Multimodal Tweet Sentiment Classification Algorithm Based on Attention Mechanism

Peiyu Zou[1](✉) and Shuangtao Yang[2](✉)

[1] Northeast Agricultural University, Harbin 150036, China
zoupeiyul2l3@gmail.com
[2] Lenovo AI Lab, Beijing 100000, China
lufiedby@gmail.com

Abstract. With the rapid development of Internet, multimodal sentiment classification has become an important task in natural language processing research. In this paper, we focus on the sentiment classification of tweets that contains both text and image, a multimodal sentiment classification method for tweets is proposed. In this method Bidirectional-LSTM model is used to extract text modality features and VGG-16 model is used to extract image modality features. Where all features are extracted, a new multimodal feature fusion algorithm based on attention mechanism is used to finish the fusion of text and image features. This fusion method proposed in this paper can give different weights to modalities according to their importance. We evaluated the proposed method on the Chinese Weibo dataset and SentiBank Twitter dataset. The experimental results show method proposed in this paper is better than models that only use single modality feature, and attention based fusion method is more efficient than directly summing or concatenating features from different modalities.

Keywords: Multimodal · Sentiment classification · Attention mechanism

1 Introduction

With the rapid development of Internet, social network become more and more popular in daily live. People begin to tell others what they are doing, what they are feeling, what they are thinking, or what is happening around them more and more through social network. As a result, how to excavate people's sentiment expression in social networks exactly has attracted more and more attention. Sentiment classification has become an important task in the natural language processing research.

Some related works have been conducted on social network (Twitter or Sina Weibo and so on) sentiment classification, such as [1–4], while these work mainly focus on the use of single modality feature for sentiment classification, most mainly use text features, and rarely use image, audio or visual feature. With the rapid development of social network, user's tweets are often multimodal, when users publish tweets, may upload an image, audio or video at the same time. If we cannot effectively use these multimodal features, we will not be able to accurately evaluate user's sentiment. So,

A. Monreale et al. (Eds.): ECML PKDD 2018 Workshops, CCIS 967, pp. 68–79, 2019.
https://doi.org/10.1007/978-3-030-14880-5_6

recently multimodal sentiment classification has attracted more and more attention, such as [5–8].

In this paper, we will focus on the sentiment classification of tweets (also include micro-blogs from Sina Weibo) that contain both text and image. A sentiment classification method for multimodal tweets classification is proposed, which use BI-LSTM model and VGG-16 model to extract text features and image features separately and then use attention mechanism to complete the fusion of these two features. The proposed method has been tested on Chinese Weibo dataset and SentiBank Twitter dataset, all showing promising results.

This paper is organized as follows: Sect. 2 introduces related works of multimodal sentiment classification. Section 3 describes the method proposed in this paper. Section 4 carries out some comparison experiment and result analysis. And conclusions are shown in Sect. 5.

2 Related Work

Multimodal Sentiment Classification can be divided into two categories: early fusion and late fusion according to modality information fusion time as described in [9, 10]. In early fusion (also named feature fusion), features of different modalities are first extracted by different feature extraction model, then these extracted features are fused, which can be finished by concatenating, summing or by some fusion models. Finally, the fusion features are fed into classifier. Similar to the early fusion strategy, in the late fusion (also named decision fusion), different models are used to extract features from different modalities, but late fusion does not fuse features directly, it will first feed features to a classifier in each modality, and finally a model will be used to combine all classification results from different modalities to get the final classification results.

