura 30 descreve a estratégia de avaliação de divisão de série temporal adotada na avaliação.

Figura 30 – Divisão de série temporal usada na configuração experimental.



Fonte: Elaborado pelo Autor.

Para a etapa de avaliação dos resultados, a medida F_1 é utilizada, em que corresponde à média harmônica de Precisão 5.3 e Recall 5.4. A Equação 5.2 define o índice F_1. Essa métrica

foi empregada devido as classes desbalanceadas em todas as divisões de avaliação.

$$F_1 = \frac{2 \times Prec \times Rec}{Prec + Rec}, \tag{5.2}$$

$$Prec = \frac{TP}{TP + FP}, \tag{5.3}$$

$$Rec = \frac{TP}{TP + FN}, \tag{5.4}$$

em que TP (Verdadeiro Positivo) refere-se ao número de documentos de uma classe na qual o algoritmo classificou corretamente, e FP (Falso) indica o número de documentos que não pertencem a uma classe que o algoritmo classificou erroneamente como pertencente. Por fim, FN (Falso Negativo) refere-se ao número de documentos de uma classe que o algoritmo classificou erroneamente como outra classe.

5.1.4 Resultados e Discussão

Avaliações experimentais são realizadas para investigar dois aspectos da supervisão fraca. No primeiro, analisou-se o impacto de cada modelo de representação textual considerando os cinco diferentes algoritmos de classificação. No segundo, avaliou-se a influência dos modelos de linguagens neurais em duas tarefas de classificação de supervisão fraca.

Quanto ao primeiro aspecto, as Tabelas 23 e 25 apresentam os resultados de classificação dos algoritmos MLP, SVM, KNN, GNB e MNB nos conjuntos de dados de milho e soja. Esta tabela abrange um cenário de avaliação PB para Milho e Soja, denominado de MBP e SBP. Cada linha representa o resultado de F_1 para um modelo específico. Em negrito destaca-se os maiores valores para cada modelo de classificação. Os valores sublinhados refletem o melhor desempenho dos modelos de representação textual (BoW, BERT e TD-BERT), e o valor entre parênteses é o melhor desempenho considerando todos os resultados.

Os modelos de linguagem neural não são processados para o MNB porque este não aceita vetores com valores negativos. No entanto, considerou-se essencial manter os resultados do MNB para as representações do BoW, a fim de compará-los com outros resultados. Analisando os valores destacados do MBP (negrito) para cada modelo de classificação, observa-se que as representações BERTimbau (B.Br), Binário (Bin) e TD-BERTimbau (TD-Br) obtiveram os melhores valores F_1. Nota-se também que o modelo BERTimbau (MLP) apresentou o maior valor entre todos os resultados (0,499). Os resultados do SBP mostraram Binário, TF-IDF e BERTImbau como os melhores valores para cada razão de classificação F_1, sendo BERTimbau (0,500) o valor mais alto entre todos os resultados do SBP. A Tabela 24 apresenta os melhores resultados de MBP e SBP em termos de precisão, recall e accurária.

Tabela 23 – Resultados da avaliação Binária Positiva. Comparação (macro F_1) de modelos BoW, linguagens neural pre-treinados e o modelo híbrido TD-BERT.

Milho – Binário Positivo (MBP)									
Mod.	**Bin.**	**TF**	**TF-IDF**	**BERT**	**Distil.**	**B.Br**	**TD-B**	**TD-D**	**TD-Br**
MLP	0.496	0.495	0.495	0.486	0.488	(**0.499**)	0.378	0.342	0.356
SVM	**0.456**	0.454	0.452	0.431	0.422	0.412	0.416	0.389	0.388
KNN	0.484	0.483	0.490	0.476	0.479	0.492	0.483	0.486	**0.495**
GNB	0.439	0.439	0.444	0.496	0.485	**0.497**	0.494	0.495	0.462
MNB	0.487	0.488	0.451	-	-	-	-	-	-
Soja – Binário Positivo (SBP)									
Mod.	**Bin.**	**TF**	**TF-IDF**	**BERT**	**Distil.**	**B.Br**	**TD-B**	**TD-D**	**TD-Br**
MLP	**0.490**	0.488	0.488	0.476	0.489	0.485	0.344	0.312	0.352
SVM	0.440	0.439	**0.442**	0.398	0.371	0.387	0.381	0.355	0.357
KNN	0.483	0.481	0.484	0.478	0.474	**0.486**	0.477	0.474	0.481
GNB	0.470	0.470	0.469	0.498	0.499	(**0.500**)	0.494	0.481	0.469
MNB	0.472	0.472	0.436	-	-	-	-	-	-

Fonte: Elaborado pelo Autor.

Tabela 24 – Métricas de avaliação referentes aos melhores resultados de classificação do SBP e MBP.

	MBP: BERTimbau (MLP)				**SBP: BERTimbau (GNB)**			
	prec.	recall	f1-scr.	support	prec.	recall	f1-scr.	support
0	0.489	0.385	0.422	2917	0.478	0.465	0.458	3358
1	0.534	0.636	0.574	3227	0.544	0.559	0.540	3836
accuracy	0.518				0.511			
macro avg	0.512	0.511	**0.499**	6144	0.511	0.512	**0.500**	7194
weighted avg	0.527	0.518	0.500	6144	0.546	0.511	0.517	7194

Fonte: Elaborado pelo Autor.

As acurária de MBP e SBP são de 0.518 e 0.51, respectivamente. No entanto, analisando para os valores da Tabela 24, é possível analisar que os rótulos fracos estão razoavelmente desequilibrados. Portanto, ao analisar os resultados para este tipo de avaliação, a Macro F_1 torna-se indicada. Em relação ao NB, a Tabela 25 apresenta dois cenários de avaliação, denominado como classificação Binária Negativa Milho (CBN) e Soja (SBN). Tabela 26 lista os melhores resultados de CBN e SBN em relação à precisão, recall e $F1$ *score*.

Tabela 25 – Resultados da avaliação Binária Negativa. Comparação (macro F_1) dos modelos BoW, liguagens neural pré-treinado, e o modelo híbrido TD-BERT.

Milho – Binário Negativo (MBN)									
Mod.	**Bin.**	**TF**	**TF-IDF**	**BERT**	**Distil.**	**B.Br**	**TD-B**	**TD-D**	**TD-Br**
MLP	0.491	0.495	0.496	0.482	**0.497**	0.493	0.358	0.343	0.362
SVM	**0.452**	0.451	0.451	0.428	0.418	0.411	0.406	0.375	0.374
KNN	0.484	0.481	0.490	0.474	0.479	0.492	0.485	0.483	**0.494**
GNB	0.439	0.439	0.443	0.497	0.490	(**0.507**)	0.493	0.494	0.469
MNB	0.491	0.490	0.446	-	-	-	-	-	-
Soja – Binário Negativo (SBN)									
Mod.	**Bin.**	**TF**	**TF-IDF**	**BERT**	**Distil.**	**B.Br**	**TD-B**	**TD-D**	**TD-Br**
MLP	0.48	0.487	**0.488**	0.474	0.481	0.485	0.339	0.288	0.344
SVM	**0.434**	0.432	0.432	0.389	0.363	0.378	0.376	0.358	0.364
KNN	0.471	0.471	0.472	0.471	0.468	**0.480**	0.47	0.47	0.478
GNB	0.451	0.452	0.453	0.500	(**0.501**)	0.496	0.485	0.486	0.461
MNB	0.474	0.473	0.429	0	0	0	0	0	0

Fonte: Elaborado pelo Autor.

Tabela 26 – Métricas de avaliação referentes aos melhores resultados de classificação CBN e SBN.

	CBN: BERTimbau (GNB)				**SBN: DistilBERT (GNB)**			
	prec.	recall	f1-scr.	support	prec.	recall	f1-scr.	support
-1	0.489	0.520	0.492	2881	0.479	0.524	0.483	3339
0	0.554	0.520	0.520	3263	0.553	0.512	0.517	3855
accuracy	0.517				0.506			
macro avg	0.521	0.520	**0.507**	6144	0.516	0.518	**0.501**	7194
weighted avg	0.534	0.517	0.511	6144	0.552	0.506	0.513	7194

Fonte: Elaborado pelo Autor.

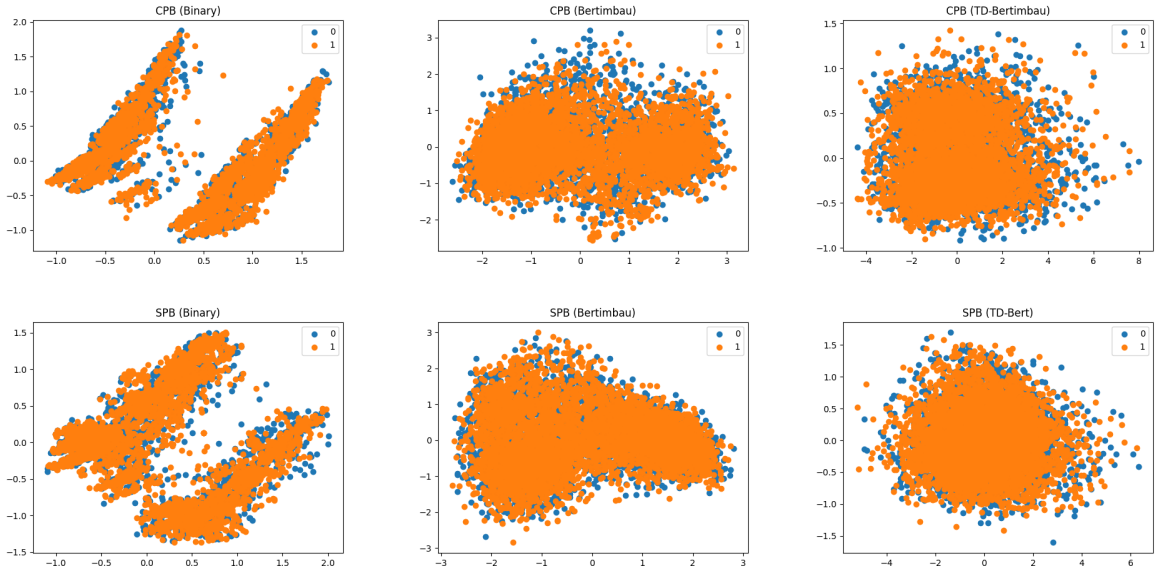
Observando os resultados do CBN, destaca-se que as representações DistilBERT (Diltil.), Binária, TD-BERTimbau e BERTimbau (B.Br) apresentaram os maiores valores (negrito) de F_1 para cada algoritmo de classificação, respectivamente. Em relação aos resultados do SBN, as representações TF-IDF, Binária, BERTimbau e DiltilBERT obtiveram os melhores resultados. Neste caso, os valores 0,517 e 0,569, entre parênteses, representam os melhores resultados do CBN e do SBN, respectivamente.

Com o objetivo de investigar o segundo aspecto da avaliação experimental, analisou-se o impacto do modelo de linguagem neural na supervisão fraca. De acordo com o resultado sublinhado de MBP e SBP na Tabela 23, a representação binária apresenta os valores F_1 mais altos, *ou seja*, 0.496 e 0.490, respectivamente. O modelo de linguagem neural BERTimbau apresenta melhor desempenho nos dois cenários com valores F_1 de 0.499 e 0.500. Por fim, as representações TD-BERTimbau e TD-BERT alcançaram resultados F_1 com valores de 0.495 e 0.494. Assim, os modelos de representações baseados em linguagem neural alcançaram melhor desempenho que os modelos BoW. Para ilustrar a distribuição vetorial dos textos, a Figura 31 apresenta um gráfico das representações textuais que obtiveram melhor desempenho em cada modelo de representação da Tabela 23 (valores sublinhados).

A técnica PCA (*Principal Component Analysis*) foi utilizada para reduzir a dimensionalidade da representação textual do conjunto de dados de commodities agrícolas. Observou-se que as manchetes classificadas como positivas (1) estão mais concentradas na distribuição do gráfico, enquanto as manchetes classificadas como neutras são um pouco mais esparsas. Além disso, as representações MBP (TF-IDF) e SBP (TF) possuem intervalos menores nos eixos do que as representações baseadas em BERT. Nesse sentido, acredita-se que esse espectro mais amplo possa abstrair mais informações semânticas dos textos.

Comparando os resultados sublinhados de CBN e SBN na Tabela 25, o TF-IDF, BERTimbau, DistilBERT, TD-DistilBERT e TD-BERTimbau obtiveram o melhor desempenho de classificação para os algoritmos de aprendizagem, respectivamente. Os modelos de linguagem neural DistilBERT e BERTimbau tiveram melhor desempenho para o método GNB. Em ambos os experimentos (PB e NB), observou-se que os melhores resultados vieram das representações baseadas em Distilbert e BERTimbau com os modelos GNB e MLP. A Figura 32 ilustra um gráfico das representações textuais que tiveram melhor desempenho em cada modelo de

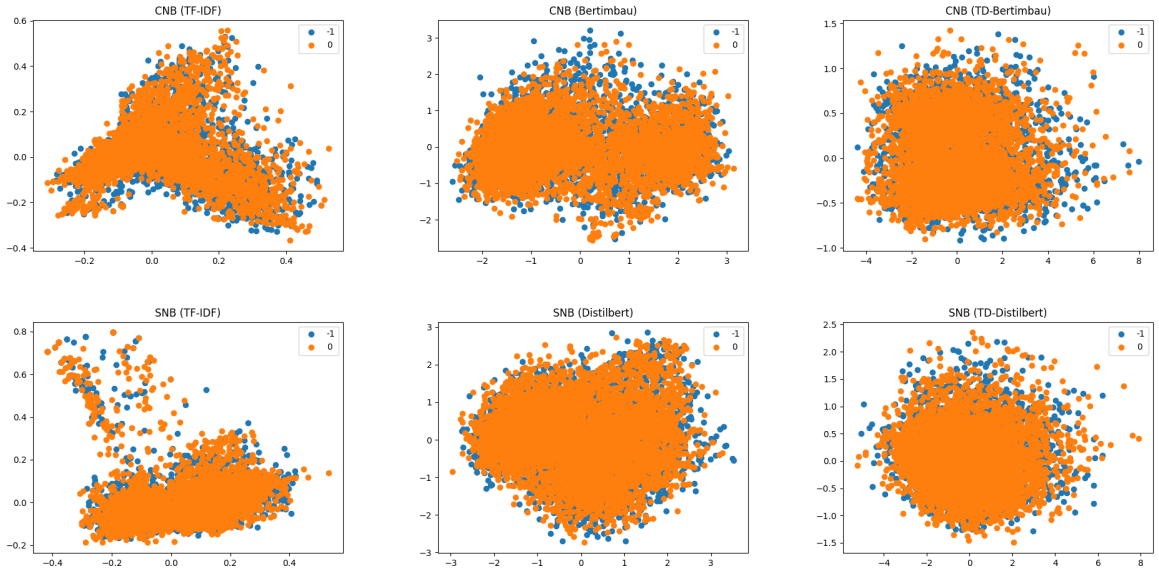
Figura 31 – MBP e SBP. Técnica PCA para plotagem de representações textuais.



