Application and Research of Prediction Model Based on ARIMA-LSTM and Multi-Sequence LSTM Algorithm - Taking Wordle as an Example

Zhengze Li*, Hui Zhang

School of Cyber Science and Engineering, Nanjing University of Science & Technology, Jiangyin, Jiangsu, 214443

*Corresponding author: leezz18@163.com

Abstract. The article aims to explore the relationship between the changing trends in the number of Wordle players and the distribution of players' attempts at the game based on word and time attributes. This paper employs wavelet analysis and ARIMA-LSTM model to predict the number of players, and uses canonical correlation analysis and multi-sequence LSTM model to respectively determine the correlation and classification results of word attributes and player attempt distribution. The results indicate that on March 1st, 2023, the number of Wordle players is 16,352. If the solution word on March 1st is 'EERIE', the distribution of player attempts will be [0,8,19,25,23,17,8].

Keywords: ARIMA-LSTM, wavelet denoising, canonical correlation analysis (CCA), Wordle puzzle

1. Introduction

Wordle, a popular puzzle currently offered daily by the New York Times, has taken the world by storm. Players try to solve the puzzle by guessing a five-letter word in six tries or less, receiving feedback with every guess. MCM provided a file of daily results of Wordle for January 7, 2022 through December 31, 2022, including number of reported results, percentage of scores etc. According to this file, we will analyze and predict the trend of number of reported results, explore the relationship between word attributes and percentage of scores, and predict the percentage of scores This article provides a reference for evaluating and improving the operation of related games, and the data prediction and classification methods used in it can also be transferred to other fields.

This paper focuses on the analysis of the prevalence and transmission patterns of Wordle. The existing epidemic spread prediction models are mainly based on two types of methods: 1. Methods based on infectious disease models (SIR, SEIR). For example, Lin et al.(2015) applied the SEIR model with saturation contact rate to study the spread diffusion process of online public opinion, and analyzed the intrinsic mechanism of the spread threshold through specific simulation experiments[1]. Zhang et al.(2021) improved the SEIR model by integrating group size, herding effect and social reinforcement effect, and studied the dynamic evolution process of public opinion dissemination in mobile social networks[2]. 2. Time-series model based on prediction methods (ARIMA, LSTM). For example, Peng et al.(2008) used ARIMA model to predict the onset of hemorrhagic fever in renal syndrome with good fitting and prediction[3]. Sham et al.(2014) applied the ARMA model to predict the trend of HFMD in Malaysia and found that the ARMA(1,4) model fitted the data well with over 90 % predictive power[4]. Ouyang et al.(2022) Constructing an early warning system based on TVP-VAR-LSTM model through the predictive role of macro-state variables such as SSE Composite Index returns and online opinion indices on financial risk[5].

Most of the above research works use a single infectious disease model or traditional time-series models to predict the epidemic spread of information. However, the infectious disease model cannot estimate the parameters such as infection rate and recovery rate accurately, and the traditional time series model cannot capture the nonlinear factors in the time series. Therefore, in this paper, wavelet analysis is firstly used to decompose the number of reported results into high-frequency components and low-frequency components, and ARIMA and LSTM are used to predict the high-frequency components and low-frequency components respectively. And on the basis of typical correlation analysis, the percentage of scores is predicted by using multisequence LSTM for classification.

2. Model I: A hybrid ARIMA-LSTM model based on Wavelet Denoising

This section is aim to analyze and forecast changes in the number of players reporting their results. To start, a visualization of player numbers over time is presented(Figure 1). As shown in the figure, between January 7, 2022 and December 31, 2022, the reported results exhibit an initial upward trend followed by a decline.

The game's influence is being disseminated via Twitter, with a substantial number of individuals who have never played the game joining in out of sheer curiosity - much akin to the propagation of an early-stage epidemic. Nonetheless, as time has elapsed, the game's allure has waned and individuals have developed a resistance to its appeal, resulting in a decrement in the number of active players. Just like humans produce antibodies after recovery from an illness. So barring any unforeseen circumstances, the downward trend in player numbers is to continue.

In summary, this process is similar to the transmission process of infectious diseases. To ensure the accuracy of prediction, we chose to decompose the original time series using wavelet analysis and use a combined LSTM-ARIMA model for prediction. as an alternative to the traditional SIR and SEIR model.

