Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/33/20230224

Deep learning-based sentiment analysis for social media: A focus on multimodal and aspect-based approaches

Bowen Feng

Li Jia high school, Chongqing, 401122, China

1910841218@mail.sit.edu.cn

Abstract. Commonly referred to as opinion mining, sentiment analysis harnesses the power of deep learning systems to discern human emotions and subjective sentiments towards a wide array of subjects. As such, it has become an integral tool in identifying and distinguishing sentences that harbor emotional biases or trends. By systematically examining sentiment-tinged data, researchers can unearth pivotal insights that not only reflect current perspectives but also predict future behaviors and trends. This process involves intricate computational models that analyze and interpret the emotional undertones embedded within a body of text. Whether these undertones are positive, negative, or neutral, sentiment analysis allows us to delve into the subtle nuances of human communication. This ability to "understand" and quantify sentiment is particularly vital in our modern digital age, where opinions and reviews shared through social media and online platforms can greatly influence public sentiment and consumer behavior. By extending beyond the literal meanings of words and phrases, sentiment analysis can provide a more comprehensive understanding of how people truly feel. It is instrumental in fields as diverse as marketing, politics, social science, and even artificial intelligence development, given its potential to gauge public opinion and predict societal trends. This paper aims to consolidate relevant research within the field of sentiment analysis conducted in recent years. Furthermore, it seeks to prognosticate the future trajectories and impacts of this rapidly evolving domain. Emphasis is placed on the role of deep learning and its transformational effects on the approach and capabilities of sentiment analysis, anticipating how its further advancement will continue to refine this intricate process of emotion recognition and interpretation.

Keywords: deep learning, sentiment analysis, social media.

1. Introduction

As society progresses, an increasingly vast amount of information in our world has metamorphosed into what is known as 'big data'. This term signifies an enormous base of data wherein the proportion of effective, useful information is relatively miniscule. Therefore, the task of memorizing and distinguishing this sort of data utilizing only human brainpower seems untenable given current social developmental trends. Conversely, machines trained by humans possess advantages such as high-speed data processing and maintaining high, stable accuracy rates when processing large volumes of data. Thus, a key development goal in the field of natural language processing has emerged - enabling machines to understand human language. The aim is for these machines to comprehend the same human language under varying circumstances and make accurate judgments based on their understanding of human language [1]. However, when machines encounter human languages, they must surmount challenges seldom faced by humans.

A primary difficulty lies in dealing with the multi-modal emotions often embedded within pictures or videos, rendering full emotion recognition in data challenging for machines, especially when the data includes obscure or ambiguous text. Moreover, the same words can hold different meanings in diverse contexts, such as the word 'book' being used in "book a room" or "read books". Due to a lack of flexibility in information processing, machines are prone to errors when confronted with such texts, a phenomenon particularly noticeable on social media platforms. In response to these challenges, researchers in this field have proposed various models and training methods aimed at equipping machines with better language understanding and sentiment analysis capabilities. This focus on improving machine interpretation of human emotions is not only a key challenge but also an exciting frontier in the domain of natural language processing and sentiment analysis.

2. Relevant Theories

2.1. Development history of sentiment analysis

The application of convolutional neural networks for sentiment analysis dates back to 2014. Initially, these networks were primarily aimed at analyzing the sentiment of straightforward text to decrease manual workload. However, due to the prevalence of complex emotions within sentences, sentiment analysis tasks became increasingly time-consuming, leading to the development of models that focused on deciphering complex and cryptic text sentiment. As of now, the challenge of conducting sentiment analysis on simple texts has largely been surmounted. The focus has shifted to more complicated text sentiment analysis, such as the construction of models capable of interpreting texts containing aspect-level sentiments and satirical sentiments. This progression demonstrates the evolution of sentiment analysis, growing from handling basic text interpretation to tackling multifaceted and nuanced emotional contexts.

2.2. Different levels of emotional analysis

By now, the field of Sentiment Analysis for Social Media based on Deep Learning can be divided into two main parts. The former is single text analysis, and the latter is Multi-modal affective analysis. For the former, in 2014 years, Yoon Kim et.al proposed the model named Text CNN (Convolutional Neural Networks for Sentence Classification) following [2], As shown in Figure 1.

