Previsão do preço futuro de commodities agrícolas: um estudo para enriquecer séries temporais.

Ivan J. Reis Filho * **Ricardo M. Marcacini** ** **Solange O. Rezende** **

* *Universidade do Estado de Minas Gerais, Frutal, MG, (e-mail: ivan.filho@uemg.br).*
** *Instituto de Ciências Matemática e de Computação, Universidade de São Paulo, São Carlos, SP (e-mail: ricardo.marcacini@icmc.usp.br) (e-mail: solange@icmc.usp.br)*

Abstract: Products derived from corn and soy are consumed on a large scale in the world. Market price fluctuations have far-reaching effects on grain consumers, criteria and indices. Thus, the forecast of future grain and grain prices has attracted the attention of investments and agribusiness companies. Estimation, forecasting models use time series to predict future values. However, external factors can originate the data in time series, such as political events, improvement patterns and the foreign exchange market. This information is not explicit in time series data and can make it difficult to predict variable values. Textual data extracted from news, forums and social media can be a source of knowledge about external factors and potentially useful for weather forecasting models. Some studies present text mining techniques to combine textual data with time series. However, existing representations have limitations, such as the curse of dimensionality and ineffective attributes. In this sense, this work proposes representations of time series enriched with textual information. The results indicate that the methods used can be an alternative to improve the prediction performance in regression tasks.

Resumo: Os produtos derivados do milho e da soja são consumidos em grande escala no mundo. As flutuações dos preços no mercado têm efeitos de longo alcance sobre os consumidores, agricultores e processadores de grãos. Assim, a previsão de preços futuros desses grãos tem atraído significativa atenção dos pesquisadores e empresas do agronegócio. Geralmente, os modelos de previsão usam séries temporais para predizer valores futuros. No entanto, fatores externos podem influenciar os dados em séries temporais, como eventos políticos, crises econômicas e o mercado de câmbio. Essas informações não são explícitas nos dados da série temporal e podem dificultar a previsão dos valores das variáveis. Os dados textuais extraídos de notícias, fóruns e redes sociais podem ser uma fonte de conhecimento sobre fatores externos e potencialmente úteis para modelos de previsão de séries temporais. Alguns estudos apresentam técnicas de mineração de texto para combinar dados textuais com séries temporais. No entanto, as representações existentes apresentam algumas limitações, como a maldição da dimensionalidade e atributos ineficazes. Nesse sentido, este trabalho propõem representações de séries temporais enriquecidas com informações textuais. Os resultados indicam que os métodos utilizados podem ser uma alternativa para melhorar o desempenho de previsão em tarefas de regressão.

Keywords: agribusiness commodities; prices forecasting; machine learning; time series.

Palavras-chaves: Commodities do Agronegócio; Previsão de Preços; Aprendizado de Máquina; Séries Temporais.

1. INTRODUÇÃO

O agronegócio brasileiro é reconhecido como um setor crucial para o crescimento econômico do país. Em 2019, a soma de bens e serviços gerados no agronegócio chegaram a R$ 1,55 trilhão ou 21,4% do PIB brasileiro [1]. Nesse cenário, dados históricos têm sido bastante utilizados no agronegócio para realizar análises, estimativas atuais e projeções futuras de mercado.

As flutuações de preços no mercado de milho e da soja têm efeitos de longo alcance sobre os consumidores, agricultores e processadores de grãos. Compreender as tendências dos preços tornam-se um pré-requisito para que os formuladores de políticas implementem diretrizes de subsídios aos produtos agrícolas e ao consumidor. Nessa perspectiva, muitos trabalhos têm sido propostos para predizer os preços futuros de commodities agrícolas (Ayankoya et al.,

[1] Segundo dados do CEPEA/USP em parceria com a Confederação da Agricultura e Pecuária do Brasil (CNA)

2016; Wang et al., 2017; Zhang et al., 2018; Puchalsky et al., 2018; Jiang et al., 2019). No entanto, realizar a predição em um cenário real de mercado é uma das aplicações mais desafiantes devido a sua natureza complexa, dinâmica e não linear Sezer et al. (2020a).

Os fatores que influenciam as commodities agrícolas incluem diversas variáveis que afetam os preços futuros. Além de questões climáticas, Venter et al. (2013) destacam os principais motivos que afetam o mercado de commodities: i) Dados históricos e recentes do mercado; ii) procura e oferta doméstica; iii) procura e oferta internacional; iv) macroeconomia; e, v) fatores políticos. Os três primeiros fatores estão geralmente contidos em séries temporais. Porém, os últimos são mais complexos e subjetivos, geralmente disponíveis de forma implícita em textos extraídos de notícias, redes sociais e reportagens de diferentes áreas do conhecimento.

