Review of text emotion detection

Ziheng Zhang *

Crestwood Preparatory College (CPC), Toronto, Canada

* Corresponding Author Email: evan.zhang@crestwood.on.ca

Abstract. Emotion is one of the essential characteristics of being human. When writing essays or reports, people will add their own emotions. Text sentiment detection can detect the leading emotional tone of a text. Text emotion detection and recognition is a new research field related to sentiment analysis. Emotion analysis detects and identifies emotion types, such as anger, happiness, or sadness, through textual expression. It is a subdomain of NLP. For some applications, the technology could help large companies' Chinese and Russian data analysts gauge public opinion or conduct nuanced market research and understand product reputation. At present, text emotion is one of the most studied fields in the literature. Still, it is also tricky because it is related to deep neural networks and requires the application of psychological knowledge. In this article, we will discuss the concept of text detection and introduce and analyze the main methods of text emotion detection. In addition, this paper will also discuss the advantages and weaknesses of this technology and some future research directions and problems to be solved.

Keywords: Emotion detection; Recurrent neural network; Convolution neural network; Related research

1. Introduction

Text emotion detection is an important Natural Language Processing (NLP) task. It is used to determine whether an article is positive, negative, or neutral. Humans have complex emotions, and texts have complex emotions. People will have joy, anger, sorrow, and joy, and there will be some delicate and subtle emotions, such as jealousy, shame, shame, pride, etc. These emotions are influenced by an interplay of factors such as mood, personality, purpose, and hormonal and neurotransmitters. Over the past 20 years, there has been a gradual increase in the study of emotions in a wide range of fields, including psychology, medicine, and computer science.

Since the beginning of the 20th century, artificial intelligence has gradually developed. Text emotion detection is a branch of artificial intelligence, and it is more important. Moreover, text emotion detection is a positive field. With some training and mental models, it will automatically look for emotional expressions in a text. Extract the opinions and ideas in the text, and text emotion detection can obtain the primary emotions of the article. This technology can play an essential role in translation systems, question answering systems, text summarization systems, and sentiment analysis.

The rise of social media has made this technology increasingly crucial to companies. People use blogs, video comments, social media, and other public platforms to talk about all kinds of things, whether a product or a service, or a news article. In these cases, the article will leave some emotional footprint on the person. Sentiment analytics can mine online expressions and emotions in the form of text and capture the speech of the target audience to understand people's attitudes towards these products. Some companies use Instagram, YouTube, Twitter, and other social media platforms to gather feedback about their products online. Getting more positive feedback through text recognition analysis shows that the product is valuable. Not only that, but it is essential to know this information before launching a product or service. Sentiment analysis can help salespeople understand consumers' opinions to make changes to products or services later.

Therefore, in the next part, we will introduce the background of text emotion detection in detail and some models and principles to complete the task. In addition, we will explain and review some of the new work on these technologies and explain what problems the current task will solve. Finally, we will speculate about the future development of text emotion detection and present some possible challenges in the current task.

2. Background of Emotion classification

2.1. Definition of Emotion Classification

Emotion classification is a way to distinguish and contrast one emotion with another. In the classification of emotions study, researchers classified emotions with two basic ideas. One is that emotions are discrete and fundamentally different structures. Second, emotions can be represented on the basis of grouping dimensions.

1) One is that emotions are discrete and fundamentally different structures. Second, emotions can be represented based on grouping dimensions. The subjective human experience is that emotions are discernible in ourselves and others. This obvious recognition tends to lead to the identification of many emotions that are fundamental to all human beings. But for experts, the debate among them questions this understanding of what emotions are. In terms of basic emotions, the activation of emotions is triggered by the brain's evaluation of the perceiver's goal or survival-related stimulus or event. It is a commontheme in many basic theories of emotion that we should be able to tell what a person's emotions are by looking at their brain activity.