Fonte: Elaborado pelo Autor.

representação da Tabela 25.

Figura 32 – CBN e SBN. Técnica PCA para plotagem de representações textuais.



Fonte: Elaborado pelo Autor.

Confrontando as estratégias investigadas, foi identificado que o uso do BoW pode reduzir o desempenho. Por outro lado, as características semânticas permitem resultados satisfatórios quando se considera um extenso conjunto de treinamento. Assim, foi comparado apenas os desempenhos das representações do BERT (BERT e Dist. B.Bau) e do modelo TD-BERT proposto (TD-Be, TD-Di e TD-Bau) quanto às avaliações do PB e do NB. Pode ser observado na Tabela 23 que entre os melhores resultados, 75% são de modelos BERT e 25% são de modelos

TD-BERT. Quanto aos desempenhos da Tabela 25, houve empate de 50% para cada modelo de representação.

Em relação aos melhores resultados dos cenários de avaliação Binário Positivo e Binário Negativo, dez entre dezesseis são representações dos modelos BERT (62,5%). Os modelos TD-BERT propostos tiveram melhor desempenho em alguns casos ao analisar modelos de representação baseados em linguagem neural. Em geral, os modelos de representação baseados em linguagem neural superaram os modelos baseados em BoW. Porém, uma limitação dos modelos TD-BERT é o tempo de processamento, e trabalhos futuros podem ser realizados para reduzir custos computacionais. Em um outro trabalho, uma avaliação aprimorada do TD-BERT foi realizada considerando seis representações vetoriais de textos para três datasets de diferentes domínios (MARTINS *et al.*, 2023). Quatro algoritmos de diferentes paradigmas de aprendizagem foram utilizados na avaliação, os resultados se mostraram eficaz e obteve desempenho similares aos modelos de linguagem neural.

A função de rotulagem projetada pode ser uma alternativa para anotar um grande volume de documentos de texto. A rotulagem automática pode ser imprecisa, mas útil quando muitos textos não estão rotulados. Neste abordagem, a limitação da análise da supervisão fraca consistiu no desequilíbrio de classe. Pesquisas futuras podem desenvolver estratégias para propagar rótulos por meio de aprendizagem semissupervisionada para reforçar a rotulagem. Além disso, por meio das abordagens conexionistas, outros fatores externos podem ser utilizados na função de rotulagem; *e.g.*, pode se considerar um coeficiente de ponderação na rotulagem das notícias.

Em uma abordagem semelhante, uma função é apresentada para rotular notícias considerando as oscilações das séries de preços da soja no mercado nacional, internacional e cotação do dólar (TRINDADE *et al.*, 2022). Em ambas avaliações, os modelos de linguagem neural demonstraram melhor desempenho e os resultados indicaram que a abordagem pode ser uma alternativa para aplicações de tempo real.

5.2 Classificação de textos usando grafos multimodais.

Diariamente, muitas notícias são publicadas em diversas fontes de informação, oferecendo relatórios técnicos e atualizações de mercado em tempo real. Essas atualizações servem como recursos para os profissionais da área, proporcionando um meio de se manterem informados sobre as mudanças do mercado. No entanto, uma parcela significativa das notícias publicadas online são irrelevantes e não oferecem informações importantes que auxiliem nos processos de tomada de decisão dos especialistas do domínio (CLAPHAM; SIERING; GOMBER, 2021). Esta disseminação de notícias irrelevantes complica a análise e prolonga o tempo necessário para investigar o mercado financeiro. Como resultado, nos últimos anos, aplicações têm sido propostas para otimizar análises e classificar notícias com base nos interesses de especialistas (MAN; LUO; LIN, 2019).

A exploração do sentimento da notícia no contexto do mercado financeiro têm recebido bastante atenção nos últimos anos, com pesquisadores investigando a análise de sentimentos para compreender o impacto das mídias sociais nas tendências do mercado (LI; WU; WANG, 2020) (ZHOU *et al.*, 2020) (KHALIL; PIPA, 2022). Alguns trabalhos visam discernir como os relatórios técnicos influenciam a dinâmica das séries de preços, enfatizando como esses relatórios moldam o comportamento do mercado (ZHONG; HITCHCOCK, 2021) (LI; WU; WANG, 2020) (YE *et al.*, 2022). Por outro lado, alguns resultados apresentados na literatura sugerem que a eficácia da análise de sentimento ou da incorporação de notícias para a previsão do mercado financeiro pode ser inferior ao esperado (SEZER; GUDELEK; OZBAYOGLU, 2020b).

Considerando os conceitos apresentados, nesta etapa é proposto a modelagem de dados textuais e de séries temporais por meio de grafos para explorar a heterogeneidade dos dados na classificação das notícias do mercado de soja como relevantes ou irrelevantes. Utilizando GNN, apresenta-se métodos de última geração para dados modelados por meio de grafos e atualmente utilizados para classificar notícias. O objetivo é determinar se as informações de séries temporais sobre os preços da soja, influenciam positivamente e significativamente a classificação de notícias do mercado da soja. Para alcançar esses resultados, as seguintes questões de pesquisa são levantadas:

1. A série temporal de preços da soja impacta a classificação das notícias sobre a soja como relevantes ou irrelevantes, utilizando a estratégia de concatenação?
2. A série histórica de preços da soja impacta a classificação das notícias sobre a soja como relevantes ou irrelevantes, usando GNN?
3. As GNNs superam os algoritmos de classificação comumente usados na literatura para classificar notícias sobre soja como relevantes ou irrelevantes?

Dados em domínios específicos, como o mercado do agronegócio, são naturalmente multimodais e podem ser representados de diversas formas. Ao integrar essas modalidades, a abordagem visa avaliar a precisão da classificação de notícias e a eficácia dos modelos de aprendizado de máquina. Dessa forma, modelagens de grafos e uma GNN heterogênea são apresentadas para classificar notícias do mercado da soja como relevantes ou irrelevantes. As representações ***Text Embedding*** (Seção 5.2.1), das **Séries Temporais** (Seção 5.2.2), as **Modelagens dos Grafos** (Seção 5.2.3) e as **Redes Neurais de Grafos** (Seção 5.2.4) são apresentados para definição da estrutura da presente abordagem. Por fim, a **Configuração Experimental** é apresentado na Seção 5.2.5 e os **Resultados e Discussão** na Seção 5.2.6.

5.2.1 Representação Text Embedding

Para representar os textos das manchetes do agronegócio, *embeddings* de texto derivados do modelo Bidirecional Encoder from Transformers (BERT) (DEVLIN *et al.*, 2018) são

utilizados. Durante o processo de treinamento, o BERT realiza duas tarefas, a primeira delas envolve completar frases preenchendo palavras mascaradas. Dada a sequência de palavras de um documento textual $\mathbf{d}_i = \{w_1, w_2, ..., w_3\}$ o modelo BERT gera uma versão corrompida de $\mathbf{d}_i$, $\hat{\mathbf{d}}_i$, em que as palavras são selecionadas aleatoriamente e substituídas por um símbolo [MASK]. Além disso, considere $\overline{\mathbf{d}}_i$ ser os *tokens* mascarados de $\mathbf{d}_i$. Então, o treinamento do BERT consiste em reconstruir as palavras mascaradas $\overline{\mathbf{d}}_i$ a partir de $\hat{\mathbf{d}}_i$ (YANG *et al.*, 2019):

$$\max_{\Theta} \log p_{\Theta}(\bar{\mathbf{d}}_i|\hat{\mathbf{d}}_i) \approx \sum_{x_j \in \mathbf{d}_i} c_{x_j} \log \frac{\exp(H_\theta(\hat{\mathbf{d}}_\mathbf{i})^T_{x_j} \mathbf{e}(x_j))}{\exp(\sum_{x'} H_\theta(\hat{\mathbf{d}}_\mathbf{i})^T_{x_j} \mathbf{e}(x'))}, \tag{5.5}$$

em que $c_{x_j} = 1$ indica que x_j está mascarado, $\mathbf{e}(x_j)$ indica a incorporação da palavra x_j, x' é $\mathbf{d}_i$ sem x_j, $H_\theta(\mathbf{d}_i) = \{\mathbf{h}_\theta(\mathbf{d}_i)_1, \mathbf{h}_\theta(\mathbf{d}_i)_2, \ldots, \mathbf{h}_{\,theta}(\mathbf{d}_i)_v\}$ é uma sequência de vetores ocultos mapeados por um Transformer.

5.2.2 Representação de Série Temporal

Uma série temporal TS de tamanho m é definida como uma sequência ordenada de observações, ou seja, $TS = (s_1, s_2, ..., s_m)$, em que $s_t \in \mathbb{R}^d$ representa uma observação s no tempo t com características f. Nesta abordagem, as características f representam séries temporais financeiras com valores de fechamento e volumes de negociação realizados no dia.

Cada documento de texto d_i está alinhado com s_i no tempo t. Para classificar d_i, foi considerado u observações anteriores a cada s_i, ou seja, a cada d_i, um subconjunto s_{i-u} é alinhado temporalmente. Para construir a modelagem do grafo é empregada KMeans de séries temporais para processar agrupamentos de todos os conjuntos $SC = (s_{1-u}, s_{2-u}, ..., s_{m-u})$. O objetivo é encontrar K centroides em $C = (c_1, c_2, ..., c_k)$ que minimizem a soma das distâncias euclidianas quadradas entre cada subconjunto de séries temporais (s_{i-u}) e os centroides do agrupamento ao qual pertence. A função objetivo K-means é dada por:

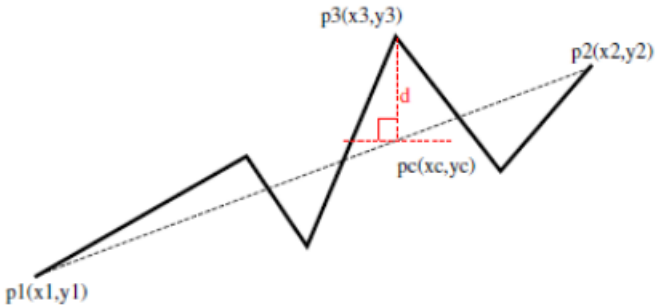
$$J(C, SC) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} ||s_{ik} - c_k||^2 \tag{5.6}$$

em que o K é o número de agrupamentos, n_k é o número de séries temporais no agrupamento k, s_{ik} é um ponto em uma série temporal no agrupamento k e c_k é o centroide do agrupamento k.

A estratégia *Perceptually Important Points* (PIP) é empregada para extrair pontos representativos de momentos cruciais da série temporal. A ideia inicial para o processo de identificação PIP foi introduzida por (CHUNG *et al.*, 2001) nos padrões de análise técnica para aplicações financeiras. Testes comparativos envolvendo três técnicas diferentes para cálculo da distância entre pontos são apresentados em (FU *et al.*, 2008): distância vertical, perpendicular e euclidiana. A Distância Vertical (VD) apresenta resultados superiores e, consequentemente, é utilizada nesta

abordagem. A Figura 33 ilustra o VD entre o ponto de teste $p3$ e a linha que conecta os dois PIPs adjacentes.

Figura 33 – Distância vertical entre dois pontos (s_1, s_m)



Fonte: (FU *et al.*, 2008).

Os pontos $p1$ e $p2$ representam PIPs previamente escolhidos e indicam $p3$ como candidato à promoção a PIP. Este cálculo é realizado para todos os pontos da série temporal, e os maiores valores n são selecionados como PIPs. Em termos práticos, a Equação 5.7 é usada para calcular as distâncias verticais dos pontos candidatos entre (s_1, s_m).

$$VD(p_3, p_c) = |y_3 - y_c| = \left| \left(y_1 + (y_2 - y_1) \left(\frac{x_3 - x_1}{x_2 - x_1} \right) \right) - y_3 \right| \tag{5.7}$$

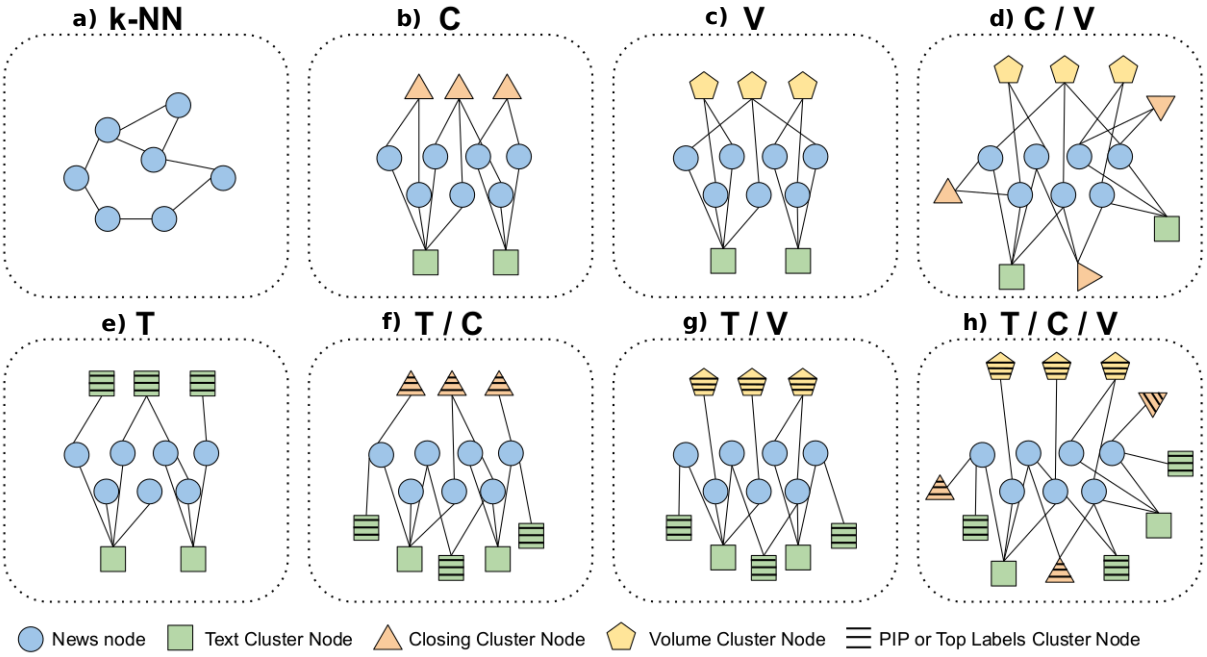
em que o VD é igual à distância entre a linha tracejada e o ponto de teste $p3$. Pontos Perceptualmente Importantes (PIPs) em séries temporais podem ser usados para identificar instâncias específicas em que ocorrem mudanças ou anomalias significativas, atraindo a atenção devido ao seu impacto perceptível no padrão geral. A presente abordagem também explora essa estratégia de representação das séries temporais, já que a *Efficient Market Hypothesis* (HEM) postula que todas as informações relevantes são rapidamente refletidas nos preços dos ativos (TIMMERMANN; GRANGER, 2004). No entanto, a identificação de anomalias nos PIPs desafia esta suposição, sugerindo casos em que ineficiências de mercado ou eventos inesperados podem influenciar os preços.

