International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:14, No:12, 2020

A Survey of Sentiment Analysis Based on Deep Learning

Pingping Lin, Xudong Luo, Yifan Fan

Abstract—Sentiment analysis is a very active research topic. Every day, Facebook, Twitter, Weibo, and other social media, as well as significant e-commerce websites, generate a massive amount of comments, which can be used to analyse peoples opinions or emotions. The existing methods for sentiment analysis are based mainly on sentiment dictionaries, machine learning, and deep learning. The first two kinds of methods rely on heavily sentiment dictionaries or large amounts of labelled data. The third one overcomes these two problems. So, in this paper, we focus on the third one. Specifically, we survey various sentiment analysis methods based on convolutional neural network, recurrent neural network, long short-term memory, deep neural network, deep belief network, and memory network. We compare their futures, advantages, and disadvantages. Also, we point out the main problems of these methods, which may be worthy of careful studies in the future. Finally, we also examine the application of deep learning in multimodal sentiment analysis and aspect-level sentiment analysis.

Keywords—Natural language processing, sentiment analysis, document analysis, multimodal sentiment analysis, deep learning.

I. Introduction

THE rapid development and popularity of the Internet inevitably lead to a significant increase in the number of online data [1]. Many of the data are about opinions that people express on public forums such as Facebook, Twitter, Microblog, Blogs, and e-commerce websites. Notably, online comment texts on e-commerce websites reflect buyers' real feelings or experiences on the quality of the purchased goods, business services, and logistics services, regarding not only satisfaction information of consumers' shopping, but also their acceptance and expectation to new products or services. The insights into online comments significantly affect consumers' desires and decisions, which in turn impacts the efficiency of e-commerce platforms. Therefore, it is crucial to quickly mine and effectively take advantage of the comments. However, it is difficult to extract valuable information from these massive online texts. Therefore, the academic community and the industry pay lots of attention to sentiment analysis [2], [3].

Generally, sentiment analysis is to detect, analyse, and extract attitudes, opinions, and emotions expressed by people in a given dataset [4]. As a result, sentiment analysis is also called *opinion mining*, *orientation analysis*, *emotion classification*, and *subjective analysis*. Sentiment analysis tasks involve many problems in the field of natural language processing, including named entity recognition, word polarity disambiguation, satire detection, and aspect extraction. The number of problems involved in a sentiment analysis task is directly proportional to the difficulties users face in their application.

The term of sentiment analysis is coined by Nasukawa and Yi [5], but it is Pang and Lee [6] who first propose the task of sentiment analysis. They define the subjective calculation process of a text as sentiment analysis and opinion mining, yet they fail to give a more detailed definition of sentiment analysis. Later on, Liu [7] defines an emotion expression as a 4-tuple of (*Holder*, *Target*, *Polarity*, *Time*), where *Holder* means the opinion holder, *Target* refers to the object to be evaluated, *Polarity* stands for expressed emotion category, and *Time* is the evaluation time. Sentiment categories involved may vary with the sentiment analysis tasks. For example, in some sentiment analysis tasks, they are *positive* and *negative* only; in other tasks, they may be *positive*, *negative*, and *neutral*; or *happy*, *anger*, *sorrow*, and *fear*; or just 1-star to 5-star.

The significance of sentiment analysis of online commentary texts is twofold: 1) *Practical significance*. Online comments could be consumers' ones that support their decisions of buying a product or a service, or people's opinions about social issues, which are concerns of a government. Moreover, the impact of decision-making, online analysis of sentiment has important practical significance. Alfaro et al. [8] illustrate the effect of sentiment analysis technology for government and public institutions. Generally, the application of sentiment analysis on massive data can help to improve Internet's potential public opinion monitoring systems, expand the company's marketing capabilities and achieve detection of world anomalies or emergencies. Moreover, it can also be applied to the research fields of psychology, sociology, and financial forecasting. 2) *Theoretical significance.* The analyses of texts with emotions need the expertise of multiple disciplines such as natural language process, machine learning, and text classification are needed. They can help establish a reliable emotion dictionary, sub-area sentiment dictionary, and abundant corpus resources. Moreover, it can help to improve the accuracy of various classification algorithms.

The traditional methods for sentiment analysis are mainly based on sentiment lexicon or machine learning (ML), and use classification, regression, and other methods to achieve feature extraction and classification. These methods suffer data-sparse problems and the word order problem as well. Moreover, ML-based methods require a large number of labelled texts.

This work was supported by the National Natural Science Foundation of China (No. 61762016) and Guangxi Key Lab of Multi-source Information Mining & Security (No. 19-A-01-01).

Pingping Lin, Xudong Luo*, and Yifan Fan are with College of Computer Science and Information Engineering, Guangxi Normal University and Guangxi Key Lab of Multi-Source Information Mining & Security, Guangxi Normal University, Guilin 541004, China (*corresponding author, e-mail: luoxd@mailbox.gxnu.edu.cn).

Self-supervised learning solves the problem of a lack of training data. However, deep learning (DL) can avoid the cumbersome feature selection process, automatically abstract the features, learn the corresponding parameters, and capture sophisticated features. DL-based methods embed hidden layers between the input and output layers to simulate intermediate representations of data that cannot be learned by other algorithms. The mechanism of DL can effectively learn deeper information from high-dimensional data. Its architecture can be adapted to many types of tasks, including sentiment analysis of texts.

Although some researchers have surveyed the deep learning-based method for sentiment analysis, our survey in this paper is different from theirs. In 2018, Zhang et al. [9] first gave an overview of deep learning and then surveyed their applications in sentiment analysis. However, unlike our survey in this paper, theirs does not cover the literature after 2017. In 2019, Prabha and Srikanth [10] reviews sentiment analysis at the sentence level and aspect level, but most of their literature were before 2018. However, in this paper, we survey the state-of-art deep learning-based sentiment analysis at all the levels (the sentence level, the aspect/target level, and the chapter level). We even cover the sentiment analysis of multimodal (text, audio, images, and videos). Unlike ours, Soleymani et al. [11] just surveyed the work on multimodal sentiment analysis before 2017. Notably, we exmine the studies published in top conferences of AAAI and IJCAI in 2018-2020.

The rest of this paper is organised as follows. Sections II and III review sentiment analysis methods based on convolutional neural network and recurrent neural network (including long short-term memory), respectively. Section IV briefs some hybrid sentiment analysis methods based on these deep learning models. Section V gives some examples of the methods based on deep neural network, deep belief network, and memory network. Section VI outlines the application of deep learning techniques in multimodal sentiment analysis tasks. Finally, Section VII concludes the paper with future work.

II. CNN based methods

This section discusses some sentiment analysis methods based on Convolutional Neural Network (CNN).

A. Pure CNN Based Methods

Stojanovski et al. [12] design a deep learning system for sentiment analysis of Twitter messages. It uses a CNN to extract the characteristics of message texts and fuses different classification algorithms for emotion classification experiments. Their system does sentiment analysis of news-related Twitter messages to provide insights into the public's responses to certain events. A specific application of Twitter messages related to several games in the 2014 FIFA World Cup to identify the emotions expressed by the audience. Kim and Yoon [13] use CNN to do sentiment analyses through text vectorisation. However, the above two methods have some disadvantages. For example, the pure CNN model gives up the relationship between contexts during the training process, so cannot solve the timing problem well, and cannot accurately analyse data such as transitional sentences, either.

B. WP-CNN Based Method

To improve the quality of input text representation, He et al. [14] construct a word, part of speech pairs-CNN (WP-CNN) model. It considers part of speech features, and use such features for word meaning elimination to improve word vector training. They employ a dual-channel for input. One channel retains the matrix input of the original text word vector. The other inputs the improved word vector. This method can enrich the input characteristics of the model and solve the problem of network training over-fitting caused by excessive input noise. The dual-channel input solves the problem of overfitting the network training caused by excessive input noise. Its effect of sentiment analysis is better on English data sets than on Chinese ones. Further work could be to apply WP-CNN to other NLP (natural language process) tasks, especially on Chinese data sets.

C. RCNN Based Method

Sun and Wang [15] propose a deep learning based method for sentiment analysis. It uses the Regional CNN (RCNN) to preserve the temporal relationship of sentences capture more semantic relationships between words. Thus, they solve the problem that the traditional neural network model has less connection between sentences and less semantic information between words when dealing with the aspect-based sentiment analysis task. RCNN training time is shorter than that of ATT-LSTM (the latter is three times the former). It is suitable for aspect-based sentiment analysis and has good adaptability.

