Document-Level Multimodal Sentiment Analysis

Subjects: Computer Science, Artificial Intelligence

Contributor: Dehong Zeng , Xiaosong Chen , Zhengxin Song , Yun Xue , Qianhua Cai

An increasing number of people tend to convey their opinions in different modalities. For the purpose of opinion mining, sentiment classification based on multimodal data becomes a major focus. Sentiment analysis at the document level aims to identify the opinion on a main topic expressed by a whole document.

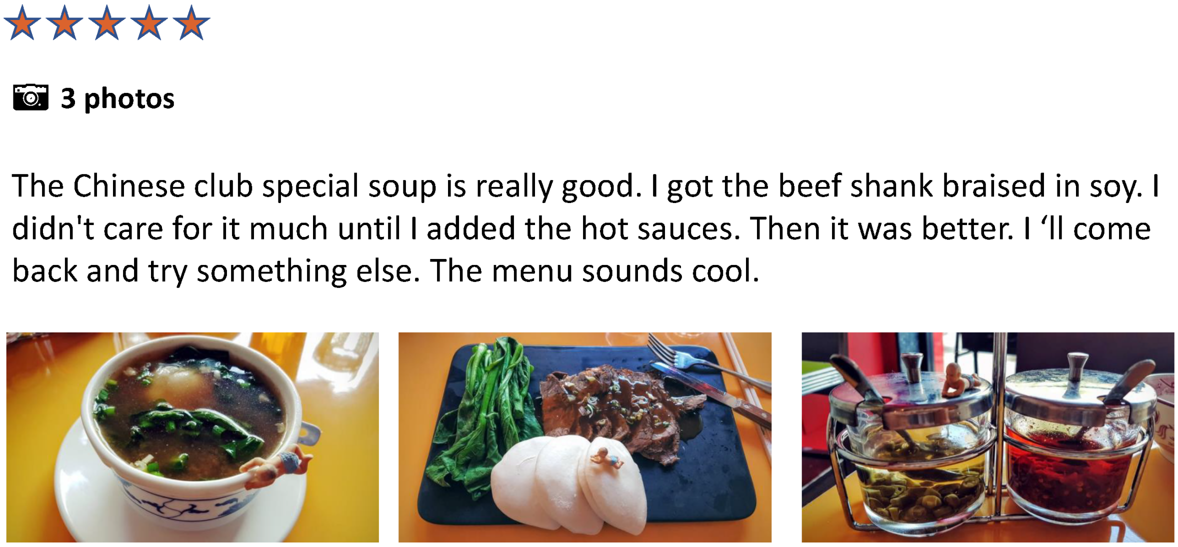
document-level multimodal sentiment classification graph convolutional networks

1. Introduction

Sentiment analysis at the document level aims to identify the opinion on a main topic expressed by a whole document. Instead of understanding the sentiment at the sentence or aspect level, document-level sentiment analysis (DLSA) tends to extract the overall sentiment of the whole document. Driven by the commercial demands, document-level sentiment analysis (DLSA), on the basis of deep learning algorithms, is currently widely employed to deal with the online product reviews [1]. That is, the general sentiment toward a product or service based on an overwhelming abundance of textual data is captured directly and classified as either positive, neutral or negative [2]. As such, DLSA is capable of delivering opinions in a way that clearly facilitates the product recommendation and sales prediction [3].

Typically, the task of DLSA mainly focuses on dealing with the textual information. In line with the flourish of deep neural networks, researchers exploit a variety of methods to extract textual features and capture the context information from the document. Zhou et al. utilize a convolutional neural network (CNN) to extract a sequence of higher-level phrase representations and feed them into a long short-term memory recurrent neural network (LSTM) to obtain the sentence representation [4]. Yang et al. propose a hierarchical attention network, which aims to extract the features at both the word and sentence level to construct the document representation [5]. The DLSA models based on graph neural networks are also developed [6]. In this model, graphs for each input text are built with significant local features extracted and the memory consumption reduced.

More recently, the widespread use of smartphones has given rise to more opportunities to express opinions via different modalities (i.e., textual, acoustic and visual modalities). On social media, the text and the image are generally taken to mutually reinforce and complement each other; see **Figure 1**. For this reason, there is an ongoing trend to devise document-level multimodal sentiment analysis (DLMSA) methods that tackle multimodal information. In practice, the major challenge of DLMSA models lies in aligning and fusing textual and visual information using data of distinguishing format and structure. On the task of multimodal sentiment analysis, Zadeh et al. work on computing the outer product between modalities to characterize the multimodal relevance [7]. However, this scheme greatly increases the feature vector dimension, which results in the difficulty and complexity of model training. Furthermore, recent publications report the multimodal fusion at the feature level. Truong et al. consider visual information as a source of alignment at the sentence level and assign more attention to image-related sentences [8]. In addition, Du et al. use image features to emphasize the text segment by the attention mechanism and take a gating unit to retain valuable visual information [9].