5.2.3 Modelagens dos Grafos

A presente abordagem apresenta modelagens de grafos com dados de séries temporais e textos, utilizando de séries financeiras do mercado da soja, em específico os valores de fechamento e volume de negociação. Nesse sentido, são construídos grafos considerando nós dos valores de fechamento e os nós do volume de negociação. Essas informações são inseridas no grafo usando nós de agrupamentos. Primeiramente, as notícias são agrupadas considerando cada tipo de informação. O número de agrupamentos corresponde ao número de nós, e cada notícia é conectada com o agrupamento correspondente. Por exemplo, pode-se agrupar as notícias com *embeddings* de texto (BERT), gerar os tipos de nós do agrupamento de texto e conectar as

notícias aos seus respectivos agrupamentos. É importante notar que todos os grafos possuem este tipo de nó de agrupamento de texto (exceto o grafo gerado k-nn). A Figura 34 ilustra as estratégias de modelagem de grafos.

Figura 34 – Tipos de nós e estratégias para modelar os grafos usando notícias e informações de séries temporais.



Fonte: Elaborado pelo Autor.

O grafo com apenas os nós notícias (modelagem k-NN) é considerado na avaliação (Figura 34 (a)). Os nós são conectados com seus k vizinhos mais próximos nesta modelagem. Em todas as modelagens subsequentes, todas consideram nós de notícias (circunferência azul) e nós de agrupamento de textos (quadrados verdes). Na modelagem ilustrada na Figura 34 (b) são conectados nós de agrupamentos (**C**) dos valores de fechamentos (triangulo). Os nós de agrupamentos de volume (**V**) de negócios são incluídos na modelagem (Figura 34 (c)). Na modelagem ilustrada na Figura 34 (d), os dois tipos de nós (**C/V**) são conectados com nós de notícias e agrupamento de textos. Essas modelagens possuem a conectividade dos nós de notícias e nós de agrupamentos de textos considerando os agrupamentos de volume e fechamento de todos os dias.

A fim de considerar momentos importantes na série temporal e de mercado, modelagens de grafos são apresentados com estratégias de agrupamentos em dias PIP e dias que possuem mais de 70% de notícias relevantes (Top Labels). Dessa forma, as quatro modelagens (**T**, **T/C**, **T/V**, **T/C/V**), respectivamente ilustradas nas Figuras 34 (e) (f) (g) (h), consideram a conexão dos nós de notícias (circunferência azul) e agrupamentos de textos (quadrados verdes) com os agrupamentos de fechamento (triângulo listrado) e de volume (pentágono listrado), somente nos dias que ocorrem PIP na séries de preços, ou em dias com Top Labels (TL).

5.2.4 Redes Neurais de Grafos

Rede Neural de Grafo (do inglês *Graph Neural Network* - GNN) é um modelo mais recente e moderno para o aprendizado de representações multimodais (DWIVEDI *et al.*, 2023; TSITSULIN *et al.*, 2023). A explicação detalhada sobre a fundamentação da GNN foi incluída no capítulo da proposta, em vez de ser inserida no capítulo de fundamentação, porque o capítulo de fundamentação focou em modelos de regressão, enquanto esta parte da tese lida com tarefas de classificação. As GNNs têm o benefício de capturar características estruturais dos grafos e agregam informações de nós vizinhos para gerar *embeddings* para os nós (WU *et al.*, 2020). Formalmente, GNNs consideram a representação estruturada de cada nó $\mathbf{o}_i \in \mathbf{O}$ e a matriz de adjacência $\mathbf{A}$ como entrada para o processo de aprendizagem de representação. Portanto, $g(\mathbf{O},\mathbf{A};\mathbf{W})$ representa uma GNN com pesos treinados $\mathbf{W} = \{\mathbf{W}^{(1)},\cdots,\mathbf{W}^{(L)}\}$ em L camadas ocultas. Formalmente, para a 1-ésima camada, a propagação GNN pode ser resumida da seguinte forma:

$$\mathbf{H}^{(l+1)} = g(\mathbf{H}^{(l)},\mathbf{A};\mathbf{W}^{(l)}), \tag{5.8}$$

em que $\mathbf{H}^{(l)}$ é a entrada para a l-ésima camada GNN, e $\mathbf{H}^{l+1}$ é a saída desta camada . As representações $\mathbf{O}$ são as entradas para a primeira camada, ou seja, $\mathbf{O} \equiv \mathbf{H}^{(0)}$. Nesse sentido, $\mathbf{H}^{(L)}$ são os *embeddings* aprendidos para cada nó.

Um GNN tem duas etapas principais durante o aprendizado. Primeiramente, a etapa de agregação visa coletar informações dos vizinhos de cada nó. A segunda é a etapa de combinação, que busca atualizar as representações do nó combinando as informações agregadas dos vizinhos com as representações existentes do nó (TANG; LIAO, 2022). A agregação é definida por (XU *et al.*, 2019; TANG; LIAO, 2022):

$$\mathbf{a}^{l}_{o_i} = AGGREGATE^{(l)}(\{\mathbf{h}^{l-1}_{o_j} : o_j \in N_{o_i}\}), \tag{5.9}$$

em que $\mathbf{a}^{l}_{o_i}$ é o resultado da agregação dos vizinhos o_i definidos como N_{o_i}, e $\mathbf{h}^{l-1}_{o_j}$ é o vetor de recursos do nó o_j na camada $l-1^{th}$. A combinação da representação agregada e da representação do nó é definida pela Equação 5.10 (XU *et al.*, 2019; TANG; LIAO, 2022). Por exemplo, o operador médio pode realizar etapas agregadas e combinadas.

$$\mathbf{h}^{l}_{o_i} = COMBINE^{(l)}(\mathbf{h}^{l-1}_{o_i},\mathbf{a}^{l}_{o_i}). \tag{5.10}$$

Existem diferentes camadas GNN quando a propagação de pesos GNN (Equação 5.8), agregação (Equação 5.9) ou combinação (Equação 5.10) são modificadas. Essa abordagem se concentrará na Rede Convolucional de Grafos (GCN) (KIPF; WELLING, 2017) e no Grafo

SAGE (HAMILTON; YING; LESKOVEC, 2017). Formalmente, um GCN pode ser definido pela Equação 5.11 (KIPF; WELLING, 2017):

$$\mathbf{H}^{(l+1)} = \sigma(\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}), \tag{5.11}$$

em que $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ é a matriz de adjacência com relações próprias, $\mathbf{I}$ é a matriz identidade, $\tilde{\mathbf{D}}$ é uma matriz diagonal com $\tilde{\mathbf{D}} = \sum_j \tilde{\mathbf{A}}_{ij}$, e σ é uma função de ativação. No GCN, as etapas de agregação e combinação são integradas da seguinte forma (XU *et al.*, 2019):

$$\mathbf{h}^l_{o_i} = \sigma(\mathbf{W}^{(l)} \cdot MEAN\{\mathbf{h}^{l-1}_{o_j} : o_j \in \{N_{o_i} \cup o_i\}\}. \tag{5.12}$$

Hamilton, Ying e Leskovec (2017) propôs o GraphSAGE criando um novo método de combinação e diferentes agregadores. O método GraphSAGE realiza uma amostragem uniforme de $N(.)$ vizinhos, o que reduz o tempo e a complexidade da memória. Esta abordagem apresenta o método de combinação GraphSAGE na Equação 5.13 e o agregador *Pooling* na Equação 5.14:

$$\mathbf{h}^l_{o_i} = \sigma\left(\mathbf{W}^l \cdot CONCAT(\mathbf{a}^l_{o_i}, \mathbf{h}^{l-1}_{o_i})\right), \tag{5.13}$$

$$AGGREGATE^l_{pool} = max(\{\sigma(\mathbf{W}_{pool}\mathbf{h}^l_{oj} + \mathbf{b}), \forall_{o_j} \in Ns(o_i)\}), \tag{5.14}$$

em que $Ns(o_i)$ é o conjunto de vizinhos o_i amostrado, $\mathbf{W}_{pool}$ são os pesos de aprendizagem da agregação e *max* denota o operador máximo elemento a elemento.

Para a classificação binária usando GNNs, as representações $\mathbf{H}^{(L)}$ e uma camada *softmax* são usadas para calcular a probabilidade de cada documento pertencer às duas classes (YAO; MAO; LUO, 2019). Considerando uma GNN de 2 camadas:

$$\mathbf{Y} = softmax(g(g(\mathbf{H}^{(0)}, \mathbf{A}; \mathbf{W}^{(0)}), \mathbf{A}; \mathbf{W}^{(1)})) \tag{5.15}$$

em que $\mathbf{Y}$ é uma matriz com duas colunas indicando a probabilidade de cada nó pertencer às classes relevantes ou irrelevantes.

5.2.5 Configuração experimental

A presente abordagem considera diferentes modelagens de grafos usando dados de séries temporais e de texto do mercado de soja. Vários modelos preditivos são usados para a tarefa de classificação, incluindo um modelo BERT pré-treinado. Ainda, diferentes paradigmas de aprendizado de máquina são utilizados para comparar com as técnicas de modelagem de grafos. Os modelos Multi-Layer Perceptron (MLP) do paradigma conexionista, o K-Nearest Neighbors (KNN) do paradigma de classificação baseado em instâncias, os probabilísticos *Gaussian Naive*

Bayes (GNB) e *Multinomial Naive Bayes* (MNB), o *Support Vector Machine* (SVM) representa o paradigma da teoria da aprendizagem estatística são empregados na avaliação. Por fim, o *Fine-tunning* BERT também é considerado para comparar com as modelagens de grafos. A métrica de acurácia é utilizada para avaliar o desempenho dos resultados.

Conjunto de dados de texto e uma série temporal relacionada à soja são selecionados para avaliação. Os dados textuais em português foram extraídos de um site de notícias agrícolas[3]. Fundada em 1997, a *Notícias Agrícolas* está entre os veículos de comunicação mais influentes do agronegócio brasileiro e também republica notícias de outras fontes. A Série Temporal refere-se às flutuações de preços na Bolsa de Chicago (CBOT). A Tabela 27 fornece uma descrição do período do conjunto de dados, o número de dias e detalhes sobre as informações do texto.

Tabela 27 – Visão geral dos dados de textos e séries temporais usado para avaliação experimental.

Commoditie	Soja
Período	02-01-2015 à 31-12-2022
Número de dias	2012
Atributos da ST	Valores (**Close**, **Volume**, Open, High, Low)
Número de manchetes de notícias	10534
Rótulos	4421 Relevante (1), 6113 Irrelevante (0)

Fonte: Elaborado pelo Autor.

Esta abordagem apresenta um conjunto de dados composto por 2.012 dias e 10.534 artigos de notícias e manchetes. O conjunto de dados foi coletado de um site, rotulado como relevante ou irrelevante, e disponibilizado em um repositório online[4]. Dois especialistas do domínio, juntamente com dez estudantes de graduação, participaram do processo de rotulagem. Durante o processo de rotulagem, foi categorizado os textos como relevantes se contivessem informações que pudessem potencialmente moldar os desenvolvimentos futuros no mercado de soja, fornecendo percepções que poderiam influenciar as decisões de negócios. Por outro lado, artigos de notícias detalhando eventos que já haviam ocorrido e não tinham capacidade de impactar o mercado foram considerados irrelevantes. A metodologia utilizada no processo de rotulação está detalhada no trabalho (FILHO *et al.*, 2024).

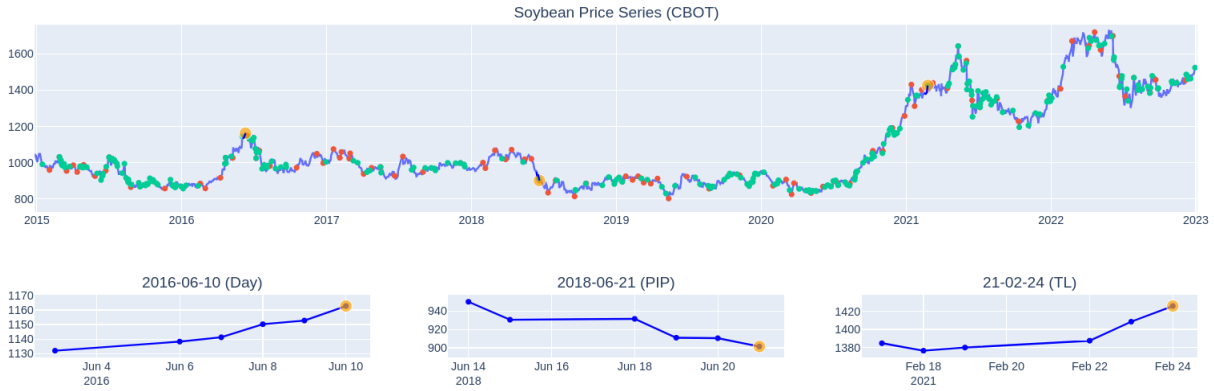
Na fase inicial de pré-processamento, *embeddings* textuais são geradas para as manchetes das notícias, usando o modelo BERT Paraphrase multilíngue pré-treinado[5]. Este modelo mapeia frases e parágrafos em um espaço vetorial denso de dimensões de 384, tornando-o aplicável a tarefas de agrupamento ou pesquisa semântica. Para representar a camada “nó de agrupamento de texto", o método KMeans é usado para gerar clusters k. A camada “nó de agrupamento de texto"é incorporada em todos os modelos. Essa camada se conecta aos nós de notícias correspondentes dentro do seu agrupamento.