In multimodal sentiment classification, some related works have been conducted. Paper [5] explored the joint use of multiple modalities for the purpose of classifying the polarity of opinions in online videos and experiment result showed that the integration of visual, audio, and textual features can improve significantly over the individual use of one modality at a time. In [6] and [11], Support Vector Machine (SVM) is used for multimodal sentiment classification, they both first extract features from different modalities, then combine features into a single vector and feed the vector into SVM classifier. In [7], a unified model (CBOW-DA-LR) was proposed, which works in an unsupervised and semi-supervised way to learn text and image representation. For video sentiment classification task, in order to capture inter-dependencies and relations among the utterances in a video, [12] developed a LSTM-based network to extract contextual features from the utterances and got better result comparing to traditional method. [8] proposed a novel method for multimodal emotion recognition and sentiment analysis, which uses deep convolutional neural networks to extract features from visual and textual modalities and then feeds such features to multiple kernel learning (MKL) classifier which is a feature selection method and it is able to combine data from different modalities effectively. Paper [13] proposed a multimodal affective data analysis framework which can extract user opinion and emotions from video content, and multiple kernel learning is also used to combine visual, audio and textual modalities. Paper [14] introduced a new end-to-end early fusion method for multimodal sentiment analysis termed Tensor Fusion Network which use a fusion layer to disentangles uni-modal, bimodal and tri-modal dynamics by modeling each of them explicitly. Paper [15] proposed a multimodal sentiment analysis model named Select-Additive Learning which attempts to prevent identity-dependent information from being learned in a deep neural network. Paper [16] proposed a model named GME-LSTM which is able to better model the multimodal structure of speech through time and perform better sentiment comprehension. GME-LSTM is composed of 2 modules: Gated Multimodal Embedding which alleviates the difficulties of fusion when there are noisy modalities; LSTM performs word level fusion at a finer fusion resolution between input modalities and attends to the most important time steps.

3 The Proposed Model

In this paper, we will focus on early fusion strategy. According to the process of early fusion strategy, we can find that it attempts to fuse the features from different modalities to obtain a global feature which contains all valuable features from every modality. In the fusion, the simplest approach is to concatenate or sum features from different modalities directly to get the global representation. Considering concatenate or sum fusion method does not reflect the importance of feature from different modalities, in this paper, we propose a multimodal feature fusion method based on attention mechanism, which will give different weights to different modalities according to their importance, and it helps to complement and disambiguate different features from different modalities. Our model is shown as Fig. 1.

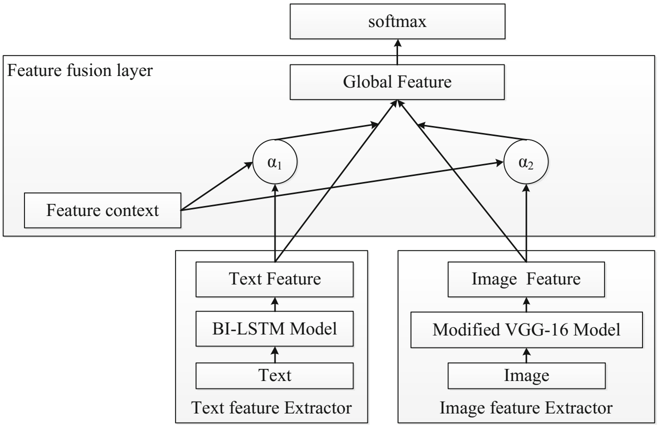


Fig. 1. Attention based multimodal sentiment classification model

3.1 Text Feature Extractor

In order to get better text modal features, it is necessary to select the appropriate text feature (lexical) representation method first. Text feature representation has been extensively studied in text categorization tasks. Mainstream text representation methods include word bag representation and word vector representation. In word-bag model word is represented in one-hot vector which cannot effectively express the relation of different words, for example “like” and “love”. In addition, the dimension of the one-hot vector needs to be consistent with the size of the vocabulary. The dimension one-hot vector is often too high and extremely sparse which is not unfriendly to computation. So, text features will be represented by word2vec in this paper.

In word2vec, all words are embedded into an N-dimension semantic space [17, 18]. In this semantic space, the semantic distance of related words will be closer, and the semantic distance is far away from unrelated words. After the words are represented into vectors, the sentence will be transformed into M * N vector, while N is the dimension of word2vec and M is the word number of the sentence. Then we can use CNN or LSTM model to encode this two-dimensional vectors to obtain effective text modal features [19].