Uma série temporal pode ser definida como uma sequência de dados numéricos, registrados sequencialmente no tempo. Modelos de previsão de séries temporais são construídos considerando a hipótese de que valores futuros podem ser estimados a partir de uma função matemática, estabelecida e parametrizada por observações passadas (Mills, 2019). Tradicionalmente, modelos paramétricos e lineares têm sido explorados para previsão de dados em séries temporais (Zou et al., 2007; Adanacioglu et al., 2012; Ahu, 2016). O Modelo Auto-Regressivo Integrado de Médias Móveis (ARIMA) têm sido uma das abordagens paramétricas mais populares para previsão de séries temporais em diferentes domínios de aplicação. Modelos não paramétricos têm sido propostos e avaliados em relação aos modelos paramétricos (Wang et al., 2017, 2019; Li et al., 2020). Redes Neurais Artificiais (RNA), Aprendizado Profundo (do inglês, *Deep Learning*) e Vetores de Suporte de Regressão (do inglês, Support Vector Regression(SVR) são exemplos de modelos não paramétricos que usam dados históricos para aprender uma dependência não linear. No modo geral, métodos de aprendizado de máquina têm mostrado resultados promissores para previsão de séries temporais (Sezer et al., 2020b; Ozb, 2020).

Considerando os recentes avanços dos modelos não paramétricos, estudos existentes na literatura têm focado nos métodos de previsão explorando as tendências e sazonalidades, análises que podem ser capturadas somente das séries temporais (Henrique et al., 2019; Sezer et al., 2020a). Contudo, algumas lacunas de pesquisa têm sido levantadas no sentido de enriquecer séries temporais usando informações externas sobre o domínio do problema (Kumar and Ravi, 2016; Pejić Bach et al., 2019). Por exemplo, no domínio de commodities agrícolas, diferentes fatores externos podem influenciar dados de séries temporais, como eventos políticos, crises econômicas, política macroeconômica do governo, desastres naturais e mercados de câmbio (Crone and Koeppel, 2014). Nesse sentido, informações de textos têm sido usados como conhecimento externo e valioso para complementar séries temporais e indicadores econômicos (Crone and Koeppel, 2014; Chen et al., 2016; Li2, 2019).

Técnicas de mineração de texto foram utilizadas em estudos para selecionar recursos de texto e incorporá-los em séries temporais (Wang et al., 2012; Chen et al., 2016). A ideia geral é extrair uma representação estruturada dos textos e associá-los as séries temporais de preços. A abordagem *Bag-of-Words* (BoW) é o método mais comum para representar texto através de um modelo de espaço vetorial, no qual as palavras são indexadas e ponderadas conforme a ocorrência da palavra no texto (Aggarwal, 2014). A representação BoW é estruturada com base em palavras independentes e não expressam o relacionamento entre as palavras, sintaxe do texto ou semântica Sin (2019). Por exemplo, o texto D1 *"O Brasil exporta 90% da soja para a China "* e o texto D2 *"A China exporta 90% da soja para o Brasil"* tem a mesma representação vetorial no modelo BoW (mesmas palavras como índices e mesma ocorrência dos termos para D1 e D2), mas indicam significados opostos. No entanto, quando considerados recursos semânticos, o texto D3 *"China importa quase toda soja do Brasil"* possui maior similaridade para D1 do que para D2.

Considerando as abordagens apresentadas, a inclusão de dados em séries temporais assume a premissa que dados não estruturados podem ser combinados em séries temporais com o propósito que seja considerado fatores externos em tarefas de previsão. Dessa forma, informações externas representam uma alternativa para enriquecer séries temporais. Neste trabalho, foi utilizado o *Support Vector Regression* (SVR) e *Long Short-Term Memory* (LSTM) para prever três representações: série temporal (TS), série temporal combinada com texto por meio da BoW (TS/Textos) e série temporal enriquecida com *texts embeddings* do *Bidectional Enconder Representations Transformes* (TS/BERT).

Este artigo está organizado da seguinte forma: A seção 2 apresenta uma revisão da literatura sobre modelos de previsão para séries temporais. O método é apresentado na Seção 3, onde é discutida a estratégia de pré-processamento para inclusão de informação textual em séries temporais e modelos de regressão. A seção 4 apresenta uma avaliação experimental em séries temporais e textos sobre agronegócio. A seção 5 discute qualitativamente os resultados, e por fim, na seção 6 a conclusão do estudo.