[3] https://www.noticiasagricolas.com.br/
[4] https://data.mendeley.com/datasets/f8fdmpp6yh/2
[5] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

A segunda fase envolveu a construção das camadas de nós para o valor de fechamento do mercado de soja (C) e o volume diário de negociação (V). Após uma série de experimentos, cada documento (d_i) é alinhado temporalmente com um subconjunto da série temporal, contendo valores dos últimos cinco dias (s_{i-5}). Assim, cada título (d_i) no momento t está alinhado com o subconjunto de valores (s_{i-5}) da série dos dias anteriores. Em seguida, os métodos *ScalerMeanVariance* e *KMeans* são aplicados em séries temporais para determinar os k agrupamentos da série s_i para títulos d_i. Como resultado, as camadas “Nó do agrupamento de fechamento"e “Nó do agrupamento de volume"são conectadas à camada de nós de notícias correspondente.

Na próxima fase, camadas de nós são construídas em pontos da série temporal em que ocorrem os PIPs. Essa estratégia envolve a identificação de documentos cruciais (d_i) e valores (s_i) para formar agrupamentos de textos (T), valores de fechamento (C) e volumes (V) para aqueles dias específicos. São extraídas 12 PIPs por ano, totalizando 72 pontos em seis anos para o treinamento e 24 pontos para as etapas de testes. Adicionalmente, é adotada a estratégia de considerar os dias em que mais de 70% das notícias são relevantes, denominados Top Labels (TL). A hipótese é que algo significativo provavelmente acontecerá no mercado financeiro em dias com maioria de notícias relevantes. Essas modelagens também incluem o alinhamento temporal dos valores dos últimos cinco dias e a aplicação dos métodos *ScalerMeanVariance* e *KMeans*. A Figura 35 ilustra os momentos das ocorrências de PIP e TL na série de preços.

Figura 35 – Ocorrências de PIP e Top Labels na séries de preços da soja.



Fonte: Elaborado pelo Autor.

A Figura 35 ilustram PIPs (pontos vermelhos) e momentos na série de preços onde ocorrem os Top Labels (pontos verdes). Além disso, a figura inclui três exemplos para demonstrar que os documentos d_i (pontos laranja) no momento t estão alinhados com a série de preços s_{i-5}. No total, são 270 TLs de 2015 a 2022. Os PIPs geralmente coincidem com picos positivos ou negativos da série, enquanto os TLs são distribuídos em períodos aleatórios. Houve 11 pontos comuns entre PIPs e TLs, indicando que estes são considerados pontos importantes na série de preços e apresentaram muitas notícias relevantes no dia.

Com base nas estratégias de agrupamento de textos, valores de fechamento e volumes de negócios no mercado financeiro, a modelagem de diferentes grafos é idealizada na avaliação

experimental. Os valores de $k = \{8, 16\}$ são escolhidos para modelar todas as camadas (C, V, T). Esses tamanhos k são selecionados como os melhores valores de agrupamento após vários experimentos. A exploração combinatória entre as camadas é realizada para os valores de k, e a combinação ótima é escolhida e apresentada na Tabela 29. Os parâmetros utilizados no GNN são detalhados na próxima seção.

Na etapa de avaliação experimental, os seguintes parâmetros são utilizados para o método proposto e os baselines.

- **Multi Layer Perceptron (MLP)**: ativação: {Relu, Logistic, Tanh}, hidden layer sizes: {[50], [150], [50,50,50], [150,150,150], [50,50,50,50,50,50], [150,150,150,150,150,150]}.
- **Support Vector Machine (SVM)**: kernel: {linear, rbf, sigmoid, poly}, gama: {scale, auto}, degree: {2, 3, 4, 5, 6}.
- **K Nearest Neighbor (KNN)**: number of neighbors: {3, 7, 11, 15}, métrica: cosseno, algoritimo: Balltree.
- **Gaussian Naive Bayes (GNB)**: variance smoothing: {1e-01, 1e-03, 1e-05, 1e-10, 1e-11, 1e-12, 1e-20, 1e-21, 1e-22, 1e-23,}
- **Random Forest (RF)**: estimadores: {50, 80, 100, 120, 150, 200}, max. depth: {5, 6, 8, 10, 12, 15}, min. samples split: { 2, 3, 5, 10}, min. samples leaf: { 1, 2, 4}, bootstrap: {True, False}.
- **Fine Tuning BERT**: epochs: {15, 30}, batch size {32, 64}, learning rate = {0.0001, 0.0005, 0.001}, maxlen: 50.
- **Graph Neural Networks (GNN):** layer types: {Convolutional, GraphSage}, architectures: {[8, 2], [16, 12]}, learning rate = {0.0001, 0.001, 0.01}, epochs: 1000, early stop with patience = {50, 100}.

Considerando o conjunto de dados, os dados de 2015 a 2020 são utilizados para treinamento e de 2021 a 2022 para testes, pois o aprendizado temporal entre treino e teste é essencial para a avaliação. A fim de escolher os melhores parâmetros, entre o conjunto de parâmetros, para treinar os modelos e prever o conjunto de testes é gerado um conjunto de validação com 10% do conjunto de treinamento. Por fim, o desempenho dos modelos são avaliados comparando a precisão (*accuracy*) do modelo.

5.2.6 Resultados e Discussão

Avaliação experimental é realizada para investigar dois aspectos. Em primeiro lugar, analisar o impacto dos modelos unimodais e multimodais na tarefa de classificação. Dessa

forma, é gerado quatro representações: i) incorporação de texto unimodal (TE); ii) três modelos multimodais nos quais são propostos diferentes concatenações da representação de texto com a representação da série temporal (valor de fechamento e volume de negócio). Assim, as *Embeddings* de Textos e valores de Fechamento (TEC), *Embeddings* e Volume de negócios (TEV), e *Embeddings* com Fechamento e Volume (TECV) são apresentada como *baseline*. A Tabela 28 apresenta os resultados para essas quatro representações considerando cinco classificadores.

Tabela 28 – Resultados para as quatro representações TE, TEC, TEV e TECV considerando cinco classificadores. Os melhores resultados para cada classificador estão em negrito. Entre parênteses o valor usado na próxima discussão comparando este resultado com os resultados do grafos.

	TE	**TEC**	**TEV**	**TECV**
MLP	(**0.708**)	0.696	0.703	0.662
SVM	**0.703**	0.697	0.698	0.687
KNN	**0.673**	0.651	0.627	0.632
GNB	0.652	**0.653**	0.651	0.652
RF	0.674	**0.675**	0.668	0.635

Fonte: Elaborado pelo Autor.

O pior desempenho foi observado com TEV utilizando KNN, e o melhor foi alcançado com TE utilizando MLP. A representação que gerou os piores resultados foi a TECV, enquanto a TE foi a melhor, seguida da TEC. Devido à ênfase apenas na representação textual, concatenar as representações para ter o benefício da multimodalidade não alcançou resultados esperados para classificar as notícias do mercado da soja. Isto sugere que ou o método de combinação das informações poderia ter sido melhor escolhido (apesar de ser utilizado principalmente na literatura) ou que as informações de preços da série temporal não influenciam a classificação das notícias como relevantes ou irrelevantes. Adicionalmente, as representações multimodais foram combinadas por meio de Grafos para uma avaliação mais abrangente envolvendo GNN.

Na próxima etapa, é proposto um GNN heterogênea para classificar as notícias como relevantes ou irrelevantes para o mercado da soja. Esta comparação também visa discernir se a concatenação das informações prejudica os modelos ou se a inclusão dos dados de preços da soja melhora a classificação das notícias. A Tabela 29 apresenta os resultados das modelagens de Grafos. Diferentes modelagens de séries temporais e dados textuais são comparados usando GNN, a fim de avaliar se essas informações extras (preço da soja) melhoram a classificação das notícias.

Na Tabela 29, a modelagem **T** utilizando a estratégia PIP teve o pior resultado GNN, enquanto a modelagem **T/V** com a estratégia Top Labels (TL) gera a maior precisão. As estratégias Daily, PIP e Top Labels obtiveram melhor acurácia com a modelagem **C**, **T/C** e **T/V**, respectivamente. Destaca-se que sete grafos modelados com texto e séries temporais alcançam maior precisão do que a modelagem *k*NN. Por outro lado, a modelagem *k*NN obtém maior precisão do que nove modelagens de grafos com informações de texto e séries temporais. Embora haja um equilíbrio entre alcançar melhores valores de precisão para o GNN com ou sem

Tabela 29 – Resultados das modelagens de grafos C, V, T e *k*-NN, considerando estratégias com série de preços diário, PIP e Top Labels. Os melhores resultados estão em negrito. Entre parênteses, o valor usado na próxima discussão comparando este resultado com o melhor classificador de linha de base.

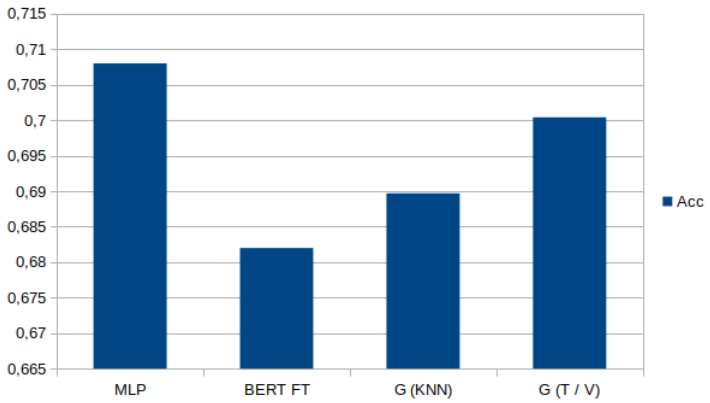
	C	**V**	**C/V**	**T**	**T/C**	**T/V**	**T/C/V**	**KNN**
Diário	**0.696**	0.693	0.685	-	-	-	-	(0.689)
PIP	0.684	0.675	0.668	0.667	**0.689**	0.683	0.687	-
TL	0.699	0.697	0.693	0.693	0.678	(**0.700**)	0.688	-

Fonte: Elaborado pelo Autor.

os dados de séries temporais, as diferenças estão localizadas na terceira casa decimal, ou seja, indicando que são diferenças pequenas e insignificantes. Além disso, é fundamental observar que não há unanimidade sobre o método ideal para modelar textos e séries temporais por meio de grafos para melhor classificar as notícias como relevantes ou irrelevantes para o mercado da soja. Esses resultados de grafos sugerem evidências de que o preço da soja, seja modelado por meio de grafos ou concatenado em representações textuais, não influencia a classificação das notícias.

Outro aspecto analisado é comparar o melhor resultado de cada modelo com o BERT Finetuning, uma vez que este método é um dos modelos estado da arte para dados de texto. Dessa forma, é comparado o melhor resultado da classificação de grafos, modelado com dados textuais (*k*-NN), dados textuais com dados de séries temporais (TL - T/V) e o MLP com a representação TE. A Figura 36 apresenta estes resultados em que o BERT Finetuning obtém a menor precisão $(0,682)$, e o melhor resultado é o MLP, seguido da modelagem de grafo (GNN) enriquecido com séries temporais. Ressalta-se que mesmo com resultados melhores e piores, a diferença entre os métodos ficou na terceira casa decimal. Este fato mostra que os métodos obtiveram resultados que não apresentam diferença significativa. Mais uma vez, esta comparação também mostra evidências de que as informações sobre os preços da soja não contribuíram significativamente para a classificação da notícia.

Figura 36 – Comparação entre BERT Finetuning (BERT FT) e os melhores resultados da Tabela 29 (Grafo *k*NN e Grafo (TL - T/V) e Tabela 28 (MLP).



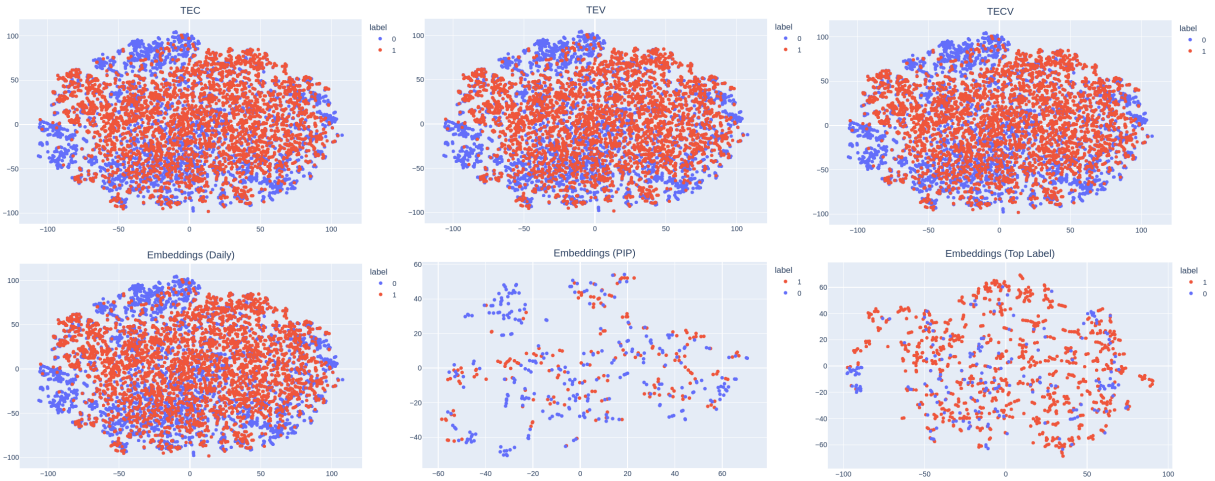
Fonte: Elaborado pelo Autor.

