Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/44/20230241

Incorporating emotional trend into multi-emotion analysis models for long-text sentiment analysis

Yu Zhang

University of Southampton, Southampton, Hampshire, England, SO17 1BJ

yu.zhang1604@gmail.com

Abstract. The role of sentiment analysis is vital in natural language processing(NLP) and has garnered significant attention across different domains. However, multi-emotion analysis in long-text is still a challenging task due to the intricate emotional nuances that are conveyed. In this paper, a novel approach for long-text multi-emotion analysis is proposed by integrating emotional trends. This integration aims to enhance the ability of the model to recognize emotions by including word-level sentiment scores as supplementary features. To achieve this, the ISEAR and IMDB datasets are leveraged to investigate the impact of sentiment scores with varying weights on three models: BiLSTM, CNN, and CNN+BiLSTM. The models are trained for 20 and 50 epochs and evaluated by accuracy, precision, recall, F1 score ROC curve and AUC value. The experimental results indicate that the incorporation can improve the processing speed of the multi-emotion analysis task while maintaining performance with a 66.7% probability. The highlighted improvement over the baseline model reduced the time by 33.42%. In the best case, the accuracy of the model increased by 2.26% and the F1 score increased by 2.16% without affecting the running speed.

Keywords: Multi Emotion Analysis, GloVe, Bidirectional LSTM, Word Features, Sentence Features.

1. Introduction

Sentiment analysis, as a crucial component of NLP, has gained immense significance owing to its ability to extract and analyze information from databases, thereby enabling accurate sentiment computation for various mediums such as text, speech, images or multi-modal corpus. Although sentiment analysis has numerous applications in marketing, music and social media, text-based sentiment analysis has become mainstream and is being investigated by more researchers [1][2][3]. Recently, the majority of research has focused on binary classification, assigning either a positive or negative tag. For multiple emotions, the amount of feature categories would be greater than for binary classification. It would complicate this classification task, resulting in a significant decrease in classification accuracy and precision [4]. Inspired by the early mood swings in depression relapse, the sentiment of each sentence in long texts may have an impact on the eventual sentiment judgment of the long text [5]. Therefore, this paper proposes a novel long-text multi-emotion analysis by adding the impact of emotional trends during analysis.

DOI: 10.54254/2755-2721/44/20230241

The contributions of this paper are as follows:

- This paper presents an innovation strategy for integrating word-level sentiment scores as supplementary attributes in long-text sentiment analysis models. By doing so, these models can effectively account for both textual characteristics and affective patterns. Furthermore, the approach is cross-validated using CNN, BiLSTM, and CNN+BiLSTM models.
- This model is tested on the International Survey on Emotion Antecedents and Reactions (ISEAR) dataset, which is a long text corpus containing eight sentiment labels. Meanwhile, it is also implemented based on the Internet Movie Database (IMDB) with two sentiment labels.
- The incorporation of affective patterns into multi-emotion analysis models enhances the models' capacity to detect emotions in the long-text sentiment analysis.

The remaining sections of this paper are organized as follows: Section II will introduce the literature review for this work. In Section III, it will present the design of emotional trends and the proposed model. Afterwards, Section IV will depict the experimental setup, datasets, and performance metrics. Section V will evaluate the results and discuss future work. Eventually, Section VI will outline the conclusion.

2. Literature review

Depending on the scale, sentiment analysis is classified into document level, sentence level, and word level. Additionally, based on the analysis method, sentiments can be estimated using rule-based and statistics-based approaches. In early investigations, most challenges in sentiment analysis were addressed through the rule-based approach under its superior interpretability of the data. In 2003, Neural Network Languages consolidated sentiment analysis and neural networks, leading to a breakthrough improvement [6]. Subsequently, Deep Neural networks (DNNs) brought statistical methods to become the dominant research direction due to their high performance and diminished dependency on handcrafted features. DNN mainly has two architectures, which are Convolutional Neural Network(CNN) and Recurrent Neural Network(RNN) [7][8]. CNN can hierarchically extract local features from supported data. RNN can deal with temporal issues and implement long-term dependence. Especially, Long Short-Term Memory (LSTM), a distinct RNN adaptation, adopted a gate mechanism to control the flow and loss of features [9]. It can enhance the performance of the sentiment analysis model by memorizing the sentiments of words in long texts. Benefiting from its characteristics, bi-directional LSTM (BiLSTM) incorporated forward and backward LSTM frameworks to simultaneously learn forward and backward features in context [10].