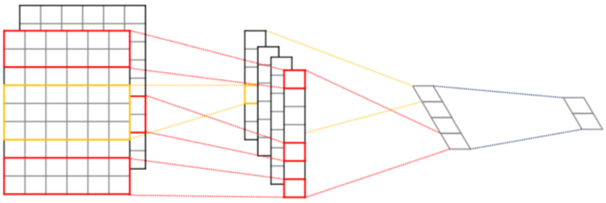


Figure 1. Neural network structure diagram (Photo/Picture credit: Original).

This model employs a convolutional neural network to extract words from a max-pooling layer, passing them through a convolutional kernel, and then projecting them into a classification space to achieve text sentiment classification. Each convolutional kernel within the model scans at a specific ratio, allowing each output space to contain two or more windows. This design improves the precision of lexical meaning screening and ensures optimal results for text classification screening. However, the "max pooling" area lacks relevant structural information. Consequently, this model can only discern the presence of complex structures within the text, including long sentences and intricate grammatical usage, but it falls short in distinguishing the emotional trends in these complexly structured sentences. In 2017, Tencent's AI Lab proposed DPCNN (Deep Pyramid Convolutional Neural Networks for Text Categorization), an improvement on TextCNN [3]. It includes two modules in each convolutional layer, making the model richer in lexical representation. Each two-level convolutional block of equal length is pooled with a max pooling of 'size=3' and 'STRIDE=2', compressing the sequence length to half of its original size. This design doubles the number of perceptible text fragments compared to previous models, hence enhancing its ability to process longer texts. Nevertheless, the DPCNN model has its limitations, especially in recognizing buzzwords or trendy slang that often deviate from traditional meanings. Also in 2017, Devamanyu Hazarika and colleagues proposed the CASCADE (Contextual Sarcasm Detection in Online Discussion Forums) model [4]. Building on TextCNN, it introduces content topics (such as forum topics) into the emotional range of the original model. By integrating user embedding and forum topic encoding with the CNN model, CASCADE forms a classification vector, which enhances its capability to process ambiguous texts. However, to pursue higher accuracy, the IOU threshold is increased in this model, raising the risk of overfitting and potentially wasting time on finding consistent interpretations. In 2023, He et al. proposed the Multi-Layer Bi-Directional LSTM with a Trapezoidal Structure [5]. This model, which is convoluted by a neural network with a trapezoidal structure and then input into the classification space, can process sentences more efficiently with fewer model parameters. However, the reduced parameters may compromise accuracy when dealing with long and complex texts, thus necessitating a balance between speed and accuracy in decision making. Much of the text content on media platforms containing images comprises text accompanied by emojis or other images. The image carries emotional tones based on color, size, etc., and the text often provides context or explanation to the image to express extended or enhanced emotions. Due to the complexity of such composite text, a model capable of analyzing multi-modal emotions is necessary for appropriate interpretation. In the image-matching module, Pontiki M et al proposed the Inception module in 2014, which has since undergone four generations of refinement [6]. As shown in Figure 2.

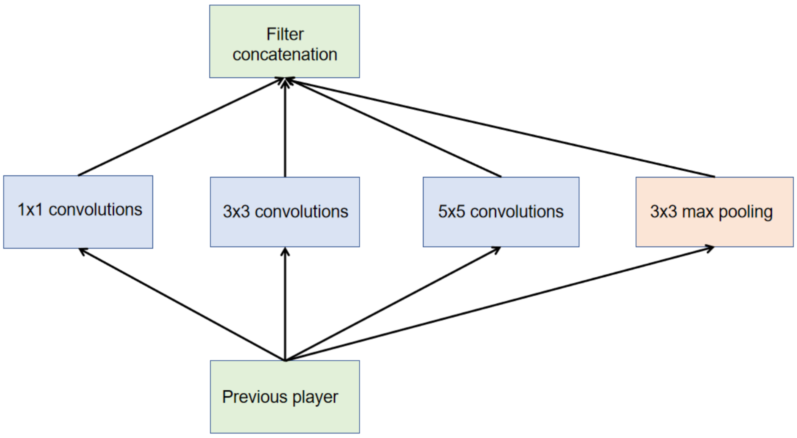


Figure 2. Inception v1 (Photo/Picture credit: Original).

