Sentiment Analysis：The State of The Art and Future Directions

YuanyuanZhang, XiaoWang, andFei-YueWang

Abstract—In recent years, sentiment analysis has become a hot subarea in the field of natural language processing, since sentiment analysis can have many practical applications. Most of existing sentiment analysis work is text-based sentiment analysis, with text data as the only source of information. However, text-based sentiment analysis requires a large amount of textual corpus-labeled data. With the development of Internet technology, in addition to text, various multimedia such as voice, image, and video also contain prolific emotional information, If this kind of multimedia information can be fully exploited, the effect of emotional analysis can be improved as much as possible when labeled texts are lacking. This kind of sentiment analysis task based on multimodal data is called multimodal sentiment analysis. In this paper, we will give an overview of text-based sentiment analysis and multimodal sentiment analysis, including introducing the background and existed research work of these two types of tasks, as well as the current challenges and prospects in the field of multimodal sentiment analysis.

***Index Terms*—Sentiment analysis, multimodal sentiment anal-ysis, text-based sentiment analysis.**

1. INTRODUCTION

With the continuous development of Intemet, not only has the Internet become a source of information, but also the best platform for everyone to express their opinions and share their personal lives. Therefore, the Internet is full of people's views on various public events, evaluations of certain commodities, and text information sharing on personal life. It is valuable to analyze these opinion-oriented data, because they may affect all aspects of social development, and help us to solve many practical problems. For example, merchants can improve their products by analyzing customers' comments on goods. Similarly, sentiment analysis can also be used in the recommendation system. By analyzing a user's historical sentiment tendency towards the products, the system could avoid recommending wrong products. On the other hand, from understanding the public's attitudes towards certain events, government could flexibly adjust some policies, which could be really helpful for citizens. In addition, sentiment analysis can also help predict some important events, such as presidential elections. Sentiment analysis can help polling agencies correctly analyze voters' attitudes towards election candidates, and increase the election prediction rate.

Sentiment analysis is a classification task in the field of Natural Language Processing (NLP) that focuses on distinguishing and analyzing one's attitude and opinion towards a special entity automatically. According to the definition of sentiment in [1], the sentiment is a long-term disposition of human-beings to specific targets, those targets can be peole, entities, topics, and so on.

Nowadays, most researches on sentiment analysis are based on texts, which aim at extracting sentiment features from text semantic information. With the development of deep learning and the emergence of large-scale pre-trained language models, such as Bert[2] and GPT-3[3], the fine-tuned models have shown good results on many sentiment analysis datasets, but those pre-trained models contain enormous amounts of parameters and request much more time and expenses to collect the labeled data. However, with the Internet, there is much more multi-media data, such as audio and video, which could be used for sentiment analysis. Compared with texts, audios and videos contain richer information, thus we can explore more detailed information from them Hence, sentiment analysis based on multimodal data could be more effective.

In the remainder of this review, Section 2 will introduce the existed work in text-based sentiment analysis and multimodal sentiment analysis, as well as related datasets, Section 3 will introduce current challenges, Section 4 will make a conclusion and look forward to multimodal sentiment analysis.

2. LITERATURE REVIEW

Multimodal sentiment analysis is a sub-task of sentiment analysis. Traditional sentiment analysis aims to explore enough sentiment information from text data, while multimodal sentiment analysis utilizes information from multiple sources. The techniques of multimodal sentiment analysis and the approaches of traditional sentiment analysis are similar in many stages, thus we first introduce the existing research work of traditional text sentiment analysis in subsec-

Yuanyuan Zhang is with the Key Laboratory of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China (e-mail: zhangyuanyuanl8s@ict.ac.cn).

Xiao Wang is with the Department of Computer Science al Purdue University, West Lafayette, IN, 49706, USA(e-mail: wang3702@purdue.edu)

Fei-Yue Wang is with the Stale Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: feiyue@ieee.org).

tion II.A, which is divided into 3 different directions: sentiment dictionaries, machine learning, and deep leaming. Then we review the multimodal sentiment analysis in subsection II. B, which includes two main categories, the approaches based on decision-fusion or feature-fusion.

2.1. Text-based Sentiment Analysis

In 2002, Pang et al.[4] compared the effects of three supervised learning methods on the task of movie review sentiment analysis and concluded that support vector machines [5] performed better than maive bayes[6] and maximum entropy Principle[7]. In 2003, Nasukawa et al. [8] first proposed the term "sentiment analysis". Then in the following years, researches in sentiment analysis developed rapidly. To sum up, all the following methods on text-based sentiment analysis could be divided into three categories:

2.1.1. Approaches based on sentiment dictionaries: The core of sentiment dictionaries[9] is how to use the constructed sentiment dictionary to extract the sentiment words and phrases in the text. On the one hand, happy, interesting, excellent are positive sentiment words, on the other hand, disappointing, terrible, ugly are negative words. Some verbs could also express the sentiment, such as like, hate. For example, "I dislike this movie, because it is so boring", in this sentence, "dislike" is a negative verb and "boring" is a negative adjective, and file 2 words would be assigned different sentiment scores. Then the next step is to calculate the sentiment score of the whole sentence, and the simplest way is regarding the difference between the positive sentiment value and the negative sentiment value of all sentiment words as the final sentiment score of a sentence. In 2014, Ghag et al. proposed SentiTFIDF model[10], which used Term Frequency Inverse Document Frequency (TF-IDF)[ll] as sentiment weighting functions.