In 2017, a method for multi-class sentiment analysis was proposed [11]. It utilized writing patterns and a unique unary grammar. To train the model, the random forest machine learning algorithm was employed on a dataset of 21,000 Twitter texts that had already been labeled. These texts encompassed seven emotions, including happiness, sadness, anger, love, hate, sarcasm, and neutrality. Meanwhile, the ultimate accuracy for the multi-class sentiment analysis task reached 56.9%. Additionally, in 2020, SG-BERT was developed to ground an effective emotional semantic graph based on word and sentence features, which can be deemed emotional labels in long-text sentiment analysis tasks [12]. Although this model can achieve better performance in accuracy, it was implemented with a substantial computational overhead and had shortcomings in ultra-long text. In the proposed system, the model will incorporate the emotion trend of the text in addition to the training data for model training. The impact of this approach on sentiment classification tasks will be explored using the ISEAR and IMDB dataset.

3. Model architecture

In this section, the cross-validated model will be demonstrated. Meanwhile, the emotional trend can be presented in detail.

3.1. Emotional trend

In this experiment, the emotional trend is straightforwardly considered as sentiment results at the word level. At the word level, the model can utilize the Sentiment Intensity Analyzer supported by the Natural Language Toolkit(NLTK) to capture the sentiment score of each word in the sentence.

DOI: 10.54254/2755-2721/44/20230241

Additionally, the calculated compound value is exploited as the comprehensive sentiment score of the words, which can represent the sentiment polarity and intensity of the text. When it is positive(>0), it indicates that the sentiment of this text is positive. If the value is closer to 1, this text will be more positive. On the contrary, if it is negative(<0), it indicates that the sentiment of this text is negative. Meanwhile, the closer the value is to -1, the more negative the text is. Finally, if the compound value is close to 0, the sentiment of this text tends to be neutral.

To incorporate its impact on the model, two approaches can be exploited: using the sentiment score as input and merging the feature matrix. For this particular investigation, this paper has opted to merge the feature matrix. This decision is rooted in the fact that the first approach would train the model by employing sentiment analysis as a distinct input, which would not interact with the text features. Conversely, the second method can integrate the sentiment score as an additional feature with the text sequence. The model can learn from both the sentiment score features and text features concurrently. It fosters a more comprehensive utilization of the interaction between those two, thereby enabling further experimentation and discussion of the potential contribution that the sentiment score could make to sentiment analysis.

3.2. Normal model architecture

Each model will commence with an embedding layer and a dropout layer, called the initial layer. The embedding layer encodes the input data into GloVe-based word embedding representations. The dropout layer can revise the input units to zero with a random probability of 0.2 to avoid overfitting. Meanwhile, it is also implemented after each functional function to counteract the effects of overfitting.

In a normal BiLSTM model, the BiLSTM layer continues to operate from the initial layer onward. It consists of 100 hidden units and can be implemented by a sequence of length 100 and feature dimension 50 as the input data. The processed data subsequently undergoes processing through a dense layer that comprises 64 neurons. Afterward, the model leverages ReLU as the activation function to amplify the nonlinearity of the network. Ultimately, the model generates multiple probability distributions employing output processing through a dense layer utilizing softmax as the activation function.

For a normal CNN model, a 1D convolution layer is selected to extract local features of the input sequence, which would operate from the initial layer onward. Its convolution kernel is 32, and its size is specified as 3. Additionally, the model exploits the Swish activation function as the convolution activation function. After convolution, the maximum pooling layer with a pooling window of 3 is leveraged by the model to downsample the feature maps output from the convolution layer. After further processing with a dropout layer, the flattening layer is implemented to flatten the multidimensional input into one dimension. Eventually, it would be implemented as the output process of a normal BiLSTM model to output the probability distribution of multiple categories. The CNN+BiLSTM model is a modification of the CNN architecture, wherein the flattening layer is replaced by a BiLSTM and dropout layer.

3.3. Model with emotion trend architecture

The structure of the model with the emotional tendency is quite similar to those of normal models. Nevertheless, this incorporates customized input rules utilizing both textual data and emotional scores. The following instance in Figure 1 can clearly demonstrate this architecture. In this CNN case, its original model mentioned before is applied to generate feature matrices of the text data. Following this, the emotional scores are subjected to a dense layer using ReLU as the activation function, and a dropout layer to obtain its feature matrix. Finally, the two feature matrices are concatenated. Additionally, the output process is implemented to derive the probability distribution of seven emotions for predicting the test data.