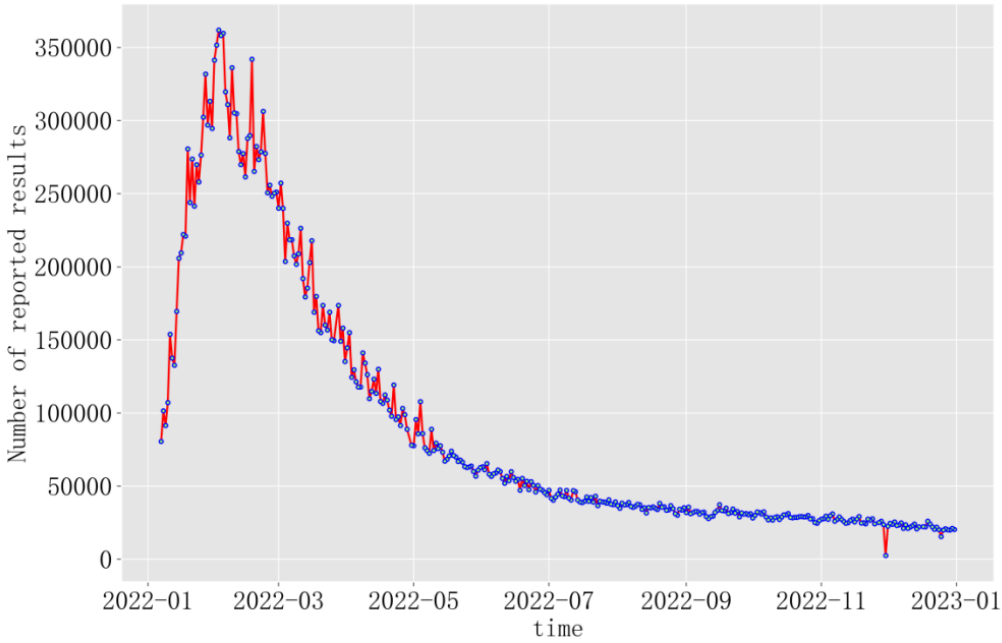

Figure 1. Trend of reported results number

2.1. Wavelet denoising

After collecting and organizing the literature[3],we found that time series forecasting is mostly based on simple ARIMA or neural networks for simple forecasting. However, there are many irregular data perturbations in the sequences we need to predict, so we adopt a method that can effectively solve such irregular perturbations.

For the first step, we divided the dataset into a training set (350 data from January 7, 2022 - December 22, 2023) and a validation set (9 data from December 23, 2022 - December 31, 2022).

In the second step, the training set is preprocessed using wavelet transform. Firstly, DB3 wavelet is used to decompose the original sequence into 4 high frequency bands $a_i(i=1\cdots4)$ and 4 low frequency bands $d_j(j=1\cdots4)$, The high-frequency part contains detailed information about the time series as well as random clutter, while the smoother low-frequency part contains the main variation patterns of the data.

We use the energy ratio as the screening standard to select low-frequency waves. The energy ratio is calculated as follows:

$$ER(k)=\frac{\sum_{i=1}^{350}a_i^2(k)}{\sum_{i=1}^{350}a_i^2(k)+\sum_{i=1}^{350}d_i^2(k)}(k=1\cdots4) \tag{1}$$

In this context, $a_i(k)$ and $d_i(k)$ denote the K-th sample values of the i-th low-frequency and high-frequency sub-bands, respectively. A higher energy ratio implies that the low-frequency sub-band carries a larger proportion of the signal energy. After comparing the available options, we chose the fourth low-frequency wavelet, referred to as a_4. Figure 2 presents the resulting images of the decomposed low-frequency wavelet a_4 and high-frequency wavelet $d_j(j=1\cdots4)$.[6-7]