2. TRABALHOS RELACIONADOS

Os modelos de previsão de série temporal são construídos na suposição de que valores futuros podem ser estimados por uma função matemática aprendida a partir de observações anteriores (Abu and Mostafa, 1996). Modelos não paramétricos têm sido propostos em trabalhos recentes e apresentado melhores desempenhos em relação aos modelos paramétricos (Wang et al., 2017; Wan, 2019; Li2, 2020). Redes Neurais Artificiais e SVR são exemplos de modelos não lineares não paramétricos que usam apenas dados históricos para aprender a dependência estocástica entre o passado e o futuro (Das and Padhy, 2018; Wang and Gao, 2018; Alameer et al., 2019; Wan, 2019; Ngu, 2020). No entanto, a maioria dos estudos geralmente aprende modelos de previsão que exploram apenas tendências, resíduos e comportamento sazonal da série histórica, sem considerar as outras informações externas valiosas sobre o domínio do problema que não são explícitas nas séries temporais (Henrique et al., 2019; Sezer et al., 2020a).

Nos últimos anos, estudos para prever o preço futuro no mercado financeiro têm sido frequentemente apresentados na literatura (Puchalsky et al., 2018; Rib, 2020; Ozb,

2020). Muitos se concentraram em modificar ou propor modelos de previsão para melhorar a precisão das previsões, usando dados quantitativos de séries temporais e diferentes métodos de validação para avaliar o modelo de previsão (Ozb, 2020). Alguns trabalhos usaram dados qualitativos em técnicas de mineração de texto para prever ou ajudar a explicar a tendência e a volatilidade das commodities, domínio financeiro e preços dos alimentos (Kumar and Ravi, 2016) Ren and Wei (2016) Dru (2019) Li2 (2020). No entanto, poucos optaram por combinar dados quantitativos com fatores externos (qualitativos) que afetam as flutuações de séries temporais em uma situação real de mercado.

Devido à variedade de trabalhos relacionados, dividem-se os trabalhos em três categorias (Zheng, 2015): i) métodos baseados apenas em informações técnicas de características de série temporal, ii) métodos baseados apenas em características textuais, e iii) métodos híbridos que combinam recursos textuais e informações técnicas de séries temporais. Estudos que usam informações técnicas de séries temporais ou apenas em características textuais, são frequentemente usados em modelos de previsão. O primeiro é normalmente investigado em tarefas de previsão de preços (Sezer et al., 2020b), enquanto o segundo é amplamente utilizado para análise de sentimento na área econômica Kumar and Ravi (2016). No escopo deste trabalho, estamos interessados em métodos híbridos, combinando séries temporais e recursos textuais para enriquecer modelos de previsão.

Considerando métodos híbridos, Wang et al. (2012) apresentaram um modelo que combina ARIMA e SVR para prever os retornos trimestrais do patrimônio líquido (ROEs) de seis empresas. A abordagem contém três etapas: i) Usa Frequência de Termo-Frequência Inversa de Documento (TF-IDF) para representar dados de texto como um vetor de recursos em um espaço vetorial de alta dimensão. ii) o modelo ARIMA analisa a parte linear da série temporal; e, iii) um modelo SVR baseado apenas no vetor de recurso textual para modelar a parte não linear. Os dados são organizados como séries temporais e um modelo híbrido de ARIMA e SVR é usado para modelar e prever essa série. Crone and Koeppel (2014) combinam artigos de notícias financeiras com conteúdo de mídia social como um preditor de retornos de séries temporais financeiras da taxa de câmbio do dólar australiano (AUD) - dólar americano (USD). O modelo híbrido examina a capacidade de descrever séries temporais financeiras e indicadores de sentimento usando uma correlação entre eles. Chen et al. (2016) apresentam um modelo que realiza uma integração de redes neurais artificiais e mineração de texto para prever os preços futuros do ouro. Fatores não quantificáveis foram processados pelo software "Clever Craft". O trabalho focaliza em identificar fatores não quantitativos para ajudar a explicar a tendência e a volatilidade dos futuros de ouro.

Picasso et al. (2019) combinaram análise técnica e fundamental para prever tendências de mercado futuro de vinte empresas mais capitalizadas listadas no índice NASDAQ100. Como resultado, o modelo classifica tendências positivas e negativas usando *Random Forrest* (RF), SVM e uma RNA feed-forward. Li et al. (2019) propuseram uma previsão baseada em texto dos preços do petróleo bruto usando uma abordagem de aprendizado profundo. Em síntese, o estudo propôs uma síntese com base em na abordagem de sentimento do tópico para construir séries temporais estruturados no modelo de Redes Neurais Convolucionais, análise de sentimento e identificação do tópico. Rodrigues et al. (2019) combinaram séries temporais e dados textuais para previsão de demanda de táxi em áreas de eventos. O estudo usou *embeddings* GloVe com abordagens denominadas de aprendizagem profunda e totalmente conectada (DL-FC) e memória de longo-curto prazo (DL-LSTM) para modelar e codificar os dados da série temporal.