Using CNN or LSTM to encode the sentence, it can effectively capture the sequence information between words in the sentence [20–23]. In this paper we will use Bi-directional LSTM to extract text feature. BI-LSTM can be considered as a composed of two different directions LSTM, mainly to compensate for the deficiency of the single direction LSTM in capturing context ability. The process of text feature extraction based on Bi-LSTM model is shown in the following Fig. 2.

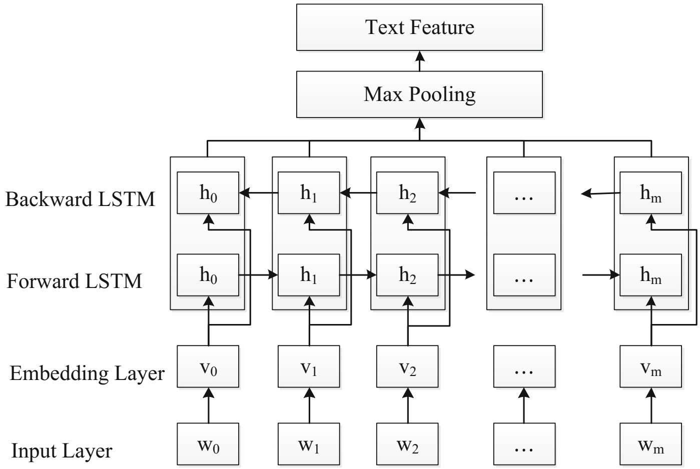


Fig. 2. Bidirectional-LSTM text feature extractor

Given a sentence with words w_t $t \in [0, m]$, we first transfer every word w_t to a N-dimension vector v_t through a pre-trained word2vec. Then we will use LSTM to encode sentence from both directions, as shown in Fig. 2.

In forward LSTM, sentence is encoded in the direction from word w_0 to word w_m. At t time-step, the hidden layer state of the LSTM is expressed as:

$$f\_h_t = forward\_lstm(v_t),\ t \in [0, m] \tag{1}$$

In backward LSTM, sentence is encoded in the direction from word w_m to word w_0. At t time-step, the hidden layer state of the LSTM is expressed as:

$$b\_h_t = backward\_lstm(v_t),\ t \in [0, m] \tag{2}$$

We will get h_t for a given word w_t by concatenating the forward hidden state $\mathrm{f\_h}_\mathrm{t}$ and forward hidden state $\mathrm{b\_h}_\mathrm{t}$.

$$h_t = concatenate(f\_h_t, b\_h_t) \tag{3}$$

Then we need run max pooling over sequence $[\mathrm{h}_0, \mathrm{h}_1, \ldots, \mathrm{h}_\mathrm{m}]$, and get the final feature. In order to make it easy to explain, we set the dimension of LSTM hidden layer in both forward LSTM and backward LSTM to d, thus for each sentence we're going to get a 256-dimensional vector as its text modality feature. Considering CNN models are also widely used in NLP tasks, we will also evaluate the performance of famous Text-CNN model [20] in our text feature extraction task.

3.2 Image Feature Extractor

In image feature extraction, CNN has become a very standard and universal model. In this paper VGG-16 pre-trained with the ImageNet dataset will be used as an image feature extractor [24, 25]. ImageNet dataset is a super large image annotation data set, and pre-training VGG-16 model with this data set is very helpful to obtain better image feature extraction capability.

The input of the VGG-16 network is 3 × 224 × 224 RGB color image. The network has 13 convolution layers and all the convolution layers have a very small receptive field which is 3 × 3. During the convolution process, max pooling is carried out after the third, fourth, seventh, tenth and thirteenth convolution layer. Max pooling is performed over a 2 × 2 pixel window with stride 2. After a stack of thirteen convolution layers, three fully-connected (FC) layers follows and the channel of the first, second FC layer is 4096, the third FC layer is 1000.

First, we need to preprocess the image before we feed image to the model. The preprocessing phase mainly adjusts the size of all the image data to 3 × 244 × 244, and we do the same preprocessing for each pixel: subtracting the mean RGB value for each pixel in the training set. The images are enhanced by random transformation in training set.