Considerando os diferentes grafos propostos, nota-se que a GNN não obtém maior

precisão do que o MLP com a representação BERT pré-treinada. Por outro lado, a rede neural de grafos contribui para a multimodalidade, heterogeneidade e interpretabilidade, uma vez que os grafos podem ser enriquecidos com informações para cobrir esses aspectos. Além disso, esses aspectos interessam a usuários e empresas que buscam obter notícias relevantes do mercado da soja, e os grafos podem oferecer esses aspectos.

Visualmente pode ser analisado que as representações (TE, TEC, TEC e TECV) e representações das manchetes de notícias em dias de PIP ou Top Labels. A Figura 37 apresenta projeções bidimensionais das representações. As representações são geradas usando *t-distributed Stochastic Neighbor Embedding* (t-SNE) para a análise (MAATEN; HINTON, 2008).

Figura 37 – Projeções bidimensionais das notícias relevantes (pontos vermelhos) e irrelevantes (pontos azuis). Considerou-se as quatro representações utilizadas e as representações do BERT nos dias PIP ou Top Labels.



Fonte: Elaborado pelo Autor.

O primeiro aspecto que observa-se é a semelhança da projeção bidimensional da representação unimodal *Embeddings* (TE) com as demais representações concatenadas, o que reforça ainda mais a falta de contribuição da informação de preços na separação das notícias em relevantes (rótulo 1) ou irrelevantes (rótulo 0). Além disso, observa-se que alguns grupos nas representações apenas nos dias do PIP, mas sempre com um misto de notícias relevantes e irrelevantes. Nos dias de Top Labels, fica claro que há uma presença mais significativa de notícias relevantes, pois este é um fator de seleção. Por outro lado, destaca-se que notícias irrelevantes também ocorrem nesses dias, e não houve separação de suas notícias apesar do predomínio de notícias relevantes. Por fim, as projeções medem a dificuldade de separar notícias relevantes e irrelevantes no mercado da soja, o que converge com os resultados de 70% de precisão.

5.3 Considerações Finais

No presente capítulo foram apresentadas duas abordagens de classificação de notícias considerando dados de séries temporais e dados de textos. Na primeira abordagem, foi apre-

sentado um modelo de rotulagem automática de texto usando informações extraídas de séries temporais. Esta abordagem inovou ao considerar uma técnica de supervisão fraca para rotular um grande volume de textos por meio de padrões obtidos da série temporal. Documentos de texto e séries de preços de commodities do agronegócio foram utilizados para avaliar o desempenho em quatro cenários de classificação (CPB, SPB, CNB e SNB). A avaliação experimental considerou nove representações textuais e diferentes paradigmas de aprendizagem. Além disso, foi proposto um novo modelo de representação vetorial de texto que mede a distância de cosseno entre Termos e Documentos a partir de modelos BERT pré-treinados (TD-BERT).

Ainda sobre a primeira abordagem, a função de rotulagem apresentada pode ser uma alternativa para anotar um grande volume de documentos de texto. A rotulagem automática pode ser imprecisa, mas útil quando muitos textos não estão rotulados. Nesta abordagem, a limitação da análise da supervisão fraca consistiu no desequilíbrio das classes. Pesquisas futuras podem desenvolver estratégias para propagar rótulos por meio de aprendizagem semissupervisionada para reforçar a rotulagem, ou modelos com base em *Graph Neural Network* podem ser consideradas para incluir novas modalidades e auxiliar na tarefa de rotulagem.

Com relação a segunda abordagem, modelagens de grafos com dados textuais e séries temporais foi apresentada, visando explorar a heterogeneidade dos dados e classificar as notícias do mercado de soja como relevantes ou irrelevantes. Essa abordagem teve como objetivo de superar limitações percebidas nos estudos na literatura, especialmente aquelas apresentadas e discutidas no mapeamento sistemático. Muitos estudos concentram-se predominantemente na classificação de notícias como positivas ou negativas, utilizando de estratégias léxicas, análise de sentimentos e aprendizados multimodais. No entanto, é importante ressaltar que o interesse do mercado financeiro vai além da polaridade das notícias, focando na análise da sua relevância. Independentemente de sua polaridade, uma notícia pode ser significativa para o mercado, influenciando decisões de investimento e tendo impacto nos preços dos ativos. Dessa forma, outra contribuição foi de propor um processo de rotulação de notícias do mercado da soja, em colaboração com profissionais do setor e estudantes de graduação, para rotular mais de onze mil notícias como relevantes e irrelevantes.

Ainda sobre a segunda abordagem, a configuração experimental inovou em propor diferentes modelagens de grafos, em que considera a conexão de diferentes agrupamentos de séries temporais com *embeddings* de textos. Respondendo as questões de pesquisa levantadas na segunda abordagem, os resultados iniciais indicaram que as representações concatenadas com a estratégia de *early fusion* não melhoram significativamente a precisão. Esse desempenho ocorre porque concatenar apenas cinco valores da série temporal em um vetor de incorporação denso de 384 posições não introduz mudanças semânticas significativas na representação textual. Em relação à segunda questão de pesquisa, a incorporação de séries temporais na modelagem de grafos também não resulta em desempenho superior em comparação com abordagens baseline. Apenas as informações de séries temporais não captam a dinâmica e a complexidade do mercado

financeiro. Por fim, embora a modelagem de grafos não tenha alcançado precisão superior em relação aos baselines, os resultados revelam seu potencial em comparação com outros modelos de aprendizagem.

A complexidade intrínseca ao mercado financeiro coloca desafios à modelagem do grafo, uma vez que factores imprevisíveis podem influenciar significativamente os resultados. Portanto, é importante reconhecer as limitações deste trabalho, nomeadamente a limitada variedade de dados disponíveis de outras modalidades de modelagem de grafos. Embora isto possa ser considerado uma restrição, abre oportunidades para pesquisas e inovações mais amplas na aplicação de tecnologia com base em grafos no domínio financeiro. Adicionalmente, a modelagem de grafos não considerou os efeitos de períodos sazonais ou cíclicos específicos no mercado de soja. Por outro lado, destaca-se que a segunda abordagem representa a primeira tentativa de modelar a relação entre agrupamentos de séries temporais e textos, contribuindo assim para o avanço do conhecimento neste domínio.

CAPÍTULO

6

CONCLUSÕES

No presente capítulo são listadas as conclusões, bem como as contribuições de pesquisas realizadas neste trabalho. Também, as publicações realizadas ao longo do trabalho de Doutorado são apresentadas, bem como os trabalhos futuros que podem dar continuidade a esta pesquisa.

6.1 Contribuições Científicas

Neste trabalho de doutorado, duas propostas foram apresentadas para a integração de séries temporais e informações textuais em tarefas de regressão e classificação. A primeira proposta envolve estudos para enriquecer séries temporais por meio de técnicas de Mineração de Texto para tarefas de regressão. Após uma série de experimentos, pesquisas e publicações, as questões de pesquisa apresentadas no Capítulo 1 serão abordadas a seguir:

QP1.1 *Quais são os modelos de representação textual mais utilizados para enriquecer séries temporais?* A resposta da QP foi obtida por meio do mapeamento sistemático realizado no Capítulo 3. Abordagens pioneiras, que majoritariamente utilizaram técnicas de *Early Fusion*, optaram predominantemente utilizar por representações independentes de contextos, como o Word2Vec, SenticNet e outros. Essas abordagens geralmente incorporam múltiplos dados em um representação única, e potencialmente optaram por essas estratégias pela simplicidade da fusão entre os dados, sejam por estratégias de concatenação linear, soma das *features* ou processamento paralelo. Nos últimos anos, com o advento dos modelos *Transformers*, GNN e técnicas de *multi-head attention*, abordagens têm amplamente considerado *text embeddings* de modelos pré-treinados da arquitetura *Transformers*. As representações *embeddings* têm sido empregadas para aprender correlações entre dados multimodais e gerar novas representações que capturam interações elas. Um benefício dessa abordagem é a possibilidade de gerar representações de dimensões mais compactas, e por isso, facilita a compreensão de padrões temporais. Por outro lado, uma desvantagem em potencial é a possível perda da interpretabilidade dos modelos de previsão.

QP1.2 *Como as representações (textos e séries) podem impactar as estratégias dos modelos de fusão nas tarefas preditivas?* A resposta da QP foi obtida por meio do mapeamento sistemático e resultados obtidos na proposta do Capítulo 4. Considerando a estratégia *Early Fusion* para enriquecer séries temporais, constatou-se que: i) as representações com base na BoW e *embeddings* de textos geram representações esparsas e extensas; ii) representações com termos específicos do domínio podem definir representações menores, mas por outro lado, não oferecem padrões semânticos que representam a complexidade de mercado. iii) representações independentes de contexto (exemplo o word2vec) podem não representar a complexidade semântica das notícias do mercado de maneira adequada. iv) a concatenação linear entre representações de séries temporais e dados textuais não possibilita uma combinação mais profunda e integrada dos dados durante o processo de aprendizado. Em relação a estratégia *Joint Fusion*, destaca-se que: i) independente da dimensão das representações, modelos com base em *Transformers* e técnicas *multi-head attention* pode ser utilizadas para encontrar correlações entre os dados multimodais e estabelecer padrões temporais; ii) estratégias podem ser empregadas para combinar os dados e gerar dimensões mais compactas para ser usada como entrada nos modelos preditivos.

QP1.3 *Quais são os modelos amplamente utilizados e que resultam em melhores resultados para predição de séries temporais?* Entre os estudos explorados no mapeamento sistemático, observa-se que a maioria dos estudos empregam o modelo LSTM e apresentam ganho de desempenho em relação aos baselines, variando a configuração das camadas e a integração da série temporal e dos dados textuais como entrada no modelo. Por outro lado, é importante ressaltar que abordagens mais recentes têm empregado técnicas de *multi-head attention* combinadas com diferentes modelos de GNN, LSTM e transformers, indicando uma tendência na pesquisa atual. Entre os modelos apresentados nesta tese, não foi identificada uma unanimidade em relação ao melhor desempenho de um modelo específico.

A segunda proposta envolve em integrar dados de séries temporais para classificação de textos, apresentadas em duas abordagens diferentes. A seguir são respondidas as questões de pesquisas levantadas no Capítulo 1.

QP2.1: *Como utilizar dados de séries financeiras para classificar notícias?* O uso de dados de séries temporais pode ocorrer de diversas formas. Uma delas é a rotulação automática de notícias, em que as variações abruptas nas séries de preços, frequentemente relatadas em textos de notícias, são identificadas e utilizadas para categorizar as notícias de acordo com seu impacto no mercado financeiro. Outra alternativa é o uso das séries temporais para enriquecer as representações semânticas de textos, considerando indicadores de mercado. Nesse contexto, as oscilações e indicadores técnicos das séries de preço são integrados aos modelos multimodais, proporcionando uma visão mais abrangente e precisa das relações entre eventos noticiosos e movimentos do mercado financeiro. Em outra perspectiva, pesquisas recentes têm demonstrado eficiência ao considerar múltiplas fontes de dados relacionadas a empresas ou mercados específicos. Por exemplo, o mercado da soja está intrinsecamente ligado a uma série de fatores,

incluindo o mercado cambial, o mercado externo e o mercado interno. Portanto, considerar dados de séries temporais que abrangem esses fatores é fundamental para classificação de notícias em que necessita compreensão mais completa das dinâmicas do mercado.

QP2.2: *Séries temporais financeiras contribuem significativamente para a classificação de textos relacionados ao mercado de commodities?* Considerar apenas os dados de séries temporais financeiras não contribui significativamente para melhorar o desempenho em tarefas de classificação. A hipótese do mercado eficiente sugere que as oscilações dos preços já refletem as informações disponíveis publicamente, o que deixa pouco espaço para uma exploração consistente desses dados para obter ganhos preditivos. No entanto, há lacunas a serem investigadas em relação à incorporação de múltiplas fontes de dados relacionadas ao mercado-alvo, o que pode contribuir significativamente para modelos multimodais mais eficazes.

QP2.3:*Quais são os modelos de aprendizado mais adequados para abordagens multimodais?* Os modelos mais adequados são aqueles capazes de lidar eficientemente com dados heterogêneos. Um exemplo são as *Graph Neural Networks* (GNNs), que podem ser projetadas para incorporar múltiplas fontes de dados por meio de diversas camadas e nós, permitindo a modelagem de relações complexas entre os diferentes tipos de informações. Essa abordagem possibilita a inclusão de *embeddings* de texto, indicadores de séries temporais de diversas fontes, entre outros dados relevantes. Outra opção viável são os modelos baseados em Transformers, que possuem camadas de encoder capazes de processar dados de texto e outros tipos de dados de forma eficiente. Esses modelos podem ser adaptados para lidar com múltiplas fontes de dados, utilizando diferentes encoders para cada tipo de informação e, em seguida, combinando essas representações para obter uma visão integrada e abrangente. Uma estratégia promissora é a utilização de técnicas de *multi-head attention*, que permitem que o modelo atente a diferentes partes dos dados de entrada simultaneamente, possibilitando a captura de relações complexas e a geração de representações combinadas mais ricas. Essas técnicas podem ser empregadas em conjunto com diferentes arquiteturas de redes neurais, como GNNs, LSTMs e Transformers, para criar modelos poderosos e flexíveis capazes de lidar efetivamente com dados multimodais.

6.2 Publicações

Durante o desenvolvimento da presente Tese, as contribuições foram divulgadas por meio de publicações em periódicos, congressos e conferências. Essas publicações são listadas a seguir, apresentando a relação de cada um com este trabalho e indicando aquelas que estão relacionadas às questões de pesquisas inicialmente estabelecidas.

Artigos publicados em periódico

REIS FILHO, I. J.; MARCACINI, R. M.; REZENDE, S. O.. On the enrichment of time series with textual data for forecasting agricultural commodity prices. **MethodsX**, v. 9, p. 101758, 2022. (Qualis A3). Neste artigo foram publicados contribuições correspondentes às questões de

pesquisa **QP1.2** e **QP1.3**.

