Star Business Value Prediction based on Sentiment Analysis

Meixuan Li

Computer Science and Technology, Harbin University of Science and Technology, Harbin, China
limeixuan0401@126.com

Abstract. The article crawls the audience comments from some videos on the YouTube platform of "Talk Show Conference Season 3", extracts relevant content about popular champions Wang Mian and Wang Jianguo for sentiment analysis, uses Google Data API to design a crawler to obtain comment content, and to crawl After the received content is preprocessed, Word2vec is used to build a word vector model, and finally an LSTM model is built for training prediction. It can be seen that the popularity of the player Wang Mian is higher than that of Wang Jianguo. The entertainment company where the two are located can adjust the artist's work according to the changes in the public's love for the players.

Keywords: Youtube Comments; Crawlers; Sentiment Analysis; Word2vec; LSTM.

1. Introduction

With the advent of the big data era, using artificial intelligence to crawl comments on social media for sentiment analysis has become one of the hot issues in the field of natural language processing. Large-scale sentiment analysis calculations on the content posted by netizens can help the government use public opinion data to adjust or deploy work. Through sentiment analysis on e-commerce platform item evaluations, it helps merchants improve their products and improve product performance. Through the collection and analysis of netizen comments on related topics, it can be applied to election prediction and competition result prediction. The application of big data and artificial intelligence for sentiment analysis has become a powerful tool and means in the information age. People's living standards have gradually improved, and young people have focused more on entertainment. Celebrities have become one of the main topics for young people. Celebrity live broadcasts, celebrity sales, and celebrity promotion promote economic development and promote economic prosperity, and celebrity reputation has become a brokerage company to arrange artists Activities adjust the important standards of the artist development plan, and the brokerage company needs to control and adjust in time.

This article uses the comments under the YouTube platform "Talk Show Season 3" video for sentiment analysis. "Talk Show Conference" was founded in 2017. Since its broadcast, the ratings have gradually increased. The successful holding of the first and second seasons has made Li Dan and Pang Bo the leading figures in the Chinese talk show industry. With the continuous development of talk show, Li Dan established Xiaoguo Culture Co., Ltd. Under the continuous development of media and industry, talk show actors began to undertake a large number of advertisements, which successively drove market consumption, and talk show culture gradually entered the vision of Chinese audiences. The broadcast of "Talk Show Conference Season 3" caused a enthusiastic response from the audience. Yang Li ridiculed men and Peking University talented woman Li Xueqin was good at self-deprecating, and Hulan complained about overtime. Once "Talk Show Conference Season 3" was launched, it caused a huge response from netizens. Netizens continue to discuss the two outstanding players Wang Mian and Wang Jianguo.

2. Current Status of Related Research

Text sentiment analysis through the analysis, processing, induction and reasoning of text content containing sentimental tendencies, and finally achieve the purpose of obtaining the textual sentiment tendencies. Text sentiment analysis can be divided into three levels from the analysis granularity: document level, which analyzes the sentiment tendency of the document; sentence level, which

analyzes the sentiment tendency of a single sentence; entity feature level, judges the sentiment of the sentence at the entity feature level tendency. There are currently two main analysis methods for text sentiment analysis: one is text sentiment analysis based on sentiment dictionary, which depends on the construction of sentiment dictionary, and the final effect depends on the perfection of sentiment dictionary. Another sentiment analysis method based on deep learning, the final effect depends on the selection of training data, the construction of the model, and the correct labeling of text emotions. Currently, the neural network structures commonly used in sentiment analysis mainly include Multilayer Perception (MLP), Recurrent Neural Network (RNN), Convolutional Neural Network (CNN) and attention mechanism .[1]

Recurrent neural network is a type of recurrent neural network that takes sequence data as input, recursively in the evolution direction of the sequence, and all nodes are connected in a chain [2] . RNN is also one of the tools of semantic analysis. Lai et al. applied RNN to text classification [3], and achieved good results. Long Short-term memory neural network (LSTM) is one of the common recurrent neural networks (RNN), which solves the problems of gradient disappearance and gradient explosion in ordinary RNNs, and is widely used in speech recognition, language modeling, sentiment analysis and text Forecast and other fields. Tai et al. [4] recognized the advantages of long and Short-term memory neural networks in the field of natural language processing, and applied the long- and Short-term memory networks to text sentiment analysis, verified them on different data sets and achieved good results.