Second, we pre-train the VGG-16 model on Image Net dataset. When we finish the pre-training, model parameters except the last full connected layer in the model are frozen. We will modify the output dimension of last layer to 2 $*$ d, and d is the dimension of LSTM hidden layer which has been introduced in Sect. 3.1. In the proposed model, image features and text features need to be mapped into same dimension space which means the dimension of the text feature should be consistent with the dimension of text feature. Therefore, we need to transform the dimension of the image features through the last FC layer which dimension should be 2 $*$ d.

3.3 Attention Base Fusion of Text and Image Feature

In early fusion, after we get text features and image features, we have to fusion them to get global features which will contain all important features from different modalities. We believe that not all features from different modalities contribute equally to the global feature. In order to get better feature fusion results, we propose an attention mechanism based fusion method for text and image features. In this method we will pay more attention to the modalities that are more important to classification which means features extracted from the modalities will contribute more in global feature.

For illustrative purposes, we record the textual features extracted by Text Feature Extractor (introduced in Sect. 3.1) $\mathrm{f_{text}}$, and mark the image features extracted by Image re Extractor (introduced in Sect. 3.2) as $\mathrm{f_{image}}$, and they have the same dimensions. First, we feed the $\mathrm{f_{text}}$ and $\mathrm{f_{image}}$ to a one-layer MLP respectively, we will get:

$$u_{text} = MLP(f_{text}) = \tanh(f_{text}W_u + b_u) \tag{4}$$

$$u_{image} = MLP\left(f_{image}\right) = \tanh\left(f_{image}W_u + b_u\right) \tag{5}$$

where $\mathrm{W_u}$ represents the weight matrix in MLP, and $\mathrm{b_u}$ represents the bias in MLP.

Then, we will measure the importance of text features and image features by calculating the similarity between the $\mathrm{u_{text}}$ and the $\mathrm{u_{image}}$ and $\mathrm{u_{context}}$, parameter $\mathrm{u_{context}}$ is randomly initialized and jointly learned during the training process. We also will use softmax function to normalize the feature weights $\mathrm{a_{text}}$ and $\mathrm{a_{image}}$. The attention score $\mathrm{a_{text}}$ for text feature is computed as formula, where $\mathrm{u_{text}^T}$ is the transpose of $\mathrm{u_{text}}$, and $\mathrm{u_{image}^T}$ is the transpose of $\mathrm{u_{image}}$.

$$a_{text} = \frac{\exp\left(u_{text}^T u_{context}\right)}{\exp\left(u_{text}^T u_{context}\right) + \exp\left(u_{image}^T u_{context}\right)} \tag{6}$$

The attention score $\mathrm{a_{image}}$ for image feature is computed as:

$$a_{image} = \frac{\exp\left(u_{image}^T u_{context}\right)}{\exp\left(u_{text}^T u_{context}\right) + \exp\left(u_{image}^T u_{context}\right)} \tag{7}$$

Then we can get global_feature as shown in formula 8.

$$global\_feature = a_{text} f_{text} + a_{image} f_{image} \tag{8}$$

4 Experiment

4.1 Datasets

(1) Chinese Weibo Dataset
We first collected lots of tweet (or named Weibo) from Sina Weibo which is a famous social network in China. Then 12000 tweets are annotated manually. Every tweet contains text and image. All the tweets are labeled as three categories: Positive, Negative and Neutral. The categories of this dataset are balanced, each category, and each category contains 4000 annotated tweets. In order to verify the effectiveness of the proposed model, we adopt a 5-fold cross validation method, which means in each validation 80% annotated tweets are used as training set and 20% are used as testing set. The training set and test set are preprocessed, for text, we mainly normalized the special fields, such as emotion tags and URL, and used Jieba (a Chinese word segmentation tool) to segment tweets content into words.

(2) SentiBank Twitter Dataset
SentiBank Twitter Dataset consists of 470 positive and 133 negative tweets with images, related to 21 topics, annotated using Mechanical Turk [26]. For this dataset, we also use 5-fold cross-validation to evaluate our model.