Os trabalhos apresentados exploram informações técnicas de domínio para combinar ou analisar observações de séries temporais. Os estudos diferem na avaliação do conjunto de teste e treinamento, representação vetorial de textos e combinados com séries temporais. Os modelos híbridos apresentam um aumento de desempenho em relação aos modelos de previsão de séries temporais. Nessa perspectiva, este estudo apresenta representações de séries temporais enriquecidas com informações textuais de domínio do agronegócio.

3. MÉTODOS

Nesta seção são apresentados os métodos usados para investigar a combinação de informação textual com séries temporais. A Figura 1 ilustra as etapas realizadas no trabalho.

3.1 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 $s_t \in R^d$ representa uma observação s no tempo s_t com $d > 1$ (ST multivariada). Na etapa de treinamento do modelo de previsão é definida uma subsequência de tamanho r da série temporal S, ou seja, uma subsequência $S_u = (s_u, s_{u+1}, ..., s_{u+r})$ em que u indica o período 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 (etapas à frente) usando a estratégia de janela deslizante. Dessa forma, cada subsequência S_u é associada com uma variável dependente y_{u+h}, gerando um conjunto de treinamento

$$X = \{(S_{u_1}, y_{(u+h)_1}), (S_{u_2}, y_{(u+h)_2}), ..., (S_{u_n}, y_{(u+h)_n})\} \quad (1)$$

em que o X representa todo o conjunto de treinamento. Em seguida, é preciso realizar um alinhamento dos documentos de textos relacionados para cada subsequência da série temporal, isto é, o conjunto de documentos que estão no período S_{u+r} e suas respectivas representações no espaço vetorial, definida na Equação 2 (*Features* de Textos). Nas avaliações experimentais foram considerados a representação vetorial BoW e *Texts Embeddings* do BERT. Tarefas para diminuir o problema da dimensionalidade e esparsidade foram realizados para a BoW, como: remoção de *stopwords*, termos específicos do domínio e a radicalização das palavras.

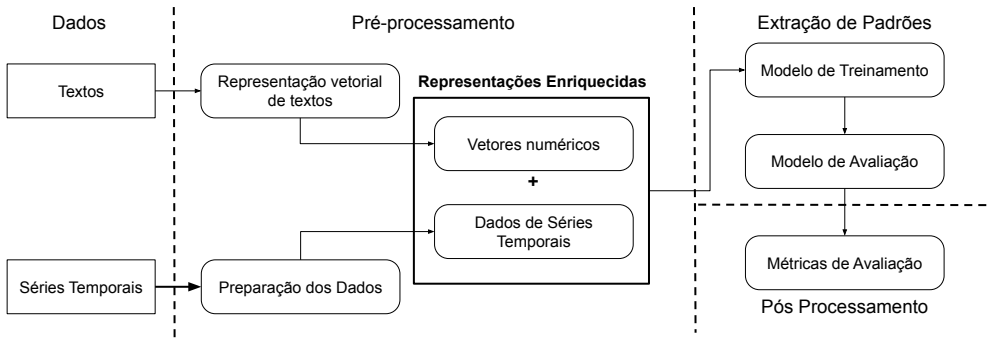

Figura 1. Etapas dos experimentos preliminares. Fonte: Autor.

$$\begin{aligned} FT(u,r) &= Q(T,u,r) \\ &= \{B(d_1), B(d_2), ..., B(d_k)\} \\ &= \{\boldsymbol{v}_{d1}, \boldsymbol{v}_{d2}, ..., \boldsymbol{v}_{dk}\} \end{aligned} \tag{2}$$

em que FT é um subconjunto (Q) de textos no tempo (T), u indica o período para a 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 expresso como um vetor $\boldsymbol{v}_{dk}$. Assim, o vetor numérico associado com a subsequência são expressos como um vetor médio dos vetores de todos os documentos de textos, conforme definido na Equação 3 (*Features* Enriquecidos).

$$FE(u,r) = \sum_{\boldsymbol{v}_d \in FT(u,r)} \frac{\boldsymbol{v}_d}{|FT(u,r)|} \tag{3}$$

A Representação Enriquecida (RE) da subsequência é formada por um vetor concatenado entre os atributos das séries temporais e os *Features* Enriquecidos, $RE(u,r) = S(u,r) \oplus FE(u,r)$, para diferentes tamanhos de r, sendo expresso como

$$\begin{aligned} X = \{&(RE_{u_1}, y_{(u+h)_1}), \\ &(RE_{u_2}, y_{(u+h)_2}), \\ &..., \\ &(RE_{u_n}, y_{(u+h)_n})\}. \end{aligned} \tag{4}$$

Diferentes representações vetoriais de textos foram consideradas como variável $FE(u,r)$ para enriquecer as séries temporais $S(u,r)$.