Therefore, this kind of method is very dependent on the accuracy of the sentiment dictionary. SentiWordNet[12] is the most common used sentiment dictionary in English. What's more, there are many other sentiment lexicons, such as WordNet[13], which was proposed by Miller in 1995, WordNet contains semantic information, and it groups all words by semantic similarity as synsets. It provides a brief and summary definition for each synset and records the semantic relationship between different synsets SentiFul[14], a lexicon extension by adding synonymy and antonymy relations, hyponymy relations, derivation, and compounding from known lexical units. In 2016, Erik further proposed SenticNet 4[15], which expanded through an ensemble of hierarchical clustering and dimensionality reduction from SenticNet 2[16].

However, if there is a sentence that does not contain any sentiment word, how should we calculate it? In 2002, Turney et al. [17] first developed an unsupervised algorithm called SOPMI. If we want to know the sentiment score of a word 'UNK', the basic idea of SO-PMI is to choose a group of praise words (Pwords) and a group of derogatory words (Nwords) as benchmark words. Then the difference could be obtained from the difference of the mutual information between the points of 'UNK' and Pwords and that of 'UNK' and Nwords.The equation of SO-PMI is as in Eq. (1).

$$SO-PMI(word) = \sum_{Pword \in Pwords} PMI(word, Pword) - \sum_{Nword \in Nwords} PMI(word, Nword) \quad (1)$$

Here SO-PMI is the overall sentiment score while PMI measures the difference between query words with Pwords and Nwords. Here, if the value of SO-PMI is bigger than 0, it means that the word is a positive word, if the value equals 0, then the word is a neural word, and if the value is smaller than 0, file word is a negative word.

2.1.2. Approaches based on machine learning: The machine learning techniques[18] are all based on supervised learning, including traditional machine learning algorithms, such as decision tree[19], support vector machine(SVM)[5] and maive bayes[6] etc. The process of machine leaming is very different from the sentiment dictionary-based methods. The main idea of machine learning sentiment analysis is how to train a robust and general classifier to analyze sentiment automatically. Therefore, there are two datasets, the training set, which is designed for training and optimizing the model and consists of texts and labels; and the test set, which is used to test the models' generalizability, and only contains texts. Above all, there are 2 main steps in machine learning-based sentiment analysis: (1) sentiment feature extraction; and (2) model training, Fig. 1 illustrates a traditional procedure for a machine learning algorithm for a sentiment analysis task.

Fig. 1 shows the general framework of classification tasks based on machine learning algorithms. The first step of such methods is to obtain a stable model on the training set. During the training process, the model will continuously adjust its own parameters according to the provided labels. After getting an optimal model parameter, the optimal model is used to make predictions on the test set.

As for feature extraction, there are several methods. The most common way is word frequency, which is counting the times of a word appearing in the texts as the feature. Another approach is TF-IDF[ll], TF-IDF is a statistical method used to evaluate the importance of a word in a document set. The importance of a word increases in proportion to the times it appears in a document, but it decreases in inverse proportion to the frequency of its appearance in the corpus.

For the machine learning model, researchers have explored several methods and achieved very promising results. In[13], Jeevanandam successfully implemented IDF [11] as feature se- lection and CART tree[20] as classifier on IMDB dataset[21]. Bhadane et a1. [22] focused on domain-specific lexicons in the aspect sentiment analysis task and used SVM as a machine learning classifier to distinguish the polarity of every aspect. Smeureanu et a1. [23] combined n-gram[24] and maive bayes to classify positive and negative class on the user movie reviews. In [25], Soelistio et al. proposed a simple model that uses Naive Bayes to analyze the emotional polarity of digital newspapers and

applied it to digital newspapers for political sentiment analysis, which obtained sentiment information about specific politicians. Dey et a1.[26] compared the performance of Naive Bayes and KNN[27] on movie reviews and hotel reviews. They found that Naive Bayes was better than KNN in movie reviews, but in hotel reviews, the accuracy of them was almost the same. Hajmohammadi et a1. [28] utilized SVM and Naive Bayes to classify film reviews with Persian language into 2 classes, positive and negative, They concluded that SVM achieved higher accuracy than Naive Bayes.

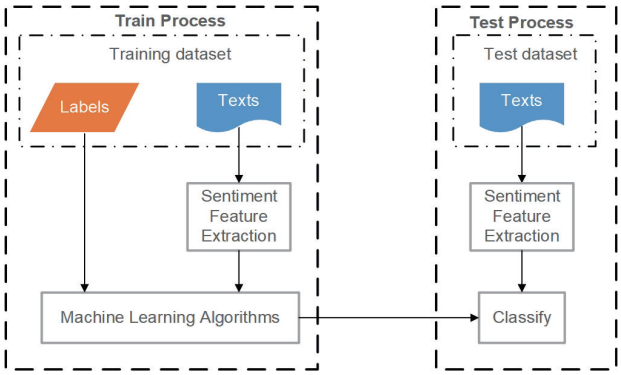

Fig. 1: The illustration of machine learning based sentiment analysis

Naive Bayes and SVM are the two most commonly used approaches in sentiment analysis tasks. Naive Bayes-based text sentiment analysis classifies sentiment by calculating probability. It is a simple algorithm and performs very well on a small dataset. However, Nffive Bayes is very sensitive to the form of the input data, since it is highly dependent on the prior probability. Consequently, it is of high error possibility for Naive Bayes when the input data is not aspecific distribution. SVM is regarded as the best machine learning-based sentiment analysis method. On the one hand, SVM has a low generalization error rate. On the other hand, the training expense of SVM is also lower than other methods. Moreover, SVM could still achieve good performance with limited data. Furthermore, kernel-based SVM enables it to process high-dimensional data. However, it is sensitive to parameters and kernel function selection.

