Forecasting future corn and soybean prices: an analysis of the use of textual information to enrich time-series.

I. J. Reis Filho[1,2], G. B. Correa[1], G. M. Freire[2], S. O. Rezende[2]

[1] State University of Minas Gerais - Frutal, Brazil
ivan.filho@uemg.br, guilherme.1092985@discente.uemg.br
[2] University of São Paulo - São Carlos, Brazil
guilhermemfreire@usp.br, solange@icmc.usp.br

Abstract. The commodities corn and soybean are products consumed on a large scale in the world. Fluctuations in market prices have far-reaching effects on consumers, farmers, and grain processors. Thus, forecasting the prices of these grains has attracted significant attention from researchers. Forecasting models generally use quantitative time-series data. However, external qualitative factors can influence data in time-series, such as political events, economic crises, and the foreign exchange market. This information is not explicit in the time-series data, and these factors can influence the prediction of the variable values. Textual data extracted from news, forums, and social networks can be a source of knowledge about external factors and potentially useful for time-series forecasting models. Some studies present text mining techniques to combine textual data with time-series. However, the existing representations have some limitations, such as the curse of dimensionality and ineffective attributes. This work applies pre-processing methods in time-series and uses representations combined with textual data to predict the future price of corn and soybeans. The results indicate that the methods used can be an alternative to improve forecasting performance in regression tasks.

CCS Concepts: • **Computing methodologies** → **Supervised learning by regression**.

Keywords: Time-series, Text mining, Forecasting, agricultural commodities

1. INTRODUCTION

Agribusiness companies are continuously affected by the world and the country's economic circumstances. Due to the influence of external factors, such as political crisis, economic globalization, climatic issues, and other uncertain elements, market prices of agricultural commodities have frequently been affected. The large number of variables related to the problem becomes a difficult task for making business decisions. For this matter, researchers and companies have adopted methods to predict indicators to plan and make market decisions in different ways [Cortazar et al. 2019].

Price fluctuations in the corn and soybean market have far-reaching effects on consumers, farmers, and grain processors. Therefore, understanding the price trend becomes a prerequisite for policymakers to implement prices in agricultural product markets. The forecast of the price of these grains is the subject of research in several studies [Wang et al. 2017][Zhang et al. 2018][Puchalsky et al. 2018][Jiang et al. 2019]. Time-series data are commonly applicable for future price predictions in most applications and researches. There are still a variety of testable strategies in applications where some models do not achieve satisfactory results.

Different external qualitative factors can influence time-series data, such as political events, economic crises, government macroeconomic policy, and foreign exchange markets [Crone and Koeppel 2014]. The information sources that implicitly incorporate these qualitative factors are texts extracted from news on web pages and headlines from different areas of knowledge. In this sense, textual data

can be used as an external and valuable knowledge base to extract relevant information that is not available in quantitative [Fung et al. 2003][Crone and Koeppel 2014][Chen et al. 2016]. For this purpose, the Bag-of-Words approach is the most common method to represent text through vectors, in which terms are indexed and weighted according to the occurrence of the word in text [Wang et al. 2012]. The inclusion of textual features into time-series, extracted from text mining techniques such as Bag-of-Words, brings new possibilities to improve performance in forecasting tasks.

This paper introduces the use of external information features for learning forecasting models. The proposed approach explores time-series and text information as a bridge to incorporate qualitative factors into observations. Thus, a composed external technical information achieved an alternative and enriched time-series representation. In this work, we use a classification technique to obtain the attributes with the highest information gain in the time-series and the Support Vector Regression (SVR) to forecasting four representations: time-series (TS), time-series combined with text (TS/Texts), TS with attributes extracted from the Decision Tree (TS-DT), and TS with extracted attributes from decision tree and combined with texts (TS-DT/Texts).

This paper is organized as follows: Section 2 presents a literature review on forecasting models for time-series. The method is presented in Section 3, where the pre-processing strategy for including textual information in time-series and regression models is discussed. Section 4 presents an experimental evaluation in time-series and texts on agribusiness.