★★★★★

3 photos

The Chinese club special soup is really good. I got the beef shank braised in soy. I didn't care for it much until I added the hot sauces. Then it was better. I 'll come back and try something else. The menu sounds cool.

Figure 1. An example of multimodal review.

2. Document-Level Sentiment Analysis

Sentiment analysis is a major focus in the field of natural language processing that has gained an increasing amount of attention. Sentiment analysis determines sentiment polarity or predicts sentiment scores from a given text. With the advancement in social media, massive user-generated texts are accessible, which has further promoted the research in sentiment analysis [10].

In general, a document consists of multiple sentences. While once restricted to processing methods, development in DLSA greatly progresses with advances in deep learning algorithms. On the basis of deep neural networks, a variety of DLSA approaches are reported [11][12]. Chen et al. train a convolutional neural network (CNN), which is applied to sentence-level sentiment analysis via pre-trained word vectors, achieving a satisfying working performance [11]. Lai et al. propose an integrated model by combining the superiorities of recurrent neural networks (RNNs) and convolutional neural networks (CNNs) [13]. That is, the context information is captured via RNNs, while the document representation and the local feature of sentence are derived via CNNs. On the other hand, RNN-based methods, integrated with attention mechanisms, also have their distinctiveness in DLSA [14][15]. Specifically, the hierarchical-structure networks are the most pronounced to process on both word and sentence levels. Yang et al. establish a hierarchical attention network that, respectively, applies attention mechanisms to word and sentence levels, which fuses more valuable information into each document [5]. Huang et al. develop a hierarchical multi-attention network to accurately assign the attentive weights on distinguishing levels [16]. Huang et al. establish a hierarchical hybrid neural network with multi-head attention to extract the global and local features of each document [17]. Due to the distinguishing contribution of each sentence to the sentiment polarity, Choi et al. propose a gating-mechanism-based method to identify the sentence importance in a document [18]. So far, there is an ongoing trend to model the document based on its hierarchical structure and thus precisely extract the document feature [19][20].

3. Document-Level Multimodal Sentiment Analysis

In the multimodal sentiment analysis domain, deep-learning based methods play a pivotal role. Previous work tends to directly fuse unimodal features to construct a multimodal representation for sentiment analysis [21][22][23]. In [21][22], feature vectors from different modalities are concatenated for multimodal integration. Soujanya et al. extract textual and visual features using CNN, concatenate the multimodal features, and classify the sentiment polarity via a multicore learning classifier. However, such approaches fail to deal with the cross-modal interaction [23]. In [7], a TFN model is proposed to use tensor outer products to dynamically model data across modalities. This approach generally results in oversized models for training.

More recently, studies have addressed multimodal interaction and information fusion, especially by using attention mechanisms [24][25][26][27]. Amir et al. develop a multi-level attention network to extract multimodal interaction by assuming the interactions of different information between modalities [24]. Xu et al. propose a visual feature guided attention LSTM model to extract words for sentiment delivery and aggregate the representation of informative words with visual semantic features, objects and scenes [25]. Since textual and visual information reinforce and complement each other, Xu et al. construct a co-memory network to iteratively interact the textual and visual information for multimodal sentiment analysis [26]. Similarly, Zhu et al. apply an image–text interaction network for multimodal analysis to explore the interaction between text and image regions through a cross-modal attention mechanism [27].

Notwithstanding, all the aforementioned work is carried out based on the one-to-one correspondence between text and images. While in practice, for most multimodal samples such as blog posts and e-commerce reviews, no conformity between text and image information is set in advance. For example, a single document can contain multiple images. As is known, the DLMSA is a more text-oriented task, and the image features are auxiliary for better analysis [8][28]. Instead of directly feeding images into sentiment classifiers, visual information is typically considered as a source on sentence-level alignment. Truong et al. exploit pre-trained VGG networks to obtain image features and then align the visual information as attention to each sentence, based on which more focus is assigned to image-related sentences [8]. Guo et al. leverage a set of distance-based coefficients for image and text alignment and learn sentiment representations of documents for online news sentiment classification [29]. Aiming to obtain the sentiment-related information, Du et al. propose a method based on a gated attention mechanism [9]. In this method, a pre-trained CNN is taken to extract fine-grained features of images, and then, the gated attention network is employed to fuse the image and text representations, based on which a better sentiment analysis result is achieved.