4.2 Result and Analysis

It needs to be pointed out that we used the pre trained word vectors, respectively. For Chinese Weibo Dataset we pre-trained a 300 dimension word2vec on a very large Chinese text corpus. For SentiBank Twitter Dataset, we also pre-trained a 300 dimension word2vec on other twitter corpus and wiki corpus. We will use micro F1-score to evaluate our model, and we calculate it as described in paper [27]:

$$micro\_precision = \frac{\sum_{j=1}^{Q} tp_j}{\sum_{j=1}^{Q} tp_j + \sum_{j=1}^{Q} fp_j} \tag{9}$$

where Q is the total number of testing case, $\mathrm{tp_j}$ and $\mathrm{fp_j}$ are the number of true positives and false positives for the label $\mathrm{j_{label}}$ considered as a binary label.

$$micro\_recall = \frac{\sum_{j=1}^{Q} tp_j}{\sum_{j=1}^{Q} tp_j + \sum_{j=1}^{Q} fn_j} \tag{10}$$

where fn_j is the number of false negatives for the label j_{label} considered as a binary label.

$$micro\_f1 = \frac{2 \times micro\_precision \times micro\_recall}{micro\_precision + micro\_recall} \tag{11}$$

First, we compared the model proposed in this paper with models those can only use single modality feature (text feature or image feature only). Experiment results are shown in Table 1.

Table 1. Result on Chinese Weibo dataset

Feature	Model	Micro F1-score
Text (only)	CBOW+SVM	71.9%
	Bi-LSTM	73.9%
	Text-CNN	72.7%
Image (only)	VGG-16-modified	70.1%
Text + image	Bi-LSTM+VGG-16	84.2%
	Text-CNN+VGG-16	82.9%

When only use text modality features, we carried out three classification experiments of CBOW+SVM model, BI-LSTM model and Text-CNN model respectively. We finally find that Bidirectional-LSTM model gets the best classification result, comparing to the Text-CNN model, there is a 1.2% increase of micro f1-score.

For image modality experiment, we evaluate VGG-16 modified model which has been described in Sect. 3.2 and get score of 71.5%. By comparing the features of the text separately and using the image features alone, we can find that the text feature contains more explicit sentiment information and is easier to get better classification results. As shown in Table 1, when using text feature and image feature both, we can find the model proposed in the paper gets the best classification result, and comparing to models that using single modality feature there is a particularly obvious improvement, and the micro f1-score is up to 84.2%. As shown in Table 1, we also use Text-CNN model as text feature extractor to replace Bidirectional-LSTM model in our model, and get average accuracy of 82.9%. Comparing to Bi-direction-LSTM version, the micro f1-score drops by 1.3% (Tables 2 and 3).

Table 2. Result on SentiBank dataset

Feature	Model	Micro-F1 score
Text (only)	CBOW+SVM	72.4%
	Bidirectional-LSTM	73.8%
	Text-CNN	73.3%
Image (only)	VGG-16-modified	71.2%
Text + Image	BI-LSTM+VGG-16	82.8%
	Text-CNN+VGG-16	80.7%

Table 3. Result on Chinese Weibo dataset

Feature	Fusion method	Micro-F1 score
Text + image	Sum	79.6%
	Concatenate	82.3%
	This paper	84.2%

For SentiBank Twitter Dataset we conducted the same comparative experiment, and the specific results are as follows:

When only using text modality feature, Bidirectional-LSTM also achieved the highest micro f1-score among CBOW+SVM model, Bidirectional-LSTM model and Text-CNN model. If text modality feature and image feature are all used, our model also achieved the best result.

In multimodal comparison experiments, we also compared the influence of different features (text feature and image feature) fusion methods on sentiment classification. The first fusion method is sum, which means the global feature is the sum of text feature f_{text} and image feature f_{image}, using this fusion method micro f1-score is 79.6%. The second fusion method is concatenate, which means the global feature is the concatenation of text feature f_{text} and image feature f_{image}, and the micro f1-score is 82.3%. The third fusion method is the attention based fusion method proposed in the paper, it achieve the highest micro f1-score which is up to 84.2%.

