A Deep Learning Based Hybrid Model for Sales Prediction of E-commerce with Sentiment Analysis

Haotian Zhu
School of Computer Science and Technology
Sichuan University
Chengdu, China
zzrichardchry@icloud.com

2021 2nd International Conference on Computing and Data Science (CDS) | 978-1-6654-0428-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/CDS52072.2021.00091

***Abstract*—The market of E-commerce has developed rapidly since the emergence of Internet. Many companies or shops of E-commerce want to seek a way to predict the sales of their products. The prediction of sales can help the merchants to formulate a sales strategy, in order to obtain a bigger profit and attract more investment. However, many studies simply use the daily sales or some very basic daily information of the product to predict the sales, without considering the effects of reviews. In this paper, a hybrid network including Bi-directional Long Short-Term Memory (BiLSTM) and Convolutional Neural Network (CNN) is proposed to solve this prediction task. Through careful selection, several attributes and comments of the products are utilized. Feature engineering is used to normalize the different kinds of data. BiLSTM is conducted to analyze all the comments. CNN is utilized to make predictions by using the data provided by feature engineering. Analysis is described to show the advantages and effectiveness of this network. By using the attributes of the product, along with the polarization of comments, CNN can predict the sales of the product.**

Keywords—CNN, E-commerce, feature engineering, sales prediction, sentiment analysis

I. INTRODUCTION

According to Amazon, over 6,000,000 sellers have sold their products on Amazon by 2018. Besides, EBay's revenue exceeded $2.7 billion in the second quarter of 2019. According to FinancesOnline, the estimated number of global digital buyers in 2019 is 92 billion, and it is expected to reach 2.14 billion by 2021. The total global retail E-commerce sales are projected to reach $4.5 trillion by 2021. The emergence of E-commerce changes the way of shopping. Undoubtedly, online shopping is becoming the main way that people shop. Therefore, online sales has become a new way of sales, which includes huge profits. The topic of prediction of the sales of E-commerce product is very attractive to the online shop holders and their investors. For instance, by predicting the sales of the products, the E-commerce companies can better decide on sales strategies to expand profits. Besides, E-commerce companies can attract more investment by showing the predicted sales data. What's more, with the prediction of the products, companies can better analyze the advantages and disadvantages of the products.

With the development of sentiment analysis and machine learning, many studies are conducted in many fields to handle many predicting tasks. In the financial field, deep learning is proved to be effective in predicting the stock market [1], and studies show that data including the polarization from the social media towards cryptocurrencies can predict the price movement of cryptocurrencies [2]. In the field of E-commerce, many studies only focus on the usage of daily sales and prices to predict the sales through machine learning [3] without using the sentiment information from the consumers towards the products. Sentiment analysis has applications in many fields and achieve high accuracy [2][4][5]. This method is widely used in prediction especially when the result has connection to the comments which indicates the polarization of the sentiment of the users. Comments are an important reference for the customers when shopping on an E-commerce platform, which indicates that the polarization of the comments can affect the sales potentially. Many studies have conducted the experiments of sentiment analysis of the comments of the products on E-commerce platform [4][6][5], which shows that it is accessible to analyze the polarization of the comments. When using sentiment analysis in predicting sales [5][8][9], the essential features of the sales of the product aren't used properly within the researches. Besides, previous studies show that Convolutional Neural Network (CNN) has a significant effect on predicting tasks [8][9][10], and CNN is proved to be effective in predicting the sales of E-commerce products [11].

Motivated by the methods mentioned above, my study is focused to utilize sentiment analysis when predicting the sales of the E-commerce product. The result from sentiment analysis along with the previous daily sales of the product, the previous daily price of the product, consumers' rating of the product, the company's rating, etc. will be calculated to form a new feature vector. This feature vector will be the input of the machine learning. In this paper, a hybrid network is developed to predict the sales. For the method of sentiment analysis, some relative works with sentiment analysis [6] simply use numerical rating to analyze the sales of the product. This paper chooses Bi-directional Long Short-Term Memory (BiLSTM) [14] as the approach to analyze the sentiment. The network combines BiLSTM and CNN to predict the sales of the products on E-commerce. With the advantage of BiLSTM, the polarization of every comments can be captured correctly, and CNN can properly utilize all the data that can be collected to do the final predicting task. All the raw data will be prepossessed through feature engineering in order to better fit the network and make the system to predict more efficiently. Using this hybrid network is a novel way to predict the future sales of a product, because of adding sentiment analysis into the input of the CNN. Along with the polarization of the comments, much basic information, like the daily price, daily sales, etc, is the input data of the CNN. CNN can utilize all these kinds of information to give a proper prediction of the sales.