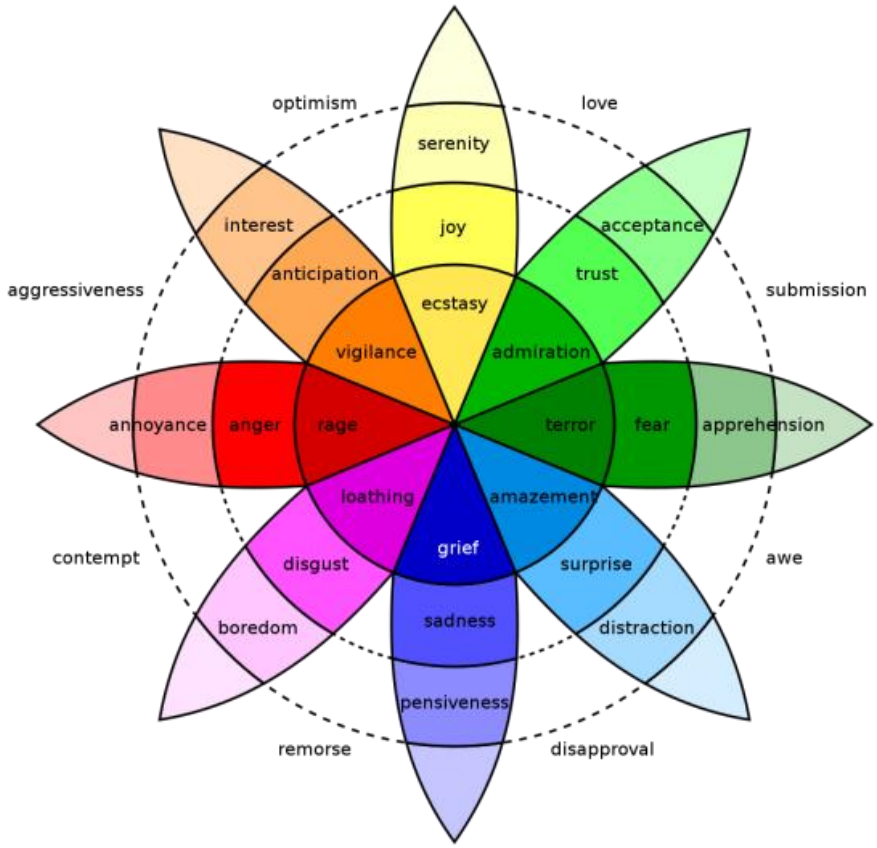

Fig 1. Robert Plutchik

For research and theory, in 1897, Wilhelm Max Wundt described emotions as having three dimensions: pleasant and unpleasant, aroused or inhibited, and tense or relaxed. In 1954, Harold Schlossberg named three dimensions: pleasure-displeasure, attention-rejection, and activation level. It was later discovered that the last two dimensions overlap. The most prominent models are the circumplex, vector, and Positive -Negative Activation (PANA) models. James Russell discovered it.Moreover, the model indicated that emotion was measured in a two-dimensional space, with arousal representing the vertical axis and the valence dimension representing the horizontal axis. Moreover, center tut stands for center-level arousal. This model is commonly used to test emotional words and facial stimuli.

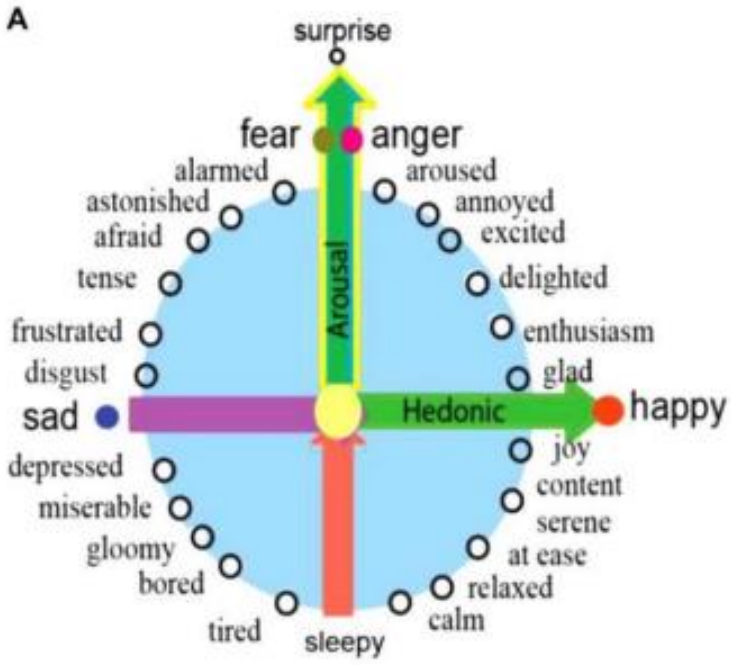

Fig 2. Circumplex Model

2) Compared with the dimensional model, the discrete model is more straightforward, so researchers are more inclined to use the discrete model to complete the classification task in text emotion recognition. However, the downside of discrete models is that they do not have a completely endless and broader range of emotional types. So dimensional models are recommended for projects that express emotional similarity.