D. Char-DCCNN Based Method

Chen et al. [16] propose a sentiment analysis method, called character embedding with a Dual-Channel Convolutional Neural Network (char-DCCNN). They divide Chinese corpus into single Chinese characters and then train them as character vectors. Sequentially, the vector-matrix representing a text is input into a dual-channel CNN. Their experiments show that their method improves the sentiment category results of Weibo Chinese short comments, and so is significant in practice. Char-DCCNN can obtain good classification performance with a small amount of labelled data and a small number of iterations. However, the character embedding increases the complexity of the distributed representation of text and increases the computational cost. In future work, it is worth making the model achieve less running time and better performance.

III. RNN Based Methods

This section discusses some sentiment analysis methods based on recurrent neural networks (RNN).

A. RNN Based MMethod

Shenoy and Sardana [17] propose a RNN based method to track the context of a conversation, interlocutor states, and the emotions conveyed by the speaker in the conversation. Their method outperforms some state-of-the-art methods on two benchmark datasets (CMU-MOSI and CMU-MOSEI) in terms of accuracy and regression metrics. Compared with the current multi-mode functional systems, their methods can utilise and capture the context of the dialogue through all mode data. In the future, it is worth increasing the number of modal data-level dialogue participants. Currently, there are very few labelled multimodal data sets available, so it is also worth constructing labelled multimodal data sets.

Hao et al. [18] construct a fine-grained opinion analysis model based on sequence annotation using Bidirectional RNN (BiRNN). By fusing the word vectors, parts of speech and text dependencies, and other linguistic features, their model can learn the modification and semantic information of texts, and meanwhile, extract the attribute entities to judge the emotional polarity of a text. In terms of accuracy, recall, and F1 value, the BiRNN method is obviously better than other RNN methods. Besides, their model can simultaneously do attribute extraction and attribute polarity analysis, which others cannot do.

B. LSTM Based Method

RNN persists information through loops (*i.e.*, using previous information to connect to the current task, and guessing the current text with past texts). Meanwhile, RNN has an obvious *long-term dependency* problem: when a historical text is too long, the effective information of the text cannot be saved. Considering the time series information between words, Pei and Wang [19] propose a model that integrates the part of speech attention mechanism with Long Short-Term Memory (LSTM) [20] (actually a special RNN). Their model can make most use of the relationship between target emotion words and emotion polar words in sentences.

C. BiLSTM Based Method

Although LSTM can solve long-term dependence problems well, it cannot predict the information followed, and cannot judge the emotion polarity transfer, either. To address the problems, Chen [21] proposes a BiLSTM model with a polar transfer. Compared with the traditional LSTM, the model effectively improves the representation ability of the model for text and improves the accuracy of emotion classification. However, it increases the computational complexity, so needs more time to train the natural network.

Peng et al. [22] propose a method for aspect sentiment triplet extraction. Particularly, their method can extract triplets (What, How, Why) from the inputs, which show WHAT the targeted aspects are, HOW their sentiment polarities are, and WHY they have such polarities. More specifically, first, they obtain many aspects with polarised emotions and a bunch of opinions (hygienic/waiter, positive/negative, clean/general/dirty/poor, good/bad attitude). Then, they match all the aspects with corresponding opinions (hygienic, clean), (waiter, bad attitude). They employ BiLSTM encoders to encode sentence-level contexts as aspects and opinions. Their method outperforms a few strong baselines adapted from state-of-the-art studies. For the problem of outputting multiple tuples, including invalid tuples, they try to set heuristic rules to constrain the pairing algorithm to only output a certain number of triples (*e.g.*, equal to the number of extracted aspects) according to the classification probability, but find that the improvements did not help. So in the future, it is worth studying how to constrain the paired triples.

D. AF-LSTM Based MMethod

Tay et al. [23] propose an aspect-based sentiment analysis method, called Aspect Fusion LSTM (AF-LSTM). It incorporates aspect information into neural models by modelling word-aspect relationships. To do this, it adds a layer of modelling in the attention layer to learn the word-aspect context, and introduce a novel association layer to model relationships so that the attention layer can focus on learning the relative importance of the fused context words. The weakness of this method is that it is difficult for the connection operator to model the relationship between aspects and context words. However, their method can effectively switch different aspects of focus words.

E. Sentic LSTM Based Method

Inspired by Xu [24], Ma et al. [25] extend LSTM, called Sentic LSTM, by integrating a hierarchical attention mechanism consisting of target-level attention and sentence-level attention. The goal-level attention learning focuses on how to recognise the target expression part with a high emotional significance, and the sentence-level attention searches for the target and aspect-related information. The novelty of this method is that commonsense knowledge and emotion-related concepts are incorporated into the deep neural network for emotion classification. Their experiments on data set SentiHood show that the attention model significantly improves classification accuracy. However, the experiments on dataset Semeval-2015 show the improvement is relatively small. The reason may be that SentiHood shields the target as a special “LOCATION”, with a small number of instances but many aspects. Therefore, they speculate that their model is unsuitable for small and sparse data sets. In the future, it is worth optimising the attention layer mechanism according to the characteristics of a data set, or keep the model unchanged but reconstruct the semantic network of a data set.

F. AA-LSTM method

Xing et al. [26] propose an Aspect-Aware LSTM (AA-LSTM) for sentiment analysis, which incorporates aspect information into LSTM cells in the context modelling stage. So, their method can select essential information about a given target and filter out irrelevant information through information flow control. So, it uses aspects to improve the information flow and can generate contextual vectors that are more efficient than classic LSTM. Their experiments on SemEval-2014

Datasets confirm its effectiveness. This is because it can dynamically generate aspect-aware context representations, retain valid information in a given aspect while filtering out useless information in a given aspect, and its resulting final emotional representation is more effective.

IV. Hybrid Methods

This section briefs some sentiment analysis methods based on multiple deep learning models. The purpose of hybrid methods are to gain better performances.

A. CNN and RNN Vased Method

Wang et al. [33] present a hybrid method based on CNN and RNN for sentiment analysis of texts. The method use CNN to generate the coarse-grained local features, and RNN to learn long-distance dependencies. Their method outperforms a state-of-the-art method on three benchmark corpora MR, SST1, and SST2. In the future, it is interesting to extend the method for sentiment analysis of long texts.

Luo and Wang [34] integrate RNN with CNN to propose a hierarchical neural network H-RNN-CNN as a general model to represent text in sentiment analysis. More specifically, because information may lose in a long text, they split the text by sentence and use them as middle layers, and then they use RNN to process sequences. They use CNN to capture the relationship between sentences. Their experiments show that their method works well on several datasets. In the future, it is worth introducing attention mechanisms to examine the impact of different sentences on the emotional sentiment of texts.

B. CNN and LSTM Based Method

Kwaik et al. [35] propose a deep learning method for dialectal Arabic sentiment analysis. They use a hybrid model based on LSTM and CNN, with more convolutional layers. In the data representation, if some words in a text have no word embedded in the pre-training model, the most similar word vector in the pre-training model is embedded as its word. It achieves an accuracy between 81% and 93% for binary classification and 66% to 76% accuracy for three-way classification. The results are better than those of the two baselines. And when facing unbalanced data sets, its high accuracy is still maintained. In the future, to improve the performance of this model, it is worth training special word embeddings for sentiment analysis, introducing attention mechanisms to increase the complexity of the model structure.

Rehman et al. [36] propose a hybrid model of LSTM and very deep CNN for sentiment analysis. More specifically, they use Word to Vector approach to translating the text strings into a vector of numeric values, calculate distances between words, and similar group words based on their meanings. Then they integrate the set of features that are extracted by convolution and global max-pooling layers with long term dependencies. On the datasets of IMDB movie reviews and Amazon movie reviews, compared with CNN based methods or LSTM based methods, their the model has higher accuracy, recall rate, and F1 value. However, their model is more suitable for small data sets with more parameters.

C. CNN and BiLSTM Based Method

Lui et al. [50] propose a hybrid model, which extracts local features of text vectors by CNN, extracts global features related to text context by BiLSTM, and fuses the features extracted by the two complementary models. More specifically, the pre-processed sentences are input into the hybrid neural network for training. The trained model can automatically classify the sentences according to emotions the sentences reflect. Their method has high accuracy and good robustness when the sample size is seriously unbalanced.