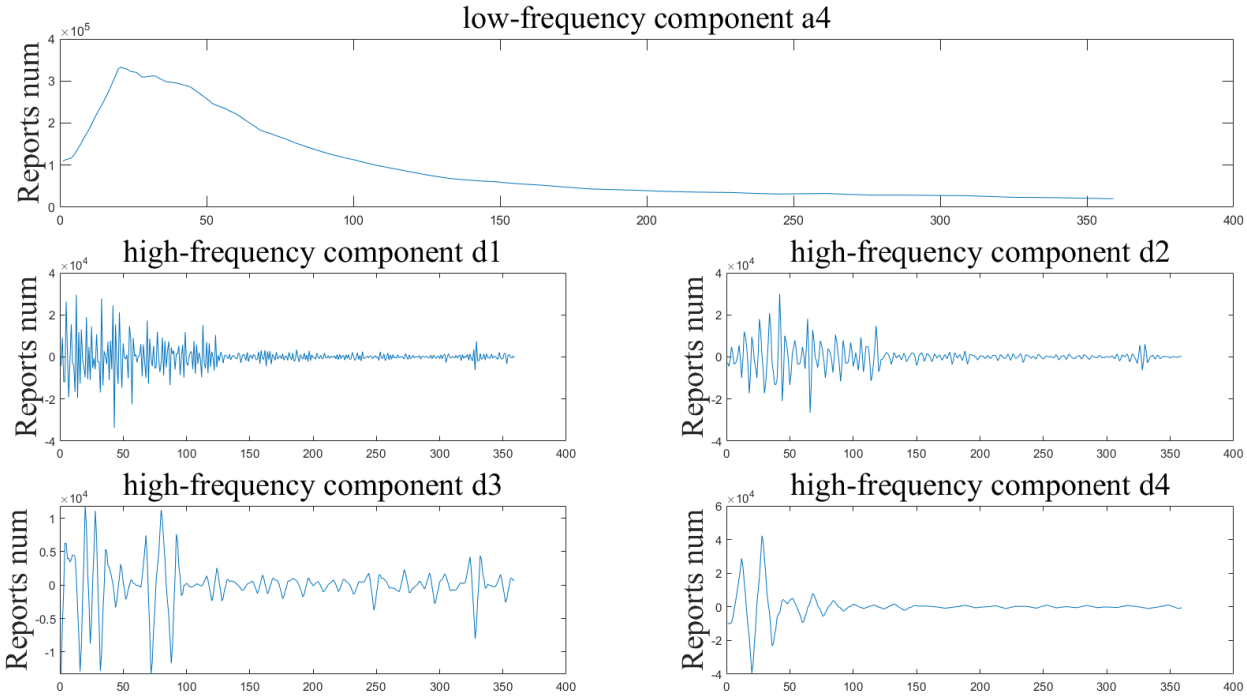

Figure 2. Wavelet decomposition results

2.2. Forecasting Methods

Given that the number of Wordle reports is influenced by a variety of non-linear factors, such as social platforms, online media, social factors etc. Therefore, we need to build a model that contains both linear and nonlinear features.

Traditional time series models can extract linear features, while neural network models have strong mapping properties for nonlinearity, so our study combines the linear time series forecasting algorithm ARIMA and the nonlinear forecasting algorithm LSTM neural network together for the number of reported results forecasting[8]. The specific structure of ARIMA-LSTM is shown in Figure 3.