3.2 Extração de Padrões

Após estruturar as representações de séries temporais enriquecidas, o processo continua para obter modelos de previsão apropriados para as representações propostas.

Support Vector Regression Em 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). O *Support Vector Regression* (SVR) é um modelo não linear derivado do *Support Vector Machine* (SVM) usado para tarefas de classificação. A função não linear usada para estimar um valor da série temporal y_i é demonstrado na Equação 5.

$$y_i = f(RE_i) = \sum_{j=1}^{N} (\alpha_j - \alpha_j^*).K(RE_j, RE_i) + b \tag{5}$$

onde b representa o período, $K(\cdot,\cdot)$ representa a função kernel que transforma os dados em um espaço característico de dimensão superior para permitir a separação linear; α_j e α_j^* são multiplicadores não negativos para cada BR_j observação (também chamadas de variáveis duais); e N o tamanho da subsequência da série temporal enriquecida. A função kernel $K(\cdot,\cdot)$ pode ser escolhida conforme as características do conjunto de dados. Os multiplicadores α_j e α_j^* são estimados no processo de otimização do SVR por meio da Equação 6, que representa a função objetivo a ser minimizada.

$$\begin{aligned} L(\alpha) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K(RE_i, RE_j) + \\ \epsilon \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) - \sum_{i=1}^{N} y_i (\alpha_i - \alpha_i^*) \end{aligned} \tag{6}$$

sujeito a

$$\begin{aligned} &\sum_{i=1}^{N} (\alpha_n - \alpha_n^*) = 0 \\ &\forall n : 0 \leq \alpha_n \leq C \\ &\forall n : 0 \leq \alpha_n^* \leq C \end{aligned} \tag{7}$$

em que 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.

Na avaliação experimental, o modelo SVR foi utilizado com três Kernels, sendo: Polinomial (P), RBF (R) e Sigmoid (S). Uma variedade de parâmetros dos kernels (grau do polinômio e valor de gama) foram estabelecidos e apresentados os resultados que obtiveram os melhores desempenhos.

Long Short-Term Memory Devido à estrutura complexa da LSTM, em que a informação propaga em diversas camadas dentro dos módulos, essa rede neural se destaca em

um cenário complexo, não linear e com dados sequenciais (Ozb, 2020). Cada módulo do LSTM possui os portões: *forget gate*, *input gate* e *output gate*. O *forget gate* é responsável por definir o que deve ser ignorado das entradas da etapa atual, definido

$$\begin{aligned} \tilde{C}_t &= tanh(W_C[h_{t-1}, RE_t] + b_C) \\ C_t &= f_t * C_{t-1} + i_t * \tilde{C}_t \\ h_t &= o_t * tanh(C_t) \end{aligned} \tag{8}$$

em que os parâmetros W e b são, respectivamente, os pesos e os ajustes por viés que ocorrem na célula da rede neural, e RE são entradas das representações enriquecidas no tempo t. Os portões F_t e I_t são aplicados de modo a combinar $\tilde{c}$ e c_{t-1} para chegar na saída c_t, enquanto o portão O_t é aplicado na saída da rede h_t, expressado na Equação 8. A função sigmoide σ usada nos portões assumem valores entre 0 e 1. Além disso, são os portões que ajudam a rede mapear as características temporais do problema, sendo possível algum valor de c_t seja propagado para várias etapas posteriores.

4. AVALIAÇÃO EXPERIMENTAL

Para realizar os experimentos iniciais, séries temporais do milho e da soja foram consideradas para comparar o desempenho dos modelos de representações enriquecidas. As séries temporais utilizadas nos experimentos foram obtidas da WAOB (do inglês, *World Agricultural Supply and Demand Estimates*) da USDA [2] (do inglês, *United States Department of Agriculture*), disponível na Kaggle [3]. A Tabela 1 descreve o período e tamanho dos *datasets*.

Tabela 1. Visão geral das séries temporais e dados textuais usados na avaliação experimental.

ST	Período	Meses	Atrib.	Textos
Milho	01/2014 a 12/2020	82	70	3671
Soja	01/2014 a 12/2020	82	112	11254

Os atributos representam informações contidas nas séries temporais, como a área plantada, a área colhida, a produção, a produtividade, a importação, a exportação, a oferta, a demanda e outras estimativas de países com produções mais significantes do milho e da soja. Os preços originais foram obtidas da CBOT (do inglês, *Chicago Board of Trade*). Neste trabalho, os preços futuros do milho e soja foi considerado o valor médio mensal. Os dados textuais foram extraídos do site *Soybean and Corn Advisor* [4]. Desde de 2009, o website fornece diariamente notícias (inglês) sobre a produção de soja e milho, relacionadas aos ciclos de crescimento da América do Sul, a situação climática, a infraestrutura e o uso da terra.