Jain et al. [37] present a method for sarcasm detection, which uses deep learning in code-switch tweets, specifically the mash-up of English with native Indian language, Hindi. More specifically, their model is a hybrid of BiLSTM with a softmax attention layer and CNN for real-time sarcasm detection. To evaluate the performance of their model, they randomly extract real-time mash-up tweets on the trending political (*government*) and entertainment (*cricket*, *bollywood*) on Twitter, which contains 3000 sarcastic and 3000 non-sarcastic bilingual Hinglish tweets. Their evaluation experiment shows that their model outperforms the baseline deep learning models with a superior classification accuracy of 92.71% and an F measure of 89.05%. Their model depends on the efficiency of online language identifiers and part-of-speech taggers, but the efficiencies of both are not high.

V. Other DL-Based Methods

Deep Learning (DL) methods include not only CNN, RNN, and LSTM but also others such as Deep Neural Network (DNN), Deep Belief Network (DBN), and Memory Network (MN). This section will give some examples of sentiment analysis methods based on these DL methods.

A. DNN Based Method

Wang et al. [27] propose a method for aspect-level sentiment analysis, which can integrate the output of the DNN with the implications of linguistic hints. The DNN has limited capabilities for modelling language prompts, and the proposed method improves this. They empirically evaluated the performance of their method on various benchmark datasets, and find that compared to the state-of-the-art DNN based method, theirs can effectively improve polarity detection accuracy by considerable margins. Since the current DNN still uses labelled training data, in the future it is worth finding a way in which there is no need for labelled training data.

B. DMN Based Method

Recently, the cold-start problem attracted attention. Similar to ABSA methods, Yang et al. [51] ignore the existing cold-start problem [52], which makes CEA more problematic than helpful in reality, and the test time is the longest. So, in the future, it is worth improving the accuracy of entity-aspect sentiment matching, and try to modify the halt conditions when updating the entity and aspect memory. They hope that to reduce the number of hops for simple samples but

TABLE I
ASPECT-LEVEL SENTIMENT ANALYSIS BASED ON DEEP LEARNING

Reference	Algorithm	Data set	Aspect	Sentiment category	Performance
Sun and Wang [15]	RCNN-BGRU-HN	SemEval-ABSA16: restaurants and laptops	Food quality, service, ambiance, price fairness, hygiene, screen, color, memory, battery time	Positive, negative, and neutral	restaurants: positive F1 88.22, negative F1 76.99, neutral F1 67.42; laptops: positive F1 81.25, negative F1 74.03, neutral F1 60.62
Tay et al. [23]	AF-LSTM	SemEval 2014 task 4 and SemEval 2015 task 12	Aspect term and aspect category	Positive and negative	Macro averaged results: 80.51%
Ma et al. [25]	Sentic LSTM	SentiHood and Semeval 2015	Aspect term and aspect category	Positive, Negative, and None	SentiHood: aspect categorization accuracy 67.43%, macro F1 78.18%; sentiment accuracy 89.32%; Semeval-2015: aspect categorisation accuracy 67.34%, macro F1 76.44%; sentiment accuracy 76.47%
Xing et al. [26]	AA-LSTM	SemEval-2014: laptop and restaurant reviews	Aspect terms: entities. Aspect category: dishes, environment, service, hygiene.	Positive, neutral, and negative	ATSA: laptop F1 Macro 68.47%, accuracy 73.20%; restaurant F1 Macro 68.71%, accuracy 79.29%; ACSA: restaurant F1 Macro 75.00%, accuracy 84.69%.
Wang et al. [27]	Linguistic Hints and DNN	SemEval 2015 task 12, SemEval 2016 task 5, including phone, camera, laptop and restaurant	Battery, screen, color, resolution, service, food quality, ambiance, and price fairness	Positive or negative	ACSA: phone accuracy 80.89%, Macro F1 80.15%; camera accuracy 88.10%, Macro F1 84.47%, laptop accuracy 85.60%, Macro F1 84.28%, restaurant accuracy 89.09%, Macro F1 85.72%; ATSA: laptop accuracy 86.19%, Macro F1 84.65%, restaurant accuracy 89.68%, Macro F1 84.12%
Aydin and Gungor [28]	Recursive and RNN	SemEval-2014 Task 4, including laptop and restaurant	color, resolution, service, hygiene, environment and so on.	Positive, negative, neutral	laptop: accuracy 76.15%; restaurant: accuracy 80.90%.
Ishaq, Asghar, and Gillani [29]	Hybridized CNN and GA	hotel reviews,[1] automobiles reviews,[2] movie reviews[3]	Environment, hygiene; appearance, color, performance; image quality, dynamics, smoothness and so on.	Positive and negative	accuracy 95.5%, precision: 94.3%, recall 91.1%, F measure 96.6%.
Jia, Bai, and Pang [30]	Hierarchical Gated -DMN	SemEval-2015: laptop and restaurant	Quality, color; price, food, hygiene and so on.	Positive, negative, and neutral	laptop: accuracy 74.82%, F1 72.53%; restaurant: accuracy 81.95%, F1 74.07%.
Liu et al. [31]	Multilingual GRCNN-HBLSTM	SemEval 2016 Task 5: cameras, laptops, restaurants, phones, hotels	Pixel, resolution; screen, performance; dishes, service; color; hygiene, service and so on.	positive, negative and neutra	cameras: positive precision 94.04%, F1 87.17%. negative precision 65.76%, F1 72.56%. Laptops: positive precision 92.28%, F1 87.16%. negative precision 75.36%, F1 78.99%; Restaurants: positive precision 97.69%, F1 92.94%, negative precision 70.68%, F1 78.45%.
Zheng et al. [32]	A neural network based on Syntax Graph	Rest14, Laptop, Twitter, Rest16	Dishes, environment, hygiene, service; Performance, battery, memory, screen, color	Positive, negative, and neutral	Rest14: accuracy 83.8%, F1 76.9%; Laptop: accuracy 78.2%, F1 74.3%; Twitter: accuracy 74.4%, F1 72.6%; Rest16: accuracy 89.6%, F1 71.2%.

increasing the number of hops for hard samples can improve both effectiveness and efficiency.

Aiming at the cold-start problem faced by the multi-entity-based sentiment analysis (ME-ABSA) task, Song et al. [52] propose a novel and scalable cold-start aware deep memory network (CADMN) framework. The basic idea of the framework is: if the model does not have enough information to create a good representation of an entity/aspect, then they will try to use representations derived from other entities/aspects to enhance it in relation to the cold-start target. The model is applied to the ME-ABSA mission. One of the challenges of the framework is how to find the target that contributes the most to the cold start target. As a result, they introduce a goal attention mechanism that aims to focus on most relevant goals, which can provide useful supplementary information to enhance the representation of cold-start goals. The model is applied to ME-ABSA. The experimental results on the English review dataset and the public Chinese review dataset show that the proposed framework is better than the benchmark of the latest technology (LSTM, CEA, and so on.) ME-ABSA task. The authors say that in the future, they will try to study the cold start problem by using other product description information.

C. DBN Based Method

Xiao et al. [53] propose an unsupervised emotion recognition algorithm based on an improved deep belief model, and integrate the algorithm with probabilistic linear discriminant analysis. They design a new type of inter-class dispersion matrix to solve the rank limitation problem in the traditional linear discriminant analysis method. Then they use the improved linear discriminant method to initialise the weight matrix between the last hidden layer and the classification layer of the deep confidence network, replacing the randomly generated weight matrix. The recognition rates of their method are 78.26% and 94.48% on the JAFFE database and the Extended Cohn-Kanade database, respectively. In the

TABLE II
SENTIMENT ANALYSIS TASKS ON TWITTER OR WEIBO

Reference	Algorithm	Social media	Data set	Sentiment category	Performance
Stojanovski et al. [12]	CNN	Twitter	News-related tweets, 2014 FIFA World Cup tweets, and social hotspots	Love, joy, surprise, anger, sad, fear, and thankful	F1 64.88%, Accuracy 58.84%.
Jain et al. [37]	CNN+ BiLSTM	Twitter	Real-time mash-up tweets are extracted on the trending political and entertainment posts on Twitter	Sarcastic and non-sarcastic	accuracy 92.71%; F measure 89.05%
Wang and Hu [38]	Attentional-GNN	Twitter	Real-world dataset concerning the 2016 presidential election of America	Hillary-for, Hillary-neutral, Hillary-against; Trump-for, Trump-neutral, Trump-against.	Trump: accuracy 0.9462, precision 0.9476, recall 0.9462, F1 0.9463; Hillary: accuracy 0.9539, precision 0.9550, recall 0.9539, F1 0.9538
Yang et al. [13]	Extended vocabulary-CNN	Weibo	Data is from NLPCC2017 competition website.	Six classes: others, like, sad, disgusting, angry, and happy	accuracy: 97.06%
Wu et al. [39]	Transformer Based MN	Weibo	Data set created by the author,which contains four aspects: traffic, service, price and environment.	Positive, negative, neutral	Accuracy: 61.67%
Ling et al. [40]	ELMo-CNN-BiLSTM	Sina Weibo	Data come from the fourth National Conference on Social Media Processing (SMP 2015)	Happy, confident, optimistic, brave, sweet, sadness, wretched, fear, frightening, annoying.	set A average: precision 88.22%, recall 89.06%, F1 88.06%; Set B average: precision 90.99%, recall 91.27%, F1 91.43%; Set C average: precision 85.69%, recall 86.33%, F1 86.71%.