This paper is organized as follows. In the next section, a brief introduction of related works is presented. Section 3 is about the elaborate description about sentiment analysis and feature engineering, followed by the detailed construction of the hybrid network. The methodology of the construction and the implementation of the model is explained in this section.

978-1-6654-0428-0/21/$31.00 ©2021 IEEE
DOI 10.1109/CDS52072.2021.00091

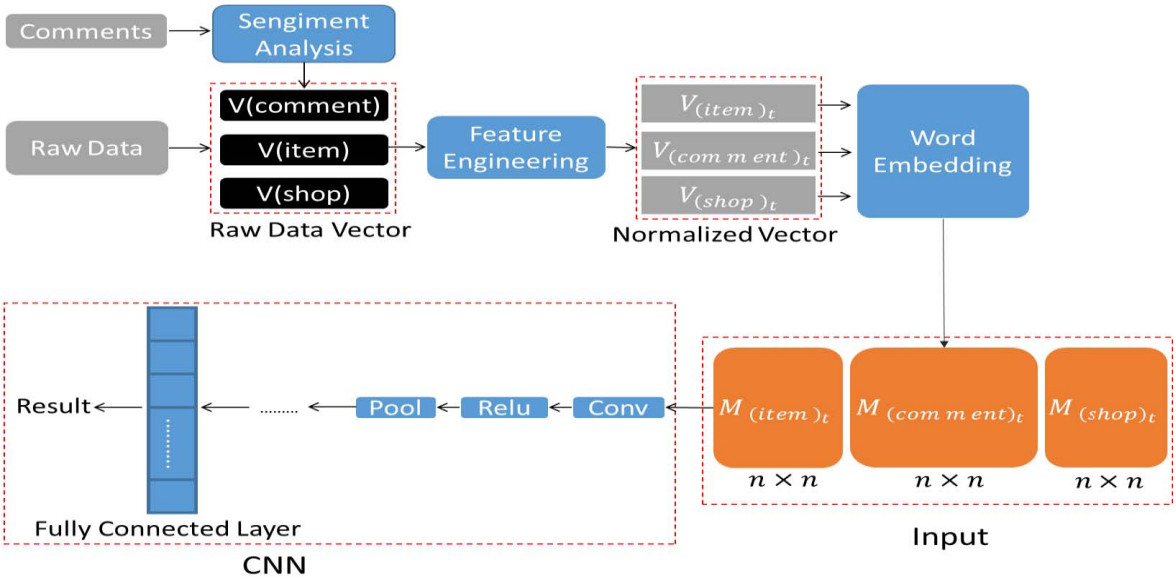

Fig. 1. The overall construction of the hybrid network

Section 4 contains the analysis of the whole system. Section 5 concludes the advantages and the disadvantages of this paper.

II. Related Works

A. Sentiment Analysis

Sentiment analysis is widely used in prediction especially the result has connection to comments, which indicates the polarization of the sentiment of the users. In E-commerce, comments are an important reference for the customers. Therefore, knowing the sentiment of the comments is beneficial to the sellers. In other fields, sentiment analysis of comments achieved a very high accuracy [5]. The utilization of sentiment analysis is of the same importance in E-commerce comments. Some of the studies used arithmetical ways to calculate the polarization [6]. As the development of machine learning, many other methods are used to do sentiment analysis [4][5]. LSTM models have been proposed to tackle the task of sentiment analysis, and BiLSTM is very suitable for the sentiment analysis [14][15].

B. Machine learning

Machine learning is an effective method to deal with E-commerce prediction tasks. Many previous works have used machine learning to predict the sales of E-commerce [3], which shows the effectiveness of machine learning in sales prediction. CNN is one of the most representative machine learning method, and many studies have proved that using CNN to make predictions is accessible and the result is much more satisfying [8][11].

Machine learning is also applied in the sentiment analysis. Many deep learning techniques are used in sentiment analysis [4][5]. LSTM models have been proposed to tackle the task of sentiment analysis, and BiLSTM is very suitable for the sentiment analysis [14][15].

III. Methodology

The hybrid network is organized as follows Figure 1. Firstly, the construction of the raw data is displayed. Then, the method of sentiment analysis is described, and the results will added into the V(comment). All the raw data will be conducted through feature engineering. At last, CNN will utilize all the data to predict the sales.