2.1.3. Approaches based on deep learning: Deep learning methods[29] adopt neural networks as feature extractors and classifiers. The mechanism of multimodal sentiment analysis with deep learning methods is shown in Fig. 2. It illustrates the general process of methods based on deep learning. For example, there is a sentence "This German horror film has to be one of the weirdest I have seen". The first step of deep learning models is to pre-process the text, which includes word segmentation, punctuation, and stop words removal, with meaningful words and phrases left. At the second step, those remaining words need to be padded to a uniform length. Third, the processed text is embedded as a vector by some pretrained language models, such as Bert[2], Glove[30] or some statistic method, such as counting frequency, TF-IDF. Finally, the neural network takes file vectors and labels as input data, and continuously adjusts the network parameters through the backpropagation algorithm to an optimal state. In recent years, deep learning works well on many classification tasks based on a large amount of data. Researchers usually use RNN[31] and CNN [32] as the main structure of a classification model for NLP tasks, whose inputs are word embeddings. In 2014, Le and Mikolov proposed a distributed text expression-word2vec [33], which successfully reduced the dimension of the distributed language model and can be widely used. Kim further developed textCNN[34] based on word2vec, in which CNN is a feature extractor. The experiment results show that textCNN can significantly improve the effect of feature expression. Another work GlobalVectors (GloVe), which is similar to word2vec, is proposed by Pennington et al. The parallelization of GloVe enables it to be trained quickly compared to word2vec. In 2016, Chen et a1.[35] compared the CNN with the GRU[36], which is one type of RNN, and found that the structure of GRU is more suitable for text sequence classification tasks.

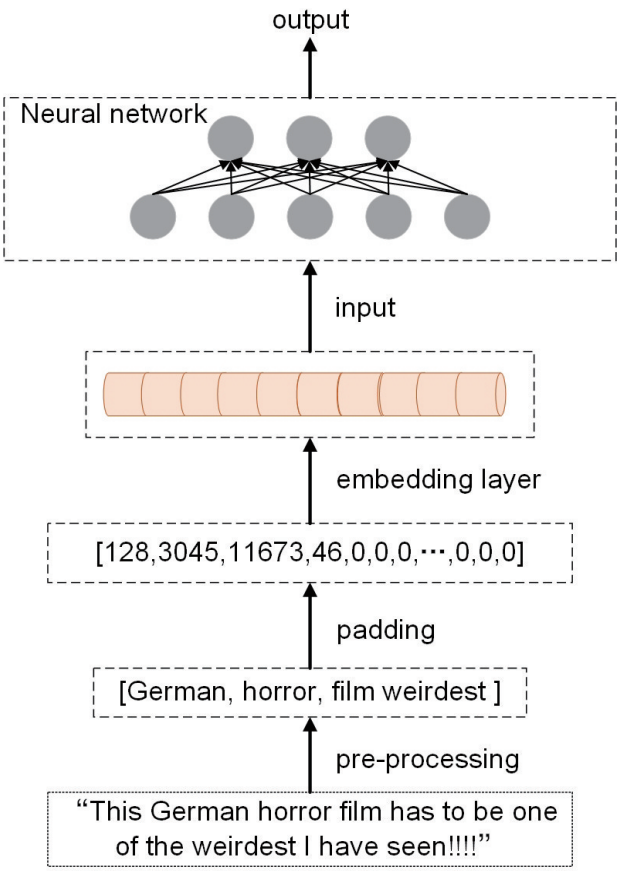

Fig. 2: The general framework of deep learning based sentiment analysis

In 2015, Severyn et a1.[37] combined the ideas of unsupervised and weakly supervised leaming on CNN, which used an unsupervised neural language model to train initial word embeddings and used distantly supervised data to furher refine the weights of the network. Ouyang et al.

proosed word2vec+CNN framework with 7 layers, which performed better than RNN and MV-RNN[38] on 5-class sentiment analysis. Vateekul et a1. [39] first implemented deep learning on the Thai twitter sentiment analysis dataset, and compared the performance of several models They drew a conclusion that DCNN outperformed other models, such as LSTM, SVM, and Naive Bayes on the Thai dataset. Wang et al. [40] proposed an attention-based LSTM model [41], which can focus on different parts of a sentence. By connecting the aspect vector to the hidden sentence representation (AE-LSTM model[40]), or embedding the aspect vector into each word input vector (ATAE-LSTM model [40]), the attention weight can be calculated. Experimental results show that the two proposed models are better than the baseline models(LSTM, TD-LSTM[42], and TC-LSTM [42]), which showed that the attention-based LSTM model can improve the performance of the trained model. In 2018, Zhang et a1. [43] developed an approach of soft attention and dynamic memory network (DMN)[44] to transform the target emotion classification task into a question and answer system. The experimental results on SemEval 2014 (laptop and restaurant reviews)[45] and Twitter dataset[46], proved that the attention-based GRU and internal attention can be used to solve the weight bias problem. In 2019, Yu et a1. [47]proposed a framework for aspect and opinion term extraction using Bi-LSTM[48] and multi-layer attention network, which achieved good results in aspect sentiment analysis tasks.

2.2. Multimodal Sentiment Analysis

This section will introduce recent works on multimodal sentiment analysis. Multimodal sentiment analysis is a new subfield of traditional text-based sentiment analysis. The first part is about multimodal sentiment analysis datasets, and the second part is about approaches.