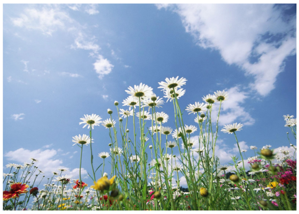
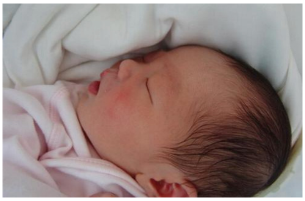
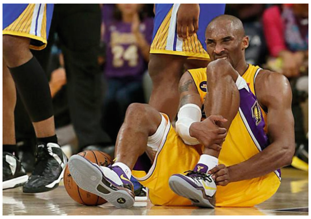
SentiBank experiments can be found in Table 4. The results also show that the feature fusion method based on attention mechanism is more effective than simple sum or concatenate features from different modalities.

Table 4. Result on SentiBank dataset

Feature	Fusion method	Micro-F1 score
Text + Image	Sum	79.6%
	Concatenate	81.2%
	This paper	82.1%

In order to better explain our method, we find some illustrative examples from Weibo Dataset where are shown in Table 5. For the first case, if we predict only based on text modality features, we get negative sentiment which is wrong. When we use text features and image features both, we get the right sentiment label, and the attention score of text modality is 0.29, which is much lower than that of image modality. In the third case, there is a man who sprained his ankle in the picture, so if we only use image modality features, we get negative sentiment, however when we focus on the text content we can make sure that the sentiment of this case is positive. As we can see, our attention based fusion model also paid more attention to text modality features and the attention score of text modality is 0.84, which is much higher than image modality.

Table 5. Illustrative examples from Weibo testset

Image and Text	Feature	Classification result
 Even life is always disappointing and frustrating, we can't stop. (Translated from Chinese)	Text	Negative Confidence = 0.82
	Image	Positive Confidence = 0.35
	Text+Image	Positive Confidence = 0.78 Text modality attention score: 0.29 Image attention score: 0.71
	Ground truth: Positive	
 Maybe god knows what kind of pain I had last night. (Translated from Chinese)	Text	Negative Confidence = 0.91
	Image	Positive Confidence = 0.61
	Text+Image	Positive Confidence = 0.76 Text modality attention score: 0.35 Image attention score: 0.65
	Ground truth: Positive	
 My hero, I hope you can recover soon, love, love and love. (Translated from Chinese)	Text	Positive Confidence = 0.96
	Image	Negative Confidence = 0.45
	Text+Image	Positive Confidence = 0.66 Text modality attention score: 0.84 Image attention score: 0.16
	Ground truth: Positive	

5 Conclusion

In this paper, we focus on the sentiment classification of tweet that contains both text and image. A multimodal sentiment classification method for tweet is proposed. In this method Bidirectional-LSTM model is used to extract text modality feature and VGG-16 model is used to extract image modality feature. After all features are extracted, a new multimodal feature fusion algorithm based on attention mechanism is used to finish the fusion of text and image features. This fusion method proposed in this paper can give different weights to modalities according to its importance. We evaluated the proposed method on the Chinese Weibo dataset and SentiBank Twitter dataset. Our, experimental results show method proposed in this paper is better than models that use single modality feature, and attention based fusion method is more efficient than directly summing or concatenating features from different modalities.