TABLE III
SENTIMENT ANALYSIS TASKS ON E-COMMERCE PLATFORM'S PRODUCT REVIEWS

Reference	Algorithm	Platform	Data set	Sentiment category	Performance
Kumar, Yadava, and Roy [41]	-	e-commerce website	EEG dataset	Positive, neutral, negative, and compound score	RMSE 0.29, Recall 0.72
Mukherjee et al. [42]	RNN, GRU, LSTM and BiLSTM	Amazon	Amazon Review dataset,[4] including Amazon Instant Video and Musical Instruments	1-5 stars	Amazon Instant Video accuracy: RNN 56.67%, GRU 60.21%, LSTM 60.73, BiLSTM 62.98%; Musical Instruments: RNN 67.17%, GRU 67.07%, LSTM 66.88, BiLSTM 68.14%.
Shrestha and Nasoz [43]	GRU-RNN	Amazon	Amazon.com Reviews [44]	1-5 stars	review embedding accuracy 81.29%, precision 58.61%, recall 40.85%; combine product embedding and review embedding accuracy 81.82%, precision 59.45%, recall 42.52%.
Chauhan et al. [45]	A analysis of adverb types	Amazon	data crawled from Amazon by using. Net crawler, including office products and DVDs	1-5 stars	general superlative adverbs F measure: 0.86, degree-wh adverbs F measure: 0.80.
Shenoy and Sardana [17]	Context-Aware RNN	-	CMU-MOSI and CMUMOSEI	Anger, disgust, fear, happy, sad, and surprise	CMUMOSEI: text + audio F1 79.88% Acc 80.18%; text + audio + video: F1 80.01%, accuracy 82.10%.

future, it is worth transplanting this network to an embedded system to test its effect.

D. MN Based Method

Shen et al. [49] propose a dual user and product memory network model to learn user-profiles and product information for reviews classification using separate memory networks. More specifically, they use the two memory networks together for sentiment analysis. On three benchmark datasets IMDB, Yelp13, and Yelp14, their model outperforms state-of-the-art unified prediction models. However, their model has some shortcomings. For example, its performance highly depends on the number of documents related to a specific user or product. When the number of documents is insufficient, the performance of the model drops sharply. The possible solution is to obtain more available documents from similar users or similar documents. Of course, the shortcomings of the model provide the opportunity for future study. The author points out two directions for future work. One is to study the contribution of user-profiles and product information in sentiment analysis tasks, and the other is to find a way to merge the knowledge base to further improve the effect of model classification.

E. GCN Based Method

Nian et al. [54] prove the use of graph convolutional networks to extract facial expression features. Zhang et al. [55] use graph neural networks to model the context of sentiment analysis tasks. It can be seen that graph convolutional networks show exciting results in various applications [56]. Benssassi et al. [57] believe that graph convolutional networks can

TABLE IV
SENTIMENT ANALYSIS TASK OF MOVIE REVIEWS, RESTAURANT REVIEWS OR HOTEL REVIEWS

Reference	Algorithm	Field	Data set	Sentiment category	Performance
Cheng et al. [46]	MC-AttCNN-AttBiGRU	Movie reviews	IMDB, Yelp 2015, MR and CR	Yelp 2015 five levels of ratings from 1 to 5; IMDB and MR: positive and negative; CR: whether an opinion is expressed	IMDB: accuracy 91.70%, Yelp 2015: accuracy 92.90%, MR: accuracy 83.89%, and CR: accuracy 87.23%.
Rehman et al. [36]	Hybrid CNN-LSTM	Movie reviews	IMDB movie review dataset and Amazon movie reviews dataset	Positive and negative	accuracy 91%
Tran, Ba, and Huynh [47]	BiLSTM-CRF	Hotel reviews	Trip Advisor[5]	1-5 star, positive, negative and neutral	ATE-Precision: 0.8693, ATE-Recall:0.8772, ATE F1 score: 0.8732.
Al-Smadi et al. [48]	Deep-RNN	Hotel reviews	SemEval-ABSA16	Positive and negative	accuracy: 87%, execution time 0.17 hour
Shen et al. [49]	Dual memory network	Movie review, restaurant review	IMDB, Yelp13 and Yelp14	-	IMDB: accuracy 0.539, RMSE 1.279, MAE 0.734; Yelp13: accuracy 0.662, RMSE 0.667, MAE 0.375; Yelp14: accuracy 00.676, RMSE 0.639, MAE 0.351.
Ishaq, Asghar, and Gillani [29]	Hybridized CNN and GA	Hotel reviews, movie reviews	hotel reviews,[6] movie reviews[7]	Positive and negative	Accuracy: 95.5%, precision: 94.3%, recall: 91.1%, F measure: 96.6%.

help sentiment analysis tasks. They proposed the use of graph convolutional neural networks for multimodal sentiment analysis tasks, and they also faced the problem of extracting features of image and video data. In this study, they chose to spike neural networks [58] to model the multimodal data interaction information, that is, to build a graph network to model the interaction of neurons, and then use graph convolution and neural network synchronization to model. Spiking neural networks can effectively extract the features of various modal data in an unsupervised learning manner, where the features are represented by the spike pattern of the gods. Their model uses unsupervised STDP learning and uses the effectiveness of SNN's neural network and graph convolution to generate better feature representations and multimodal data interactive information modelling.

On the eNTERFACE05 dataset, Ryerson Audio-Visual Database achieves 98.3% and 96.82% accuracy rates. The experiment indirectly shows that the neural synchronisation calculation with spike timing and stimulation can integrate audio and video data, and modelling the interactive information of different modal data helps improve the performance of the model. The robustness of Graph Convolutional Networks on more massive data sets needs to be discussed. Another issue worth considering is that for the data representation problem after fusion, graphics can maintain the relationship between different data modalities in time and space.

F. Discussion

The current mainstream deep learning models based on aspect-based sentiment analysis (ABSA) are LSTM, CNN, RNN and DNN, and have achieved good performance. Table I makes a detailed analysis and comparison. Tables II-IV compare the sentiment analysis effects of deep learning-based mthods in social media (Weibo, Twitter), e-commerce platforms, and entertainment consumption. Most of ABSA's research is on sentiment analysis in areas such as notebooks (performance, memory), restaurants (Food quality, service, ambience, price fairness), and hotels (services, environment, hygiene), and finally obtain positive, negative and neutral polarity. More fine-grained sentiment such as happy, angry, satisfied, disappointed, and so on need to further research. For implicit aspects, deep learning models may need to combine rule-based, semantic similarity and machine learning-based methods to achieve good results. In addition, mixed models may bring negative effects, and we should think about how to improve the negative effects of mixed models on model performance. We believe that defining rules and applying a model that complements the defined rules will be a solution. The source, size and label of the data set have a great impact on the performance of the model. We should continue to seek the applicable data set and data set size for different models. More importantly, other methods of deep learning technology also have significant results in natural language processing tasks. In the future, we can try to apply fuzzy logic rules or reinforcement learning in ABSA.

There are many open topics, such as domain adaptation in sentiment analysis, pre-trained models. The state-of-the-art research shows that based on adversarial learning can effectively solve the domain adaptation problem. Xue et al. [59] use deep adversarial mutual learning to solve the domain adaptation problem. Dai et al. [60] choose the method of adversarial training, Lin et al. [61] propose multi-source sentiment generative adversarial Network, for visual sentiment classification. Wan et al. [62] propose a new neural network model based on the pre-trained language model BERT for joint detection of target-aspect-sentiment triples. Pre-training models will play an essential role in aspect-level sentiment analysis tasks. The multi-label classification task in sentiment analysis is also an important research topic. Fei et al. [63] propose a latent sentiment memory network for multi-label sentiment classification in a single sentence. The memory network can write and learn the distribution of emotion features without external knowledge. However, there are few types of research on sentiment analysis of multiple labels.

VI. Multimodal Sentiment Analysis

The methods mentioned in previous sections are concerned with texts only. However, sometimes sentiment analysis is concerned with not only texts but also images because a picture is worth a thousand words. This section will discuss the studies of this kind.