2.2. Benchmarks

1) Text sentiment analysis is different from text mining and text classification because emotion is abstract. Therefore, it is complicated to process literal information directly for a long time, and it is possible to get the wrong result. The main tasks of emotion analysis include emotion information extraction, emotion classification, emotion retrieval, and induction.

Emotional information extraction refers to extracting valuable emotional information from the target text and finding out the tendentious elements in the text, such as the emotional expression, the evaluation object, and the emotional viewpoint. To extract opinion holders (ee), Kim proposed a recognition method of EE based on semantic role, which identifies EE through characteristics such as target word or phrase type and maximum entropy classification. The accuracy of this method is up to 78.7%. Carstens then proposed an opinion holder recognition system based on multiple models, thus improving the system's universality. Compared with another method, Support Vector Machine (SVM), the accuracy rate is improved by 5.6%. The evaluation object refers to the object described in the text, and it is also the carrier of emotion expressed by the emotional expression. Eirini proposed an extraction method based on adjective rating. This method extracts objects from nouns described by adjectives. As for inspirational words, they are words with emotional orientation. At present, the extraction of emotional words is mainly based on emotional dictionaries and rules. This kind of method is to obtain the evaluation words by analyzing the meaning relation between words, and then select the emotion words by comparing with WordNet thesaurus. In the film review data set experiment, the accuracy of this method can reach 77.17%.

Sentiment categorization, also known as sentiment orientation analysis, identifies subjective viewpoints of formulated texts. It can judge the positive and negative tendencies of text emotion. Paltoglou uses an emotion lexicon-based emotion classification method, using a variety of linguistic predictive functions such as negative words, capital letters, enhancement and attenuation of emotion, and polarity of emotion. In experiments on Twitter, MySpace, Digg, and other social media sites, the

method's accuracy reached 86.5%. Later, Qiu developed an emotion classification method based on syntactic analysis and an emotion dictionary, identifying emotional sentences from advertisements and extracting consumers' attitudes according to themes and keywords. Structural correspondence Learning (SCL) is used to reduce the noise of machine translation. The accuracy of this method is 85.4%. There are two kinds of emotion analysis methods based on machine learning: supervised learning and semi-supervised emotion analysis. The supervised learning approach regards effective classification as a common pattern classification problem for labeled documents. Pang applies supervised learning methods to emotion classification by comparing unitary features, binary features, and adjectives and has an accuracy rate of 82.9% in film reviews. The semi-supervised learning method uses a small number of labeled samples and many unlabeled samples for training. Ortigosa applied attitude analysis to the semi-supervised learning emotion classification method, optimized the attitude subjectivity of publishers and other indicators, and classified the emotions in sentences with an accuracy rate of 54%. Socher added tag information of emotion categories and proposed a recursive self-coding semi-supervised learning emotion analysis model based on this method. This method improved the accuracy of emotion prediction to 86.4%.

2) Rule Construction Approach: This approach lists the main grammatical and logical rules and makes it easier to detect emotion in text. For a few documents, this method can easily follow standard rules. However, large amounts of documentation can be complex to create logical rules. Therefore, the rule construction method includes keyword recognition (KR) and lexical affinity. Keyword recognition is the ability to find keywords in a document from its existing dictionaries, such as Wordnet-affect and DepecheMood. When an emotional word in a sentence is identified, it will be assigned a label that marks the emotion of the sentence. Morphological affinity (LA) adds a keyword recognition method. Because there are so many emotions, lexical affinities tend to place negative emotions, such as sadness, in the harmful category. This will reduce the time used for keyword identification. However, using this method may lead to some problems. For example, in a sentence, the author will write a positive emotion word, such as excellent, ironically, but the sentence's meaning is to show that this thing is not good. However, by both methods, the sentence will be labeled as "positive" because KR recognizes the word "excellent."