2. LITERATURE REVIEW

Time-series analysis methods have investigated the application of traditional statistical approaches, such as the integrated autoregressive moving average (ARIMA) [Darekar and Reddy 2017] and Seasonal (SARIMA) [Adanacioglu et al. 2012]. In recent years, machine learning (ML) approaches, such as the Support Vector Regression [Das and Padhy 2018], the Artificial Neural Network have proposed promising forecasting results in several domains [Baruník and Malinská 2016][Wang and Gao 2018][Puchalsky et al. 2018][Alameer et al. 2019].

Regression models aim to estimate the relationship between the dependent (predictive) and independent (features) variables. The independent variable is observed data and is often a vector X_i (i denotes the data row), dependent variables are the observed data and denoted as scalar y_i. Vector β often denotes unknown parameters, and scalar e_i denotes observed error terms in data. The behavioral relation between y_i and X_i variable must be established by a function $f(X)$. The Equation model $y = f(X)$ is considered simple when it involves a causal relationship between two variables, and multivariate when it involves two or more variables (Equation 1). Some regression models propose that y_i is a function of X_i and β, with e_i representing an error term that may stand in for non-modeled values of y_i or random statistical noise.

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e_i \tag{1}$$

where X_i are samples of the regression model, y_i the predicted value. The challenge of regression models in machine learning is learning a function $f(X)$ that best fits the data and results in the most accurate value of y_i, as shown in Equation 2:

$$Y_i = f(X_i, \beta) + e_i \tag{2}$$

The time-series data or a combination of time-series and text attributes can represent the regression equation independent variable [Drucker et al. 1997]. Recently, some research has investigated improvements in forecasting models when combining texts with time-series. Fung *et al.* (2003) proposes an integration of the text mining approach and time-series for stock prediction, and the results show that it is possible to correlate news and time-series data. Wang *et al.* (2012) proposed the use of textual information to aid the financial time-series forecasting and an approach combining ARIMA and

SVR. Crone and Koeppel (2014) presented an empirical evaluation used text mining and multilayer's perceptrons (MLP) to predict exchange rates (FOREX) combined with sentiment indicators. Chen *et al.* (2016) presents an integration of artificial neural networks and text mining to forecast future gold prices, where the results demonstrated that appropriate training times had higher forecasting accuracy than that of ARIMA models.

Text mining techniques have been used in studies to select text features and incorporate them into time-series [Fung et al. 2003][Wang et al. 2012][Chen et al. 2016]. Literature studies that explore textual features combined with time-series in forecasting models use a well-known textual representation called Bag-of-Words, where terms are indexed and weighted. BoW identifies the main terms in each text, and the occurrence of each word is used as a feature to train training a model. Labeled weights specified by binary measures, TF, or TF-IDF, represents extracted terms from texts [Aggarwal and Zhai 2012]. BoW's weighted values can complement the resources extracted from time-series and be input data for regression models. Thus, this work presented an integrated representation of time-series and derived attributes from texts.

3. METHODS

This section presents the methods used to investigate the combination of text resources with time-series data to improve forecasting models.

3.1 Pre-processing

The Decision tree method was used to obtain the attributes with time-series information gain. Two numerical labels were assigned in each month to represent the variation in the average monthly price of corn and soybeans. Label 0 represents that the price was neutral or below, and label 1 represents the price rising by at least 1% about the previous month. The decision tree was processed ten times, and the attributes discarded in the construction of all models were also discarded in the regression model. Thus, in this work, we considered two sizes for time-series: all the attributes of the time-series (TS) and attributes of the time-series extracted from the decision tree (TS-DT), presented Table I.

Table I. Overview of time-series applyed in experiment evaluation.

time-series	**Period**	**Months**	**All Attributes (TS)**	**Attributes (TS/DT)**	**Texts**
Corn	Jan 2014 to Feb 2020	73	112	44	3671
Soybean	Jan 2014 to Feb 2020	73	70	22	11254

A time-series S of size m is defined as an ordered sequence of observations, i.e., $S = (s_1, s_2, ..., s_m)$, where $s_t \in \mathbb{R}^d$ represents an observation s at time t. Moreover, time-series analysis can be univariate ($d = 1$) or multivariate ($d > 1$). In the learning stage of a forecasting model, to consider a size r subsequences, extracted from time-series S, is defined a subsequence $S_u = (s_u, s_{u+1}, ..., s_{u+r})$, where u indicates the time period of the first observation of the subsequence, with $1 \leq u \leq m - r$. Given a predefined subsequence size r, it is extracted several subsequences from S by using a sliding window strategy. Each subsequence S_u is associated with a forecasting target value y_u (e.g. future observations), thus generating a training set $X = \{(S_{u_1}, y_{u_1}), (S_{u_2}, y_{u_2}), ..., (S_{u_n}, y_{u_n})\}$, with n training instances.