2.2.1. Datasets: In 2011, Morency et a1. [49] first proposed the task of multimodal sentiment analysis, and the first dataset of this task, Youtube, which contains three kinds of modalities, texts, audios, and videos. In 2013, W611mer et al.[50] put the ICT-MMMO forward, regarding the English reviews' videos in the social network., ICT-MMMO contains 370 videos, and each video lasts for 1 to 3 minutes. All of the videos come from Expo and Youtube and the ratio of positive, neutral, and negative is 228: 23: 119. In the same year, P~rez Rosas et al. [51] collected another new dataset, Multimodal Opinion Utterances Dataset (MOUD), which aims at sentiment analysis in Spanish videos. MOUD consists of 498 utterances, 182 of them are positive, 82 of them are neutral and the rest of them are negative. In 2016, Zadeh et a1.[52] build MOSI dataset in videos related to movie reviews from Youtube. The total duration of those videos is 2 hours 37 minutes, and they are clipped into 2199 utterances and labeled into 7 categories, which referred to the intensity of sentiment. In 2018, an extended dataset MOSEI[53] based on MOSI was constructed. Compared to MOSI, MOSEI contains more videos and is further labeled in emotion, the duration of MOSEI is 66 hours and clipped into 23500 utterances. In detail, as for sentiment annotations, MOSEI could be a 2-class or 5-class dataset, and as for emotion, MOSEI provides annotations: happiness, sorrow, anger, fear, disgust, and surprise. The detailed information about those datasets is shown in TABLE I.

2.2.2. Methods: Since 2011, many researchers have switched to multimodal sentiment analysis, In the early period, researchers always fused the features from different modalities at the last stage of a model, which is called decision-fusion. The process is demonstrated in Fig. 3, which contains two types of fusion. The first is called decision-level fusion. Each modality performs the feature extraction and sentiment classification independently, and the prediction results of each modality will be combined in the last stage. The advantage of this type of method is that the processing steps of each modality are independent from each other, which makes it to be more robust. The second is concatenation fusion. Here each modality is independent only in the feature extraction stage. Then all the vectors are concatenated as one fusion vector, and one classifier is trained based on the fusion vector. The advantage of such methods is that the classifier could learn the relationships between different modalities, while it is not so robust.

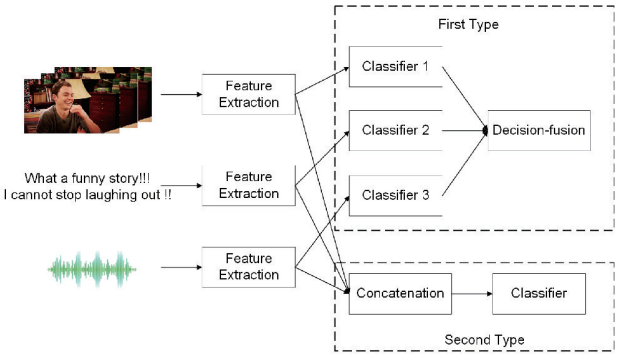

Fig. 3: Two types process of decision-fusion

Morency et al. have applied Hidden Markov Models[54] as a classifier on Youtube dataset that utilized all trimodal data as input. W/511mer et al.[50] adopted two Bi-LSTM to capture contextual features in audios and videos of ICT-MMMO, then utilized support vector machines to detect sentiment in audios and videos. Rosas et al. did a concatenation of all single modality features as a fused representation, then used SVM as the final classifier. The detailed framework of the BiLSTM+SVM model is shown in Fig. 4. Poria et a1.[55] added an additional step of feature selection, using two different feature extractors: (1). the cyclic correlationbased feature subset selection (CFS), (2). principal component analysis(PCA[56]) with top K features. In short, those methods are based on decision-level fusion or simple vector concatenations and cannot fully take advantage of the complementary traits of modalities.

In recent years, researchers are more willing to fuse modalities at an earlier stage, known as feature-fusion, in which those different modalities could learn information from each other at the step of feature extraction. Fig. 5

TABLE I: Multimodal sentiment datasets statistics.

Dataset	Size	Speaker	Number of Modality	Sentiment Label	Emotion Label	Duration
Youtube	300	50	3	Yes	No	30min
ICT-MMMO	370	200	3	Yes	No	14h
MOUD	498	101	3	Yes	No	1h
CMU-MOSI	2199	89	3	Yes	No	2h37min
CMU-MOSEI	23500	1000	3	Yes	Yes	66h

shows the framework of feature fusion in multimodal sentiment analysis.

Poria designed a novel feature-level fusion way, the SPF GMKL[55], a multi-kernel learning method that could be used to tackle heterogeneous data. With the emergence of attention mechanisms, more researchers noticed that multimodal information utilization can be improved through mutual learning. In 2017, Zadeh et a1.[57] achieved integrated learning between features with tensor fusion network, which aggregates interactions of unimodal, bimodal, and trimodal. However, the tensor fusion network is based on the product operation between multiple matrices, which could cause the problem of dimensional explosion easily, making the model too large and hard to be trained. Chen et a1.[58] proposed the Gated MultiEmbedding LSTM with Temporal Attention (GME-LSTM(A)), which performed modality fusion at the word level. GME-LSTM(A) could better model the temporal multi-stmcture. In 2018, Zadeh et al. conceived a creative model called Multi-attention Recurrent Network(MARN) [59], which takes the temporal information into account. The main structure of MARN consists of Long-short Term Hybrid Memory (LSTHM) which is for mulated to capture sequential information and Multiattention Block (MAB), which is used to calculate the relative weight of each modal. Later, Zadeh improved MARN by removing noise that interferes with each other and proposed Memory Fusion Network(MFN)[60]. Compared to MARN, at every timestep, MFN would not depend on the encoded information at the former step, therefore, MFN integrates multimodal information independently, which implements "Delta-memory attention" and "Multi-View Gated Memory" to complete timestep capture and modality interaction. In 2018, Zadeh et al. build the CMUMOSEI dataset[53] and proposed a new fusion method, which combined multiple modalities in a hierarchical graph structure, called DFG. There are 2 steps in DFG: 1. Iteratively modality information updating. First, they used unimodal to model the bimodal interactive information. Then the unimodal and bimodal interactive information are combined to jointly model the trimodal relations. Finally, the final modality combination consists of the three types of interactive information from those three modalities. 2. The fusion weights of the previous time series are explored to guide unimodal feature extraction of the next time step. Hence, the importance of each modality in each temporal sequence can efficiently be used in adjusting the structure of the graph dynamically. Ghosal et al. [61] adopted the idea of conversation, which models the relations between pairs of modalities, such as image and text, text and audio, audio and image, and then concatenate those pairs together as the final fusion vector.