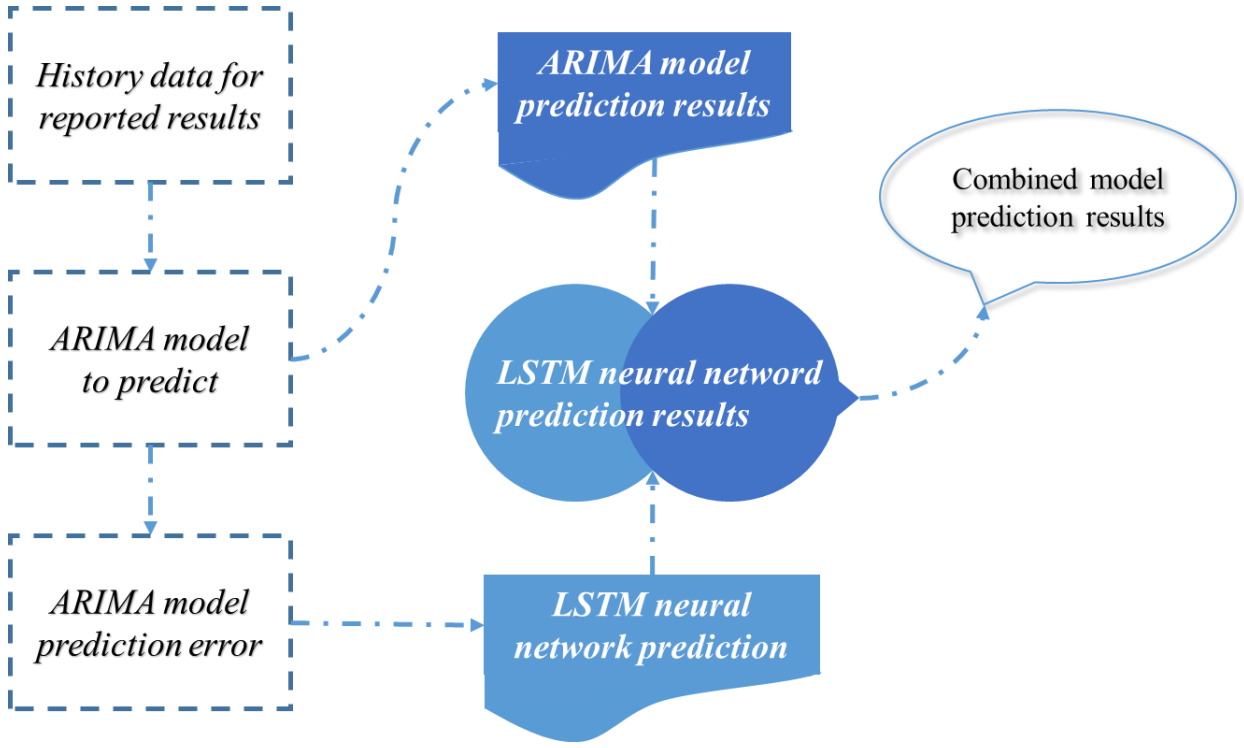

Figure 3. Model ARIMA-LSTM Overview

(1) ARIMA Linear Prediction Model

The ARIMA(p,d,q) model is known as the Autoregressive Integrated Moving Average Model, where is p the autoregressive term; d is the number of differences when the time series is stationary; q is the number of moving average items. This model is a combination of autoregressive(AR) and moving average(MA), which can transform a non-stationary time series into a stationary time series, and then regress the lagged values of the dependent variable, the present and lagged values of the random error term to the model established. The formula is given in equation 2.

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{2}$$

Where y_t is the current value; μ is the constant term; γ_i is the autocorrelation coefficient; ε_t is the error term; θ_i is the coefficient of error term.

(2) LSTM Nonlinear Prediction Model

As LSTM models excel in capturing complex nonlinear features in time series data, and ARIMA models are effective in predicting stable trends, we applied an error compensation approach to combine the strengths of both models. Specifically, we used the LSTM model to predict the decomposed low-frequency wavelet, and the reconstructed high-frequency wavelet D_{sum} was predicted using an ARIMA model with $p=2$, $d=0$, $q=2$. We then combined the predicted results of the low-frequency and high-frequency components to obtain the estimated validation set for December 23-31, 2022, as illustrated in Figure 4.

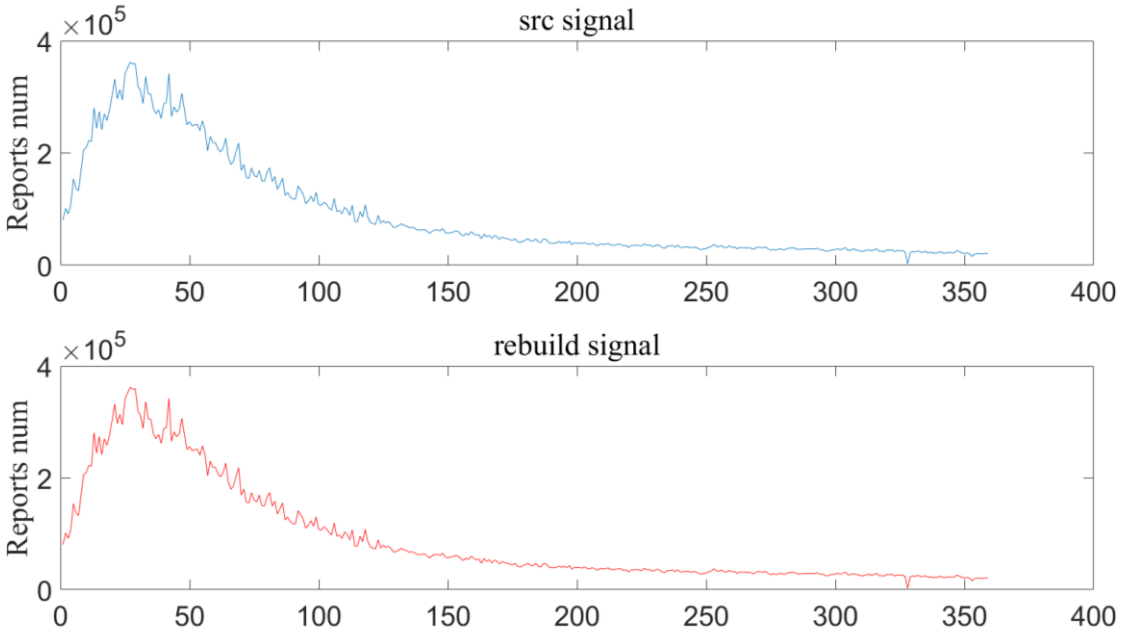

Figure 4. Prediction results of the validation set

In order to verify the superiority of the ARIMA-LSTM-based asset reported results forecasting model, we use Root Mean Squared Error (RMSE) and Mean Absolute Percent Error (MAPE) as the model performance evaluation indexes.