Kusen compared the performance of three emotional words and identified the best emotional words for a rule-centric approach to identifying emotions in social media texts. Using social software such as Twitter, they compared the performance of emotional words in the NRC, EmosenticNet, and DepecheMood, even though the words labeled in each dictionary had different emotions. They used ISEAR's data to conduct a survey. In this study, they asked individuals to assign their emotions to Twitter and then compared the Ground Truth score with scores from each of the three dictionaries. These results show that the classification performance of these three dictionaries is better than that of most dictionaries. The NRC is better at detecting anger, fear, and happiness. DepecheMood did a better job of detecting sadness. However, because the comparison method required human manipulation, they could not calculate more data through artificial allocation. The deficiency of these data will affect the performance of the measured model.

ML Approach: The text is classified from different emotional categories by the ML algorithm, which solves the ED problem. Detection is usually performed using some supervised or unsupervised ML technique. Supervised ML algorithms are widely implemented in text-based ED problems and provide better detection rates than unsupervised ML techniques. The supervised deep learning model has been widely used in text emotion detection in recent years. Deep learning techniques are more robust, and their depth can extract text that may carry text-based ED problems. These technologies are also considered superior to TUML [4].

2.3. Evaluation metrics

There are two main performance metrics for social emotion classification. One is micro-average F1 score, and the other is Pearson correlation coefficient. For MicroF1, d is e_{top}^{d}, and the predicted

emotion as $\hat{e}_{top}^{d}$. And these two values are interchangeable if they have the same probability in actual emotion distribution. The performance metric as follow:

$$\text{MicroF1} = \frac{\sum D \in D_{test}^{I_d}}{|D_{test}|} \tag{1}$$

Where:

$$I_d \begin{cases} 1, \hat{y}_{top} = y_{top} \\ 0, otherwise \end{cases} \tag{2}$$

And $|D_{test}|$ is the testing dataset. The bigger MicroF1 is, it also means that the system performs better at predicting mood tags. Pearson correlation coefficient (AP) is computed as follow:

$$\text{AP} = \frac{\sum D_{test} r(\hat{y}, y)}{|D_{test}|}, \tag{3}$$

Where:

$$r(\hat{y}, y) = \frac{\text{cov}(\hat{y}, y)}{\sqrt{var(\hat{y}) var(y)}} \tag{4}$$

r is for Pearson correlation coefficient between the predicted emotion distribution $\hat{y}$, and the ground truth distribution y. cov denotes the covariance operation [5].

3. Related neural network models

3.1. Recurrent Neural Network

RNN(Recurrent Neural Network) plays an essential role in text recognition. This neural model introduces the concept of time sequence into network structure design and makes it more adaptable to text sentiment analysis. In the large field of RNN, the LSTM model (Long Short-Term Memory) makes up for the defects existing in many other models, such as gradient disappearance, insufficient long-term Memory ability, and other problems, so that the recurrent neural network can use information more efficiently.

In a given sequence $x = (x_1, x_2, ..., x_n)$, using a standard RNN model, a hidden layer sequence $h = (h_1, h_2, ..., h_n)$ and an output sequence $y = (y_1, y_2, ..., y_n)$ can be computed iteratively.

In:

$$h = f_a(W_{xh} x_t + W_{hh} h_{t-1} + b_h) \tag{5}$$

$$y_t = W_{hy} h_t + b_y, \tag{6}$$

W is the loop coefficient matrix, for example, W_{xh} is the loop coefficient matrix from the input layer to the hidden layer. b is the offset vector, like b_h is the offset vector for the hidden layer. f is the activation function, like the tanh function. T is the time.

Traditional RNN can deal with nonlinear time series effectively, but there are still some problems. Because the gradient disappears, RNN cannot process time series with too long delay. Second, training RNN model needs to prioritize the length of delay window, but in practical application, it is difficult to automatically obtain the optimal value of secondary parameters. LSTM model solves these

problems well. In the LSTM model, RNN cells of the hidden layer were transformed into LSTM cells, which enabled them to have long-term memory ability. The most extensive LSTM model cell calculation method, z is the input module, can be expressed as:

$$\begin{aligned} i_t &= \sigma(W_{xi}x_i + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \\ f_t &= \sigma(W_{xf}x_i + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \\ c_t &= f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\ o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \\ f_t &= \sigma_t \tanh(b_t) \end{aligned} \tag{7}$$

In these formulas, i, f, c, o are input gate, oblivion gate, cell state gate and output gate. W and b are the corresponding weight coefficient matrix and bias term respectively. σ and tanh are sigmoid and hyperbolic tangent activation functions.