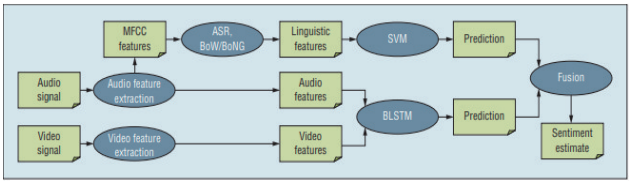

Fig. 4: System architecture for fusion of audio-visual and linguistic information mentioned in[50]

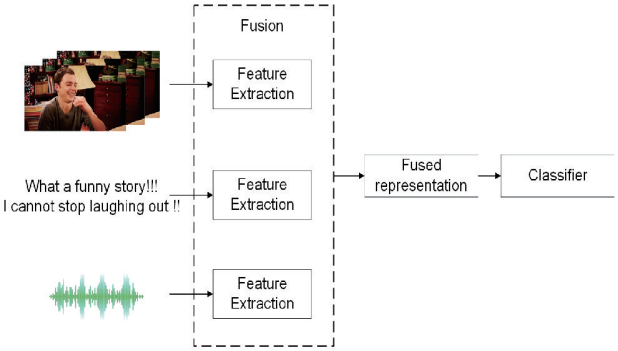

Fig. 5: The overall framework of feature-fusion

TABLE II: The results of some models in CMU-MOSI and CMU-MOSEI.

Model	Dataset			
	CMU-MOSI		CMU-MOSEI	
	Acc	F1-score	Acc	F1-score
RF[57]	56.4	56.3	-	-
CNN-MKL[55]	73.1	75.2	-	-
GME-LSTM(A)[58]	76.5	73.4	-	-
DHF[62]	76.9	76.9	-	-
Graph-MFN[53]	76.9	77	-	-
MARN[59]	77.1	77	-	-
TFN[57]	77.1	77.9	-	-
BC-LSTM[64]	80.3	-	-	-
DCCA[65]	80.6	80.57	**83.62**	**83.75**
Multilogue-Net[66]	81.19	80.1	82.1	80.01
MMMU-BA[61]	82.31	-	79.8	-
MulT[67]	81.1	81	82.5	82.3
B2+B4[63]	83.91	**81.17**	81.17	78.53
Human[59]	**85.7**	-	-	-

In 2019, Georgiou et a1.[62] proposed a hierarchical fusion mechanism, a method named DHF which fuses the representations of multi modalities on the network layer and fee~fforwards them to achieve hierarchical deep integration. It's a mechanism that can be extended to a neural network with any depth. In 2020, Kumar et a1.[63] used the self-attention mechanism and gating mechanism to capture the long-term dependence and leam the relative importance of different modalities, respectively. TABLE II shows the details of comparison results of various methods on the CMU-MOSI and CMU-MOSEI datasets.

3. CHALLENGES

This section will introduce existing challenges in the multimodal sentiment analysis task, which includes problems in datasets and difficulties in approaches. Although there have been many techniques in multimodal sentiment analysis, there are also various limitations in this area. First of all, those methods are highly dependent on high-quality annotated data. Since the multimodal sentiment analysis task requires data from different sources, including text, image, and audio, the alignment of multimodal data is very challenging. To alleviate this problem, current multimodal datasets labeled all the related information on video level. However, when dealing with practical multimodal sentiment analysis problems, the data may not be presented in the form of video, but maybe a collection of discrete pictures and texts. On that kind of dataset, a model trained on a video-labeled dataset cannot be generalized. Therefore, how to make the model of multimodal sentiment analysis not stick to video format data is an urgent problem to be solved. Second, from the existing dataset, we find that all labels are either positive or negative, while in practical social media users are more likely to present neutral opinions. Therefore, it is imperative to explore new datasets which can detect neutral sentiment. Third, because of the same information carried by various data, redundant information is mixed in the mutual fusion learning process. How to remove the interference of this redundant information is a big challenge. Different levels of noise are another concern, considering the noises generated by various modalities are different. The next challenge is how to combine data from heterogeneous sources, such as text is symbolic, pictures are RGB matrixes. Here video is a time-series RGB matrix, and the sound needs to be sampled into a one-bit array.

4. CONCLUSION

In this paper, we carefully review the text-based and multimodal sentiment analysis task, including its definition, datasets, various methods, and challenges. However, there are some problems with text-based sentiment analysis in many practical applications. For example, the text sequence is short, thus there are many colloquial or network or a lot of abbreviations in the text. In these situations, the model trained based only on the texts cannot be very effective. Since the model can't capture the true semantic information expressed in the texts. Therefore, other multimedia information is needed to complement the semantics to correct such errors. In addition, with the development of social media, such as lnstagram and Twitter, people are more willing to share pictures and short videos to express their opinions because sometimes audio and images could express more prolific emotions. Therefore, it is necessary to explore new approaches to multimodal sentiment analysis. Moreover, the applications of this task are broad. For instance, it can be applied to video classification and spotlight clips of video. Beyond practical applications, multimodal sentiment analysis tasks would assist other research tasks. For example, through the emotional identity between multidata, connections between heterogeneous data, such as text and pictures, can be established. In addition, because multimodal sentiment analysis could exploit more information than unimodal sentiment analysis, it could provide more accurate results.

