HILL PUBLISHING

Advances in Computer and Communications, 2023, 4(2), 74-79
https://www.hillpublisher.com/journals/acc/
ISSN Online: 2767-2875

Sentiment Analysis of Medium and Long Text Based on Feature Fusion Model

Changchun Yang, Cong Cai*, Tongguang Ni

College of Computer and Artificial Science, Changzhou University, Changzhou, Jiangsu, China.

How to cite this paper: Changchun Yang, Cong Cai, Tongguang Ni. (2023) Sentiment Analysis of Medium and Long Text Based on Feature Fusion Model. *Advances in Computer and Communication*, 4(2), 74-79. DOI: 10.26855/acc.2023.04.002

Received: April 15, 2023
Accepted: May 12, 2023
Published: June 8, 2023

***Corresponding author:** Cong Cai, College of Computer and Artificial Science, Changzhou University, Changzhou, Jiangsu, China.

Abstract

At present, the research on microblog short text has reached saturation, while the middle and long text sentences are complex and have many emotional words, which makes it difficult to classify the whole sentence. In order to solve the problem that the emotional feature extraction is not sufficient in the current micro-blog long text sentiment analysis task, which leads to the inability to extract the text sentiment semantics comprehensively, a multi-feature fusion sentiment analysis method (MBEA) combining capsule network, multi-layer bidirectional long short-term memory network and residual network is proposed. This method uses the Word2vec model to generate word vectors, and then inputs the word vectors into the residual network, capsule network and multi-layer bidirectional long-term and short-term memory network to obtain their vector feature representations respectively. Finally, the fully connected layer is input, and the emotion discrimination is performed by the softmax activation function. Experiments on COVID Dataset and Financial Dataset verify the accuracy and effectiveness of the model compared with other baseline models of sentiment analysis.

Keywords

Residual network, capsule network, multi-layer bidirectional long short-term memory network, emotional analysis

1. Introduction

With the continuous progress of society, WeChat, Taobao, Weibo and other social software emerge in an endless stream. People can express their views, questions and opinions through these software, resulting in a large amount of data that can be mined. By mining, processing and analyzing network data, we can generally understand the emotional tendency of users and predict the topic public opinion that a social problem may lead to. In the past two years, the outbreak of COVID-19 has disrupted people's normal life, and the 'virtual world' of the Internet has also produced complicated texts. Different from other types of texts, Weibo epidemic text sentences are lengthy, and the expression is more colloquial. The semantic information is highly related, and the phenomenon of the same word disagrees often occurs. How to extract the characteristics of information and analyze the emotional changes and views of netizens in time is of great practical significance to help the government control public opinion and scientifically prevent and control it.

2. Related Work

Sentiment analysis is an important task in the field of natural language processing. This task mainly helps users to obtain, organize and analyze data, and analyze and process emotional text. At present, the methods of sentiment analysis mainly include sentiment analysis methods based on sentiment dictionary, sentiment analysis methods based

on machine learning, and sentiment analysis methods based on deep learning. Many researchers have done research on the method of early sentiment dictionary. The method of sentiment dictionary is simple and does not require additional resources. It mainly relies on the sentiment dictionary library, matches the processed vocabulary with the dictionary library, and then calculates the sentiment score and performs sentiment classification. Yan et al. [1] proposed a sentiment dictionary word set screening method by constructing a domain sentiment dictionary, and used the representation to calculate the distance between vectors. The validity of the model was verified in the tourism online review data set. Zhu et al. [2] proposed a combination of financial dictionary and attention mechanism for the lack of sufficient corpus, and achieved an accuracy of 74 % on the constructed financial data set. However, the method based on dictionary and rules is simple, but it requires a lot of manual intervention, and the labor cost is high. In fact, the classification effect is not accurate.