A. Data

All the raw data will be collected from the E-commerce platform and are supposed to be obtained in daily units. Basically, all the data can be divided into three categories: item, comment and shop.

Item vector includes the following data: the prices (PR), the courier fee (CF), the sales (S), the number of views (VN), the collection volume of the product (CV), the number of searches (SN), the number of people adding this product to the shopping cart (SCN), the number of times shared (TS), the sales ranking of this product (RK), the number of discounts and the number of promotional activities (DN).

Comment vector includes the following data: the number of comments (CN), the rating from each comments (RT), the sentiment of each comments (SEC) and the number of pictures in the reviews (PN).

Shop vector includes the following data: the number of the searches of the shop (SS), the number of collection of the store (SC), the number of fans of the store and the credit score of the shop (FN) and the credit rating of the shop (CD).

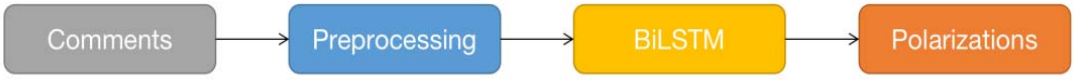

Fig. 2. The overall construction of the hybrid network

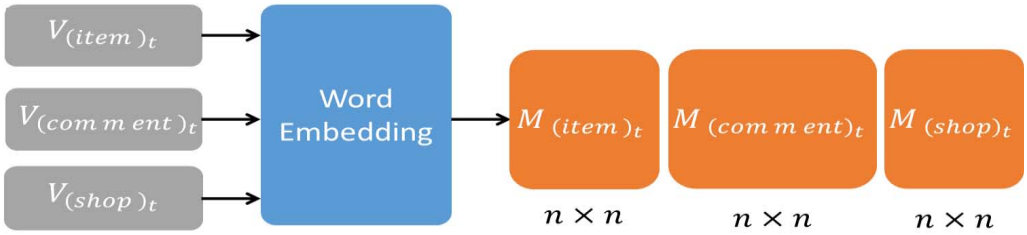

Fig. 3. Word embedding process of the raw data vectors

B. Sentiment Analysis

For the comments of the product are short messages, there are probably emoticons, target, emojis and hashtags in the comments. Emoticons are facial expressions represented by letters and punctuation, like :). Targets are a symbol that refers to other users, like @. Emojis are ideograms and smileys used in electronic messages and web pages. Hashtags are symbols that users use to show their comments which are related to some specific topics. All hese symbols will be dropped before sentiment analysis.

The processed data will be the input of the Bidirectional LSTM model, like Fig. 2.

BiLSTM combines bidirectional recurrent neural network models and LTSM units is used to capture the context information [16]. The model is composed of two LSTM, a forward LSTM and a backward LSTM. The results of the two LSTM will be connected to a softmax layer. The output is [0,1], and the closer to 0, the more negative the polarization of the comment is. The closer to 1, the more positive the polarization of the comment is.

By applying BiLSTM to sentiment analysis, the polarization of each comments will be collected, which will be used while doing feature engineering.

C. Feature Engineering

The data in a time period $T = [t, t + i]$ needs to be normalized in order to guarantee that features are on a similar scale, where i means the length of one training period.

X_{max} is the maximum value of X in T and X_{min} is the minimum value of X in T. X can be any attributes in the vector. In order to normalize all the values, max-min normalization is adopted, X^* is the normalized number of X, the equation is as (1).

$$X^* = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

For the item vector $V_{(item)_t} = [P^*, S^*, VN^*, CV^*, SN^*, SCN^*, TS^*, RK^*, DN^*]$ at t. By using (1), P^* is the normalized price of P, and P is calculated as (2).

$$P = PR + CF \tag{2}$$

S^*, VN^*, CV^*, SN^*, SCN^*, TS^*, RK^* and DN^* are the normalized values of S, VN, CV, SN, SCN, TS, RK and DN.

For the comment vector: $V_{(comment)_t} = [CN^*, RT^*, SEC^*, PN^*]$ at t. By using (1), CN^*, PN^*, RT^* and SEC^* are the normalized values of CN, PN, RT′ and SEC ‘. RT′ and SEC ‘ are calculated as (3) and (4).

$$RT' = \frac{\sum RT}{CN} \tag{3}$$

$$SEC' = \frac{\sum SEC}{CN} \tag{4}$$

For the shop vector: $V_{(shop)_t} = [SS^*, SC^*, FN^*, CD^*]$ at t time. By using (1), SS^*, SC^*, FN^* and CD^* are the normalized values of searches of SS, SC, FN and CD.