In this work, the u value of the subsequences indicates a date. The date is used to delimit a period of time. For example, given a monthly granularity of a time-series, if u represents the month "Jan 2020" of a subsequence of size $r = 3$, then the time periods involved in the subsequence are $(u, r) = \{$"Jan 2020", "Feb 2020", "Mar 2020"$\}$. This time period is important to compose a time alignment function between the time-series and the textual knowledge base T. Let $Q(T, u, r) = T_{(u,r)} = \{d_1, d_2, ..., d_k\}$ an alignment function that returns the set of k documents $T_{(u,r)}$ given time interval in (u, r), by using the publication date of each document d.

The $d \in T$ documents are represented in a space-vector model. In this work, it's used resources extracted from the bag of words with n-gram = 2, excluding terms (words) with occurrence below 5% and above 95% in the texts. Text mining methods available in [Aggarwal and Zhai 2012] are used, where the function $B(d) = \vec{v}_d = \{w_1, w_2, ..., w_b\}$, with b defining the size of the BoW space. The function $B(d)$ maps text in natural language (contained in d) to a b-dimensional vector representation.

The proposed approach to obtain a new representation for the time-series is presented, which considers the features from BoW, defined as time-series combined with text (TS/Texts), and TS with attributes extracted from decision tree and combined with texts (TS-DT/Texts). So far, a subsequence set represents the time-series data. Given a S_u subsequence of size r, a vector of features (BoW) enriches the subsequence. Firstly, it is identified via time alignment all textual documents related to the subsequence and their respective representations in the BoW, as defined in Equation 3. Then, features representation associated with the subsequence is computed as an average vector from the documents vectors, as described in Equation 4.

$$\begin{aligned} BS(u,r) &= Q(T,u,r) \\ &= \{B(d_1), B(d_2), ..., B(d_k)\} \\ &= \{\vec{v_{d_1}}, \vec{v_{d_2}}, ..., \vec{v_{d_k}}\} \end{aligned} \tag{3}$$

The combined representation of subsequence is formed by a vector concatenation between the features of the subsequence and the features of BoW, $BR(u) = S_u \oplus BF_u$, for different values of r. Thus, training set $X = \{(BR_{u_1}, y_{u_1}), (BR_{u_2}, y_{u_2}), ..., (BR_{u_n}, y_{u_n})\}$ are used in regression models.

$$BF(u,r) = \sum_{\vec{v}_d \in BS(u,r)} \frac{\vec{v}_d}{|BS(u,r)|} \tag{4}$$

3.2 Regression Model

The Non-linear regression model is assumed more appropriate, due to the chaotic nature of the time-series that requires the addition of textual information to reduce uncertainty. In this sense, the Regression Vector Support model obtains promising results in several time-series forecasting works. In Equation 5, the nonlinear SVR [Drucker et al. 1997] forecast function used to estimate a time-series value y_i from the BR_i input is presented:

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

where (b) is the period, $K(,)$ represents the kernel function that transforms the data into a higher dimensional characteristic space to allow linear separation; and α_j and α_j^* are non-negative multipliers for each BR_j observation (also called dual variables). The characteristics of the data set influences in the appropriate chosen kernel function $K(,)$. The most common kernels are Polynomial, RBF, and Sigmoid, as presented in Table II. The SVR optimization process through Equation 6 estimates multipliers α_j and α_j^*, which represents the minimized objective function.

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

subject to

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

Table II. Kernel Functions

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

where K is the kernel, ϵ defines a margin of tolerance where there is no given penalty for forecasting errors; and C is a previously defined positive constant that controls the penalty for observations that exceed the ϵ margin; which also helps to avoid excessive overfitting. The quadratic programming optimization techniques solve the minimization problem of the nonlinear SVR. In this work, minimum sequential optimization (SMO), available in LIBSVM [Drucker et al. 1997], solves SVR related problems.