REIS FILHO, I. J. R.; COLETI, J. C.; MARCACINI, R. M.; REZENDE, S. O. Dataset: Annotated Soybean Market News Articles. **Data in Brief**, p. 110545, 2024. Contribuições para subsidiar estudos recentes e auxiliar nas questões **QP2.2** e **QP2.3**.

Artigo recentemente submetido em periódico

FILHO, I. J. R.; GOLO, M. P. S.; MARCACINI, R. M.; REZENDE, S. O. How do financial time series enhance the detection of news significance in market movements? a study using graph neural networks with heterogeneous representations. **Neural Computing and Applications**, Springer (Qualis A1). Contribuições para responder às questões de pesquisa **QP2.2** e **QP2.3**.

Trabalho publicados em anais de eventos

REIS FILHO, I. J. et al. Forecasting future corn and soybean prices: an analysis of the use of textual information to enrich time-series. In: **Anais do VIII Symposium on Knowledge Discovery, Mining and Learning**. SBC, p. 113-120. 2020. (Qualis B3). Contribuições para responder às questões de pesquisa **QP1.2** e **QP1.3**.

REIS FILHO, I. J. et al. Sequential short-text classification from multiple textual representations with weak supervision. In: **Brazilian Conference on Intelligent Systems. Cham: Springer International Publishing**, 2022. p. 165-179. (Qualis A4). Contribuições para responder às questões de pesquisa **QP2.1** e **QP2.2**.

CARMO, P.; REIS FILHO, I. J.; MARCACINI, R. Commodities trend link prediction on heterogeneous information networks. In: **Anais do IX Symposium on Knowledge Discovery, Mining and Learning**. SBC, p. 81-88. 2021. (Qualis B3). Este trabalho foi apresentado no KDMile pelo autor principal, Paulo Carmo. Esse trabalho foi convidado para ser publicado em uma revista por ser considerado um dos “best papers". O Autor desta tese colaborou na coleta e o pré-processando dos dados utilizados no trabalho.

REIS FILHO, I. J.; MARCACINI, Ricardo M.; REZENDE, Solange O. Previsão do preço futuro de commodities agrícolas: um estudo para enriquecer séries temporais. In: **Simpósio Brasileiro de Automação Inteligente-SBAI**. 2021. (Qualis B4). Contribuições para responder às questões de pesquisa **QP1.2** e **QP1.3**.

CARMO, P.; REIS FILHO, I. J.; MARCACINI, R. TRENCHANT: TRENd PrediCtion on Heterogeneous InformAtion NeTworks. **Journal of Information and Data Management**, v. 13, n. 6, 2022. (Qualis B1). Este trabalho é uma versão estendida do KDMile por ter sido considerado um dos “best papers"do evento. O Autor desta tese colaborou na coleta e o pré-processando dos dados utilizados no trabalho.

TRINDADE, R. N, MARTINS, L. H., CORREA, G. N., REIS FILHO, I. J.. Using a labeling function for automatic classification of agribusiness news: A weak supervisory approach. In: **Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional**. SBC, p. 73-82.

2022. (Qualis B4). Este trabalho foi apresentado no ENIAC pelo autor principal, o Rodrigo Trindade. O autor principal foi bolsista de iniciação científica no curso de Sistemas de Informação da Universidade do Estado de Minas Gerais - Frutal. O Autor desta tese atuou como orientador do trabalho.

6.3 Trabalhos Futuros

Com relação à previsão de séries temporais enriquecidos com dados semânticos de textos, novas representações podem ser propostas utilizando técnicas de *multi-head attention* para combinar as múltiplas fontes de dados de forma mais rica e gerar representações mais completas. Essas representações podem ser empregadas nos modelos de regressão, permitindo que se aprendam padrões temporais entre múltiplos conjuntos de dados. Nesse sentido, trabalhos futuros podem ser propostos incluindo fontes de dados relacionado ao mercado alvo e variando os métodos de combinação, como a utilização de diferentes tipos de atenção e arquiteturas de redes neurais. Além disso, investigações adicionais podem se concentrar em avaliar o impacto dessas novas representações na precisão e na robustez dos modelos de previsão.

Quanto a classificação de notícias usando séries temporais, abordagens podem ser apresentadas para considerar não apenas os dados relacionados ao mercado alvo, mas também os dados de mercados relacionados. Modelos de Graph Neural Networks (GNN) com mecanismos de atenção podem ser utilizados para considerar múltiplas fontes de dados na modelagem do grafo e utilizar técnicas de *multi-head attention* para a propagação do grafo. Além disso, novas modelagens podem ser apresentadas com mais camadas para capturar representações mais complexas e abstratas dos dados, garantindo assim uma melhor capacidade de generalização e predição. Essas novas abordagens podem contribuir significativamente para melhorar a precisão e a robustez dos modelos de classificação de notícias em cenários financeiros.

REFERÊNCIAS

ACADEMY, D. S. **Deep Learning Book Online**. 2021. Disponível em: <https://www.deeplearningbook.com.br/>. Acesso em: 04 de abril 2021. Citado na página 57.

ADANACIOGLU, H.; YERCAN, M. *et al.* An analysis of tomato prices at wholesale level in turkey: an application of sarima model. **Custos e Agronegócio Online**, v. 8, n. 4, p. 52–75, 2012. Citado na página 24.

AGGARWAL, C. C. **Data Classification: Algorithms and Applications**. 1. ed. [S.l.]: Chapman & Hall/CRC, 2014. Citado nas páginas 23, 46, 48, 50 e 87.

AGGARWAL, C. C.; AGGARWAL, C. C. **Machine learning for text: An introduction**. [S.l.]: Springer, 2018. Citado nas páginas 50 e 110.

AHUMADA, H.; CORNEJO, M. Forecasting food prices: The case of corn, soybeans and wheat. **International Journal of Forecasting**, v. 32, n. 3, p. 838 – 848, 2016. ISSN 0169-2070. Citado na página 24.

AVRAMELOU, L.; NOUSI, P.; PASSALIS, N.; TEFAS, A. Deep reinforcement learning for financial trading using multi-modal features. **Expert Systems with Applications**, Elsevier, v. 238, p. 121849, 2024. Citado nas páginas 69 e 72.

AWAD, A. L.; ELKAFFAS, S. M.; FAKHR, M. W. Stock market prediction using deep reinforcement learning. **Applied System Innovation**, v. 6, n. 6, p. 106, 2023. Citado na página 68.

BENGIO, Y.; DUCHARME, R.; VINCENT, P. A neural probabilistic language model. **Advances in neural information processing systems**, v. 13, 2000. Citado nas páginas 51 e 52.

BOECKING, B.; NEISWANGER, W.; XING, E.; DUBRAWSKI, A. Interactive weak supervision: Learning useful heuristics for data labeling. **arXiv preprint arXiv:2012.06046**, 2020. Citado na página 106.

BOJANOWSKI, P.; GRAVE, E.; JOULIN, A.; MIKOLOV, T. Enriching word vectors with subword information. **Transactions of the association for computational linguistics**, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , v. 5, p. 135–146, 2017. Citado na página 53.

BOX, G. E.; JENKINS, G. M.; REINSEL, G. C.; LJUNG, G. M. **Time Series Analysis: forecasting and control**. [S.l.]: John Wiley & Sons, 2015. Citado nas páginas 37 e 38.

CHANG, Z.; ZHANG, Z. Judging stock trends according to the sentiments of stock comments in expert forums. **Electronics**, MDPI, v. 12, n. 3, p. 722, 2023. Citado nas páginas 73 e 77.

CHATFIELD, C.; XING, H. **The Analysis of Time Series: an introduction with R**. [S.l.]: CRC press, 2019. Citado nas páginas 23, 24 e 36.

CHEN, L.-M.; XIU, B.-X.; DING, Z.-Y. Multiple weak supervision for short text classification. **Applied Intelligence**, Springer, p. 1–16, 2022. Citado na página 106.

CHEN, Y.-F.; HUANG, S.-H. Sentiment-influenced trading system based on multimodal deep reinforcement learning. **Applied Soft Computing**, Elsevier, v. 112, p. 107788, 2021. Citado nas páginas 73 e 74.

CHENG, D.; YANG, F.; XIANG, S.; LIU, J. Financial time series forecasting with multi-modality graph neural network. **Pattern Recognition**, Elsevier, v. 121, p. 108218, 2022. Citado nas páginas 24, 73 e 74.

CHO, K.; MERRIËNBOER, B. V.; BAHDANAU, D.; BENGIO, Y. On the properties of neural machine translation: Encoder-decoder approaches. **arXiv preprint arXiv:1409.1259**, 2014. Citado na página 45.

CHU, A. K. G.; ACLAN, A. J.; EL-ALFY, H.; PARDABAEV, O.; PATEL, D. M. Stock price prediction using machine learning: A survey of recent techniques. In: **22nd International Symposium on Communications and Information Technologies (ISCIT)**. [S.l.: s.n.], 2023. p. 278–283. Citado na página 67.

CHUNG, F. L. K.; FU, T.-C.; LUK, W. P. R.; NG, V. T. Y. Flexible time series pattern matching based on perceptually important points. In: **Workshop on Learning from Temporal and Spatial Data in International Joint Conference on Artificial Intelligence**. [S.l.: s.n.], 2001. Citado na página 117.

CLAPHAM, B.; SIERING, M.; GOMBER, P. Popular news are relevant news! how investor attention affects algorithmic decision-making and decision support in financial markets. **Information Systems Frontiers**, Springer, v. 23, p. 477–494, 2021. Citado na página 115.

CRYER, J. D.; CHAN, K.-S.; CHAN, K.-S. **Time series analysis: with applications in R**. [S.l.]: Springer, 2008. v. 2. Citado na página 32.

DAI, E.; SHU, K.; SUN, Y.; WANG, S. Labeled data generation with inexact supervision. In: **Conference on Knowledge Discovery and Data Mining (SIGKDD)**. [S.l.: s.n.], 2021. p. 218–226. Citado na página 106.

DARADKEH, M. K. A hybrid data analytics framework with sentiment convergence and multi-feature fusion for stock trend prediction. **Electronics**, MDPI, v. 11, n. 2, p. 250, 2022. Citado nas páginas 25, 73 e 75.

DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. **ArXiv Preprint ArXiv:1810.04805**, 2018. Citado nas páginas 56, 58 e 116.

DEY, R.; SALEM, F. M. Gate-variants of gated recurrent unit (gru) neural networks. In: **International Midwest Symposium on Circuits and Systems**. [S.l.: s.n.], 2017. p. 1597–1600. Citado na página 45.

DRUCKER, H.; BURGES, C. J.; KAUFMAN, L.; SMOLA, A. J.; VAPNIK, V. Support vector regression machines. In: **Advances in Neural Information Processing Systems**. [S.l.: s.n.], 1997. p. 155–161. Citado nas páginas 39, 40 e 41.

DWIVEDI, V. P.; JOSHI, C. K.; LUU, A. T.; LAURENT, T.; BENGIO, Y.; BRESSON, X. Benchmarking graph neural networks. **Journal of Machine Learning Research**, v. 24, n. 43, p. 1–48, 2023. Citado na página 120.

EL-SAPPAGH, S.; SALEH, H.; SAHAL, R.; ABUHMED, T.; ISLAM, S. R.; ALI, F.; AMER, E. Alzheimer's disease progression detection model based on an early fusion of cost-effective multimodal data. **Future Generation Computer Systems**, Elsevier, v. 115, p. 680–699, 2021. Citado na página 25.

ESLAMIEH, P.; SHAJARI, M.; NICKABADI, A. User2vec: A novel representation for the information of the social networks for stock market prediction using convolutional and recurrent neural networks. **Mathematics**, MDPI, v. 11, n. 13, p. 2950, 2023. Citado nas páginas 73 e 76.

FARIMANI, S. A.; JAHAN, M. V.; FARD, A. M.; HAFFARI, G. Leveraging latent economic concepts and sentiments in the news for market prediction. In: **8th International Conference on Data Science and Advanced Analytics (DSAA)**. [S.l.: s.n.], 2021. p. 1–10. Citado nas páginas 25, 73 e 74.

FARIMANI, S. A.; JAHAN, M. V.; FARD, A. M.; TABBAKH, S. R. K. Investigating the informativeness of technical indicators and news sentiment in financial market price prediction. **Knowledge-Based Systems**, Elsevier, v. 247, p. 108742, 2022. Citado nas páginas 69 e 71.

FILHO, I. J. dos R.; COLETI, J. de C.; MARCACINI, R. M.; REZENDE, S. O. Dataset: Annotated soybean market news articles. **Data in Brief**, Elsevier, p. 110545, 2024. Citado nas páginas 105 e 122.

FILHO, I. J. R.; CORREA, G. B.; FREIRE, G. M.; REZENDE, S. O. Forecasting future corn and soybean prices: an analysis of the use of textual information to enrich time-series. In: **VIII Symposium on Knowledge Discovery, Mining and Learning**. [S.l.: s.n.], 2020. p. 113–120. Citado nas páginas 72, 83 e 86.

FILHO, I. J. R.; GOLO, M. P. S.; MARCACINI, R. M.; REZENDE, S. O. How do financial time series enhance the detection of news significance in market movements? a study using graph neural networks with heterogeneous representations. **Neural Computing and Applications**, Springer, 2024. Citado na página 105.

FILHO, I. J. R.; MARCACINI, R. M.; REZENDE, S. O. Previsão do preço futuro de commodities agrícolas: um estudo para enriquecer séries temporais. In: **Simpósio Brasileiro de Automação Inteligente (SBAI)**. [S.l.: s.n.], 2021. v. 1, n. 1. Citado nas páginas 83 e 91.

_____. On the enrichment of time series with textual data for forecasting agricultural commodity prices. **MethodsX**, Elsevier, v. 9, p. 101758, 2022. Citado nas páginas 69, 83 e 96.

FILHO, I. J. R.; MARTINS, L. H.; PARMEZAN, A. R.; MARCACINI, R. M.; REZENDE, S. O. Sequential short-text classification from multiple textual representations with weak supervision. In: **Brazilian Conference on Intelligent Systems**. [S.l.: s.n.], 2022. p. 165–179. Citado nas páginas 49, 105 e 108.