D. Hybrid Network

This hybrid network aims to use the log data which is collected from a specific time to predict the sales of the product in the future.

After the feature engineering, the log data will be embedded into three matrices. The output of embedding are These three matrices have the same length, height and dimension, which is demonstrated in Fig. 3.

After word embedding, the convolutional layer is adopted. The activation function is applied after convolution, and then maximization pooling operations are adopted. From convolution to maximization, the data will be calculated through these three layers for three times. At last, the data will

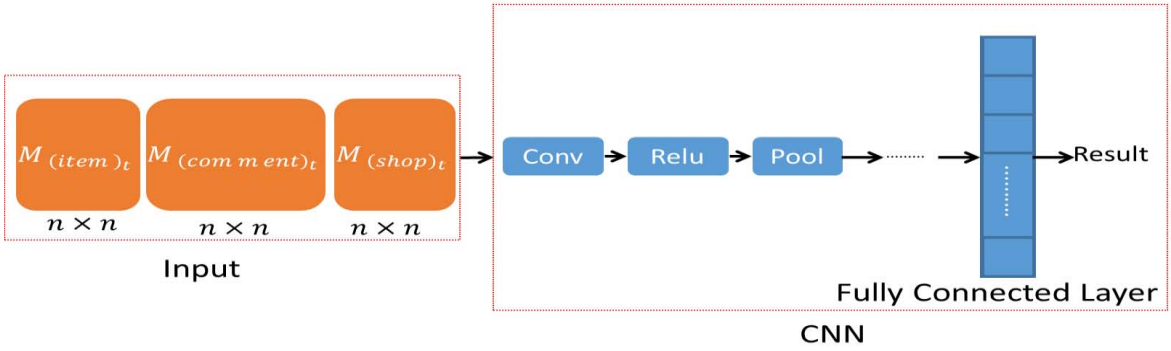

Fig. 4. The framework of sales forecasting with CNN

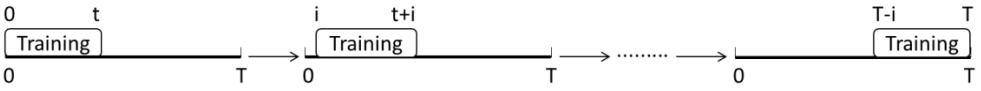

Fig. 5. The training process of the model

be connected to a fully connection layer, and then give out the results of forecasting Fig. 4. The output of CNN is a regression model.

E. Implementation Issues

The convolutional neural network can better utilize the data to handle predicting tasks. CNN achieve big success in many other fields, and [11] proves that CNN can effectively use the data and gives satisfying results. Adding the polarization of the comments into the input data can make the prediction more accurate. The sliding modeling also improve the accuracy of prediction.

In order to train a model to predict the sales of T+1 day Fig. 5, the model needs to be trained from [0, t] to [T-t, T]. t is the length of one training period, and i is the length that the model moves between every training period. After trained from [0, t], the model moves x days to the next epoch to be trained. At last, use the model trained from [T-t, T] to predict the target day: T+1.

IV. Analysis

In this paper, a novel hybrid network is proposed to solve the problem of sales prediction in E-commerce, which is to add the result of sentiment analysis into the input of the CNN.

The reason of choosing the reviews of the product to do sentiment analysis is that the reviews can show the satisfaction of the consumer who bought this product before. The polarization of the reviews will be considered by the next consumer who hesitates to buy this product. If a product is filled with positive comments, the following customers is more probably to purchase this item, and vice versa, negative comments can lower the sales. This is the reason why sentiment analysis of the comments is crucial to this system. Therefore, the sentiment of the whole reviews is of great importance in predicting the future sales of a product. BiLSTM is a very useful model to conduct the sentiment analysis. The model can obtain the information from forward and backward, which predicts the polarization more accurately. What's more, the comments are usually short messages, and BiLSTM is very suitable for such short messages. By applying BiLSTM to sentiment analysis in this case, the polarization of the comments can be predicted more accurately.

For the item vector, this vector contains the basic information of the product. The price of the product will be considered by the customers. Generally speaking, the cheaper the product is, the more likely the customer will by this product. The daily sales is also included in this vector because people tend to buy a product which was bought by many people, so that this attribute can also affect the sales of the product. The other remaining information in item vector, like the number of views, the collection volume and etc, is about the popularity of the product.