4. EXPERIMENT EVALUATION

This section presents experiments evaluations using three regression models to compare the performance of the time-series forecasting combined with texts. The experiments used a sliding window method for the learning (training) and validation (test) phases. For matters of assessing model performances and validity, it was used the Mean Absolute Percentage Error (MAPE) statistical indicator.

4.1 Datasets

For the experiments, different forecasting models compared the performance between soybean and corn. The time-series data source used in this experiment is from the World Agricultural Supply and Demand Estimates (WAOB) from the United States Department of Agriculture (USDA), available on the Kaggle[1] website. Table I describes the period and size of the datasets. The attributes represent several time-series features, such as planted area, harvested area, yield, imports, supply, demand, and other estimates from countries with the largest corn and soybean production. From the Chicago Board of Trade (CBOT) are obtained the original price data, available at CME[2] Group's website. The CBOT is a designated contract maker for the CME Group for future exchange for trades of the agricultural commodity contracts, and the prices charged at CBOT are a benchmark in worldwide prices. In this work, corn and soybean future price prediction consider monthly average value.

4.2 Experiments and Results

For the proposed model evaluation, it is used the Mean Absolute Percentage Error (MAPE), presented in Equation 8.

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - x_i}{x_i} \right| * 100 \tag{8}$$

where n is the number of testing samples, *x(i)* is the actual value of each dataset, and *y(i)* is the forecasting value of the corresponding future price data.

Figure 1 illustrates how the method was applied in this work. As presented in Equation 1, the consider X the training window, Y dependent variable (forecasting value), and (n) the number of tests. In the experimental evaluation, six "windows" were used to assess the accuracy of the forecast, where $Y = 1$, $X_i = (2, 6, 12, 24, 36, 48)$ are the set of windows defined for training, and $n = (71, 67, 61, 49, 37, 25)$ the number of tests (forecast), respectively. As the size of the windows were

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

increased in training, the number of tests were decreased. Variable y_i shown in Equation 8 are Y results in each test (Figure 1).

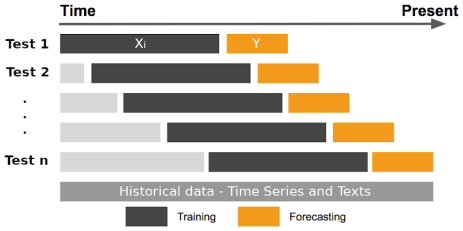

Fig. 1. Sliding window

The experiments analyzed the lowest MAPE value in each window. The representations TS, TS/Texts, TS(DT), and TS(DT)/Texts were applied to three regression models, the polynomial function, RBF, and sigmoid. The Table III presents the results for the corn time-series forecasting. The value in bold represents the lowest MAPE value in each window, and the underline the lowest value obtained in the representation in each regression model. The polynomial regression model using data from the TS obtained two results with less MAPE, the TS(DT) obtained five results, and the TS(DT)/Texts obtained one results, and TS/Texts representation did not achieve satisfactory performance.

Table III. Corn - MAPE Results

Model	Training	TS	TS/Texts	TS(DT)	TS(DT)/Texts
Polynomial	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%
RBF	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%
Sigmoid Function	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%

Table IV. Corn - Best results

Model	TS	TS/Texts	TS(DT)	TS(DT)/Texts
Polynomial	3.74%	3.79%	**3.69%**	3.72%
RBF	3.48%	**3.23%**	3.45%	3.27%
Sigmoid Function	3.99%	**3.98%**	3.99%	3.99%

The best result of the polynomial regression was the TS(DT), with a period of 48 months of training with a lower MAPE of 3.69%. The RBF regression model using data from TS representation did not achieve satisfactory performance, TS/Texts obtained three results as the best result, the TS(DT) received none best result, and TS(DT)/Texts representation got five results as the best result. The period with 48 months of training with TS/Texts reached the lowest MAPE with 3.23%. Sigmoid Function using TS data achieved one result with the lowest MAPE, the TS/Texts obtained two, the TS(DT) got three, and TS(DT)/Texts achieved one result. However, the period with 12 months of

training of representation TS/Texts obtained the lowest MAPE with 3.98%. Table IV presents the results that obtained the lowest MAPE of each regression model for corn time-series.