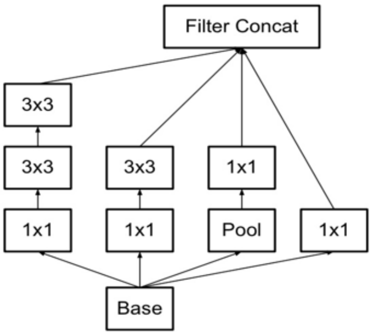
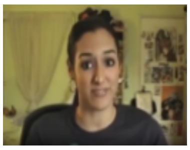
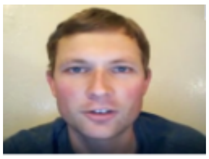


Figure 3. Inception v2 (Photo/Picture credit: Original).

From adding a BN layer to a V2 model that reduces changes in the internal distribution of the data of the internal neuron; To decompose convolution cores into several smaller one-dimensional volumes thus further accelerating the model's computational speed and improving the accuracy of the overall judgment of the V4 model. Reducing the number of trunk tasks, splitting task items, and reducing the scope of task categorization are the main reasons Inception models have faster computation speed. This also gives Inception models a comparative advantage over multi-modal sentiment analysis tasks. The main purpose of each change is to increase the number of brunches, but this make the number of uncertainty factors get more. It means that if there are some problems founded in this model, it is more inconvenient for people to find and correct errors. As shown in Figure 3. 2019 years, Huang etal put forward the DMAF (deep multimodal attention fusion) model, which inherits the advantages of the former model, judges and captures the features of multiple modes and matches them with the text, enhanced the ability to judge multi-modal synthetic text [7]. To be more specific, it divides the whole information in texts into two parts. The former is about the human's features in the picture, while the latter is the relevant texts. By analyzing the details of the features such as human's appearances or their expressions and looking for the key word in texts, it is easier for machine to find out the whole emotional

trend in the text. On the other hand, this model cannot keep a high accuracy when it is processing sentences that has a large number of words and few key words. Generally, the details of emotional trends is limited. It means that this model needs more improvements in task features and text matching.

Emotional polarity	T-GIF		CMU-MOSI		CMU-MOSEI	
	video	text	video	text	video	text
Positive	[1]	A man with a short mustache is smiling at another person.	 [4]	It was really really funny.	 [7]	May the grace of Jesus be with you.

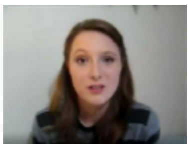
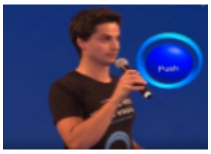
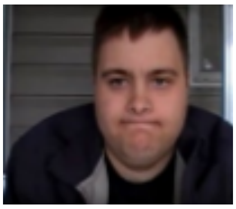
Neutral	 [2]	A young woman is fixing her while standing next to a car.	 [5]	That was really boring.	 [8]	It is so natural to see something and click it with his hand.
Negative	 [3]	A man is jumping backwards, trying to avoid white stuff.	 [6]	I am sorry and disappointed to say this i did not like it.	 [9]	It has never been something that i want us to forget.

Table 1. Dataset Table.

2023 years, Liao et al put forward ITIGNN (Image-text interaction graph neural network) model, which adds the steps of dividing and refining various factors on the basis of the original model [8]. This model increase the workload of each brunches. It divides the two main parts (text and picture) into several pieces. Each feature in the image is paired with the emotional words in the text. The matching teams were then divided again into different emotion segments and analyzed for the proportion of emotion in each segment, which was then summarized for output. This change makes the accuracy about the recognition of this model higher than original models. However, this model can only analyze the single text with one picture. Also, the text with obscure information can not be tested correctly by using this model. Both of these information are mentioned by the author of that paper (ITIGNN) [9].