A. Fusion of Text and Image

How to effectively use multimodal representations (such as images, audio, and video) for sentiment prediction is still a core research challenge in multimodal field [72]. Xu et al. [73] propose a method for aspect based multimodal sentiment analysis to capture the impact of aspects on texts, images, and their interactions. The method consists of two interactive memory networks to supervise the textual and visual information for a given aspect. The method learns not only the interactive influences between multimodal but also the self influences in single-modality data. They construct new publicly available multimodal data sets from ZOL.com (one of the leading IT information and business portals in China). Their method outperforms representative text-level sentiment analysis methods and the latest variant of multimodal sentiment analysis that can capture multiple correlations on aspect-based multimodal sentiment data. However, the accuracies of traditional baselines and their method are both lower than 62%. So, in all traditional baseline methods with texts and images, there is much room for improving their accuracies. Their work falls into the intersection of aspect level and multimodal sentiment analysis. However, So far, there is not much work done on the intersection of aspect-level and multimodal sentiment analysis. Therefore, in the future, it is worth doing more research on the intersection of aspect-level and multimodal sentiment analysis.

Cai and Xia [74] propose a sentiment analysis method based on CNN graphic fusion. They use the features of image and text both to construct a CNN model, then analyse the semantic features of different levels (words, phrases, and sentences). On Flickr (a photo-sharing social networking site where users can tag and describe uploaded photos) datasets and Twitter (pictures and texts on microblog) data show that compared with traditional CNN models, their model increases the accuracy by 5.6%. In the future, it is worth optimising the model by selecting different combinations of features.

B. Fusion of text, mage, and video

Recently, multimedia contents are no longer limited to text and images, but also include videos and GIFs, which have become very popular. However, how to effectively use multimodal representations for sentiment prediction remains a challenge of multimodal sentiment analysis [72]. Thus, Cao et al. [75] develop a sentiment analysis system for cross-media microblog. More specifically, the system first performs textual sentiment analysis and visual sentiment analysis separately, and then selects a fusion strategy to aggregate the single analysis results. The system mainly uses DNN to extract features from texts and images, and then uses different fusion methods for joint sentiment analysis. Yu et al. [76] use CNN to analyse the sentiment in Chinese microblogs from both textual and visual contents. More specifically, they train a CNN on top of pre-trained word vectors for textual sentiment analysis and use a deep CNN with generalised dropout for visual sentiment analysis. They use a data set collected from Sina Weibo to evaluate their method, and find it is better than the sentiment analysis method of a single text or a single image. In the future, it is worth extending the system to recognise facial expression and character to enhance it in cross-media sentiment analysis.

Truong and Lauw [77] propose a visual aspect attention network for sentiment analysis of texts and images, called VistaNet. VistaNet treats visual information as a source of sentence-level alignment to use visual attention [78] to point out important sentences in a document. They argue that the sentiment analysis of the photos is only a supplementary rather than an independent one because the photos cannot tell the whole story alone. So, to focus on the most prominent sentence or "aspect" in a text, they suggest not to use photos directly used as features, but it is fine to use them as a visual means. VistaNet has a three-tier architecture, which summarises representations from words to sentences (Word Encoder, Word Attention, and Sentence Encoder), to image-specific document representations (*i.e.*, visual aspect attention), and finally to the final document representation. VistaNet has been tested on restaurant reviews, but in the future, its application can be extended to other types of web documents, such as blog posts, tweets, or any document containing images. Liu et al. [79] propose a multi-modal sentiment analysis method based on context-enhanced LSTM. Firstly, they capture single-modal information based on context features. Secondly, they merge independent information of single modals to obtain the interactive information between them, forming a multi-modal feature representation. Finally, they use the maximum pool strategy to reduce the dimensionality of multi-modal features. The trained classifier is used to complete sentiment analysis tasks. Their experiment conclusions show that the proper introduction of contextual information can significantly enhance the effect of sentiment analysis, but too much contextual enhancement information may be counterproductive. The accuracy of their method on dataset MOSI is 75.3%, which is 8.1% higher than the traditional SVM method.

Besides joint text-visual multi-modal sentiment analysis, there is another multi-modal sentiment analysis, *i.e.*, video sentiment analysis. Video processing must first process speech. They commonly use speech sentiment analysis method is to convert speech into text, and then treat a task of text sentiment analysis However, some researchers do not use speech-to-text methods. Kaushik et al. [80] use automatic speech to recognise audio sentiment and detect the speaker's opinions or emotions from natural audio. Verma et al. [81] propose a multi-modal (visual, text, and sound) sentiment analysis model with universality and uniqueness. First, they use a deep convolution tensor network to extract common information from the multi-modal representation.

TABLE V
MULTIMODAL SENTIMENT ANALYSIS BASED ON DEEP LEARNING

Reference	Algorithm	Modality	Data set	Sentiment category	Performance
Huang et al. [64]	Deep multimodal attentive fusion	Text, image	Getty Image, Twitter, Flickr.	Positive and negative	Getty Image: Precision -0.882, recall 0.851, F1 0.866, accuracy 0.869; Twitter: Precision -0.778, recall 0.760, F1 0.769, accuracy 0.763; Flickr: precision 0.855, recall 0.845, F1 0.850, accuracy 0.859.
Mahesh et al. [65]	BiLSTM	Text, image, and audio	IEMOCAP28 and CMU-MOSI29	IEMOCAP28: angry, happiness, neutral and sadness; CMU-MOSI29: positive and negative	IEMOCAP28: accuracy 82.69%, Weighted feature Ensemble-97.73s; CMU-MOSI29: accuracy 82.42%, Weighted feature Ensemble-33.01s.
Mahesh et al. [66]	Attention-BiRNN	Video	CMU-MOSI, IEMOCAP	IEMOCAP: happiness (excitement), sadness, anger and neutral; CMU-MOSI: positive and negative.	CMU-MOSI: accuracy 83.33%; IEMOCAP: accuracy 83.87%.
Zhang et al. [67]	QT-LSTM	Text, audio, and video	MELD, IEMOCAP	MELD: positive, negative of neutral, anger, disgust, fear, joy, neutral, sadness or surprise; IEMOCAP: anger, happiness, sadness, neutral, other.	MELD: precision 0.742, recall 0.755, F1 0.729, accuracy 0.756; IEMOCAP: precision 0.631, recall 0.647, F1 0.623, accuracy 0.648.
Harish et al. [68]	Attention-based DNN	Text, audio, and video	CMU-MOSI	positive and negative	Accuracy: 9.3%; F1 Score: 83.8%.
Chen et al. [69]	Multi-head attention mechanism	Text, audio, and video	MOUD, CMU-MOSI	Positive and negative	MOUD: accuracy 90.43%; CMU-MOSI: accuracy 82.71%.
Kim et al. [70]	MA-RN	Text, audio, and video	CMU-MOSI	Positive or negative	CMU-MOSI: accuracy 84.31%.
Mittal et al. [71]	M3ER: learning-based method	Text, image, and audio	IEMOCAP , CMU-MOSEI	IEMOCAP: angry, happy, neutral, sad; CMU-MOSEI: anger, disgust, fear, happy, sad, surprise.	IEMOCAP: mean accuracy 82.7%; CMU-MOSEI: mean accuracy 89.0%.