FU, R.; ZHANG, Z.; LI, L. Using lstm and gru neural network methods for traffic flow prediction. In: **Youth Academic Annual Conference of Chinese Association of Automation**. [S.l.: s.n.], 2016. p. 324–328. Citado na página 46.

FU, T.-c.; CHUNG, F.-l.; LUK, R.; NG, C.-m. Representing financial time series based on data point importance. **Engineering Applications of Artificial Intelligence**, Elsevier, v. 21, n. 2, p. 277–300, 2008. Citado nas páginas 117 e 118.

GERS, F.; CUMMINS, F.; FERNANDEZ, S.; BAYER, J. **Understanding LSTM Networks**. 2015. Disponível em: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/#fnref1>. Citado na página 45.

GIUDICE, P. L.; MUSARELLA, L.; SOFO, G.; URSINO, D. An approach to extracting complex knowledge patterns among concepts belonging to structured, semi-structured and unstructured sources in a data lake. **Information Sciences**, Elsevier, v. 478, p. 606–626, 2019. Citado na página 24.

GONG, Z.; TANG, Y.; LIANG, J. Patchmixer: A patch-mixing architecture for long-term time series forecasting. **arXiv preprint arXiv:2310.00655**, 2023. Citado na página 46.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [S.l.]: MIT Press, 2016. <http://www.deeplearningbook.org>. Citado na página 44.

GREFF, K.; SRIVASTAVA, R. K.; KOUTNÍK, J.; STEUNEBRINK, B. R.; SCHMIDHUBER, J. Lstm: A search space odyssey. **IEEE - Transactions on Neural Networks and Learning Systems**, v. 28, n. 10, p. 2222–2232, 2016. Citado na página 45.

GU, J.; SHUKLA, S.; YE, J.; UDDIN, A.; WANG, G. Deep learning model with sentiment score and weekend effect in stock price prediction. **SN Business & Economics**, Springer, v. 3, n. 7, p. 1–20, 2023. Citado nas páginas 69 e 72.

HAMILTON, J. D. **Time series analysis**. [S.l.]: Princeton university press, 2020. Citado na página 31.

HAMILTON, W.; YING, Z.; LESKOVEC, J. Inductive representation learning on large graphs. In: **Proceedings of the Advances in Neural Information Processing Systems**. Los Angeles, USA: MIT, 2017. p. 1024–1034. Citado na página 121.

HAO, P.-Y.; KUNG, C.-F.; CHANG, C.-Y.; OU, J.-B. Predicting stock price trends based on financial news articles and using a novel twin support vector machine with fuzzy hyperplane. **Applied Soft Computing**, Elsevier, v. 98, p. 106806, 2021. Citado nas páginas 69 e 71.

HASSANI, H.; BENEKI, C.; UNGER, S.; MAZINANI, M. T.; YEGANEGI, M. R. Text mining in big data analytics. **Big Data and Cognitive Computing**, MDPI, v. 4, n. 1, p. 1, 2020. Citado na página 23.

HAYKIN, S. **Redes Neurais: Princípios e Prática**. [S.l.]: Bookman Editora, 2007. Citado nas páginas 41 e 42.

HAYKIN, S. S. *et al.* **Neural Networks and Learning Machines**. [S.l.]: New York: Prentice Hall,, 2009. Citado nas páginas 42 e 43.

HELMSTETTER, S.; PAULHEIM, H. Collecting a large scale dataset for classifying fake news tweets using weak supervision. **Future Internet**, MDPI, v. 13, n. 5, p. 114, 2021. Citado na página 106.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural computation**, MIT press, v. 9, n. 8, p. 1735–1780, 1997. Citado na página 45.

HUANG, K.; LI, X.; LIU, F.; YANG, X.; YU, W. Ml-gat: A multilevel graph attention model for stock prediction. **IEEE Access**, IEEE, v. 10, p. 86408–86422, 2022. Citado nas páginas 73 e 75.

HUANG, S.-C.; PAREEK, A.; SEYYEDI, S.; BANERJEE, I.; LUNGREN, M. P. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. **NPJ digital medicine**, Nature Publishing Group, v. 3, n. 1, p. 1–9, 2020. Citado nas páginas 25, 59 e 60.

HYNDMAN, R. J.; ATHANASOPOULOS, G. **Forecasting: principles and practice**. [S.l.]: OTexts, 2018. Citado na página 61.

HYNDMAN, R. J.; KOEHLER, A. B. Another look at measures of forecast accuracy. **International Journal of Forecasting**, Elsevier, v. 22, n. 4, p. 679–688, 2006. Citado na página 61.

ISHII, R. P.; RIOS, R. A.; MELLO, R. F. Classification of time series generation processes using experimental tools: a survey and proposal of an automatic and systematic approach. **International Journal of Computational Science and Engineering**, Inderscience Publishers, v. 6, n. 4, p. 217–237, 2011. Citado na página 32.

ISLAM, M.; SIVAKUMAR, B. Characterization and prediction of runoff dynamics: a nonlinear dynamical view. **Advances in water resources**, Elsevier, v. 25, n. 2, p. 179–190, 2002. Citado na página 37.

JANEV, V.; PUJIĆ, D.; JELIĆ, M.; VIDAL, M.-E. Survey on big data applications. In: **Knowledge Graphs and Big Data Processing**. [S.l.]: Springer, Cham, 2020. p. 149–164. Citado na página 23.

JI, Z.; WU, P.; LING, C.; ZHU, P. Exploring the impact of investor's sentiment tendency in varying input window length for stock price prediction. **Multimedia Tools and Applications**, Springer, p. 1–35, 2023. Citado nas páginas 25, 73 e 76.

KARPATHY, A. **The Unreasonable Effectiveness of Recurrent Neural Networks**. 2015. Disponível em: <http://karpathy.github.io/2015/05/21/rnn-effectiveness/>. Acesso em: 03 de abril 2021. Citado na página 44.

KENDALL, M. Review of box and jenkins (1970). **Journal of the Royal Statistical Society**, v. 134, p. 450–453, 1971. Citado na página 24.

KHALIL, F.; PIPA, G. Is deep-learning and natural language processing transcending the financial forecasting? investigation through lens of news analytic process. **Computational Economics**, Springer, v. 60, n. 1, p. 147–171, 2022. Citado nas páginas 69, 71 e 116.

KIPF, T. N.; WELLING, M. Semi-supervised classification with graph convolutional networks. In: **International Conference on Learning Representations**. [S.l.: s.n.], 2017. Citado nas páginas 120 e 121.

KITCHENHAM, B.; CHARTERS, S. Guidelines for performing systematic literature reviews in software engineering. Technical report, ver. 2.3, 2007. Citado na página 63.

LE, Q.; MIKOLOV, T. Distributed representations of sentences and documents. In: **International Conference on Machine Learning**. [S.l.: s.n.], 2014. p. 1188–1196. Citado nas páginas 52 e 53.

LI, X.; SHANG, W.; WANG, S. Text-based crude oil price forecasting: A deep learning approach. **International Journal of Forecasting**, Elsevier, v. 35, n. 4, p. 1548–1560, 2019. Citado nas páginas 69 e 70.

LI, X.; WANG, J.; TAN, J.; JI, S.; JIA, H. A graph neural network-based stock forecasting method utilizing multi-source heterogeneous data fusion. **Multimedia Tools and Applications**, Springer, v. 81, n. 30, p. 43753–43775, 2022. Citado nas páginas 73 e 75.

LI, X.; WU, P.; WANG, W. Incorporating stock prices and news sentiments for stock market prediction: A case of hong kong. **Information Processing & Management**, Elsevier, v. 57, n. 5, p. 102212, 2020. Citado nas páginas 25, 69, 70, 79 e 116.

LI, Y.; FU, K.; ZHAO, Y.; YANG, C. How to make machine select stocks like fund managers? use scoring and screening model. **Expert Systems with Applications**, Elsevier, v. 196, p. 116629, 2022. Citado nas páginas 69 e 71.

LIANG, W.; XIAO, L.; ZHANG, K.; TANG, M.; HE, D.; LI, K.-C. Data fusion approach for collaborative anomaly intrusion detection in blockchain-based systems. **Internet of Things Journal**, IEEE, 2021. Citado nas páginas 25 e 59.

LIM, B.; ARIK, S. Ö.; LOEFF, N.; PFISTER, T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. **International Journal of Forecasting**, Elsevier, v. 37, n. 4, p. 1748–1764, 2021. Citado nas páginas 39 e 46.

LIM, B.; ZOHREN, S. Time-series forecasting with deep learning: a survey. **Philosophical Transactions of the Royal Society A**, The Royal Society Publishing, v. 379, n. 2194, p. 20200209, 2021. Citado na página 24.

LIM, T.-Y.; ANSARI, A.; MAJOR, B.; FONTIJNE, D.; HAMILTON, M.; GOWAIKAR, R.; SUBRAMANIAN, S. Radar and camera early fusion for vehicle detection in advanced driver assistance systems. In: **Machine Learning for Autonomous Driving Workshop at the 33rd Conference on Neural Information Processing Systems**. [S.l.: s.n.], 2019. v. 2, p. 7. Citado na página 25.

LIN, C.-T.; WANG, Y.-K.; HUANG, P.-L.; SHI, Y.; CHANG, Y.-C. Spatial-temporal attention-based convolutional network with text and numerical information for stock price prediction. **Neural Computing and Applications**, Springer, v. 34, n. 17, p. 14387–14395, 2022. Citado nas páginas 73 e 75.

LISON, P.; HUBIN, A.; BARNES, J.; TOUILEB, S. Named entity recognition without labelled data: A weak supervision approach. **arXiv preprint arXiv:2004.14723**, 2020. Citado na página 106.

LIU, P.; ZHANG, Y.; BAO, F.; YAO, X.; ZHANG, C. Multi-type data fusion framework based on deep reinforcement learning for algorithmic trading. **Applied Intelligence**, Springer, v. 53, n. 2, p. 1683–1706, 2023. Citado nas páginas 73 e 76.

LIU, Z.; ZHU, Z.; GAO, J.; XU, C. Forecast methods for time series data: a survey. **IEEE Access**, IEEE, v. 9, p. 91896–91912, 2021. Citado na página 24.

LUCY, L.; GAUTHIER, J. Are distributional representations ready for the real world? evaluating word vectors for grounded perceptual meaning. **ArXiv Preprint ArXiv:1705.11168**, 2017. Citado na página 53.

MA, Y.; MAO, R.; LIN, Q.; WU, P.; CAMBRIA, E. Multi-source aggregated classification for stock price movement prediction. **Information Fusion**, Elsevier, v. 91, p. 515–528, 2023. Citado nas páginas 73, 76 e 79.

MAATEN, L. Van der; HINTON, G. Visualizing data using t-sne. **Journal of machine learning research**, v. 9, n. 11, p. 2579–2605, 2008. Citado na página 127.

MAN, X.; LUO, T.; LIN, J. Financial sentiment analysis (fsa): A survey. In: **International Conference on Industrial Cyber Physical Systems (ICPS)**. [S.l.: s.n.], 2019. p. 617–622. Citado na página 115.

MARTINS, L. H. D.; TRINDADE, R. N.; CORREA, G. N.; CARVALHO-HEITOR, C. C.; FILHO, I. J. dos R. Avaliação do td-bert com diferentes modelos de representação textual para tarefas de classificiação de textos. **Revista de Tecnologias (RETEC)**, v. 16, n. 1, p. 40–52, 2023. Citado na página 115.

MCCANN, B.; BRADBURY, J.; XIONG, C.; SOCHER, R. Learned in translation: Contextualized word vectors. In: **Advances in Neural Information Processing Systems**. [S.l.: s.n.], 2017. p. 6294–6305. Citado na página 53.

MCCLELLAND, J. L.; RUMELHART, D. E.; GROUP, P. R. *et al.* **Parallel Distributed Processing**. [S.l.]: MIT press Cambridge, MA, 1986. v. 2. Citado na página 43.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient estimation of word representations in vector space. **ArXiv Preprint ArXiv:1301.3781**, 2013. Citado nas páginas 51, 52 e 53.

MINSKY, M.; PAPERT, S. An introduction to computational geometry. **Cambridge Tiass**, 1969. Citado na página 42.

MONDAL, P.; SHIT, L.; GOSWAMI, S. Study of effectiveness of time series modeling (arima) in forecasting stock prices. **International Journal of Computer Science, Engineering and Applications**, Citeseer, v. 4, n. 2, p. 13, 2014. Citado nas páginas 24 e 36.

MONTGOMERY, D. C.; JENNINGS, C. L.; KULAHCI, M. **Introduction to time series analysis and forecasting**. [S.l.]: John Wiley & Sons, 2015. Citado nas páginas 30, 31 e 33.

NIE, Y.; NGUYEN, N. H.; SINTHONG, P.; KALAGNANAM, J. A time series is worth 64 words: Long-term forecasting with transformers. **arXiv preprint arXiv:2211.14730**, 2022. Citado na página 46.

NTI, I. K.; ADEKOYA, A. F.; WEYORI, B. A. A novel multi-source information-fusion predictive framework based on deep neural networks for accuracy enhancement in stock market prediction. **Journal of Big data**, SpringerOpen, v. 8, n. 1, p. 1–28, 2021. Citado nas páginas 25, 69 e 70.

OLAH, C. **Understanding LSTM Networks**. 2021. Disponível em: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>. Acesso em: 01 de abril 2021. Citado na página 45.

PARMEZAN, A. R. S.; SOUZA, V. M.; BATISTA, G. E. Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model. **Information Sciences**, Elsevier, v. 484, p. 302–337, 2019. Citado nas páginas 30, 32, 36, 42 e 43.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: **Conference on empirical methods in natural language processing (EMNLP)**. [S.l.: s.n.], 2014. p. 1532–1543. Citado na página 53.

PETERS, M. E.; NEUMANN, M.; IYYER, M.; GARDNER, M.; CLARK, C.; LEE, K.; ZETTLEMOYER, L. Deep contextualized word representations. In: **Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**. [S.l.]: Association for Computational Linguistics, 2018. v. 1, p. 2227–2237. Citado na página 54.

PICASSO, A.; MERELLO, S.; MA, Y.; ONETO, L.; CAMBRIA, E. Technical analysis and sentiment embeddings for market trend prediction. **Expert Systems with Applications**, Elsevier, v. 135, p. 60–70, 2019. Citado nas páginas 25, 69, 70 e 79.