4.1 Experimentos e Resultados

Para a avaliação do modelo proposto foi utilizado o Erro Médio Porcentagem Absoluto (MAPE), apresentado na Equação 9.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| * 100 \tag{9}$$

em que o n é o número de amostra de teste, y_i é o valor atual e $\hat{y}_i$ é o valor predito. O resultado MAPE é um percentual que relaciona o valor predito com o valor real.

Este trabalho apresentam duas representações de séries temporais enriquecidas: i) Séries Temporais enriquecidas com BoW (TS/BoW); e, ii) Séries Temporais enriquecidas 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 foram utilizadas para o conjunto de Treino (12, 24, 36, 48 e 60), enquanto que o teste é uma única etapa a frente, ou seja, a previsão do preço no próximo mês. Em relação aos dados textuais, manchetes foram utilizadas em vez de notícias. Em relação à BoW, a representação vetorial de textos foi 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.

Para avaliar o desempenho de previsão das representações foram utilizados os modelos de regressão SVR e LSTM. As Tabelas 2 e 3 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 2. 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%

Analisando os resultados da previsão do preço do milho, o modelo SVR (P) com séries temporais (TS) obteve dois resultados com menor MAPE, o TS/BoW obteve uma e a representação TS/BERT atingiu dois resultados. O melhor desempenho do modelo polinomial foi o TS/BERT

[2] https://www.usda.gov/oce/commodity/wasde
[3] https://www.kaggle.com/ainslie/usda-wasde-monthly-corn-soybean-projections
[4] http://www.soybeansandcorn.com

Tabela 3. 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%

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 melhor resultado, a representação TS/BoW obteve dois e o TS/BERT não alcançou 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, TS/BoW obteve um e o TS/BERT também apenas um. O menor valor de MAPE do modelo LSTM foi no período de 12 meses usando a TS/BoW com valor 4.28%. Observando o 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.

5. DISCUSSÃO

Analisando os resultados das Tabelas 2 e 3 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 2 e 3, 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 2 e 3, 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 2 e 3, 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 2 e 3), 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 4 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 2 representam a data das manchetes da Tabela 4.

As polaridades das manchetes da Tabela 4 apresentam simultaneidade com as variações dos preços do milho e da soja. Como exemplo, na data de 13-06-2016 (rótulo 3), 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.

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

Tabela 4. Notícias do site Soybean & Corn Adivsor.

Rótulo	Data	Manchetes
1	10-07-2014	*Farmers in Argentina have finished harvesting their 2013/14 soybean crop, but they have been very slow sellers of the crop with only about 45% of the crop sold.*
1, 2	10-07-2014	*As corn prices continue to decline, farmers in Mato Grosso are looking for alternative crops to grow for the second crop after their soybeans are harvested.*
3	06-06-2016	*Soybean prices are expected to reach a record high in recent years, the high price of R $ 100 per bag (approximately $ 13.00 per bushel) soon.*
4	13-06-2016	*Corn Priced Decline in Brazil due to Initial Harvest Pressure*
5, 6	26-06-2018	*Grain Trade in Brazil has been Minimal for more than Three Weeks.*
6	05-06-2018	*Analyst Lower their 2017/18 Brazilian Corn Estimates*

Fonte: Soybean & Corn Advisor.

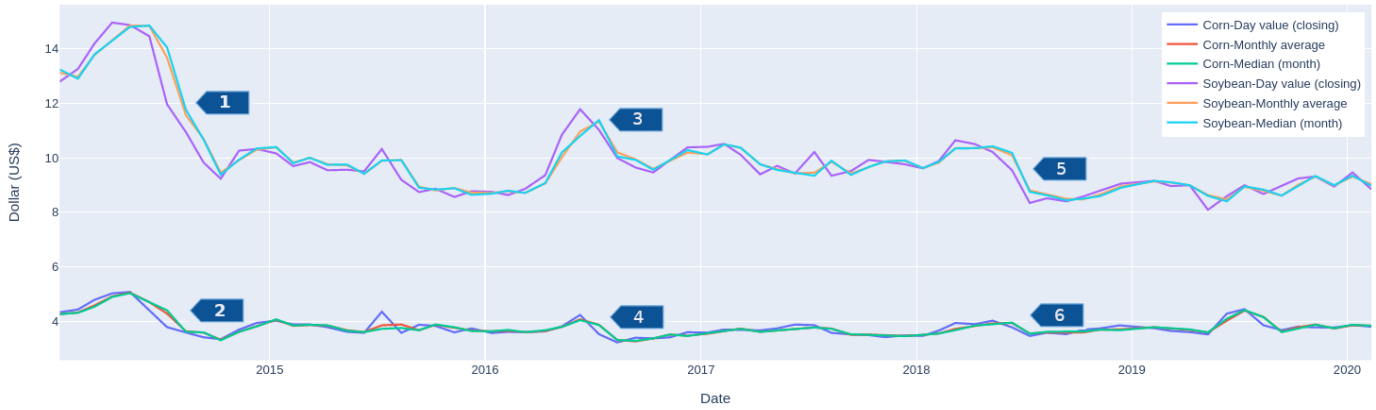

Figura 2. Preço histórico do milho e da soja de 2014 à 2020 cotados no CBOT.