DOI: 10.54254/2755-2721/44/20230241

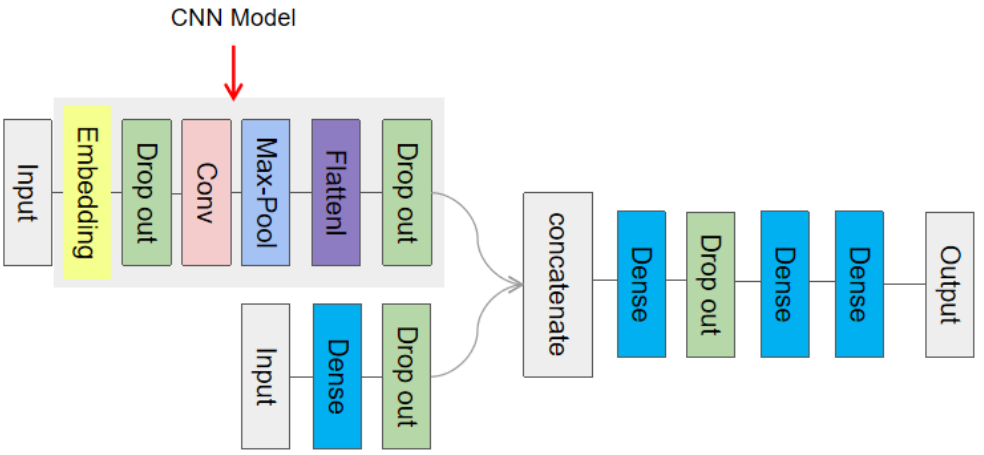

Figure 1. CNN model structure with emotion trend.

4. Experiment

In this section, the detailed parameters, the ISEAR dataset, the IMDB dataset and performance metrics are demonstrated. In addition, experimental steps are also displayed below.

4.1. ISEAR Dataset

In this paper, the ISEAR dataset will be trained and tested. It mainly contains eight emotions, including joy, fear, anger, sadness, disgust, shame, guilt, and guile [13]. Each emotion has an average of 1050 long text sentences. Since the emotion of guilt only has one data point, this experiment would investigate seven other emotions.

4.2. IMDB Dataset

The IMDB dataset provides 50,000 highly polar movie reviews for long-text binary sentiment classification [14]. The longest review has more than 1400 words. In this experiment, this dataset will be the baseline to explore differences in the impact of the proposed model on multi-emotions and binary sentiment classification.

4.3. Experimental Setup

This experiment will be implemented on ISEAR and the IMDB datasets. The data is preprocessed by the model to remove numbers, punctuation, and stop words, and transform all texts into lowercase. Furthermore, the model leverages the GloVe embedding to obtain the vector representation of each word. Afterwards, it processes one-hot encoding to represent each extracted sentiment label as a sparse binary vector, which strengthens the model to capture their differences more accurately. For the sentences in the dataset, the SentimentIntensityAnalyzer function from NLTK can be leveraged to calculate the compound value of each word as the word's sentiment score. Subsequently, these scores would be attributed to divergent weight values, specifically 0.1, 0.5, 1.0, 1.5 and 2.0. Upon completion of these procedures, the resultant data is partitioned into two sets, namely the training set and the test set, wherein the former comprises 80% of the data and the latter comprises 20%.

The experimental procedure entails subjecting all models to 20 and 50 epochs while maintaining a constant batch size of 64. Moreover, the output layers of the ISEAR and IMDB datasets will, respectively, generate 7 and 2 probability distributions. For the model with an emotional trend, the training process incorporates input data consisting of word emotions with variable weights and training data. Notably, the parameters of this model remain consistent with the previously established standards.