REFERENCES

[1] J. Deonna and F. Teroni, *The emotions: A philosophical introduction.* Roudedge, 2012.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pretraining of deep bidirectional fransformers for language understanding" *arXiv preprint arXiv:1810.04805*, 2018.

[3] T B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, Ih Dhariwal, A. Neelakuntan, Ih Shyam, G. Sastry, A. Askell et al., "Language models are few-shot learners" *arXiv preprint arXiv: 2005.14165,* 2020.

[4] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up2 sentiment classification using machine learning techniques, " *arXiv preprint cs/ 0205070*,2002.

[5] L. Wang, *Support vector machines: theory arid applications*. Springer Science & Business Media, 2005, vol. 177.

[6] I. Rish et al., "Ail empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence,* vol. 3, no. 22, 2001, pp. 41-46.

[7] A. Berger, S. A. Della Pietra, and V. J. Della Pietra, "A maximum entropy approach to natural language processing" Computational lin guistics, vol. 22, no. 1, pp. 39-71, 1996.

[8] 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.

[9] T. Hardeniya and D. A. Borikar, "Dictionary based approach to sentiment analysis-a review," *International Journal of Advanced Engineering, Management arm Science*, vol. 2, no. 5, p. 239438, 2016.

[10] K. Ghag and K. Shah, "Sentitfi~ff sentiment classification using relative term frequency inverse document frequency," *International Journal of Advanced Computer Science arm Applications,* vol. 5, no. 2, 2014.

[11] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, 1972.

[12] S. Baccianella, A. Esuli, and F Sebastiani, "Sentiworchnet 3.0: all enhanced lexical resource for sentiment analysis and opinion mining." in *Lrec*, vol. 10, no. 2010, 2010, pp. 2200-2204.

[13] G.A. Miller, "Wordnet: a lexical database for english," *Cornmunications oftheACM*, vol. 38, no. 11, pp. 39-41, 1995.

[14] A. Neviarouskaya, H. Prendinger, and M. Ishizuka, "Sentiful: A lexicon for sentiment analysis, " *IEEE Transactions on Affctive Computing*, vol. 2, no. 1, pp. 22 36, 2011.

[15] E. Cambria, S. Poria, R. Bajpai, and B. Schuller, "Senticnet 4: A semantic resource for sentiment analysis based on concepmul primitives, " in *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*, 2016, pp. 2666-677.

[16] E. Cambria, C. Havasi, and A. Hussain, "Senticnet 2: A semantic and affective resource for opinion mining and sentiment analysis," in *Twenty- Fifth international FLAIRS conference*, 2012.

[17] P.D. Tumey, "Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews," *arXiv preprint cs/0212032*,2002.

[18] M. Ahmad, S. Aftab, S. S. Muhammad, and S. Ahmad, "Machine learning techniques for sentiment analysis: A review," *Int. J. Multidiscip. Sci. Eng,* vol. 8, no. 3, p. 27, 2017.

[19] J. R. Quinlan, "Indudion of decision trees," *Machine learning*, vol.

1, no. 1, pp. 81 106, 1986.

[20] R. J. Lewis, "An introdudion to classification and regression tree (cart) analysis," *in Annual meeting of the society for academic emergency medicine in Sari Francisco, California*, vol. 14, 2000.

[21] A. Maas, R. E. Daly, Ih T Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies,* 2011, pp. 142-150.

[22] C. Bhadane, H. Dalai, and H. Doshi, "Sentiment analysis: Measuring opinions" *Procedia Computer Science,* vol. 45, pp. 808 814, 2015.

[23] I. Smeureanu and C. Bucur, "Applying supervised opinion mining techniques on online user reviews," *Informatica economica,* vol. 16, no. 2, p. 81, 2012.

[24] Ih F. Brown, V. J. Della Pietra, th V. Desouza, J. C. Lai, and R. L. Mercer, "Class-based n-gram models of natural language" *Computational linguistics,* vol. 18, no. 4, pp. 467-480, 1992.

[25] Y.E. Soelistio and M. R. S. Surendra, "Simple text lnining for sentiment analysis of political figure using naive bayes classifier method," *arXiv preprint arXiv:*1508.05163, 2015.

[26] L. Dey, S. Chakraborty, A. Biswas, B. Bose, and S. Tiwari, "Sentiment analysis of review datasets using naive bayes and k-lin classifier," *arXiv preprint arXiv*:1610.09982, 2016.

[27] S. A. Dudani, "The distance-weighted k-nearest-neighbor rule," *IEEE Transactions on System& Man, arid Cybernetics,* no. 4, pp. 325-327, 1976.

[28] M. S. Hajmohammadi and R. Ibrahim, "A sore-based method for sentiment analysis in persian language" in *International Conference on Graphic arid Image Processing (ICGIP 2012),* vol. 8768. International Society for Optics and Photonics, 2013, p. 876-838.

[29] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey" *Wiley Interdisciplinary Reviews: Data Mining arm Knowledge Discovery,* vol. 8, no. 4, p. e1253, 2018.

[30] J. Pennington, R. Socher, and C. D. Maiming, "Glove: Global vectors for word representation," in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532-1543.