As séries temporais utilizadas nos experimentos possuem uma dependência temporal mensal. Dessa forma, para manter o alinhamento temporal entre séries temporais e o conjunto de notícias, a média mensal foi considerada a variável dependente do modelo preditivo. As representações propostas com os dados disponíveis possuem algumas características que precisam ser ajustadas. Por exemplo, considere um conjunto de notícias do mês de janeiro para prever o valor do preço no mês de fevereiro. Para determinadas situações essa configuração pode ser válida, no entanto, não é praticado em contexto real na tomada de decisão. Nesse sentido, acredita-se que o desempenho de previsão seria melhor se realizada predições de séries temporais intra-dia e não intra-mês.

6. CONCLUSÃO

Este estudo propõe representações de séries temporais enriquecidas com dados extraídos de textos para o aprendizado de modelos de previsão. Padrões ocultos e fatores externos contido em notícias podem corresponder aos movimentos de mercado das commodities. Os modelos híbridos existentes demonstraram um ganho de precisão na previsão de séries temporais. Muitos estudos não consideram fatores externos, como sentimento do mercado, política e outros aspectos. Com intuito de investigar alternativas para superar essas limitações, este trabalho propôs uma análise no modelo de previsão usando duas representações de séries temporais enriquecidas: TS/BoW e TS/BERT. Os resultados indicam que as representações enriquecidas não alcançam ganho considerável de desempenho. Mesmo utilizando modelos preditivos robustos e atributos semânticos, os resultados das representações propostas não apresentaram bom desempenho em relação ao modelo de previsão de ST. Devido a carência de dados textuais relacionados ao domínio, uma limitação do trabalho é a utilização de somente uma fonte de dados textuais, utilizados para enriquecer séries temporais. Trabalhos futuros podem ser realizados para extrair informações de textos e incluir técnicas de mineração de eventos para considerar aspectos semânticos das notícias. Dessa forma, pretende-se extrair mais detalhes dos textos, como entidades nomeadas e relações causais para melhorar as representações de séries temporais enriquecidas.

AGRADECIMENTOS

Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) e Centro de Inteligência Artificial (C4AI) [Processos 2019/25010-5 e 2019/07665-4] e Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [processo nº 426663/2018-7]

REFERÊNCIAS

(2016). Forecasting food prices: The case of corn, soybeans and wheat. *International Journal of Forecasting*, 32(3), 838 – 848.

(2019). Artificial bee colony-based combination approach to forecasting agricultural commodity prices. *International Journal of Forecasting*.

(2019). Knowledge-enhanced document embeddings for text classification. *Knowledge-Based Systems*, 163, 955 – 971.

(2019). A survey of the applications of text mining for agriculture. *Computers and Electronics in Agriculture*, 163, 104864.

(2019). Text-based crude oil price forecasting: A deep learning approach. *International Journal of Forecasting*, 35(4), 1548 – 1560.

(2020). Comparison of forecast models of production of dairy cows combining animal and diet parameters. *Computers and Electronics in Agriculture*, 170, 105258.

(2020). Deep learning for financial applications : A survey. *Applied Soft Computing*, 93, 106384.

(2020). Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series. *Applied Soft Computing*, 86, 105837.

(2020). A novel text-based framework for forecasting agricultural futures using massive online news headlines. *International Journal of Forecasting*.

Abu, Y. and Mostafa, A.A. (1996). Introduction to financial forecasting-applied intelligence.

Adanacioglu, H., Yercan, M., et al. (2012). An analysis of tomato prices at wholesale level in turkey: an application of sarima model. *Custos e Agronegócio Online*, 8(4), 52–75.

Aggarwal, C.C. (2014). *Data Classification: Algorithms and Applications*. Chapman & Hall/CRC, 1 edition.

Alameer, Z., Abd Elaziz, M., Ewees, A.A., Ye, H., and Jianhua, Z. (2019). Forecasting gold price fluctuations using improved multilayer perceptron neural network and whale optimization algorithm. *Resources Policy*, 61, 250–260.

Ayankoya, K., Calitz, A.P., and Greyling, J.H. (2016). Real-time grain commodities price predictions in south africa: a big data and neural networks approach. *Agrekon*, 55(4), 483–508.