TABLE VI
PUBLICLY AVAILABLE MULTIMODAL DATA RESOURCES

Name	Data	Description	Access
Multi-ZOL	Text and image	A multimode sentiment data set,which is suitable for text sentiment analysis or multimode sentiment analysis tasks.	https://github. com/xunan0812/MIMN
Flickr	Video	From Flickr social networking site, users can tag and describe uploaded photos.	https://www.flickr.com
Yelp	Text and image	Covering 5 cities in the United States, there are 44305 reviews and 244569 pictures.	Https://www.yelp.com/search?cflt=restaurantsfind_loc=San+Francisco%2C+CA
MOSI	Text, audio, and video	Contains 93 video blogs (vlog) of YouTube movie comment expression videos. The emotional annotation of this data set is not the viewer's feeling,but the emotional tendency of the commentator in the video.	Https://download.csdn. net/download/qq_44186838/12060616
CMU-MOSI	Text, image, and audio	Coming from YouTube, it contains 3228 videos with both personal and emotional annotations. Emotion annotation is divided into six aspects: happy, sad, angry, fear, disgust, and surprise.	Https://download.csdn. net/download/qq_44186838/12060616
ICT-MMMO	Text, audio, and video	There are 370 movie review videos from social media sites YouTube and ExpoTV on movie reviews, including 228 positive reviews, 23 neutral reviews and 119 negative reviews	-
POM	Videos	Contains YouTube videos for movie reviews	-
MVSA-multiple	Text and image	New MVSA dataset including more tweets and annotations. In the new dataset, each tweet is annotated by three annotators. The original MVSA, where each tweet only has one label.	Http://www.mcrlab.net/research /mvsa-sentiment-analysis-on-multi-view-social-data/
Tumblr	Text and image	From Tumblr, the data set is a text and picture tweet containing fifteen emotion tags. There are 256,897 multimodal tweets with emotions marked as fifteen emotions, including joy, sadness, and disgust.	Http://yahoolabs.tumblr.com/post/89783581601/one-hundred-million-creative-commons-
Twitter-15 and Twitter-17	Text and image	A multimodal data set containing text and pictures corresponding to the text. The data set is marked with the target entity and the sentiment tendency expressed in the graphic.	Https://download. csdn.net/download/qq_40930675/10829937
MELD	Text, image, and audio	The positive example is an English tweet that contains a picture topic-specific tags (such as sarcasm, etc.), and the negative and some example is an English tweet with an image but no such tags. The annotation of the data set is satire / not satire.	https://zhuanlan. zhihu.com/p/86777515
IEMOCAP	Text, image, and audio	A total of 4787 impromptu sessions and 5255 scripted sessions are included. The final data annotations are emotional annotations. There are ten categories, including fear and sadness.	The data set needs to be applied. The data is about 18G. If you have any questions, you can contact this person Anil Ramakrishna (akramakr@usc.edu)
SEMAINE	Text, image, and audio	SEMAINE is a dimensional model data set, including several dimensions of valence, arousal, expectancy, power	https://semaine-db.eu/

Then, they use a unique subnet to obtain modal-specific information. Finally, the fusion layer fuses the two layers of information to improve the generalisation performance of the multi-modal system and makes up for the loss caused by the information. They propose a novel multi-peak data fusion architecture, called DeepCU (with universality and uniqueness) (the source code of DeepCU can be found at https://github.com/sverma88/DeepCU-IJCAI19) and its performance surpassed the existing technology. In the future, it is significant to introduce the attention network to efficiently fusing the information obtained by the two subnets.

Yu and Jiang [82] construct a multimodal pre-trained model based on BERT [83]. They first generated a text representation with the help of BERT and then designed an attention mechanism to match text and images to generate a text-image representation. In order to model the interactive information in the text-image, they used a multi-layer attention mechanism to capture the interactive information. On the basis of BERT, the hyperparameters were adjusted, and the BERT was retrained using multimodal data sets. They experiments on the multimodal datasets Twitter15 and Twitter17 achieved accuracy rates of 77.15% and 70.50%, respectively. The experiment further shows that the pre-training model performs best in the target-level classification, and the model can be improved in the way of pre-training in the future. The limitation of multimodal BERT is that it is not sensitive to the target in the image representation. Therefore, a lot of text-independent noise is input in the process of text and image matching, which leads to the degradation of model performance. The pre-training model has many parameters. Increasing the number of network layers can learn a lot of parameters along with the deep feature representation. The parameter problem was revealed in their experiments. When the multimodal BERT performed well, the further increase in the number of transformer layers caused the performance of the model to decrease.

C. Discussion

Table V compares some studies of multimodal sentiment analysis based on deep learning and Table VI summarises some available multimodal data sets. From Table V, we can see that the latest work is almost always on coarse-grained sentiment classification, and few concerns fine-grained sentiment classifications. The performance of the coarse-grained sentiment analysis model based on deep learning is good. The result of coarse-grained multimodal sentiment analysis is always better than that of fine-grained multimodal sentiment analysis. The reason may be that fine-grained sentiment analysis is rarely studied in the multimodal field, and faces such as lack of label data, sentiment Difficulties such as multi-label prediction and information fusion. Multi-label prediction problem is also an exciting research direction in the field of multimodal sentiment analysis. Most of the existing multimodal data sets are coarse-grained emotional labile. At present, researchers are more urgent for fine-grained labelled data sets. Of course, we can create specialized fine-grained data sets.

Multimodal information fusion is the core challenge in the field of multimodal sentiment analysis. The current mainstream multimodal information fusion methods are mainly divided into two types: feature-level fusion [84] and decision-level fusion [85][86]. In addition, the combination of the two is another effective fusion mechanism explored by scholars. The feature-level and decision-level hybrid fusion mechanism [68] retains not only the advantages of the two but also overcomes their shortcomings. Recently, some researchers have explored new fusion mechanisms on this basis, such as tensor fusion mechanism [87], quantum interference-inspired multimodal decision fusion method [67], end-to-end fusion method with transformer [88]. In the future, we should continue to explore efficient multimodal information fusion mechanisms. Many researchers also deal with multimodal tasks based on the attention mechanism and multi-head attention mechanism, which indeed proves that the attention mechanism is machine-effective in single-modal feature extraction and information fusion tasks [64] [89]. We think attention mechanism also plays a significant role in multimodal tasks.

Besides, pre-training models (MLM [90], BERT [83], XLNet [91], GPT-3 [92]) are a new paradigm in the field of natural language processing. Pre-training models play an important role in multimodal tasks and multilingual tasks. Future multimodal sentiment analysis tasks should make full use of pre-training models to improve the performance of the model.

VII. Conclusion

Sentiment analysis is very important. In this paper, we survey various sentiment analysis methods based on deep learning. These deep learning methods are convolutional neural networks, recurrent neural networks, long short-term memory, deep neural networks, deep belief networks and memory networks. We analyse their advantages and disadvantages and point out which methods can be used to improve these methods in the future. We also compare the sentiment analysis effects of deep learning models in the fields of social media, e-commerce platform and entertainment consumption.

The general directions for future improvement include to collect high-quality data, optimise feature extraction, properly combine features, reduce training time, and improve classify accuracy. It is also worth integrating deep learning based methods with fuzzy rules, language similarity, machine learning, and even reinforcement learning in sentiment analysis. Besides, it is worth paying more attention to the application of deep learning in multimodal sentiment analysis and aspect-level sentiment analysis. In the future, pre-training models will play an important role in the field of sentiment analysis.

References

[1] K. Coffman and A. Odlyzko, *Internet Growth: Is there a "Moore's law" for data traffic?* Handbook of Massive Data Sets, 2002, pp. 47–93.

[2] S. F. Pengnate and F. J. Riggins, "The role of emotion in p2p microfinance funding: A sentiment analysis approach," *International Journal of Information Management*, vol. 54, p. 102138, 2020.

[3] C.-M. Yu, "Mining opinions from product review: Principles and algorithm analysis," *Information Studies: Theory & Application*, vol. 32, no. 7, pp. 124–128, 2009, (In Chinese).

[4] B. Liu and L. Zhang, *A Survey of opinion mining and sentiment analysis*. Mining Text Data, 2012, pp. 415–463.

[5] T. Nasukawa and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," in *Proceedings of the 2nd International Conference on Knowledge Capture*, 2003, pp. 70–77.

[6] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

[7] B. Liu, “Sentiment analysis and opinion mining,” *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1–167, 2012.

[8] C. Alfaro, J. Cano-Montero, J. Gomez, J. M. Moguerza, and F. Ortega, “A multi-stage method for content classification and opinion mining on weblog comments,” *Annals of Operations Research*, vol. 236, no. 1, pp. 197–213, 2016.

[9] L. Zhang, S. Wang, and B. Liu, “Deep learning for sentiment analysis: A survey,” *Data Mining and Knowledge Discovery*, vol. 8, no. 4, pp. 1–25, 2018.

[10] I. Prabha M and G. Umarani Srikanth, “Survey of sentiment analysis using deep learning techniques,” in *Proceedings of the 1st International Conference on Innovations in Information and Communication Technology*, 2019, pp. 1–9.

[11] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, “A survey of multimodal sentiment analysis,” *Image and Vision Computing*, vol. 65, pp. 3–14, 2017.

[12] D. Stojanovski, G. Strezoski, G. Madjarov, and et al., “Deep neural network architecture for sentiment analysis and emotion identification of twitter messages,” *Multimedia Tools Applications*, vol. 77, no. 24, pp. 32 213–32 242, 2018.

[13] X.-L. Yang, S.-J. Xu, H. Wu, and R.-F. Bie, “Sentiment analysis of weibo comment texts based on extended vocabulary and convolutional neural network,” in *Proceedings of the 2018 International Conference on Identification, Information and Knowledge in the Internet of Things*, 2018, pp. 9.361–368.