Word embedding or distributed vector is a technology that converts words expressed in natural language into a vector or matrix form that the computer can understand [5]. Word2vec is a widely used word vector model. It has two network structures, namely the continuous bag-of-words model CBOW and the skip-gram model Skip-gram. The training input of the CBOW model is the word vector corresponding to the context-related word of a certain feature word, and the output is the word vector of this specific word. The Skip-Gram model is the opposite of CBOW, that is, the input is the word vector of a specific word, and the output is the context word vector corresponding to the specific word. [6]

3. Long Short-term Memory Neural Network LSTM

Compared with RNN, the LSTM model contains two transmission states, a cell state and a hidden state. The LSTM model mainly contains four stages. The first is the forgetting stage. This stage is completed through a forgetting gate and can decide what information to discard from the cell state. The second stage is to determine the updated information. This stage determines what new information is retained in the cell state. The third stage is to update the cell state, determine the output value, and finally output information. [7]

Figure 1. Four stages of the LSTM model

4. Experimental Design and Conclusion

4.1 Data Set

Write a program to use Google's own YouTube Data API to crawl the comments under the seventh, eighth, ninth, and tenth videos of "Talk Show Conference Season 3" released by Tencent Video, a total of 5607 items, including Article 1647 of the tenth video.

4.2 Pretreatment

The crawled comment data contains comments made by netizens on other players and comments that do not include player reviews. A total of 862 evaluations about player Wang Mian are extracted, including 352 prediction data and 764 evaluations about player Wang Jianguo, 230 forecast data. After manual labeling, there were 481 positive reviews and 473 negative reviews. At the same time, the crawled comments contain a large number of irregular text expressions, and jieba word segmentation is used for word segmentation to obtain the preprocessed text.

Table 1. Jieba word segmentation method

Player name	Number of forecast data	Number of favorable data	Number of negative reviews	Total number
Wang Mian	352	235	192	862
Wang Jianguo	230	246	281	764

4.3 Text Word Vector Representation

After preprocessing, the words expressed in natural language need to be converted into vectors or matrices that the computer can understand, so Word2Vec is used to train word vectors. Word2Vec is an application of deep learning in the field of natural language processing. This article implements Word2Vec word embedding training through gensim.

4.4 Building an LSTM Model

After building the word vector model, use Keras to build a deep learning model. Establish the Sequential model, add the embedding layer to it first, then add the LSTM layer, set the Dropout to 0.5, add the Dense layer to aggregate its dimensions to 1, the activation function uses sigmoid, and the loss function is set as the cross-entropy function. After the model is built, use the training samples obtained before for training and save the model.

4.5 Result Analysis

After the model training was completed, sentiment analysis was performed on the comments made by Wang Mian and Wang Jianguo under the tenth video. 59.41% was praised by Wang Mian, and 53.18% was praised by Wang Jianguo. In the final ranking of the two in the third season of the talk show, Wang Mian finally won the first place and Wang Jianguo finally won the second place. It can be seen that the final ranking meets the audience's expectations. Player Wang Mian won more popularity and audience love in the third season. His agency can arrange business cooperation and business promotion for him. Player Wang Jianguo's agency can also comprehensively plan artist development based on Wang Jianguo's word-of-mouth changes. long term planning.

5. Conclusion

This article uses a crawler to crawl the comment content of popular contestants Wang Mian and Wang Jianguo under the "Talk Show Conference Season 3" under the YouTube platform by writing a program, using jieba word segmentation to process the crawled text, using Word2Vec to build a word vector model, and finally Constructed and trained the LSTM model, analyzed the sentiment of the comments, and got the proportion of good comments from the two players. According to the

proportion of the two players' praise, the audience's favorability of the two contestants can be obtained. The agency can arrange artist activities according to the changes in the audience's favorability of the artist, such as live broadcast to drive the economy, undertake promotion to drive the market, and change the direction of the artist's development, planning the artist's long-term development plan, etc.

Using deep learning methods to conduct sentiment analysis on content related to celebrities posted by social media Internet users can effectively control public opinion and understand the emotional changes of Internet users. Brokerage companies can adjust measures based on the analysis results to better cater to the public and arrange artist jobs.

References

- [1] Chen Qian, Che Miaomiao, Guo Xin, Wang Suge. A method of text sentiment classification based on cyclic convolutional attention model[D]. Taiyuan. School of Computer and Information Technology, Shanxi University. 2020.
- [2] Goodfellow, I., Bengio, Y., Courville, A. Deep learning (Vol. 1): Cambridge: MIT Press, 2016: 367-415.
- [3] Lai, S., Xu, L., Liu, K. and Zhao, J., 2015. Recurrent Convolutional Neural Networks for Text Classification. In AAAI (Vol. 333, pp. 2267-2273).
- [4] Tai K S, Socher R, Manning C D. Improved Semantic Representations from Tree- Structured Long Short-term Memory Networks [J]. Computer Science, 2015,5 (1):1-36.
- [5] L_R_H000 natural language processing (NLP) of Word Embedding [EB / OL] (2018-08-03) [2020-04-10]. https://blog.csdn.net/L_R_H000/article/details/81320286, 2018-08-03.
- [6] Wang Ling. Implementation of text classification based on Word2Vec word embedding and two-way long and Short-term memory network [D]. Guangzhou. Guangdong Vocational and Technical College of Posts and Telecommunications. 2020.
- [7] Zhu Xiaohu Xiaohu (Neil) Zhu. Understanding LSTM network. <https://www.jianshu.com/p/9dc9f41fb29>, 2015.08.28.