Chen, H.H., Chen, M., and Chiu, C.C. (2016). The integration of artificial neural networks and text mining to forecast gold futures prices. *Communications in Statistics - Simulation and Computation*, 45(4), 1213–1225.

Crone, S.F. and Koeppel, C. (2014). Predicting exchange rates with sentiment indicators: An empirical evaluation using text mining and multilayer perceptrons. In *IEEE - Conference on Computational Intelligence for Financial Engineering & Economics*, 114–121.

Das, S.P. and Padhy, S. (2018). A novel hybrid model using teaching–learning-based optimization and a support vector machine for commodity futures index forecasting. *International Journal of Machine Learning and Cybernetics*, 9(1), 97–111.

Drucker, H., Burges, C.J., Kaufman, L., Smola, A.J., and Vapnik, V. (1997). Support vector regression machines. In *Advances in Neural Information Processing Systems*, 155–161.

Henrique, B.M., Sobreiro, V.A., and Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications*, 124, 226–251.

Jiang, F., He, J., and Zeng, Z. (2019). Pigeon-inspired optimization and extreme learning machine via wavelet packet analysis for predicting bulk commodity futures prices. *Science China Information Sciences*, 62(7), 70204.

Kumar, B.S. and Ravi, V. (2016). A survey of the applications of text mining in financial domain. *Knowledge-Based Systems*, 114, 128–147.

Li, J., Li, G., Liu, M., Zhu, X., and Wei, L. (2020). A novel text-based framework for forecasting agricultural futures using massive online news headlines. *International Journal of Forecasting*.

Li, X., Shang, W., and Wang, S. (2019). Text-based crude oil price forecasting: A deep learning approach. *International Journal of Forecasting*, 35(4), 1548–1560.

Mills, T.C. (2019). *Applied Time Series Analysis: A Practical Guide to Modeling and Forecasting*. Academic Press.

Pejić Bach, M., Krstić, Ž., Seljan, S., and Turulja, L. (2019). Text mining for big data analysis in financial sector: A literature review. *Sustainability*, 11(5), 1277.

Picasso, A., Merello, S., Ma, Y., Oneto, L., and Cambria, E. (2019). Technical analysis and sentiment embeddings for market trend prediction. *Expert Systems with Applications*, 135, 60–70.

Puchalsky, W., Ribeiro, G.T., da Veiga, C.P., Freire, R.Z., and dos Santos Coelho, L. (2018). Agribusiness time series forecasting using wavelet neural networks and metaheuristic optimization: An analysis of the soybean sack price and perishable products demand. *International Journal of Production Economics*, 203, 174–189.

Ren, H. and Wei, Z. (2016). Study on sentiment analyzing of internet commodities review based on word2vec. *Computer Science*, 43(S1), 387–389.

Rodrigues, F., Markou, I., and Pereira, F.C. (2019). Combining time-series and textual data for taxi demand prediction in event areas: A deep learning approach. *Information Fusion*, 49, 120–129.

Sezer, O.B., Gudelek, M.U., and Ozbayoglu, A.M. (2020a). Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied Soft Computing*, 90, 106181.

Sezer, O.B., Gudelek, M.U., and Ozbayoglu, A.M. (2020b). Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied Soft Computing*, 90, 106181.

Venter, M., Strydom, D., and Grové, B. (2013). Stochastic efficiency analysis of alternative basic grain marketing strategies. *Agrekon*, 52(sup1), 46–63.

Wang, B., Huang, H., and Wang, X. (2012). A novel text mining approach to financial time series forecasting. *Neurocomputing*, 83, 136–145.

Wang, C. and Gao, Q. (2018). High and low prices prediction of soybean futures with lstm neural network. In *International Conference on Software Engineering and Service Science*, 140–143.

Wang, D., Yue, C., Wei, S., and Lv, J. (2017). Performance analysis of four decomposition-ensemble models for one-day-ahead agricultural commodity futures price forecasting. *Algorithms*, 10(3), 108–116.

Wang, J., Wang, Z., Li, X., and Zhou, H. (2019). Artificial bee colony-based combination approach to forecasting agricultural commodity prices. *International Journal of Forecasting*.

Zhang, D., Zang, G., Li, J., Ma, K., and Liu, H. (2018). Prediction of soybean price in china using qr-rbf neural network model. *Computers and Electronics in Agriculture*, 154, 10–17.

Zheng, Y. (2015). Methodologies for cross-domain data fusion: An overview. *IEEE - Transactions on Big Data*, 1(1), 16–34.

Zou, H., Xia, G., Yang, F., and Wang, H. (2007). An investigation and comparison of artificial neural network and time series models for chinese food grain price forecasting. *Neurocomputing*, 70(16-18), 2913–2923.