DOI: 10.54254/2755-2721/44/20230241

4.4. Performance Metrics

Based on the confusion matrix, six distinct evaluation metrics are utilized to evaluate the performance of the model. The detailed functions of accuracy, precision, recall and F1 score are described below. In addition, ROC curves and AUC values are utilized to evaluate the performance and predictive value of the classifier.

$$Accuracy = \frac{True\_Positive+True\_Negative}{Total} \tag{1}$$

$$Precision = \frac{True\_Positive}{True\_Positive+False\_Positive} = \frac{True\_Positive}{Total\_Predicted\_Positive} \tag{2}$$

$$Recall = \frac{True\_Positive}{True\_Positive+False\_Negative} = \frac{True\_Positive}{Total\_Actual\_Positive} \tag{3}$$

$$F1\_score = 2 \times \frac{Precision*Recall}{Precision+Recall} \tag{4}$$

5. Evaluation

After the finalization of the model, parameter, and metrics configurations, the experiment would commence. In this experiment, there are two baselines to evaluate performance. Implementing models on the same dataset, the normal model is the baseline to compare the influence with the model's combined emotion trend. For the same model on different datasets, the binary classification task would be the baseline to highlight that the incorporated model can lead to more impacts on the multi-emotion classification.

5.1. Results

According to the results in the IMDB dataset, it can be observed that incorporating sentiment scores into the model can potentially improve the computational speed by 66.7% while maintaining good performance. The CNN model achieved the fastest running speed by incorporating a sentiment score weight of 0.5 and running for 20 epochs. It showed a 33.42% improvement in running speed compared to the baseline. Additionally, the proposed model had a 36.67% chance of improving its performance. The highest improvement increased the accuracy by 2%, whose value was 80.17%.

From the results in the IESAR data, it was the evident that it can enhance the model's performance within a similar computational time as the baseline. Although it only accelerated the model computation in 30% of the experiments, it could maximize the model's running speed by 29.42%. Additionally, it can also provide a significant improvement in the model's performance. In the experiment where the BiLSTM model was trained with a weight value of 1.5 and ran for 50 epochs, the accuracy of the model increased by 2.6%, and the F1 score improved by 2.16%. Those values were respectively 58.55% and 58.31%.

Additionally, it effectively accelerates the computation of the model. From the ROC curve in Figure 2, the model's performance is significantly better than random guessing. Moreover, the AUC values obtained from the experiments are all greater than 0.8. It is particularly noteworthy that the maximum AUC value reaches 0.93. It represents that the model can more precisely classify texts in this emotion. Therefore, properly setting the weight values for emotion scores can enhance the predictive value of the model.

DOI: 10.54254/2755-2721/44/20230241

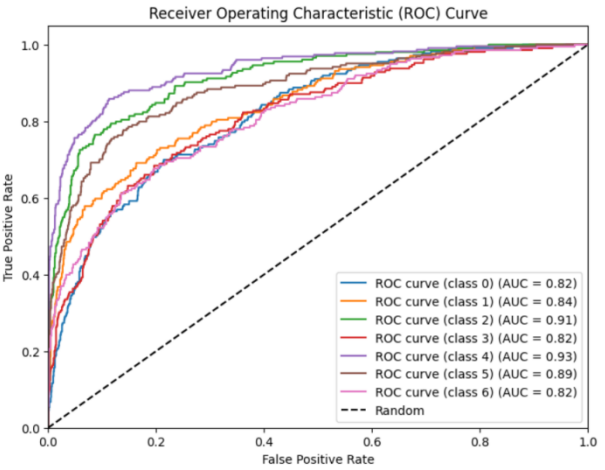

Figure 2. BiLSTM model with 1.5 weighting on ISEAR data.

5.2. Discussion

From all the experimental results, the proposed model can demonstrate a 60% improvement in the experimental outcomes of the multi-emotion classification compared to the baseline model. This advancement has been validated in 80% of the experiments conducted for binary emotion classification. Moreover, even in these experiments where the model's computational speed remained unchanged, there is still a notable enhancement of 40% in its performance. According to these, it can be concluded that the model incorporating emotion trends can enhance its computational speed and performance in the emotion classification task. However, the most improvement of the model lies in its computational speed. In 66.7% of the binary classification experiments, the model's performance has been enhanced to various extents. In addition, the model's performance only exhibits a maximum accuracy improvement of 2.6%, which is not a significant change. Furthermore, the AUC values of the experimental results indicate that the model demonstrates strong generalisation ability in classifying diverse emotions on the test date. By assigning appropriate weights, the proposed model further enhances its generalisation ability.