$$MAPE = \frac{1}{9}\sum_{t=1}^{9}\frac{abs\left(num_t^{est} - num_t^{true}\right)}{num_t^{true}} \times 100\% \tag{3}$$

$$RMSE = \sqrt{\frac{1}{N}\sum^{N}\left(num^{true} - num^{est}\right)^2} \tag{4}$$

Where num^{est} is the predicted reported results number, num^{true} is the reported results number of the true value.

According to the calculation, the RMSE value of the ARIMA-LSTM model is 0.322% and the MAPE value is 8.78%, which indicates a more significant improvement in its prediction accuracy. This is because the ARIMA-LSTM combines the advantages of the ARIMA model and the LSTM model to fit the variation pattern of the reported outcome numbers more accurately.

3. Model II:Effects of word attributes on percentage of scores based on CCA

3.1. Construction of word attribute index system

The solutions for the Wordle game vary from day to day, and our firsthand experience has shown that some words are relatively easy, requiring only 2 to 3 attempts to guess correctly, while others are more difficult, with even 6 attempts being insufficient to guess correctly. To sum up, this article carefully takes three main factors into account. (1) Overall frequency of letter use (2) Number of occurrences of repeated letters (3) The notability of words.

Factor 1: Given that each English letter has a different frequency of use, such as the letter 'e' having a usage rate of 12.7% while 'z' has a usage rate of only 0.07%, words with more commonly used letters are often easier to guess in fewer attempts. This article uses $alpha\_fre_{overall}$ to measure the influence of letter frequency on the difficulty of the word. The expressions of $alpha\_fre_{overall}$ are as follows:

$$alpha\_fre = \prod_{i=1}^{5} alpha\_fre_i$$
$$alpha\_fre_{overall} = \frac{alpha\_fre - mean(alpha\_fre)}{std(alpha\_fre)} \quad (5)$$

Factor 2: For words with two or more repeated letters, people often have the cognitive bias that a letter will not repeat after it has already appeared once. Therefore, when a word contains multiple repeated letters, its difficulty level will increase significantly. We use $repeat$ to measure the impact of the frequency of repeated letters on the difficulty of a word.

Factor 3: In real life, the frequency of usage for distinct words varies considerably. For instance, the term "paper" is prevalent in both spoken and written language, whereas some words may be more specialized and challenging for ordinary people to utilize. Hence, We employed Ucrel's frequency corpus to filter the frequency of words that fulfill the file's word characteristics (5 letters) and denoted it as $word\_fre$.

3.2. Model building

In order to facilitate the analysis of the impact of word attributes on the report score, we integrated the above three word attributes and percentage of scores reported that were played in Hard Mode into a data table(Table 1), and performed canonical correlation analysis on the data table[9].

Table 1. Word attributes and corresponding report percentage distributions

word	alpha_fre	repeat	word_fre	1 try	2 tries	3 tries	4 tries	5 tries	6 tries	7 or more tries
manly	0.010536846	0	0.89	0	0.02	0.17	0.37	0.29	0.12	0.02
molar	0.035546737	0	0.46	0	0.04	0.21	0.38	0.26	0.09	0.01
havoc	0.010165469	0	0.92	0	0.02	0.16	0.38	0.3	0.12	0.02
impel	0.016529104	0	0.89	0	0.03	0.21	0.4	0.25	0.09	0.01

Assume $U = \begin{pmatrix} alpha\_fre \\ repeat \\ word\_fre \end{pmatrix}$, $V = \begin{pmatrix} 1\ try \\ \vdots \\ 7\ or\ more \end{pmatrix}$, where $alpha\_fre$ is Overall frequency of letter use in word, $repeat$ is the number of occurrences of repeated letters, $word\_fre$ is word popularity.