2.3. Vedio text analysis

As the most comprehensive modality in sentiment analysis, video content emotional analysis necessitates the consideration of a broader spectrum of factors. Beyond imagery and text information, video sentiment analysis must also account for the intrinsic data of the video itself. There are often instances where these three forms of information do not align, yielding diverse wholes. This divergence is evident in instances like thematic videos similar to films or non-thematic content akin to film commentary videos. Thus, sentiment analysis of video content requires matching and judgment of two combined variables: image-text and video-text. The matching principle for image-text in video content essentially mirrors the logic flow of image-text sentiment analysis.

Regarding video-text matching, Xiao Yunhong and colleagues [10] in 2022 assimilated a T-gif dataset created by Li et al. [11] and a CMU-MOSI dataset, as well as a CMU-MOSEI dataset charted by Zadeh et al. [12]. Both the CMU-MOSEI and CMU-MOSI datasets underwent raw data pre-processing for video classification, ensuring consistent and effective quality for each video segment and maintaining equal video length. In a speech scenario, effective quality may encompass aspects like stable audio, clear facial expressions, and distinct movements and scenes. After video emotion analysis, text content is matched with the video to ensure it aligns with the sentiment expressed by the subject in the video. This model primarily aims to secure more stable video and text resources to facilitate accurate discernment and judgment. However, a high frequency of noise that impacts the overall video quality may compromise the model's stability [13]. As shown in Table 1.

Addressing the noise issue, Hussain et al. proposed a framework for emotion recognition from multi-channel signals [14]. This approach segments each piece of information, filters the content of each segment to discard irrelevant data, and recombines the remaining sections to create a "noise reduction" effect. Consequently, this improves the model's accuracy in predicting the overarching emotional trend of backwash videos.

3. System analysis and application research

3.1. Diagnosis of the deaf group

For the application of multimodal sentiment analysis, Yang Yi et alproposed a relevant model. In the experiment, multimodal emotion analysis is used to capture people's facial features and brain signals, so as to judge the emotions of special groups (such as deaf people). The research uses 15 deaf students as a sample, lets the students watch positive, neutral and negative movies, and then uses the SAM (Self-assessment manikin) model for emotional data collection [15]. Finally, the proportion of each emotional data is analyzed, and the overall emotional tendency of the deaf students is estimated. This study accurately points out the facial emotional expression parts of the deaf population, which plays a good directional role in distinguishing special groups and judging the emotional field of special groups, and has a certain positive effect on the Defend the group’s own rights and interests.

3.2. Solving errors in public opinion networks

As a relatively common phenomenon in recent years, most of the people who express their views online are just judging the text in the topic, and do not pay attention to or even ignore the analysis of the content displayed in the picture. This kind of habit probably leads to frequent directional errors in the masses' participation in online public opinion, resulting in consequences of varying severity. For this problem, Fan Tao et al conducted multimodal sentiment analysis on events containing network public opinion based on the VGG19 model and Xception model pre-trained in the ImageNet dataset [16]. The model focuses on the research of sentiment analysis of pictures, pairs visual emotions with the emotions expressed in the text, and integrates multiple factors such as facial expressions and body movements with the text for emotional analysis, so as to play the effect of emotional joint decision-making. Such experiments improve the interpretability of the overall emotion, help to solve the problem of online public opinion, and help purify cyberspace.

3.3. Assisting individuals with mental illness

In the field of medicine, Ghosh Anay et al conducted a series of experiments on the aggressiveness of people in the event of pain [17]. Based on human tissue injuries, this experiment implements the task of classifying text emotions. It first divides the general emotional tendencies (such as non-aggression, implicit aggression, and overt aggression); Then, the sentiment prediction and analysis of the emotional factors based on the picture are carried out. Finally, the image is factor-matched to the text and the option with the highest consistency is selected to make an overall emotional judgment. The results of this experiment are helpful in the medical field for the diagnosis and treatment of people with mental illnesses (such as those with anxiety, depression, etc.), and the results of the experiment are also used to determine the level of pain in the future A certain auxiliary role.