The experimental results achieved from soybean time-series were similar to the results obtained from corn time-series. The representations TS/Texts and TS(DT)/Texts, received the five lowest MAPE values in the parameter setting of the regression model, as shown in Table V, thus providing evidence of the effectiveness of features from texts for enriching forecasting models. Table VI presents the results that obtained the lowest MAPE of each regression model. In the next subsection, the results collected from this work are discussed and analyzed.

Table V. Soybean - MAPE Results

Model	Training	TS	TS/Texts	TS(DT)	TS(DT)/Texts
Polynomial	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%
RBF	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%**
Sigmoid Function	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%**

Table VI. Soybean - Best Results

Model	TS	TS/Texts	TS(DT)	TS(DT)/Texts
Polynomial	3.90%	4.45%	**3.87%**	4.46%
RBF	4.48%	4.04%	4.49%	**3.79%**
Sigmoid Function	4.50%	4.49%	4.49%	**4.48%**

4.3 Discussion

According to the results of Table III and V related to the polynomial regression model, the representations of TS and TS(DT) obtained the lowest MAPE values. The 48-month training period using the polynomial model and TS(DT), achieved the best forecast accuracy in both experiments, with MAPE 3,69.% for corn and 3.87% for soybean. The results indicate that a greater size training window obtains a large volume of texts, and the text features are more discriminating information than small window size.

In comparison with the results of the RBF regression model of the Tables III and V, the TS/Texts and TS(DT)/Texts representations obtained the lowest MAPE values. The results of corn performed better with bigger windows for TS/Text, and TS(DT)/Texts achieved better forecasting accuracy for a training period of 24 months. TS/Texts for Corn, with 48 months of training, received the best forecast accuracy with MAPE 3.23%, and 3.79 for soybean with 24% months of training.

Analyzing the results of the sigmoid function from Table III and V, the four representations obtained the lowest MAPE value in different training periods. The TS/Texts and TS(DT)/Texts representations achieved lower MAPE in some scenarios. In this experiment, the best performance of the periods of corn and soybeans did not match. For corn, the 12-month training period obtained the lowest MAPE value with 3.99%, while for soybeans, the 2-month period obtained the best performance with 4,48%.

5. CONCLUSIONS

Existing models have demonstrated a gain accuracy in predicting time-series. Many studies do not consider external factors, such as market sentiment, politics, and other aspects. To assess these limitations, this work proposed an analysis in the forecast model using four representations of integrated time-series with attributes extracted from related news reports. The proposed representations were used to build the matrix of value attributes and concatenate with the time-series data. Experimental results showed that representations combined with texts improve the forecasting performance compared to models that consider only time-series. This analysis demonstrates that representations combined with texts offer an alternative to improve the accuracy of price forecasts in regression tasks. Future work can be carried to extract more details from the texts, such as named entities and causal relationships, and improve the textual representation, techniques to consider semantic aspects also can be applied.

ACKNOWLEDGMENTS: the authors would like to thank FAPESP and C4AI for their financial support.

REFERENCES

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

Aggarwal, C. C. and Zhai, C. *Mining text data*. Springer Science & Business Media, 2012.

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

Baruník, J. and Malinská, B. Forecasting the term structure of crude oil futures prices with neural networks. *Applied energy* vol. 164, pp. 366–379, 2016.

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

Cortazar, G., Millard, C., Ortega, H., and Schwartz, E. S. Commodity price forecasts, futures prices, and pricing models. *Management Science* 65 (9): 4141–4155, 2019.

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

Darekar, A. and Reddy, A. Predicting market price of soybean in major india studies through arima model. *Journal of Food Legumes* 30 (2): 73–76, 2017.

Das, S. P. and Padhy, S. 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, 2018.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A. J., and Vapnik, V. Support vector regression machines. In *Advances in neural information processing systems*. pp. 155–161, 1997.

Fung, G. P. C., Yu, J. X., and Lam, W. Stock prediction: Integrating text mining approach using real-time news. In *2003 IEEE International Conference on Computational Intelligence for Financial Engineering, 2003. Proceedings.* IEEE, pp. 395–402, 2003.

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

Puchalsky, W., Ribeiro, G. T., da Veiga, C. P., Freire, R. Z., and dos Santos Coelho, L. 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* vol. 203, pp. 174–189, 2018.

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

Wang, C. and Gao, Q. High and low prices prediction of soybean futures with lstm neural network. In *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*. IEEE, pp. 140–143, 2018.

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

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