[31] J. L. Ehnan, "Finding structure in time" *Cognitive science, vol.* 14, no. 2, pp. 179-211, 1990.

[32] Y' LeCun, Y. Bengio et al" "Convolutional networks for images, speech, and time series," *The handbook of brain theory arm neural networks,* vol. 3361, no. 10, p. 1995, 1995.

[33] Q. Le and T. Mikolov, "Distributed representations of sentences and documents" in *Interncaional conference on machine learning*. PMLR, 2014, pp. 1188-1196.

[34] Y. Kim, "Convolutional neural networks for sentence classification," *Eprint Arxiv*, 2014.

[35] T Chen, R. Xu, Y. He, Y. Xia, and X. Wang, "Learning user and product distributed representations using a sequence model for sentiment analysis, " *IEEE Computational Intelligence Magazirne,* vol. 11, no. 3, pp. 34-44, 2016.

[36] J. Chmng, C. Gulcebre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling" *arXiv preprint arXiv*:1412.3555, 2014.

[37] A. Severyn and A. Moscbitti, "Twitter sentiment analysis witb deep convolutional neural networks;' in *Proceedings of the 38th interrnational ACM SIGIR cornferernce on research arid development itl information retrieval,* 2015, pp. 959-462.

[38] R. Socber, B. Huval, C. D. Manning, and A. Y. Ng, "Semantic com positionality tbrough recursive matrix-vector spaces," in *Proceedirngs of the 2012 joint conference on empirical methods itl natural language processing arid computational natural language learning,* 2012, pp.1201-1211.

[39] R Vateekul and T. Koomsubba, "A study of sentiment analysis using deep learning techniques on tbai twitter data," in *2016 13th International Joint Conferernce on Computer Science arid Software Engineering (JCSSE).* IEEE, 2016, pp. 1 6.

[40] Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based lstm for aspect-level sentiment classification;' in *Proceedings of the 2016 conference on empirical methods itl natural language processing,* 2016, pp. 606-615.

[41] S. Hocbreiter and J. Schmiclhuber, "Long short-term memory," *Neural computation,* vol. 9, no. 8, pp. 1735-1780, 1997.

[42] D. Tung, B. Qin, X. Feng, and T. Liu, "Effective lstms for target-dependent sentiment classification; *'arXiv preprint arXiv*: 1512.01100, 2015.

[43] Z. Zhang, L. Wang, Y. Zou, and C. Gan, "The optimally designed dynamic memory networks for targeted sentiment classification," *Neu rocomputing*, vol. 309, pp. 36-45, 2018.

[44] A. Kmnar, O. Irsoy, E Ondmska, M. Iyyer, J. Bradbury, I. Gulrajani, V. Zhong, R. Paulus, and R. Socher, "Ask me anything: Dynamic memory networks for natural language processing" in *International conference on machine learning.* PMLR, 2016, pp. 1378- 1387.

[45] J. Wagner, E Arora, S. Cortes, U. Barman, D. Bogdanova, J. Foster, and L. Tounsi, "Dcu: Aspect-based polarity classification for semeval task 4;' 2014.

[46] L. Dong, F. Wei, C. Tun, D. Tang, M. Zhou, and K. Xu, "Adaptive recursive neural network for target-dependent twitter sentiment classification," in *Proceedirngs of the 52rd annual meeting of the association for computational linguistics (volume 2: Shortpnpers),* 2014, pp. 49-54.

[47] J. Yu, J. Jiang, and R. Xia, "Global inference for aspect and opinion terms co-extraction based on multi-task neural networks; ' *IEEE/ACM Trnsactions on Audio, Speech, arid Language Processing,* vol. 27, no. 1, pp. 168-177, 2018.

[48] A. Graves and J. Scbmidhuber, "Framewise phoneme classification witb bidirectional lstm and otber neural network architectures, " *Neural networks,* vol. 18, no. 5-6, pp. 602-510, 2005.

[49] L.-P. Morency, R. Mihalcea, and P Doshi, "Towards multimodal sen timent analysis: Harvesting opinions from tbe web;' in *Proceedings of the 13th international conference on multimodal interfnces*, 2011, pp. 169-176.

[50] M. W61hner, F. Weninger, T Knaup, B. Scbuller, C. Sun, K. Sagae, and L.-P Morency, "Youtube movie reviews: Sentiment analysis in an audio-visual context," *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 46-53, 2013.

[51] V. R Rosas, R. Mihalcea, and L.-P Morency, "Multimodal sentiment analysis of spanish online videos," *IEEE Intelligent Systems*, vol. 28,no. 3, pp. 38-45, 2013.

[52] A. Zadeb, R. Zellers, E. Pincus, and L.-P Morency, "Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos;' *arXiv preprint arXiv:*1606.06259, 2016.

[53] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P Morency, "Multimodal language analysis in tbe wild: Cmu-mosei dataset and inlterpretable dynamic fusion graph," *in Proceedings of the 56#1 Annual Meeting of the Association for Computational Lirguistics (Volume 1: Long Papers),* 2018, pp. 2236-2246.

[54] L. Rabiner and B. Juang, "An introduction to hidden markov models;' *ieee assp magazine*, vol. 3, no. 1, pp. 4 16, 1986.

[55] S. Poria, E. Cambria, and A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," in *Proceedings of the 2015 conference on empirical methods itl natural language processing*, 2015, pp. 2539-2544.

[56] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* vol. 61, no. 3, pp. 611-622, 1999.

[57] A. Zadeb, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis;' *arXiv preprint arXiv*:1707.07250, 2017.

[58] M. Chen, S. Wang, P. P. Liang, T. Baltrusaitis, A. Zadeb, and L.-P.