4. Challenges

The difficulty of sentiment analysis is mainly reflected in the long time-consuming and low accuracy of long text analysis and implicit text analysis, and the difference between the “complex program, high accuracy” model and the “simple process, not accurate enough” model. The contradiction (how to balance the relationship between the two, that is, to have a high accuracy of sentiment analysis while ensuring the running speed) is one of the main reasons why this phenomenon has continued to this day. The difficulty in sentiment analysis of cryptic texts is mainly reflected in the mining of vocabulary that contains multiple meanings, irony, and function words (containing strong emotional factors). Multi-text sentiment analysis includes the difficulty of analyzing cryptic texts. It also has difficulties in compound sentence patterns. This type of experiment usually needs to use a model with enough quantization pool spaces and convolutional layers to analyze step by step, so this usually leads to the emergence of shortcomings such as high research requirements and long research periods; and if you choose a simpler model, in the experiment Sometimes there will be "content accumulation problems such as a large backbone workload of the model", which increases the risk of gradient explosion or gradient disappearance.

5. Conclusion

In terms of emotion analysis, research on singular text and simple picture-text combinations have largely matured. However, there remains considerable scope for development in multimodal affective analysis and the interpretation of obscure texts. Analyzing the emotional content of video-like texts proves challenging due to numerous confounding factors. Additionally, the training cycles of these models tend to be long, making it challenging to maintain stable accuracy within a short timeframe. Nonetheless, emotional analysis isn't limited to the media sector. The field holds tremendous potential and is fast becoming a critical requirement for future statistical fields and analytic classes. Owing to its numerous merits, it promises a vast range of potential markets.

References

[1] Li, W., Mei, H., Li, Y., et al. (2022). Survey of Multimodal Sentiment Analysis Based on Deep Learning.

[2] Kim, Y., et al. (2014). Convolutional Neural Networks for Sentence Classification. arXiv:1408.5882v2.

[3] Johnson, R., & Zhang, T. (2017). Deep pyramid convolutional neural networks for text categorization. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 562-570.

[4] Li, H., Lin, Z., Shen, X., Brandt, J., & Hua, G. (2015). A convolutional neural network cascade for face detection. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 5325-5334.

[5] He, Z., Dumdumaya, C., & Machica, I. K. D. et al. (2023). Text Sentiment Analysis Based on Multi-Layer Bi-Directional LSTM with a Trapezoidal Structure. Intelligent Automation & Soft Computing, 37(1), 639-654.

[6] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2014). Going deeper with convolutions. arXiv:1409.4842v1.

[7] Huang, F., Zhang, X., Zhao, Z., et al. (2019). Image-text sentiment analysis via deep multimodal attentive fusion. Knowledge-Based Systems, 167, 26-37.

[8] Liao, W., Zeng, B., Liu, J., Wei, P., & Fang, J. (2021). Image-text interaction graph neural network for image-text sentiment analysis.

[9] Xiao, Y., et al. (2022). Research on Emotional Analysis of Annotated Short Videos Based on Vision and Text.

[10] Li, Y., Song, Y., Cao, L., et al. (2016). TGIF: A New Dataset and Benchmark on Animated GIF Description. IEEE.

[11] Zadeh, A., Zellers, R., Pincus, E., et al. (2016). MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos.

[12] Zadeh, A., Liang, P., Poria, S., et al. (2018). Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. Thirty-Second AAAI Conference on Artificial Intelligence. ACL.

[13] Hussain, M. D. S., Calvo, R. A., & Pour, P. A. (n.d.). Hybrid fusion approach for detecting affects from multichannel physiology. International Conference on Affective Computing and Intelligent Interaction. Springer, Berlin: Heidelberg.

[14] Xiao, Y., et al. (2022). Sentiment Recognition for Annotated Short Videos Based on Vision and Text.

[15] Yang, Y. (2022). Multimodal Emotion Analysis based on EEG and facial expression images of deaf students.

[16] Tao, F., Wang, H., Lin, K., & Liu, Y. (2022). A Probe into Netizen Sentiment Analysis in Vision-based Internet Public Opinion Events.

[17] Ghosh, A., Dhara, B. C., Pero, C., & Umer, S. (2023). A multimodal sentiment analysis system for recognizing person aggressiveness in pain based on textual and visual information. Journal of Ambient Intelligence and Humanized Computing, 14, 4489–4501.