[14] H.-Y. He, J. Zheng, and Z.-P. Zhang, “Text sentiment analysis combined with part of speech features and convolutional neural network,” *Computer Engineering*, vol. 44, no. 11, pp. 209–214, 2018.

[15] Z.-F. Sun and J. Wang, “Rcnn-bgru-hn network model for aspect-based sentiment analysis,” *Computer Science*, vol. 46, no. 9, pp. 223–228, 2018, (In Chinese).

[16] S. Chen, Y. Ding, Z. Xie, S. Liu, and H. Ding, “Chinese Weibo sentiment analysis based on character embedding with dual-channel convolutional neural network,” in *Proceedings of 2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis*, 2018, pp. 107–111.

[17] A. Shenoy and A. Sardana, “Multilogue-net: A context aware rnn for multi-modal emotion detection and sentiment analysis in conversation,” in *Proceedings of the 2020 Computing Research Repository*, 2020, pp. 1–9.

[18] Z.-F. Hao, H. Huang, R.-C. Cai, and W. Wen, “Fine-grained opinion analysis based on multi-feature fusion and bidirectional RNN,” *Computer Engineering*, vol. 44, no. 7, pp. 199–2049, 2018, (In Chinese).

[19] S.-W. Pei and L.-L. Wang, “Text sentiment analysis based on attention mechanism,” *Computer Engineering & Science*, vol. 41, no. 02, pp. 343–353, 2019, (In Chinese).

[20] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, p. 17351780, 1997.

[21] G.-H. Chen, “Text sentiment analysis based on polarity transfer and bidirectional long-short term memory,” *Information Technology*, no. 2, pp. 149–152, 2018.

[22] H.-Y. Peng, L. Xu, L.-D. Bing, F. Huang, W. Lu, and L. Si, “Knowing what, how and why: A near complete solution for aspect-based sentiment analysis,” in *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020, pp. 1–9.

[23] Y. Tay, L.-A. Tuan, and S.-C. Hui, “Learning to attend via word-aspect associative fusion for aspect-based sentiment analysis,” in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, vol. 18, 2018, pp. 5956–5963.

[24] Z. Xu, B. Liu, B. Wang, S. C., and X. Wang, “Incorporating loose-structured knowledge into lstm with recall gate for conversation modeling,” in *Proceedings of the 2016 Computing Research Repository*, 2016, pp. 1–10.

[25] Y.-K. Ma, H.-Y. Peng, and E. Cambria, “Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm,” in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018, pp. 5876–5883.

[26] B.-W. Xing, L.-J. Liao, D.-D. Song, J.-G. Wang, F.-Z. Zhang, and H.-Y. Huang, “Earlier attention? aspect-aware lstm for aspect-based sentiment analysis,” in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019, pp. 5313–5319.

[27] Y. Wang, Q. Chen, M. Ahmed, Z. Li, W. Pan, and H. Liu, “Joint inference for aspect-level sentiment analysis by deep neural networks and linguistic hints,” *IEEE Transactions on Knowledge and Data Engineering*, 2019.

[28] C. Aydin and T. Gungor, “Combination of recursive and recurrent neural networks for aspect-based sentiment analysis using inter-aspect relations,” *IEEE Access*, vol. 8, pp. 77 820–77 832, 2020.

[29] A. Ishaq, S. Asghar, and S. Gillani, “Aspect-based sentiment analysis using a hybridized approach based on CNN and GA,” *IEEE Access*, vol. 8, pp. 135 499–135 512, 2020.

[30] Z.-B. Jia, X.-X. Bai, and S.-M. Pang, “Hierarchical gated deep memory network with position-aware for aspect-based sentiment analysis,” *IEEE Access*, vol. 8, pp. 136 340–136 347, 2020.

[31] G.-F. Liu, X.-Y. Huang, X.-Y. Liu, and A.-Z. Yang, “A novel aspect-based sentiment analysis network model based on multilingual hierarchy in online social network,” *Computer Journal*, vol. 63, no. 3, pp. 410–424, 2020.

[32] y.-W. Zheng, R.-C. Zhang, S. Mensah, and Y.-Y. Mao, “Replicate, walk, and stop on syntax: An effective neural network model for aspect-level sentiment classification,” in *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020, pp. 9685–9692.

[33] X. Wang, W. Jiang, and Z. Luo, “Combination of convolutional and recurrent neural network for sentiment analysis of short texts,” in *Proceedings of the 26th International Conference on Computational Linguistics*, 2016, pp. 2428–2437.

[34] F. Luo and H.-F. Wang, “Chinese text sentiment classification by h-rnn-cnn,” *Acta Scientiarum Naturalium Universitatis Pekinensis*, vol. 54, no. 3, pp. 459–465, 2018.

[35] K. Kwaik, M. Saad, S. Chatzikyriakidis, and S. Dobnik, “LSTM-CNN deep learning model for sentiment analysis of dialectal Arabic,” in *Arabic Language Processing: From Theory to Practice*, ser. Communications in Computer and Information Science, vol. 1108, 2019, pp. 108–121.

[36] A. Rehman, A. Malik, B. Raza, and W. Ali, “A hybrid CNN-LSTM model for improving accuracy of movie reviews sentiment analysis,” *Multimedia Tools and Applications*, vol. 78, no. 18, pp. 26 597–26 613, 2019.

[37] D. Jain, A. Kumar, and G. Garg, “Sarcasm detection in mash-up language using soft-attention based bi-directional lstm and feature-rich cnn,” *Applied Soft Computing*, vol. 91, p. 106198, 2020.

[38] M.-D. Wang and G.-M. Hu, “A novel method for twitter sentiment analysis based on attentional-graph neural network,” *Information*, vol. 11, no. 2, p. 92, 2020.

[39] Z.-L. Wu, J. Ming, and M. Zhang, “Transformer based memory network for sentiment analysis of chinese weibo texts,” in *Proceedings of the 2019 International Conference on Mobile Computing, Applications, and Services*, 2019, pp. 44–56.

[40] M.-J. Ling, Q.-H. Chen, Q. Sun, and Y.-B. Jia, “Hybrid neural network for sina weibo sentiment analysis,” *IEEE Transactions on Computational Social Systems*, vol. 7, no. 4, pp. 983–990, 2020.

[41] S. Kumar, M. Yadava, and P. Roy, “Fusion of eeg response and sentiment analysis of products review to predict customer satisfaction,” *Information Fusion*, vol. 52, pp. 41–52, 2019.

[42] A. Mukherjee, S. Mukhopadhyay, P. Panigrahi, and S. Goswami, “Utilization of oversampling for multiclass sentiment analysis on amazon review dataset,” in *Proceedings of the IEEE 10th International Conference on Awareness Science and Technology*, 2019, pp. 1–6.

[43] N. Shrestha and F. Nasoz, “Deep learning sentiment analysis of amazon.com reviews and ratings,” *International Journal on Soft Computing, Artificial Intelligence and Applications*, vol. 8, no. 1, pp. 1–15, 2019.

[44] X. Fang and J. Zhan, “Sentiment analysis using product review data,” *Journal of Big Data*, vol. 2, no. 1, p. 5, 2015.

[45] U. Chauhan, M. Afzal, A. Shahid, M. Abdar, M. Basiri, and X.-J. Zhou, “A comprehensive analysis of adverb types for mining user sentiments on amazon product reviews,” *World Wide Web*, vol. 23, no. 3, pp. 1811–1829, 2020.

[46] Y. Cheng, L.-B. Yao, G.-X. Xiang, and et al., “Text sentiment orientation analysis based on multi-channel CNN and bidirectional GRU with attention mechanism,” *IEEE Access*, vol. 8, pp. 134 964–134 975, 2020.

[47] T. Tran, H. Ba, and V. Huynh, “Measuring hotel review sentiment: An aspect-based sentiment analysis approach,” in *Proceedings of the 2019 International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making*, 2019, pp. 393–405.

[48] M. Al-Smadi, O. Qawasmeh, M. Al-Ayyoub, Y. Jararweh, and B. Gupta, “Deep recurrent neural network vs. support vector machine for aspect-based sentiment analysis of arabic hotels’ reviews,” *Journal of Computational Science*, vol. 27, pp. 386–393, 2018.

[49] J. Shen, M. Ma, R. Xiang, Q. Lu, E. P. Vallejos, G. Xu, C.-R. Huang, and Y. Long, "Dual memory network model for sentiment analysis of review text," *Knowledge-Based Systems*, vol. 188, p. 105004, 2020.

[50] Z.-x. Liu, D.-g. Zhang, G.-z. Luo, M. Lian, and B. Liu, "A new method of emotional analysis based on CNN–BiLSTM hybrid neural network," *Cluster Computing*, pp. 1–13, 2020.