At present, the mainstream method is based on deep learning. The deep learning method can adaptively learn the semantic information of the text, so as to extract text features, reduce the complexity of feature construction, and have a significant effect on sentiment analysis tasks. Kim [3] and others have achieved excellent results in sentence-level sentiment analysis tasks by combining word vectors with convolutional neural networks, but the pooling operation in CNN will bring about the loss of information. Li et al. [4] proposed the Dual Channel Convolutional Neural Network (DCCNN) model, which uses two channels for convolution operation, the former channel is a word vector, the latter is a word vector, and uses multiple sizes of convolution kernels to extract sentence features. The experimental results show that the accuracy and F1 value of this model have been greatly improved compared with the traditional method, reaching more than 95 %. Chen et al. [5] proposed to combine LSTM and CNN, in which LSTM performs context word embedding and CNN performs model feature training. Wang et al. [6] divided the text into regions, combined LSTM and RCNN to predict sentiment fine-grained analysis, and set the weights according to the importance of the text. LSTM can obtain long-distance semantic dependencies. Yang et al. [7] used word2vec to form word vectors, and integrated attention mechanism into long-term and short-term memory networks, thus improving the accuracy of text classification. With the rise of capsule network, many researchers have studied on this basis. Zhao et al. [8] applied capsule networks to natural language processing for the first time, capturing noisy capsules through dynamic routing. Through many experiments, it is proved that the effect is significant on multiple text sentiment analysis datasets. Zhang et al. [9] proposed a capsule network method based on attention mechanism to extract relationships, and its performance has been significantly improved.

In summary, the method based on sentiment lexicon and rules relies too much on manual dictionaries, which is time-consuming and laborious. The method based on machine learning mainly constructs text features manually, and the real-time performance is weak. In addition, compared with the first two methods, the method based on deep learning can enable the computer to automatically read the semantic features in the vocabulary of this article, which is accurate and convenient. But usually a single network is difficult to extract the context information and semantic location information of the text.

3. Related Methods

3.1 MBEA model

In order to solve the problem of medium and long text classification, the MBEA sentiment analysis model is proposed. The specific model structure is as follows. The MBEA model is mainly divided into two parts: 1 Feature extraction of the initial text. Firstly, the word2vec model is used for word embedding. The text is transformed into a matrix form, and then sent to the residual network, the capsule network and the bidirectional long short-term memory network. Second, the features of each sub-model are spliced based on feature fusion, and then sentiment classification is performed based on the fused features.

Sen_i represents emotional text, and Multi-Stacked text superimposes the samples of one batch. B represents the batch size, S is the length after Sen _ i is PAD, D is the dimension of word segmentation vector Embedding, and word segmentation is initialized by truncated normal distribution. That is, for the current N random samples, a three-dimensional matrix is obtained: B × S × D is directly used as the input of the MBEA model.

For the two sub-networks of Resnet and CapsNet, Resnet's convolution Block convolves the input matrix, and the features after convolution are then convoluted in several layers to obtain features. The basic convolution of CapsNet convolves the original input, then continues the convolution of the feature map and takes the vector after the route as the text feature. The input of each BiLstm of Multi-BiLstm is shown in Formula 1.

$$\begin{cases} I_k = In[k:k+1, M, S, D] \\ 0 \le k < B \\ H_k \leftarrow shape[I_{k+1}][1] \end{cases} \tag{1}$$

The feature matrix generated by Resnet, CapsNet and BiLstm multi-channel network is fused, and the feature fusion formula 2 is shown.

$$\text{merged}_{\text{fea}} = \text{Feature}_1 \uplus \text{Feature}_2 \uplus \text{Feature}_3 \tag{2}$$

Finally, the activation function is sent to output the emotional polarity, as shown in Formula 3.

$$l = soft\max(merged_{fea}) \tag{3}$$

Among them, I_k represents the input of the kth cascade BiLstm, In represents the input matrix of the MBEA algorithm, and the number of hidden layer nodes H_k of the I_k layer BiLstm is the sample word segmentation length of the I_{k+1} layer input matrix. $\uplus$ means that the features obtained by all sub-networks continue to be superimposed on the last dimension. The MBEA algorithm uses three network structures to model the features of long texts. Then, the features obtained by the three networks are superimposed as the target features of the input text, and the sentiment classification of the current long text is performed.