RATNER, A.; BACH, S. H.; EHRENBERG, H.; FRIES, J.; WU, S.; RÉ, C. Snorkel: Rapid training data creation with weak supervision. In: NIH PUBLIC ACCESS. **International Conference on Very Large Data Bases**. [S.l.], 2017. v. 11, p. 269. Citado na página 106.

REZENDE, S. O. **Sistemas inteligentes: fundamentos e aplicações**. [S.l.]: Editora Manole Ltda, 2003. Citado na página 47.

RODRIGUES, F.; MARKOU, I.; PEREIRA, F. C. Combining time-series and textual data for taxi demand prediction in event areas: A deep learning approach. **Information Fusion**, Elsevier, v. 49, p. 120–129, 2019. Citado na página 25.

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. **Psychological review**, American Psychological Association, v. 65, n. 6, p. 386, 1958. Citado na página 41.

ROSSI, R. G. **Classificaçao automática de textos por meio de aprendizado de máquina baseado em redes**. Tese (Doutorado) — Universidade de São Paulo, 2016. Citado na página 50.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. **Nature**, Nature Publishing Group UK London, v. 323, n. 6088, p. 533–536, 1986. Citado na página 53.

SALEHINEJAD, H.; SANKAR, S.; BARFETT, J.; COLAK, E.; VALAEE, S. Recent advances in recurrent neural networks. **arXiv preprint arXiv:1801.01078**, 2017. Citado na página 44.

SANTOS, B. N. d.; MARCACINI, R. M.; REZENDE, S. O. Multi-domain aspect extraction using bidirectional encoder representations from transformers. **IEEE Access**, IEEE, v. 9, p. 91604–91613, 2021. Citado na página 24.

SAWHNEY, R.; WADHWA, A.; MANGAL, A.; MITTAL, V.; AGARWAL, S.; SHAH, R. R. Modeling financial uncertainty with multivariate temporal entropy-based curriculums. In: **UAI**. [S.l.: s.n.], 2021. p. 1671–1681. Citado nas páginas 69 e 71.

SEZER, O. B.; GUDELEK, M. U.; OZBAYOGLU, A. M. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. **Applied Soft Computing**, Elsevier, v. 90, p. 106181, 2020. Citado nas páginas 24 e 46.

____. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. **Applied Soft Computing**, Elsevier, v. 90, p. 106181, 2020. Citado na página 116.

SHU, K.; ZHENG, G.; LI, Y.; MUKHERJEE, S.; AWADALLAH, A. H.; RUSTON, S.; LIU, H. Leveraging multi-source weak social supervision for early detection of fake news. **arXiv preprint arXiv:2004.01732**, 2020. Citado na página 106.

SINOARA, R. A.; CAMACHO-COLLADOS, J.; ROSSI, R. G.; NAVIGLI, R.; REZENDE, S. O. Knowledge-enhanced document embeddings for text classification. **Knowledge-Based Systems**, v. 163, p. 955 – 971, 2019. ISSN 0950-7051. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0950705118305124>. Citado nas páginas 46, 47, 49 e 50.

SMOLA, A. J.; SCHÖLKOPF, B. A tutorial on support vector regression. **Statistics and Computing**, Springer, v. 14, n. 3, p. 199–222, 2004. Citado na página 40.

SORJAMAA, A.; HAO, J.; REYHANI, N.; JI, Y.; LENDASSE, A. Methodology for long-term prediction of time series. **Neurocomputing**, Elsevier, v. 70, n. 16-18, p. 2861–2869, 2007. Citado nas páginas 34 e 35.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: pretrained bert models for brazilian portuguese. In: **9th Brazilian Conference Intelligent Systems (BRACIS)**. [S.l.: s.n.], 2020. p. 403–417. Citado na página 110.

TADPHALE, A.; SARASWAT, H.; SONAWANE, O.; DESHMUKH, P. Impact of news sentiment on foreign exchange rate prediction. In: IEEE. **3rd International Conference on Intelligent Technologies (CONIT)**. [S.l.], 2023. p. 1–8. Citado nas páginas 69 e 72.

TAIEB, S. B.; BONTEMPI, G.; ATIYA, A. F.; SORJAMAA, A. A review and comparison of strategies for multi-step ahead time series forecasting based on the nn5 forecasting competition. **Expert systems with applications**, Elsevier, v. 39, n. 8, p. 7067–7083, 2012. Citado na página 34.

TAIEB, S. B.; HYNDMAN, R. Boosting multi-step autoregressive forecasts. In: PMLR. **International Conference on Machine Learning**. [S.l.], 2014. p. 109–117. Citado nas páginas 35 e 36.

TAN, J.; DEVECI, M.; LI, J.; ZHONG, K. Asset pricing via fused deep learning with visual clues. **Information Fusion**, Elsevier, v. 102, p. 102049, 2024. Citado nas páginas 73 e 77.

TANG, J.; LIAO, R. Graph neural networks for node classification. **Graph Neural Networks: Foundations, Frontiers, and Applications**, Springer, p. 41–61, 2022. Citado na página 120.

TANG, Y.; SONG, Z.; ZHU, Y.; YUAN, H.; HOU, M.; JI, J.; TANG, C.; LI, J. A survey on machine learning models for financial time series forecasting. **Neurocomputing**, Elsevier, v. 512, p. 363–380, 2022. Citado na página 24.

TIMMERMANN, A.; GRANGER, C. W. Efficient market hypothesis and forecasting. **International Journal of forecasting**, Elsevier, v. 20, n. 1, p. 15–27, 2004. Citado na página 118.

TRINDADE, R. N.; MARTINS, L. H.; CORREA, G. N.; FILHO, I. J. dos R. Using a labeling function for automatic classification of agribusiness news: A weak supervisory approach. In: SBC. **XIX Encontro Nacional de Inteligência Artificial e Computacional**. [S.l.], 2022. p. 73–82. Citado nas páginas 105 e 115.

TSAY, R. S.; CHEN, R. **Nonlinear time series analysis**. [S.l.]: John Wiley & Sons, 2018. v. 891. Citado na página 37.

TSITSULIN, A.; PALOWITCH, J.; PEROZZI, B.; MÜLLER, E. Graph clustering with graph neural networks. **Journal of Machine Learning Research**, v. 24, n. 127, p. 1–21, 2023. Citado na página 120.

USMANI, S.; SHAMSI, J. A. Lstm based stock prediction using weighted and categorized financial news. **Plos one**, Public Library of Science San Francisco, CA USA, v. 18, n. 3, p. e0282234, 2023. Citado nas páginas 73 e 76.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. Attention is all you need. **Advances in neural information processing systems**, v. 30, 2017. Citado nas páginas 54, 55, 56 e 57.

VENTER, M.; STRYDOM, D.; GROVÉ, B. Stochastic efficiency analysis of alternative basic grain marketing strategies. **Agrekon**, Taylor & Francis, v. 52, n. sup1, p. 46–63, 2013. Citado na página 24.

VERMA, S.; SAHU, S. P.; SAHU, T. P. Stock market forecasting with different input indicators using machine learning and deep learning techniques: A review. **Engineering Letters**, v. 31, n. 1, 2023. Citado nas páginas 25 e 39.

WANG, H.-C.; HSIAO, W.-C.; LIOU, R.-S. Integrating technical indicators, chip factors and stock news for enhanced stock price predictions: A multi-kernel approach. **Asia Pacific Management Review**, Elsevier, 2023. Citado nas páginas 69 e 72.

WANG, J.; LI, X.; JIA, H.; PENG, T. A graph-based approach to multi-source heterogeneous information fusion in stock market. **Plos one**, Public Library of Science San Francisco, CA USA, v. 17, n. 8, p. e0272083, 2022. Citado nas páginas 73 e 74.

WANG, Y.; YANG, W.; MA, F.; XU, J.; ZHONG, B.; DENG, Q.; GAO, J. Weak supervision for fake news detection via reinforcement learning. In: **Conference on Artificial Intelligence (AAAI)**. [S.l.: s.n.], 2020. v. 34, p. 516–523. Citado na página 106.

WANG, Z.; HU, Z.; LI, F.; HO, S.-B.; CAMBRIA, E. Learning-based stock trending prediction by incorporating technical indicators and social media sentiment. **Cognitive Computation**, Springer, v. 15, n. 3, p. 1092–1102, 2023. Citado na página 68.

WEN, Q.; ZHOU, T.; ZHANG, C.; CHEN, W.; MA, Z.; YAN, J.; SUN, L. Transformers in time series: A survey. **arXiv preprint arXiv:2202.07125**, 2022. Citado na página 24.

WINDSOR, E.; CAO, W. Improving exchange rate forecasting via a new deep multimodal fusion model. **Applied Intelligence**, Springer, v. 52, n. 14, p. 16701–16717, 2022. Citado nas páginas 25, 73 e 74.

WU, S.-F.; LEE, S.-J. Employing local modeling in machine learning based methods for time-series prediction. **Expert Systems with Applications**, Elsevier, v. 42, n. 1, p. 341–354, 2015. Citado na página 37.

WU, Z.; PAN, S.; CHEN, F.; LONG, G.; ZHANG, C.; PHILIP, S. Y. A comprehensive survey on graph neural networks. **Transactions on neural networks and learning systems**, IEEE, v. 32, n. 1, p. 4–24, 2020. Citado na página 120.

XU, H.; CHAI, L.; LUO, Z.; LI, S. Stock movement predictive network via incorporative attention mechanisms based on tweet and historical prices. **Neurocomputing**, Elsevier, v. 418, p. 326–339, 2020. Citado nas páginas 24, 25, 73 e 74.

XU, K.; HU, W.; LESKOVEC, J.; JEGELKA, S. How powerful are graph neural networks? In: **International Conference on Learning Representations**. New Orleans, USA: Open Review, 2019. Citado nas páginas 120 e 121.

YANG, X.; LOUA, M. A.; WU, M.; HUANG, L.; GAO, Q. Multi-granularity stock prediction with sequential three-way decisions. **Information Sciences**, Elsevier, v. 621, p. 524–544, 2023. Citado nas páginas 73 e 76.

YANG, Z.; DAI, Z.; YANG, Y.; CARBONELL, J.; SALAKHUTDINOV, R. R.; LE, Q. V. Xlnet: Generalized autoregressive pretraining for language understanding. **Advances in Neural Information Processing Systems**, v. 32, p. 5753–5763, 2019. Citado na página 117.

YAO, L.; MAO, C.; LUO, Y. Graph convolutional networks for text classification. In: **Conference on Artificial Intelligence**. Honolulu, Hawaii, USA: Association for the Advancement of Artifcial Intelligence, 2019. Citado na página 121.

YE, Z.; WU, Y.; CHEN, H.; PAN, Y.; JIANG, Q. A stacking ensemble deep learning model for bitcoin price prediction using twitter comments on bitcoin. **Mathematics**, MDPI, v. 10, n. 8, p. 1307, 2022. Citado nas páginas 69, 71 e 116.

YEGNANARAYANA, B. **Artificial neural networks**. [S.l.]: PHI Learning Pvt. Ltd., 2009. Citado na página 39.

YI, J.; CHEN, J.; ZHOU, M.; HOU, C.; CHEN, A.; ZHOU, G. Analysis of stock market public opinion based on web crawler and deep learning technologies including 1dcnn and lstm. **Arabian Journal for Science and Engineering**, Springer, v. 48, n. 8, p. 9941–9962, 2023. Citado nas páginas 73 e 75.

ZENG, A.; CHEN, M.; ZHANG, L.; XU, Q. Are transformers effective for time series forecasting? In: **Proceedings of the AAAI conference on artificial intelligence**. [S.l.: s.n.], 2023. v. 37, n. 9, p. 11121–11128. Citado na página 46.

ZHANG, Q.; QIN, C.; ZHANG, Y.; BAO, F.; ZHANG, C.; LIU, P. Transformer-based attention network for stock movement prediction. **Expert Systems with Applications**, Elsevier, v. 202, p. 117239, 2022. Citado nas páginas 73 e 74.

ZHANG, Q.; YANG, L.; ZHOU, F. Attention enhanced long short-term memory network with multi-source heterogeneous information fusion: An application to bgi genomics. **Information Sciences**, Elsevier, v. 553, p. 305–330, 2021. Citado nas páginas 69 e 70.

ZHANG, Q.; ZHANG, Y.; BAO, F.; LIU, Y.; ZHANG, C.; LIU, P. Incorporating stock prices and text for stock movement prediction based on information fusion. **Engineering Applications of Artificial Intelligence**, Elsevier, v. 127, p. 107377, 2024. Citado nas páginas 73 e 77.

ZHANG, X.; LI, Y.; WANG, S.; FANG, B.; YU, P. S. Enhancing stock market prediction with extended coupled hidden markov model over multi-sourced data. **Knowledge and Information Systems**, Springer, v. 61, p. 1071–1090, 2019. Citado nas páginas 69 e 70.

ZHANG, X.; ZHANG, Y.; WANG, S.; YAO, Y.; FANG, B.; PHILIP, S. Y. Improving stock market prediction via heterogeneous information fusion. **Knowledge-Based Systems**, Elsevier, v. 143, p. 236–247, 2018. Citado nas páginas 69, 70 e 79.

ZHONG, S.; HITCHCOCK, D. B. S&p 500 stock price prediction using technical, fundamental and text data. **arXiv preprint arXiv:2108.10826**, 2021. Citado nas páginas 24, 25, 59 e 116.

ZHOU, Z.; GAO, M.; LIU, Q.; XIAO, H. Forecasting stock price movements with multiple data sources: Evidence from stock market in china. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 542, p. 123389, 2020. Citado nas páginas 25, 69, 70 e 116.

ZHOU, Z.-H. A brief introduction to weakly supervised learning. **National Science Review**, Oxford University Press, v. 5, n. 1, p. 44–53, 2018. Citado na página 106.

ZOU, H.; XIA, G.; YANG, F.; WANG, H. An investigation and comparison of artificial neural network and time series models for chinese food grain price forecasting. **Neurocomputing**, Elsevier, v. 70, n. 16-18, p. 2913–2923, 2007. Citado na página 24.


ICMC
USP
SÃO CARLOS