[51] J. Yang, Y.-Q. Yang, C.-J. Wang, and J.-Y. Xie, "Multi-entity aspect-based sentiment analysis with context, entity and aspect memory," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018, pp. 6029–6036.

[52] K.-S. Song, W. Gao, L.-J. Zhao, C.-L. Sun, and X.-Z. Liu, "Cold-start aware deep memory network for multi-entity aspect-based sentiment analysis," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019, pp. 5179–5203.

[53] Y. Xiao, D.-Y. Wang, and L.-G. Hou, "Unsupervised emotion recognition algorithm based on improved deep belief model in combination with probabilistic linear discriminant analysis," *Personal and Ubiquitous Computing*, vol. 23, no. 3-4, pp. 553–562, 2019, (In Chinese).

[54] F. Nian, X. Chen, S. Yang, and G. Lv, "Facial attribute recognition with feature decoupling and graph convolutional networks," *IEEE Access*, vol. 7, pp. 85 500–85 512, 2019.

[55] M. Zhang, Y. Liang, and H. Ma, "Context-aware affective graph reasoning for emotion recognition," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2019, p. 151156.

[56] Z.-H. Wu, S.-R. Pan, F.-W. Chen, and et al., "A comprehensive survey on graph neural networks," in *arXiv:1901.00596*, 2020.

[57] E. Mansouri-Benssassi and J. Ye, "Synch-graph: Multisensory emotion recognition through neural synchrony via graph convolutional networks," in *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020, pp. 1351–1358.

[58] ——, "Speech emotion recognition with early visual cross-modal enhancement using spiking neural networks," in *Proceedings of 2019 International Joint Conference on Neural Networks*, 2019, pp. 1–8.

[59] Q.-M. Xue, W. Zhang, and H.-Y. Zha, "Improving domain-adapted sentiment classification by deep adversarial mutual learning," in *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020, pp. 9362–9369.

[60] Y. Dai, J. Liu, X.-C. Ren, and Z.-L. Xu, "Adversarial training based multi-source unsupervised domain adaptation for sentiment analysis," in *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020, pp. 7618–7625.

[61] C. Lin, S.-C. Zhao, L. Meng, and T.-S. Chua, "Multi-source domain adaptation for visual sentiment classification," in *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020, pp. 2661–2668.

[62] H. Wan, Y.-F. Yang, J.-F. Du, and et al., "Target-aspect-sentiment joint detection for aspect-based sentiment analysis," in *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020, pp. 9122–9129.

[63] H. Fei, Y. Zhang, Y.-F. Ren, and D.-H. Ji, "Latent emotion memory for multi-label emotion classification," in *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020, pp. 7692–7699.

[64] F.-R. Huang, X.-M. Zhang, Z.-H. Zhao, and et al., "Image-text sentiment analysis via deep multimodal attentive fusion," *Knowledge- Based Systems*, vol. 167, pp. 26–37, 2019.

[65] G. Mahesh, S. S. Huddar, and V. S. R. Sannakki, "Multi-level feature optimization and multimodal contextual fusion for sentiment analysis and emotion classification," *Computational Intelligence*, vol. 36, no. 2, pp. 861–881, 2020.

[66] ——, "Multi-level context extraction and attention-based contextual inter-modal fusion for multimodal sentiment analysis and emotion classification," *International Journal of Multimedia Information Retrieval*, vol. 9, no. 2, pp. 103–112, 2020.

[67] Y.-Z. Zhang, D.-W. Song, X. Li, and et al., "A quantum-like multimodal network framework for modeling interaction dynamics in multiparty conversational sentiment analysis," *Information Fusion*, vol. 62, pp. 14–31, 2020.

[68] A. Harish and F. Sadat, "Trimodal attention module for multimodal sentiment analysis," in *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020, pp. 13 803–13 804.

[69] X. Chen, G.-M. Lu, and J.-J. Yan, "Multimodal sentiment analysis based on multi-head attention mechanism," in *Proceedings of the 4th International Conference on Machine Learning and Soft Computing*, 2020, pp. 34–39.

[70] T. Kim and B. Lee, "Multi-attention multimodal sentiment analysis," in *Proceedings of the 2020 on International Conference on Multimedia Retrieval*, 2020, pp. 436–441.

[71] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," in *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020, pp. 1359–1367.

[72] T. Baltrusaitis, C. Ahuja, and L. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[73] N. Xu, W.-J. Mao, and G.-D. Chen, "Multi-interactive memory network for aspect based multimodal sentiment analysis," in *Procedings of The Thirty-Third AAAI Conference on Artificial Intelligence*, 2019, pp. 371–378.

[74] G.-Y. Cai and B.-B. Xia, "Multimedia sentiment analysis based on convolutional neural network," *Journal of Computer Applications*, vol. 36, no. 2, pp. 428–431, 2016, (In Chinese).

[75] D. Cao, R. Ji, and D. Lin, "A cross-media public sentiment analysis system for microblog," *Multimedia Systems*, vol. 22, no. 4, pp. 479–486, 2016.

[76] Y. Yu, H. Lin, and J. Meng, "Visual and textual sentiment analysis of a microblog using deep convolutional neural networks," *Algorithms*, vol. 9, no. 2, pp. 41–51, 2016.

[77] Q. Truong and H. Lauw, "Vistanet: Visual aspect attention network for multimodal sentiment analysis," in *Procedings of The Thirty-Third AAAI Conference on Artificial Intelligence*, 2019, pp. 305–312.

[78] L. Itti and C. Koch, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 20, no. 11, p. 1254, 1998.

[79] Q.-Y. Liu, D. Zhang, L.-Q. Wu, and S.-S. Li, "Multi-modal sentiment analysis with context-augmented lstm," *Computer Science*, vol. 46, no. 11, pp. 181–185, 2019, (In Chinese).

[80] L. Kaushik, A. Sangwan, and J. Hansen, "Automatic sentiment detection in naturalistic audio," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 8, pp. 1668–1679, 2017.

[81] S. Verma, C. Wang, L.-M. Zhu, and W. Liu, "Deepcu: Integrating both common and unique latent information for multimodal sentiment analysis," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019, pp. 3627–3634.

[82] J.-F. Yu and J. Jiang, "Adapting bert for target-oriented multimodal sentiment classification," in *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020, pp. 5408–5414.

[83] J. Devlin and et al., "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 North American Chapter of the Association for Computational Linguistics*, 2019, p. 41714186.

[84] V. Perez-Rosas, R. Mihalcea, and L. Morency, "Utterance-level multimodal sentiment analysis," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 2013, pp. 973–982.

[85] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," *Neurocomputing*, vol. 174, pp. 50–59, 2016a.

[86] W.-M. Yu, H. Xu, F.-Y. Meng, and et al., "Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3718–3727.

[87] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1103–1114.

[88] Z.-L. Wang, Z.-H. Wan, and X.-J. Wan, "Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis," in *Proceedings of the 29th World Wide Web*, 2020, pp. 2514–2520.

[89] C. Xi, G.-M. Lu, and J.-J. Yan, "Multimodal sentiment analysis based on multi-head attention mechanism," in *Proceedings of the 4th International Conference on Machine Learning and Soft Computing*, 2020, pp. 34–39.

[90] X. Wu, T. Zhang, L.-J. Zang, and et al., "Mask and infill: Applying masked language model for sentiment transfer," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019, pp. 5271–5277.

[91] Z. Yang and et al., "Xlnet: Generalized autoregressive pretraining for language understanding," in *Proceedings of 33rd Conference on Neural Information Processing Systems*, 2019, pp. 5754–5764.

[92] T. B. Brown and et al., "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

Pingping Lin is currently a master student at Guangxi Normal University, China.

Dr. Xudong Luo currently is a distinguished professor of Artificial Intelligence at Guangxi Normal University, China. He published one book and more than 160 papers including 2 in top journal *Artificial Intelligence*, one of which has been highly cited by, for example, MIT, Oxford, and CMU research groups. Prof. Luo has an internationally recognised reputation: co-chair and (senior) members of PC of more than 100 international conferences or workshops, including major conferences IJCAI and AAMAS, and referees for many international journals such as top journal *Artificial Intelligence*. He is also invited to make a presentation of his work in more than ten universities internationally, including Imperial College. His research focus is on the areas of agent-based computing, fuzzy sets and systems, decision theory, game theory, knowledge engineering, and natural language process. Prof. Luo has supervised or co-supervised more than 40 master students, PhD students, and research fellows.