Based on the feature fusion method, the emotion prediction of long text is obtained. In this method, different sub-network feature extraction capabilities are different, and the effect on emotion classification is not the same. This paper sets the trainable feature contribution vector μ, which is used to do the inner product with the fused, and then used for classification. The emotion classification label is shown in formula 4.

$$\begin{cases} L_i \leftarrow f(\text{merged}_{\text{fea}} \uplus \mu | \theta, \mu) \\ 0 \le i \le N \end{cases} \tag{4}$$

3.2 Capsule network module

The first layer of the MBEA model uses a capsule network, which consists of four parts: convolution layer, main capsule layer, convolution capsule layer and fully connected capsule layer. The first convolution is a standard convolution layer, which extracts different features at different positions of the microblog statement. The main capsule of the second block is to replace the scalar output with the vector output, in order to retain the word order of the text and the information between the sentences. The third block is the convolutional capsule layer. The capsule calculates the relationship between the child capsule and the parent capsule, and then calculates the capsule of the upper layer according to the dynamic routing. The fourth part of the fully connected capsule, through the transformation matrix, finally generates the final capsule to calculate the corresponding probability of each classification.

For example, a word in the Weibo data set of this article is x_i, after Word2vec becomes a word vector $v(x_i)$, then the word vector matrix is, and the capsule network extracts the local features of the text. The specific calculation formula is as follows:

$$C_i^a = f(X_{i:j+k1-1} * W^0 + b_0) \tag{5}$$

Among them: f is the activation function, W_0 is the weight function, b_0 is the bias term, $a = i:j+k1-1$ is the moving window slides from i to $k1-1$, where a= 1,2,... n, represents the capsule network to extract local features.

3.3 Muti-BiLSTM module

The second sub-layer of the MBEA model is the Muti-LSTM layer. LSTM is a variant of RNN, which is improved on the basis of RNN and can learn the structural information and sentence correlation of microblog sentences unsupervised. RNN processing long sentences may have the problem of gradient explosion or gradient disappearance, while LSTM can learn the semantic information of sentences of different lengths unsupervised. However, the LSTM model can only deal with the problem of close dependence, and the corpus of the microblog data set in this paper is long. The LSTM model may forget the long-distance information, and cannot capture the sentence information from back to front, and cannot obtain the interaction between emotional words. The BiLSTM model can model the microblog text well. The input of the model is a sequence structure. The model structure includes forward propagation LSTM and backward propagation LSTM. The sub-layer Muti-BILSTM of the MBEA model proposed in this paper is

equivalent to the improvement of the ordinary LSTM. The hierarchical connection of multiple BILSTMs can effectively learn the information between the levels, so that the feature information is extracted for the microblog sentiment analysis task. The specific model architecture is similar to the standard LSTM.

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{6}$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{7}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{8}$$

$$c_t = tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{9}$$

$$h_t = o_t * \tanh(c_t) \tag{10}$$

Muti-BiLSTM includes cell state C_t at time t, input word X_t, temporary cell state $\widehat{C_t}$, hidden layer state h_t, forgetting gate, memory gate and output gate. The input state is the previous moment h_{t-1} and the current input X_t, the input gate i_t, f_t is the forgetting gate, which is changed by X_t and h_{t-1} through the sigmoid activation function, and controls the influence ratio of the information in the hidden layer at the previous moment. o_t controls the output vector, and c_t is the current cell update unit. W and U are the weight matrices corresponding to each gate, and b is the offset term corresponding to each gate. Both σ and tanh are activation functions.