References

1. Saif, H., He, Y., Alani, H.: Semantic sentiment analysis of Twitter. In: Cudré-Mauroux, P., et al. (eds.) ISWC 2012. LNCS, vol. 7649, pp. 508–524. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-35176-1_32
2. Gautam, G., Yadav, D.: Sentiment analysis of Twitter data using machine learning approaches and semantic analysis. In: International Conference on Contemporary Computing, pp. 437–442. IEEE (2014)
3. Zhou, H., et al.: Rule-based Weibo messages sentiment polarity classification towards given topics. In: Eighth SIGHAN Workshop on Chinese Language Processing, pp. 149–157 (2015)
4. Jiang, L., Yu, M., et al.: Target-dependent Twitter sentiment classification. In: Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 151–160. Association for Computational Linguistics (2011)
5. Morency, L.P., Mihalcea, R., Doshi, P.: Towards multimodal sentiment analysis: harvesting opinions from the web. In: International Conference on Multimodal Interfaces, pp. 169–176. ACM (2011)
6. Zadeh, A., Zellers, R., Pincus, E., Morency, L.P.: Multimodal sentiment intensity analysis in videos: facial gestures and verbal messages. IEEE Intell. Syst. **31**(6), 82–88 (2016)
7. Baecchi, C., Uricchio, T., Bertini, M., Bimbo, A.D.: A multimodal feature learning approach for sentiment analysis of social network multimedia. Multimedia Tools Appl. **75**(5), 2507–2525 (2016)
8. Poria, S., Chaturvedi, I., Cambria, E., et al.: Convolutional MKL based multimodal emotion recognition and sentiment analysis. In: IEEE, International Conference on Data Mining, pp. 439–448. IEEE (2017)
9. Atrey, P.K., Hossain, M.A., Saddik, A.E., Kankanhalli, M.S.: Multimodal fusion for multimedia analysis: a survey. Multimedia Syst. **16**(6), 345–379 (2010)
10. Gallo, I., Calefati, A., Nawaz, S.: Multimodal classification fusion in real-world scenarios. In: IAPR International Conference on Document Analysis and Recognition. IEEE (2018)
11. Pérez-Rosas, V., Mihalcea, R., Morency, L.P.: Utterance-level multimodal sentiment analysis. Association for Computational Linguistics (ACL) (2013)
12. Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., Morency, L.P.: Context-dependent sentiment analysis in user-generated videos. In: Meeting of the Association for Computational Linguistics, pp. 873–883 (2017)
13. Poria, S., Peng, H., Hussain, A., Howard, N., Cambria, E.: Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis. Neurocomputing **26**, 217–230 (2017)
14. Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.P.: Tensor fusion network for multimodal sentiment analysis (2017)
15. Wang, H., Meghawat, A., Morency, L.P., Xing, E.P.: Select-additive learning: improving cross-individual generalization in multimodal sentiment analysis, pp. 949–954 (2016)

16. Chen, M., Wang, S., Liang, P.P., Baltrušaitis, T., Zadeh, A., Morency, L.P.: Multimodal sentiment analysis with word-level fusion and reinforcement learning. In: ACM International Conference on Multimodal Interaction, pp. 163–171. ACM (2017)
17. Mikolov, T., et al.: Efficient estimation of word representations in vector space. In: International Conference on Learning Representations, pp. 1–12 (2013)
18. Mikolov, T., Sutskever, I., Chen, K., et al.: Distributed representations of words and phrases and their compositionality. In: International Conference on Neural Information Processing Systems, pp. 3111–3119. Curran Associates Inc. (2013)
19. Bengio, Y.: Learning deep architectures for AI. Found. Trends® Mach. Learn. **2**(1), 1–127 (2009)
20. Kim, Y.: Convolutional neural networks for sentence classification. Eprint Arxiv (2014)
21. Zhang, Y., Wallace, B.: A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. Computer Science (2015)
22. Zhou, P., Qi, Z., Zheng, S., Xu, J., Bao, H., Xu, B.: Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling (2016)
23. Chen, T., Xu, R., He, Y., et al.: Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. Expert Syst. Appl. **72**, 221–230 (2016)
24. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Computer Science (2014)
25. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Li, F.F.: ImageNet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 248–255. IEEE (2009)
26. Bifet, A., Frank, E.: Sentiment knowledge discovery in twitter streaming data. In: Proceedings of the Discovery Science - International Conference, DS 2010, Canberra, Australia, 6–8 October 2010, pp. 1–15. DBLP (2010)
27. Madjarov, G., Kocev, D., Gjorgjevikj, D., Deroski, S.: An extensive experimental comparison of methods for multi-label learning. Pattern Recogn. **45**(9), 3084–3104 (2012)