Respectively using X and Y ,we calculate two linear combinations, U and V .

$$X_i = a_{i1} \times alpha\_fre + a_{i2} \times repeat + a_{i3} \times word\_fre \quad (6)$$

$$Y_j = b_{j1} \times (1\ try) + b_{j2} \times (2\ tries) + \cdots + b_{j7} \times (7\ or\ more\ tries) \quad (7)$$

Then calculate the variance $\mathrm{var}(X_i)$ of X_i, the variance $\mathrm{var}(Y_j)$ of Y_j, and their covariance $\mathrm{cov}(X_i, Y_j)$

Using feature decomposition method, we ultimately solve for the maximization ρ^* of correlation between U_i and V_j .

$$\rho^* = \frac{\mathrm{cov}(X_i, Y_j)}{\sqrt{\mathrm{var}(X_i)\,\mathrm{var}(Y_j)}} \quad (8)$$

4. Model III: Prediction model based on multi-sequence LSTM

4.1. The Structure of the multi-sequence LSTM Model

In this paper, we assume that the percentage of correct answers in different attempts in the wordle game is related to the difficulty of the words, and add the attribute "number of vowels" to the three basic attributes to better describe the relative difficulty of words.

In order to predict the distribution of future reported results, this paper models the distribution pattern of reported results using a Long Short-Term Memory network (LSTM), a variant of recurrent neural network (RNN) for classification, processing, and prediction based on time-series data. A common LSTM unit consists of a cell, an input gate, an output gate and a forgetting gate. The cell memorizes values at arbitrary intervals and the three gates regulate the information flow into and out of the cell. Our multi-series LSTM model can be divided into three separate modules: the input module, the LSTM module and the output module[10]. The process is shown in Figure 5.