3.4 Residual network module

The third part of the MBEA model is the application of the residual network layer. The microblog comment data set is unbalanced and the sentences are partial to medium and long sentences. The introduction of the residual network can greatly strengthen the jump layer connection of the residual unit, that is, the input and output are directly connected, and finally reactivated. It includes a forward propagation and backward propagation to embed the matrix generated by word2vec word into the residual mechanism. First, each sentence is vectorized by word2vec to generate a word embedding vector matrix.

$$T_{ij} = \{v(x_1), v(x_2), ..., v(x_n)\} \tag{11}$$

With the increase of the number of network layers, the objective function is easy to get the local optimal solution, but it will bring the problem of gradient disappearance, especially the activation function, so that the network parameters of the output layer cannot be effectively learned. The residual network can effectively alleviate the problem of gradient disappearance and improve the effective training depth of the network. The residual network includes multiple residual units. These residual units are composed by stacking. The input data of the first residual unit is, then the output of the next residual block is calculated as

$$y_i = h(x_i) + F(x_i, w_i) \tag{12}$$

$$x_{i+1} = f(y_i) \tag{13}$$

In the formula, y_i represents the output of a residual block in layer i, w_i is the weight function, F is the residual function, f is the relu function and h is the input function.

The residual function is specifically expressed as

$$F(x_i, w_i) = w_i \bullet \sigma(B(w_i) \bullet \sigma(B(x_i))) \tag{14}$$

The ' · ' in the formula is a convolution kernel, where the convolution kernel is $\sigma = \max(x, 0)$, w_i is a weight matrix.

4. Experimental Results and Analysis

4.1 Dataset

In this paper, the data set data comes from the micro-blog social software. By using the octopus collector to crawl the micro-blog about the epidemic topic comments and financial texts, and then make them into data sets COVID Dataset and Financial Dataset respectively. The micro-blog epidemic data set crawled 20,000, after removing stop

words, removing duplicates, there are still 10,000 left, which can be divided into three types of emotions, positive, negative and neutral. Among them, 1 represents positive emotion, 0 represents neutral, and 2 represents negative emotion. The training set and test set are divided by 7 : 3. Financial related text, after data cleaning, there are more than 5000 pieces of data left. This text is divided into three categories, using positive, negative and nature to express positive, negative and neutral emotions respectively.

4.2 Evaluation indicators

In order to evaluate the performance of the verification model, this paper selects the indicators commonly used in sentiment analysis tasks.

The evaluation indexes used in this experiment include precision (P), recall (R), F1 value (F1). Precision refers to the proportion of the number of TP N_{TP} in the sum of NTP and FP N_{FP}. The calculation formula is

$$p = \frac{N_{TP}}{N_{TP} + N_{FP}} \tag{15}$$

Recall rate refers to the proportion of TP in the sum of NTP and FN. The calculation formula is

$$R = \frac{N_{TP}}{N_{TP} + N_{FN}} \tag{16}$$

F1 value is the harmonic mean of precision and recall, which is equivalent to the comprehensive evaluation index of precision and recall. The calculation formula is

$$F_1 = \frac{2PR}{P + R} \tag{17}$$

4.3 Experimental results

In order to verify the effectiveness of the model proposed in this paper, multiple sets of experiments are set up for comparison, including TextCNN, CNN-TCN, Bilstm-textcnn and the model proposed in this paper. The accuracy rate, recall rate and F1 value are compared respectively.

From the above table, we can see that the model in this paper has achieved good results on both data sets. The accuracy, F1 value, and recall rate on the COVID Dataset reached 96.6 %, 96.1 %, and 97.1 %, respectively, and on the Financial Dataset reached 95.8 %, 97.9 %, and 96.3 %. From the experimental results, it can be seen that the accuracy and F1 value of the proposed model on the COVID Dataset are 19 % and 25 % higher than those of the CNN-TCN model. Compared with the TextCNN model, the indicators increased by about 2 %. Compared with the TextCNN-Bilstm model, the COVID Dataset increased by 2.1 %, and the F1 value increased by 1 %. Compared with CNN + TCN on Financial Dataset, the MBEA model in this paper improves the accuracy by 1.1 % and the F1 value by 1.9 %. Compared with the TextCNN model, the accuracy and F1 value increased by 1.8 % and 4.0 %, respectively. Compared with the TextCNN-BiLSTM model, the accuracy and F1 value increased by 13.8 % and 16.8 %, respectively. Therefore, compared with other sentiment analysis algorithm models, MBEA model has better effect on sentiment classification task.