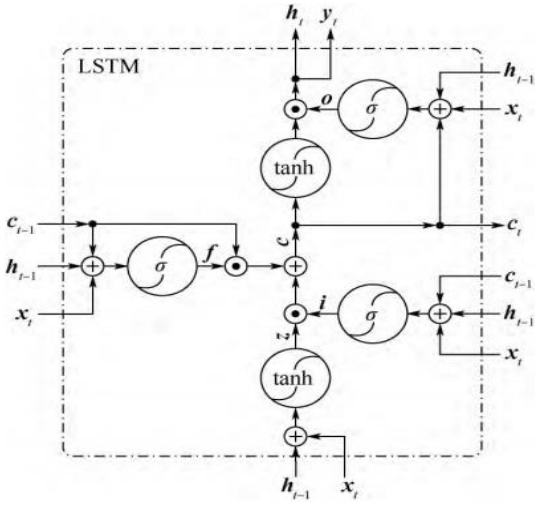

Fig 3. LSTM cell structure in hidden layer

BPTT algorithm, which is similar to BP (Back Propagation) algorithm, was used in the training process of the LSTM model. The algorithm can be divided into four steps. One is to calculate the output value of LSTM cells according to the forward calculation method. Second, the error value of each LSTM cell was calculated backward. Third, calculate the gradient of each weight according to the corresponding error direction; Fourth, the optimization algorithm based on gradient updates the weight [9].

3.2. Convolutional Neural Network

The basic structure of CNN consists of input layer, convolution layer, pooling layer, full connection layer and output layer. There will be several convolution layers and pooling layers on one side, and alternate Settings are adopted. One convolution layer is connected to one pooling layer, followed by another convolution layer. Since each neuron on the output feature plane in the convolution layer is locally connected with its input, the weighted sum of the corresponding connection weight and the local input and the bias value are added to obtain the input value of the modified neuron. This process is similar to the convolution process, hence the name CNN.

1) the convolution layer

The convolution layer is composed of multiple feature planes, and each feature plane is also composed of multiple neurons. CNN's convolution layer extracts different features of input through convolution operation. For example, the first convolution layer extracts low-level features, such as borders and lines. The higher the convolution layer, the more advanced features are extracted. The TWO-DIMENSIONAL and three-dimensional CNN can be expanded in this way. The top layer of

one-dimensional CNN is the pooling layer, the middle layer is the convolution layer, and the bottom layer is the input layer of the convolution layer.

In each convolution layer, each neuron will be connected to the local area of the feature plane of the next layer through a set of weights and then transferred to a nonlinear function such as the ReLU function through local weighting. From this point, the output of each neuron in the convolution layer can be obtained. In using one input feature surface and the same output feature surface, the weights of CNN share. Add and set the sliding step of the convolution kernel at the upper layer as one and the size of the convolution kernel as 1*3. In CNN, the size of the input feature surface of each convolution layer oMapN is:

$$oMapN = (\frac{(iMapN - CWindow)}{CInterval} + 1) \tag{8}$$

Among them, iMapN represents the size of each input feature surface. CWindow is the size of the convolution kernel. CInterval represents the sliding step size of the convolution kernel in the previous layer. Generally speaking, it is necessary to ensure that the formula is divisible, otherwise additional processing is required for the CNN. The number of parameters that can be trained for each convolutional layer, CParams, is:

$$CParams = (iMap \times CWindow + 1) \times oMap \tag{9}$$

oMap is the number of output feature surfaces for each convolutional layer, and iMap is the number of input feature surfaces. 1 is the offset, and the offset is also shared in the same output feature surface.