References

1. Rhanoui, M.; Mikram, M.; Yousfi, S.; Barzali, S. A CNN-BiLSTM model for document-level sentiment analysis. Mach. Learn. Knowl. Extr. 2019, 1, 832–847.
2. Chambers, A. Statistical Models for Text Classification and Clustering: Applications and Analysis; University of California: Irvine, CA, USA, 2013.
3. Jiang, D.; He, J. Text semantic classification of long discourses based on neural networks with improved focal loss. Comput. Intell. Neurosci. 2021, 2021.
4. Zhou, C.; Sun, C.; Liu, Z.; Lau, F. A C-LSTM neural network for text classification. arXiv 2015, arXiv:1511.08630.
5. Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; Hovy, E. Hierarchical attention networks for document classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 1480–1489.
6. Huang, L.; Ma, D.; Li, S.; Zhang, X.; Wang, H. Text level graph neural network for text classification. arXiv 2019, arXiv:1910.02356.
7. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor fusion network for multimodal sentiment analysis. arXiv 2017, arXiv:1707.07250.
8. Truong, Q.T.; Lauw, H.W. Vistanet: Visual aspect attention network for multimodal sentiment analysis. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 305–312.
9. Du, Y.; Liu, Y.; Peng, Z.; Jin, X. Gated attention fusion network for multimodal sentiment classification. Knowl.-Based Syst. 2022, 240, 108107.
10. Xiong, H.; Yan, Z.; Zhao, H.; Huang, Z.; Xue, Y. Triplet Contrastive Learning for Aspect Level Sentiment Classification. Mathematics 2022, 10, 4099.
11. Chen, Y. Convolutional Neural Network for Sentence Classification. Master’s Thesis, University of Waterloo, Waterloo, ON, Canada, 2015.
12. Zhang, X.; Zhao, J.; LeCun, Y. Character-level convolutional networks for text classification. Adv. Neural Inf. Process. Syst. 2015, 28, 1–9.
13. Lai, S.; Xu, L.; Liu, K.; Zhao, J. Recurrent convolutional neural networks for text classification. In Proceedings of the AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015; Volume 29.

14. Liu, P.; Qiu, X.; Huang, X. Recurrent neural network for text classification with multi-task learning. arXiv 2016, arXiv:1605.05101.
15. Rao, G.; Huang, W.; Feng, Z.; Cong, Q. LSTM with sentence representations for document-level sentiment classification. Neurocomputing 2018, 308, 49–57.
16. Huang, Y.; Chen, J.; Zheng, S.; Xue, Y.; Hu, X. Hierarchical multi-attention networks for document classification. Int. J. Mach. Learn. Cybern. 2021, 12, 1639–1647.
17. Huang, W.; Chen, J.; Cai, Q.; Liu, X.; Zhang, Y.; Hu, X. Hierarchical Hybrid Neural Networks with Multi-Head Attention for Document Classification. Int. J. Data Warehous. Min. (IJDWM) 2022, 18, 1–16.
18. Choi, G.; Oh, S.; Kim, H. Improving document-level sentiment classification using importance of sentences. Entropy 2020, 22, 1336.
19. Sinha, K.; Dong, Y.; Cheung, J.C.K.; Ruths, D. A hierarchical neural attention-based text classifier. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 817–823.
20. Liu, F.; Zheng, J.; Zheng, L.; Chen, C. Combining attention-based bidirectional gated recurrent neural network and two-dimensional convolutional neural network for document-level sentiment classification. Neurocomputing 2020, 371, 39–50.
21. Wang, H.; Meghawat, A.; Morency, L.P.; Xing, E.P. Select-additive learning: Improving generalization in multimodal sentiment analysis. In Proceedings of the 2017 IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 10–14 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 949–954.
22. Anastasopoulos, A.; Kumar, S.; Liao, H. Neural language modeling with visual features. arXiv 2019, arXiv:1903.02930.
23. Poria, S.; Chaturvedi, I.; Cambria, E.; Hussain, A. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In Proceedings of the 2016 IEEE 16th International Conference on Data Mining (ICDM), Barcelona, Spain, 12–15 December 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 439–448.
24. Zadeh, A.; Liang, P.P.; Poria, S.; Vij, P.; Cambria, E.; Morency, L.P. Multi-attention recurrent network for human communication comprehension. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
25. Xu, N.; Mao, W. Multisentinet: A deep semantic network for multimodal sentiment analysis. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Singapore, 6–10 November 2017; pp. 2399–2402.

26. Xu, N.; Mao, W.; Chen, G. A co-memory network for multimodal sentiment analysis. In Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, Ann Arbor, MI, USA, 8–12 July 2018; pp. 929–932.
27. Zhu, T.; Li, L.; Yang, J.; Zhao, S.; Liu, H.; Qian, J. Multimodal sentiment analysis with image-text interaction network. IEEE Trans. Multimed. 2022, 1.
28. Tian, Y.; Sun, X.; Yu, H.; Li, Y.; Fu, K. Hierarchical self-adaptation network for multimodal named entity recognition in social media. Neurocomputing 2021, 439, 12–21.
29. Guo, W.; Zhang, Y.; Cai, X.; Meng, L.; Yang, J.; Yuan, X. LD-MAN: Layout-driven multimodal attention network for online news sentiment recognition. IEEE Trans. Multimed. 2020, 23, 1785–1798.