Morency, "Multimodal sentiment analysis witb word-level fusion and reinforcement learning," *in Proceedings of the 19th ACM International Conference on Multimodal Interaction,* 2017, pp. 163-171.

[59] A. Zadeh, R P. Liang, S. Poria, R Vij, E. Cambria, and L.-P Morency, "Multi-attention recurrent network for human communication comprehension," *in Proceedings of the AAAI Conference on Artificial Intelligence,* vol. 32, no. 1, 2018.

[60] A. Zadeb, R P. Liang, N. Maztumder, S. Poria, E. Cambria, and L.-P Morency, "Memory fusion network for multi-view sequential learning," *in Proceedings of the AAAI Conference on Artificial Intelligence,* vol. 32, no. 1, 2018.

[61] D. Ghosal, M. S. Akhtar, D. Chauhan, S. Poria, A. Ekbal, and R Bhat-tacharyya, "Contextual biter-modal attention for multi-modal sentiment analysis," *in proceedings of the 2018 conference on empirical methods itl natural language processing,* 2018, pp. 3454-3466.

[62] E. Georgiou, C. Papaioannou, and A. Potamianos, "Deep hierarchical fusion witb application in sentiment analysis." *in INTERSPEECH,* 2019, pp. 1646-1650.

[63] A. Kumar and J. Vepa, "Gated mechanism for attention based multi modal sentiment analysis," *in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech arid Signal Processing (ICASSP). IEEE,* 2020, pp. 4477-4481.

[64] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeb, and L. -R Morency, "Context-dependent sentiment analysis in user-generated videos;' *in Proceedings of the 55th annual meeting of the association for computational Iinguistics (volume 1: Long papers),* 2017, pp. 873-883.

[65] Z. Sun, P K. Sarma, W. Setbares, and E. R Bucy, "Multi-modal sentiment analysis using deep canonical correlation analysis," *arXiv preprint arXiv*:1907.08696, 2019.

[66] A. Shenoy and A. Sardana, "Muhilogue-net: A context aware rnn for multi-modal emotion detection and sentiment analysis in conversation," *arXiv preprint arXiv:*2002.08267, 2020.

[67] Y.-H. H. Tsai, S. Bai, R P Liang, J. Z. Kolter, L.-P Morency, and R. Salakhutclinov, "Multimodal transformer for unaligned multimodal language sequences;' *in Proceedings of the conference. Associcaion for Computational Linguistics. Meeting,* vol. 2019. NIH Public Access, 2019, p. 6558.

Yuanyuan Zhang was born in Chengdu, China. Now Yuanyuan Zhang is a master's student of Key Laboratory of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences. Yuanyuan Zhang got her bachelor's degree in Engineering in tbe Department of computer science at Sicbuan University in 2018. Since 2018, her research interests are natural language processing and deep learning.

Xiao Wang received his B.S. degree from Departmeat of Computer Science in Xl'an Jiaotong University, Xi'an, China in 2018. He is currently pursuing his Ph.D. degree in tbe Department of Computer Science at Purdue University, West Lafayette, IN, USA. His research interests include deep learning, computer vision, bioinformatics and intelligent systems.

Fei-Yue Wang (S'87 M'8%SM'94 F'03) receivedtile Ph.D. degree in computer and systems engineerLng from tile Rensselaer Polytechnic Institute, Troy, NY, USA, in 1990. He joined The University of Arizona in 1990 and became a Professor and tile Director of tile Robotics and Automation Laboratory and tile Program in Advanced Research for Complex Systems. In 1999, he founded tile Intelligent Control and Systems Engineering Center at tile Institute of Automation, Chinese Academy of Sciences (CAS), Beijing, China, under tile support of the Outstanding Chinese Talents Program from the State Planning Council. In 2002, he was appointed as the Director of the Key Laboratory of Complex Systems and Intelligence Science, CAS. In 2011, he became the State Specially Appointed Expert and the Director of the State Key Laboratory for Management and Control of Complex Systems. His cmrent research focuses on methods and applications for parallel intelligence, social computing, and knowledge automation. He is a fellow of the INCOSE, IFAC, ASME, and AAAS. In 2007, he received the National Prize in Natural Sciences of China and became an Outstanding Scientist of ACM for his work in intelligent control and social computing. He received the IEEE ITS Outstanding Application and Research Awards in 2009 and 2011, respectively. In 2014, he received the IEEE SMC Society Norbert Wiener Award. Since 1997, he has been serving as the General or Program Chair of over 30 IEEE, INFORMS, IFAC, ACM, and ASME conferences. He was the President of the IEEE ITS Society from 2005 to 2007, the Chinese Association for Science and Technology, USA, in 2005, the American Zhu Kezhen Education Foundation from 2007 to 2008, the Vice President of the ACM China Council from 2010 to 2011, and the Vice President and the Secretary General of the Chinese Association of Automation from 2008 to 2018. He was the Founding Editor-in-Chief (EiC) of the International Journal of Intelligent Control and Systems from 1995 to 2000, the IEEE ITS Magazine from 2006 to 2007, the IEEE/CAA JOURNAL OF AUTOMATICA SINICA from 2014 to 2017, and the China's Journal of Command and Control from 2015 to 2020. He was the EiC of the IEEE INTELLIGENT SYSTEMS from 2009 to 2012 and the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS from 2009 1o 2016. He has been the EiC of the IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS since 2017, and the Founding EiC of China's Journal of Intelligent Science and Technology since 2019. He is cmrently the President of the CAA's Supervision Council, IEEE Council on RFID, and the Vice President of the IEEE Systems, Man, and Cybernetics Society.