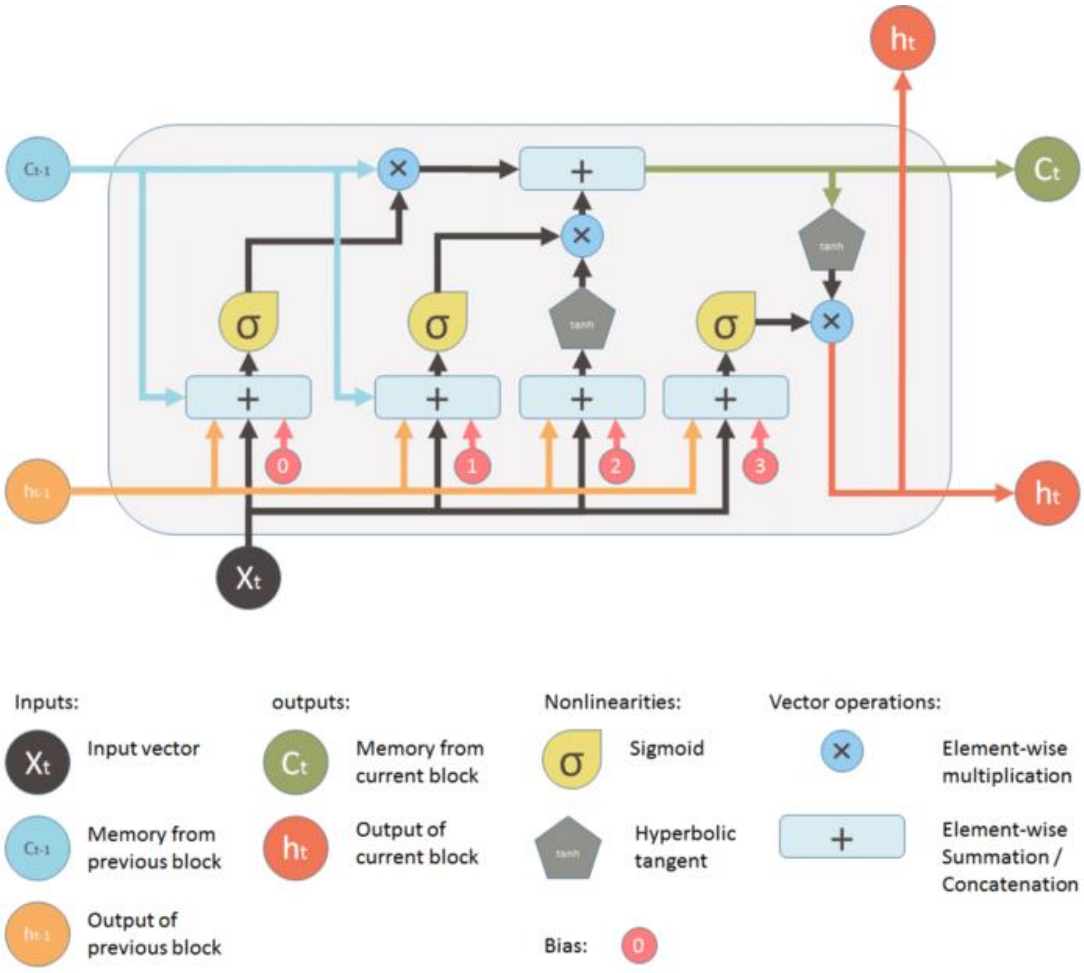

Figure 5. multi-sequence LSTM model

(1) Input Module

For a specific word text on a specific date, This article first enter the word into the embedding system to extract the feature vectors needed for the subsequent model. More information about the feature vectors is shown in Table 2.

Input module: This input vector is composed of the time stamp and the four word attributes mentioned above.For a particular timestamp, its corresponding word feature vector is used as the input to the model in this paper. In order to eliminate the influence of the magnitude on each dimension of the initial input feature vector, each dimension is first normalized here, and then the feature vector converted to 200 dimensions is used as the input to the LSTM layer through the fully connected layer.

Table 2. Feature Definition of Vector v_f

Dimension	Definition	Domain
x_1	Time stamp	{738528~738886}
x_2	Overall frequency of letter use in words	{ -0.6103~7.307}
x_3	Number of occurrences of repeated letters	{0~3}
x_4	Frequency of word use	{0~0.97}
x_5	Number of vowel letters	{0,1,2,3,4}

Tips: dimension x_2 has been normalized

All the features above are already defined in the previous sections. Therefore, the input vector for the LSTM can be formulated as:

$$\begin{cases} v_s = \text{WordEmbed}\left(\text{review}_{\text{headline}} + \text{review}_{\text{body}}\right) \in \mathbb{R}^{200} \\ v_f = \left[x_1, x_2, x_3, x_4, x_5, x_6\right] \in \mathbb{R}^6 \\ x = v_{\text{in put}} = \left[v_s, v_f\right] \in \mathbb{R}^{206} \end{cases} \tag{9}$$

(2) LSTM Module

After the input vector is transformed in the input layer, the features contained in it are learned using a 200-dimensional LSTM layer. Since LSTM does not have features that forget long term memory compared to RNN, it can learn long term parametric features without forgetting the context even for data with a large time span.

(3) Output Module

In the output module, we simply feed the hidden state of the upper LSTM into a linear layer to make predictions on the feature vector, which has the same structure as v_{input} in the input module.

$$v_{\text{pred}} = Wv_h + b \tag{10}$$

4.2. Analysis on Potential Success of Products

Before we show the analysis results, there are several training details to be stated:

(1) Loss function

During the training, we use mean square error (MSE) as the loss function:

$$\text{MSE}(y, \hat{y}) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2 \tag{11}$$

where y is the true target vector and $\hat{y}$ is the predicted feature vector and N is the number of dimension.

(2) Normalization

Due to different domains of each dimension in the input vector v_{input} a normalization technique is needed. Before we feed the input vector into the LSTM module, we apply the following per-feature normalization operation:

$$x_i = \frac{x_i - \mu_i}{\sigma_i} \tag{12}$$

where μ_i and σ_i are the mean and standard deviation of the i^{th} feature.

(3) Training data

To avoid potential gradient explosion, the gradient threshold of this model is set to 1. When the gradient is greater than this threshold, the gradient will be forced to cut to 1. In the case of smooth convergence, the segmented learning rate is used to avoid oscillation of the model near the optimal point.

To ensure the stability of the prediction results, the average of the prediction results generated by multiple operations is chosen here as the final output of the prediction results.

5. Results

The conclusion of hybrid model of ARIMA-LSTM based on wavelet denoising is shown in Figure 6. The number of number of reported results by 1 March 2023 is between 16,214 and 16,490 cases.