2) The pooling layer

The pooling layer also consists of multiple feature surfaces after the convolutional layer. But it does not change the number of convolutional surfaces, because each of its feature surfaces only corresponds to one feature surface of the previous layer. The pooling layer is designed to obtain features that are not deformed by reducing the resolution of the feature surface, and it plays the role of secondary feature extraction. Commonly used pooling methods include all values averaging and random pooling.

Random pooling has the advantage of max pooling because it avoids overfitting because of the randomness. Different neurons of the same feature surface of the pooling layer do not overlap the local receptive fields of the previous layer. However, overlapping pooling can also be used. Krizhevsky et al. used the overlapping pooling layer framework to reduce the top-1 and top-5 error rates by 0.4% and 0.3%, respectively, and the generalization ability was stronger. Let the output value of the n output feature surface l and neuron in the pooling layer be t_{nl}^{out}:

$$t_{nl}^{out} = f_{sub}(t_{nq}^{in}, t_{n(q+1)}^{in}) \tag{10}$$

t_{nq}^{in} represents the output value of the q neuron of the n input feature surface of the pooling layer. f_{sub} can be the function of taking the maximum value and the function of taking the mean value. The sliding window of the pooling layer on the upper layer also becomes the pooling core. CNN's convolution kernel and pooling kernel are equivalent to the implementation of Hubel-Wiesel model. The convolutional layer is the simple cell of this model, and the pooled layer simulates the complex cell. The size of each output feature surface of each pooling layer in CNN, DoMapN, is:

$$DoMapN = (\frac{oMapN}{DWindow}) \tag{11}$$

The pooled core has a size of DWindow. Pooling layer reduces the computation of network model by reducing the number of connections between convolutional layers.The pooled core has a size of DWindow. Pooling layer reduces the computation of network model by reducing the number of connections between convolutional layers [8].

4. Related researches

Most studies have considered multimodal analysis in recent years, but little attention has been paid to visual emotional information generated by the fusion of audio-visual emotional information at the feature or decision level. So Guangxia Xu, Weifeng Li, and Jun Liu did this study to capture users' emotions by extracting audio and text from videos. Moreover, this study also proposes a multimodal sentiment classification framework to capture users' emotions in social networks. Guangxi Xu said that a large amount of text on the Internet makes it difficult to analyze, and it is also challenging to find authentic comments on social media. So they tried to determine the person's emotion by looking at the words spoken and their facial expressions. When a person uses more vocal adjustments to express a personal opinion, the audio data often contains clues to the majority opinion. However, although multimodal analysis has been considered, little attention has been paid to visual, emotional information generated by the fusion of audio-visual emotional information at the feature level. Guangxi wants to solve the above problems by developing a multimodal fusion framework. No matter which method people use to express their emotions, they can effectively extract opinions from the user's conversation and accurately identify his emotions. In this study, they propose a new framework for multimodal emotion analysis and design a 3DCLS (3D convolution-long-short-term Memory) hybrid model to classify visual emotions, And a CNN-RNN hybrid model to analyze emotions in text. They used deep computer learning and a ConvLSTM (recurrent convolutional Long - and short-term Memory) neural network to build information about emotional and motion-recognition tasks. This study makes for the lack of emotion recognition in video and motion and makes emotion analysis more accurate [1].

Chang Wang and Bang Wang propose an end-to-end theme-enhanced self-attention network (TESAN) that codes documents SE Mantics and extracts document topics. It has been proved that document features and topic features can improve the performance of social emotion classification. However, there is still no research on extracting and utilizing these features more effectively. Chang Wang et al. also proposed a theme-enhanced self-attention mechanism that encodes semantic and topic information into document vectors. After that, the fusion gate combines the document vector with the topic embedding to form the document representation for the emotion classification cation. Experiments on three public data sets found that TESAN had higher classification accuracy and a higher Average Pearson correlation coefficient. Not only that, but they also found that TESAN ranked highest in computational efficiency in several data sets and could generate more coherent topics [2].

Xiangsheng Li and Yanghui Rao et al. developed a new semantically rich hybrid neural network model. They believe that the recent online comments are sparse, making it challenging to analyze the corresponding text sentiment. In addition, although a deep neural network can transform sparse low-level features into dense high-level features, which has been proved to be adequate for speech recognition and image analysis tasks, more research is needed to see whether it can achieve higher efficiency in emotion classification. Their model uses an unsupervised teaching model to integrate semantic domain knowledge into a neural network to improve its reasoning ability and interpretability. Xiangsheng Li studied and verified that their proposed method is superior to other advanced sentiment classification methods through three accurate media comment datasets [6].