For the shop vector contains the essential information about the shop. The number, the collection and the fans of the shop stand for the popularity of the shop. The credit score of the shop gives a general estimation about the quality of the products and services of this shop. The shop vector can reflect a shop's general performance. Therefore, the data can reflect the sales of the products in this shop.

Besides, a proper way to do the data mining of all the data that can be collected of a product is needed. However, the raw log data has a lot of noise and useless texts. Thus, feature engineering is conducted to the raw data. After feature engineering, the log data has less noise and the attributes in the log are more relative, which can be used more effectively by the CNN.

The convolutional neural network can better utilize the data to handle predicting tasks. CNN achieve big success in many other fields, and[11] proves that CNN can effectively use the data and give satisfying results. Adding the polarization of the comments into the input data can make the prediction more accurate. The sliding modeling also improve the accuracy of prediction.

V. Conclusion

This paper proposes a hybrid network to predict the sales on E-commerce. The log data includes not only the basic information of the product, but also the polarization of the comments from the product. Through analysis, using the results from sentiment analysis is effective to predict price or sales movements. Along with the polarization from sentiment analysis, the log data will be embedded into three matrices. After that, layers of CNN are adopted. At last, the data will be connected to a fully connection layer, and then give out the results of forecasting. In analysis, the reason for review and selected features are described to show the advantage and effectiveness of the propose method in this paper, respectively. Meanwhile, the role of the deep learning in the sales prediction is demonstrated as well.

Due to the limits of this paper, there are still some future work needs to be done. In feature engineering, the normalization may cause the weight of different attributes to be unbalanced. The methods of normalization should be improved in the process of experiments.

References

[1] Nabipour M , Nayyeri P , Jabani H , et al. Deep learning for Stock Market Prediction[J].

[2] Valencia F, Gómez-Espinosa A, Valdés-Aguirre B. Price movement prediction of cryptocurrencies using sentiment analysis and machine learning[J]. Entropy, 2019, 21(6): 589.

[3] Qi Y , Li C , Deng H , et al. A Deep Neural Framework for Sales Forecasting in E-Commerce[C]// the 28th ACM International Conference. ACM, 2019.

[4] Yang L, Li Y, Wang J, et al. Sentiment analysis for E-commerce product reviews in chinese based on sentiment lexicon and deep learning[J]. IEEE Access, 2020, 8: 23522-23530.

[5] Kumar H M, Harish B S, Darshan H K. Sentiment Analysis on IMDb Movie Reviews Using Hybrid Feature Extraction Method[J]. International Journal of Interactive Multimedia & Artificial Intelligence, 2019, 5(5).

[6] Li X , Wu C , Mai F . The Effect of Online Reviews on Product Sales: A Joint Sentiment-Topic Analysis[J]. Information & Management, 2018, 56(2):172-184.

[7] Liu Y , Huang X , An A , et al. ARSA: a sentiment-aware model for predicting sales performance using blogs[C]// International Acm Sigir Conference on Research & Development in Information Retrieval. ACM, 2007.

[8] Pryzant R, Chung Y, Jurafsky D. Predicting Sales from the Language of Product Descriptions[J]. eCOM@ SIGIR, 2017, 2311.

[9] Winata G I, Cahyawijaya S, Liu Z, et al. Learning fast adaptation on cross-accented speech recognition[J]. arXiv preprint arXiv:2003.01901, 2020.

[10] Liberis E, Veličković P, Sormanni P, et al. Parapred: antibody paratope prediction using convolutional and recurrent neural networks[J]. Bioinformatics, 2018, 34(17): 2944-2950.

[11] Pham D H, Le A C. Learning multiple layers of knowledge representation for aspect based sentiment analysis[J]. Data & Knowledge Engineering, 2018, 114: 26-39.

[12] Qiu X, Suganthan P N, Amaratunga G A J. Fusion of multiple indicators with ensemble incremental learning techniques for stock price forecasting[J]. Journal of Banking and Financial Technology, 2019, 3(1): 33-42.

[13] Pan H, Zhou H. Study on convolutional neural network and its application in data mining and sales forecasting for E-commerce[J]. Electron. Commer. Res., 2020, 20(2): 297-320.

[14] Feng X, Qin B, Liu T. A language-independent neural network for event detection[J]. Science China Information Sciences, 2018, 61(9): 092106

[15] Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks[C]//2013 IEEE international conference on acoustics, speech and signal processing. IEEE, 2013: 6645-6649

[16] Xu G, Meng Y, Qiu X, et al. Sentiment analysis of comment texts based on BiLSTM[J]. Ieee Access, 2019, 7: 51522-51532