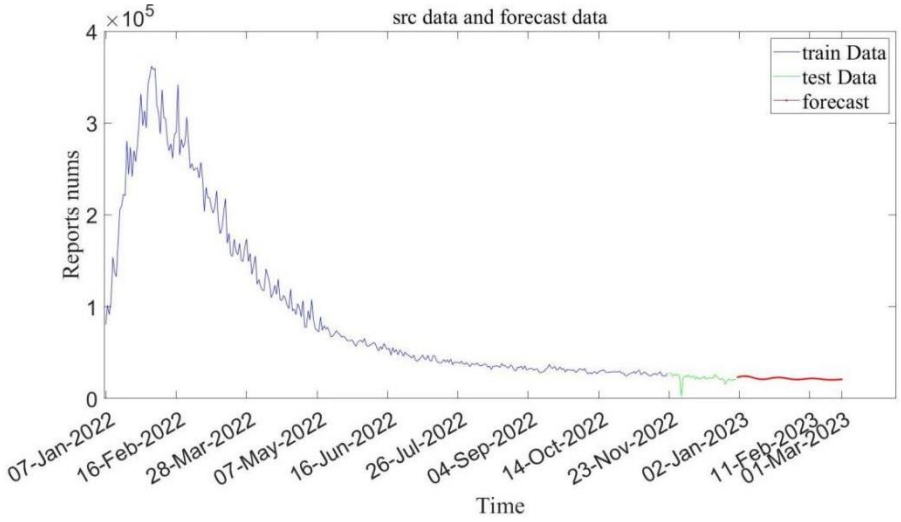

Figure 6. predicted results

The results of correlation analysis show that the three attributes of vocabulary have strong correlation with the distribution of percentage of scores, of which the canonical correlation coefficient is $\rho^* = 0.63$.

When the solution word is ’EERIE’ on March 1, 2023, The fitting result of multi-sequence LSTM model for the report distribution is [0,8,19,25,23,17,8], which basically accords with the distribution rule of the original data.

6. Conclusions

(1) This article builds a hybrid ARIMA-LSTM model based on wavelet denoising. The overall accuracy of the model is over 90%.And the number of players is 16,352 on March 1, 2023.

(2) Based on typical correlation analysis, the correlation between the attributes of the words themselves and the percentage of scores was found to be 0.63, with a strong overall correlation.

(3) By building a multi-sequence LSTM model, the distribution of percentage of scores was predicted to be [0,8,19,25,23,17,8] when the solution word of March 1, 2023 was 'EERIE'.

References

[1] Ling Xiaojing, Zhuang Yaming, Sun Liling. Research on SEIR Online Opinion Dissemination Model with Saturated Contact Rate[J]. Journal of Intelligence, 2015, 34(3): 150-155.

[2] Zhang Jidong, Jiang Liping. A study on the dynamic evolution model of mobile social network opinion dissemination incorporating user group behavior[J]. Journal of Modern Information,2021, 41(5): 159-166+177.

[3] Peng Zhihang, Bao Changjun, Zhao Yang, et al. ARIMA multiplicative seasonal model and its application in predicting the incidence of infectious diseases[J/OL]. Journal of Applied Statistics and Management, 2008(2): 362-368.

[4] SHAM N M, KRISHNARAJAH I, SHITAN M, et al. Time series modelon hand, foot and mouth disease in Sarawak, Malaysia[J/OL]. Asian Pacific Journal of Tropical Disease, 2014, 4(6): 469-472.

[5] Ouyang Zisheng, Lu Ming, Zhou Xuewei. Study on Risk Spillover and Early Warning in China's Financial Sector Based on TVP-VAR-LSTM Model[J]. Journal of Statistics and Information, 2022, 37(10): 53-64.

[6] Li, H. J., &Yang, J. M., &Deng, T. T., &Zong, J. L. (2014). Research on Turbine vibration prediction based on wavelet analysis. Computer engineering and applications, 12, 263-265+270.

[7] Lou, L., &Liu, L., &Liu, X. J., &Shi, S. Z. (2021). ARIMA-LSTM hybrid model based on wavelet denoising and its prediction of a stock price index. Journal of Changchun University of Science and Technology, 02, 119-123.

[8] Lai Xiaoying, Qian Jun. The study of ARIMA-LSTM-XGBoost weighted combination model in predicting the incidence trend of pulmonary tuberculosis [J]. Modern Preventive Medicine, 2021,48(01):5-9.

[9] Yu Mingjie, Guo Peng. Research on the relationship between input and output of regional innovation system based on canonical correlation analysis [J]. Science of Science and Management of S.& T, 2012,33(06):85-91.

[10] Qu Zongxi, Sha Yongzhong, Li Yutong. Research on ensemble prediction of major infectious diseases based on grey wolf optimization and multi-machine learning - taking COVID-19 epidemic as an example[J]. Data Analysis and Knowledge Discovery, 2022,6(08):122-133.