In recent years, the classification method based on a neural network has improved the classification performance of mance compared with the previous lexical-affective method. However, it does not fully use some essential words and sentences to judge text features. Chang Wang, Bang Wang, and Minghua Xu propose a new neural network structure that utilizes syntactic information of sentences and topic distribution of documents. Based on the syntactic dependency Tree of sentences, the architecture constructs a Tree structure long and short memory (TreeLSTM) network and obtains a sentence vector. For multi-sentence documents, they use the chain-LSTM network to extract the document representation from the hidden state of the sentence. They also designed a topic-based attention mechanism with two attention levels. Studies on three public data sets show that the method is superior to the most advanced algorithms in terms of higher average Pearson correlation coefficients and MicroF1 performance [3].

5. Challenges and Futures

In the current research on text emotion recognition, there are still some problems to overcome or some areas that have not been studied. Research in this field is mainly divided into two stages, language representation, and classification. Context extraction is critical in linguistic representation because it improves classification accuracy. Although it is a mainstream rule construction method in the KR method, some errors may occur when dealing with some inspirational words. This method divides good words into positive ones and negative emotions into negative ones. Because the emotion words are directly classified to the two extremes, regardless of the context, this will lead to some exceptional cases of the emotion classification is not marked. For example, when the author chooses to "satirize" something, he will use some positive words but express antisense. So the significant problem is to introduce a technique to extract this context information from text. There is currently a converter-based embedded reference to indicate the extraction of context information. However, there are some limitations to using this Transformer. For example, when vocabulary limits are exceeded or complexity increases when analyzing multiple articles, Transformer is a significant disadvantage.

Some use of text-based ED has not been fully utilized in some critical applications, that is, in everyday life. For example, in the field of crime, information can be identified and analyzed by analyzing what the victim says to mitigate the crime. In medicine, the technology could also make some contributions, such as analyzing patients' information and speech to determine their psychological status and provide timely support [7].

6. Conclusion

This paper mainly discusses the concept of text detection and introduces and analyzes the main methods of text sentiment detection. This paper summarizes the development and new progress of text emotion detection in recent years and some related Network models, such as the Convolution Layer and Recurrent Neural Network. In addition, the advantages and disadvantages of this technology are discussed, as well as some future research directions and problems to be solved. In the future, text mood reduction can be more prone to civilian aspects, such as testing the patient's psychological and emotional state.

References

[1] Guangxia Xu, Weifeng Li, Jun Liu. "A social emotion classification approach using multi- model fusion", Future Generation Computer Systems, 2020.

[2] Chang Wang, Bang Wang. "An End-to-end Topic-Enhanced Self-Attention Network for Social Emotion Classification", Proceedings of The Web Conference 2020.

[3] Chang Wang, Bang Wang, Minghua Xu. "Tree- Structured Neural Networks With Topic Attention for Social Emotion Classification",IEEE Access, 2019.

[4] Nourah Alswaidan, Mohamed El Bachir Menai, A survey of state-of-the-art approaches for emotion recognition in text (2020).

[5] Abdullah Alsaedi, Phillip Brooker, Floriana Grasso and Stuart Thomason, A Survey of Social Emotion Prediction Methods.

[6] Xiangsheng Li, Yanghui Rao, Haoran Xie, Raymond Yiu Keung Lau, Senior Member, IEEE, Jian Yin, and Fu Lee Wang, Bootstrapping Social Emotion Classification with Semantically Rich Hybrid Neural Networks, 2017.

[7] Francisca Adoma Acheampong, Chen Wenyu and Henry Nunoo-Mensah, Text-based emotion detection: Advances, challenges, and opportunities. 2019.

[8] Feiyan Zhou, Linpeng Jin and Jun Dong, Review of Convolutional Neural NetWork, 2017.

[9] WANG X，WU J，LIU C，et al．Exploring LSTM based recurrent neural network for failure time series prediction［J]. Journal of Bejing iversity of Aeronautics and Astronautics, 2018, 44(4): 772-784