From the information obtained from the experiment, for the COVID Dataset, from the perspective of the accuracy rate curve, regardless of TextCNN, CNN-TCN, Bilstm-TextCNN or the model MBEA proposed in this paper, the accuracy rate increases with the number of epochs, first rapidly increases, then gradually stabilizes, and finally smooths as a whole. The curve changes of these models are basically consistent with the sentiment analysis task. As the model iterates, the predictions in the vocabulary used are loaded, and the accuracy component is improved. In addition, it can be seen from the change curve that in the initial stage, the accuracy, recall rate and F1 value of the MBEA model are obviously lower than those of the BILSTM-TextCNN model. This is because the initial number of network layers of the model is the same, and the learning ability of the BILSTM-TextCNN model is better than that of the model in this paper. However, with the introduction of the residual network, the depth of the model has increased, the correlation degree of the learned information has increased, and the feature extraction ability of the model has been further improved. Obviously, the improvement rate of each index of the MBEA model proposed in this paper is much higher than that of other models. By comprehensive comparison, MBEA model has better superiority.

From the perspective of two different data sets, micro-blog texts usually have a medium length of each data, while financial texts usually have a long length of each data. The model MEBA model proposed in this paper has better effect than the other three models on both micro-blog data sets and financial data sets, and it is also fully proved that the model proposed in this paper has better universality.

5. Conclusions

In view of the single and irregular language and strong subjectivity of microblog comment text, the effect of previous machine learning methods is not ideal, and the context information of the text cannot be fully extracted. A multi-model fusion sentiment analysis model integrating capsule network, multi-layer long short-term memory network and residual network is proposed. The capsule network is used to replace CNN to extract the local features of text semantics, and Muti-BILSTM is used to extract the global features of microblog long text. In addition, the residual network is used to deepen the model depth to further extract the text sentiment semantics. Experiments on two datasets of COVID Dataset and Financial Dataset improve the classification accuracy of the model, which proves that the model has a good effect in the field of medium and long text. The next work will enrich the semantic representation of Weibo corpus by using some other features of Weibo, such as emoticons and emojis, so that the model proposed in this paper is better.

References

[1] Yan Zhongpei, Lu Wenxing, Shuang Juan, Wang Bingyou. A method for constructing an emotion dictionary for online travel-oriented reviews [J]. Computer Application Research, 2019, 36(06):1660-1664.

[2] Zhu Qinglin, Liang Bin, Xu Ruifeng, Liu Yuhan, Chen Yi, Mao Ruibin. Fine-grained sentiment analysis combining sentiment lexicon and attention mechanism in financial domain [J]. Chinese Journal of Information, 2022, 36(08):109-117.

[3] Kim Y. Convolutional neural networks for sentence classification [C]//Proceedings of The 2014 Conference on Empirical Method Natural Language Processing (EMNLP). Strouds burg, USA: Association for Computational Linguistics, 2014:1746-1751.

[4] Li Ping, Dai Yueming, Wu Dinghui. Application of dual-channel convolutional neural network in text sentiment analysis [J]. Computer application, 2018, 38(6): 1542-1546.

[5] Chen P, Xu B, Yang M, et al. Clause sentiment identification based on convolutional neural network with context embedding [C]//12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), IEEE, 2016:1532-1538.

[6] Wang J, Yu L C, Lai K R, et al. Dimensional sentiment analysis using a regional CNN-LSTM model[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2016: 225-230.

[7] Yang Yujuan, Yuan Huanhuan, Wang Yongli. A sentiment analysis method for review text [J]. Journal of Nanjing University of Science and Technology, 2019, 43(03): 280-285 + 291.

[8] Zhao W, Ye J, Yang M, et al. Investigating capsule networks with dynamic routing for text classification [J]. ArXiv Preprint ArXiv: 1804. 00538, 2018. Version 1.

[9] Zhang N, Deng S, Sun Z, et al. Attention-based capsule networks with dynamic routing for relation extraction [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2018:986-992.