Therefore, this approach of combining sentiment scores can be effectively applied to sentiment classification tasks on large-scale data. It allows the model to alleviate the computational burden caused by enormous data while maintaining model performance. Moreover, with the assistance of sentiment scores, there is a certain probability of improving the accuracy of model predictions, particularly for lengthy texts.

5.3. Limitations & Future Work

From the information provided, it can be seen that the overall performance of this model is lower than the baseline model. This could be due to the fact that the sentiment score in the model is more likely to be learned as one of the feature values among many other features. Future work can attempt to make the sentiment score the standard for determining overall sentiment, allowing it to play a more decisive role in the final results. Furthermore, the word-level emotion trend lacks contextual information, resulting in less significant improvements. To explore the emotional characteristics of paragraphs or articles, future research can extend this analysis to the sentence-level emotion trend. This can provide a better combination of contextual information to calculate more accurate emotional trends. Eventually, adaptive learning of emotion score weights can be attempted in future work. The adaptive weights can be automatically learned by the model to adjust weight values based on the characteristics of the text and the emotional score's importance. This will allow the model to achieve maximum performance improvement.

DOI: 10.54254/2755-2721/44/20230241

6. Conclusion

In this paper, the concept of integrating emotional trends into multi-emotion analysis models for long-text is proposed and investigated for their association. This research employs ISEAR and IMDB as the dataset for the models. The training data is augmented with word-level sentiment scores using varying weights. Afterward, three models, namely BiLSTM, CNN and CNN+BiLSTM, are trained for 20 and 50 epochs. Meanwhile, each experiment would leverage six metrics to evaluate the model's performance and computation speed, respectively accuracy, precision, recall, F1 score ROC curve and AUC value. The findings reveal that the inclusion of sentiment scores can improve the computation speed of 66.67 models. The maximum improvement compared to the baseline model is 33.42%. Additionally, it can also enhance the performance of the models to some extent. The precise impact depends on the connectivity parameters, such as dataset and weight values. In the best case, the model's accuracy increased by 2.26% and the F1 score improved by 2.16%, without affecting execution speed.

References

[1] M. Rambocas and B. G. Pacheco. 2018. "Online sentiment analysis in marketing research: a review," Journal of Research in Interactive Marketing.

[2] D. Edmonds and J. Sedoc. 2021. "Multi-emotion classification for song lyrics," in Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 221–235.

[3] L. Yue, W. Chen, X. Li, W. Zuo, and M. Yin. 2019. "A survey of sentiment analysis in social media," Knowledge and Information Systems, vol. 60, pp. 617–663.

[4] W. N. Chan and T. Thein. 2017. "Multi-tier sentiment analysis system in big data environment," International Journal of Computer Science and Information Security (IJCSIS), vol. 15, no. 9, pp. 204–221.

[5] J. D. Teasdale, R. G. Moore, H. Hayhurst, et al. 2002. "Metacognitive awareness and prevention of relapse in depression: empirical evidence." Journal of consulting and clinical psychology, vol. 70, no. 2, p. 275.

[6] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. 2003. "A neural probabilistic language model. journal of machine learning research, vol. 3, no".

[7] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. 1998. "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324.

[8] J. L. Elman, 1990. "Finding structure in time", Cognitive science, vol. 14, no. 2, pp. 179–211.

[9] S. Hochreiter and J. Schmidhuber, 1997. "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780.

[10] M. Bouazizi and T. Ohtsuki. 2016. "Sentiment analysis: From binary to multi-class classification: A pattern-based approach for multi-class sentiment analysis in twitter," in 2016 IEEE International Conference on Communications(ICC). IEEE, pp. 1–6.

[11] M. Schuster and K. K. Paliwal. 1997. "Bidirectional recurrent neural networks," IEEE transactions on Signal Processing, vol. 45, no. 11, pp. 2673–2681.

[12] L. Zhang, Y. Lei, and Z. Wang. 2020. "Long-text sentiment analysis based on semantic graph," in 2020 IEEE International Conference on Embedded Software and Systems (ICESS). IEEE, pp. 1–6.

[13] K. R. Scherer and H. G. Wallbott, 1994. "Evidence for universality and cultural variation of differential emotion response patterning." Journal of personality and social psychology, vol. 66, no. 2, p. 310.

[14] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, 2011, "Learning word vectors for sentiment analysis," in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, Oregon, USA: Association for Computational Linguistics, pp. 142–150. [Online]. Available: http://www.aclweb.org/anthology/P11-1015