Multimodal Sentiment Analysis via RNN variants

Ayush Agarwal
Biometric Research Laboratory, Department of Information Technology
Delhi Technological University, Bawana Road
New Delhi, India
ayush286@gmail.com

Ashima Yadav
Biometric Research Laboratory, Department of Information Technology
Delhi Technological University, Bawana Road
New Delhi, India
ashimayadavdtu@gmail.com

Dinesh Kumar Vishwakarma
Biometric Research Laboratory, Department of Information Technology
Delhi Technological University, Bawana Road
New Delhi, India
dvishwakarma@gmail.com

***Abstract*—Multimodal sentiment analysis involves the classification of sentiment by using different forms of data together, namely, text, audio, and video. Previously sentiment analysis was implemented only on textual data. Multimodal sentiment analysis relies mainly on identifying the utterances present in the video and using them as a basis for sentiment classification. In this paper, we proposed four different variants of RNN, namely, GRNN, LRNN, GLRNN and UGRNN for analyzing the utterances of the speakers from the videos. Experimental results on CMI-MOSI dataset demonstrates that our approach is able to achieve better sentiment classification accuracy on individual modality (text, audio, and video) than existing approaches on the same dataset. Moreover, our method also gave decent results after fusing the individual modality using Attention networks for multimodal sentiment analysis.**

Keywords—Attention networks, Deep learning, Long Short-Term Memory, Multimodal Sentiment analysis, Recurrent neural networks (RNNs).

I. INTRODUCTION

With the evolving landscape of social media, there has been a huge drift towards a large amount of multimedia data in the form of audios, videos, text, images, and emoticons. This has created a prospect for investigating the sentiment or the opinions from different modalities. Sentiment analysis is described as the study of people's emotions and opinions. It is majorly implemented at three levels: Document level, Sentence level, and Aspect level, where document-level sentiment analysis refers to a generalization of the sentiment of a document as a whole. Sentence level sentiment analysis deals in classifying a sentence into subjective type or objective type. As is obvious from the name, this analysis is only concerned with the sentiment or opinion described in a single sentence. Aspect level sentiment analysis generally refers to the sentiment described in different parts of the same entity. This is very commonly observed in most reviews where a user may praise some features of the product and then criticize some other features, all in a single review. It is generally seen that these levels may not be as discrete as discussed and may overlap with each other.

Sentiment analysis has many real-world applications. It is popularly applied for studying the reviews or opinions about any product on e-commerce websites to analyze the collective views of the users on a product. It is also used to mine the opinion of people towards political parties and their candidates. This is generally performed using the data received through social networking websites in the form of blogs, tweets, posts, and comments. This analysis is used by media companies to understand the voter demographic trend over time. It is also used by various banks and trading firms to understand and then predict the behavior of various segments of a market. This presents a huge financial opportunity as market prediction gives one of the most lucrative incentives for the people.

Earlier Sentiment Analysis has been used prolifically for deriving the sentiment from the text. With the advent of social networking websites and the massive amount of digital devices used for recording audio and video data, we have started to realize the significance of analyzing the multimedia data. Moreover, analyzing individual modality doesn't give us good results as the tone or pitch of the person cannot be predicted from the structure of a sentence. Hence, using multimodal data together can help us to narrow down the exact sentiment expressed by the person. This category of sentiment analysis is known as multimodal sentiment analysis. Hence, this paper uses with deep learning based approaches which include Recurrent neural networks (RNNs) along with its variants, Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs) for performing multimodal sentiment analysis on popular multimodal dataset called CMU-MOSI.

The remaining script is arranged as follows. Section 2 summarizes the previous work done in sentiment analysis with the prime focus on multimodal sentiment analysis. Section 3 discusses the proposed approach. Section 4 describes the experimental results and finally, Section 5 concludes the paper.

II. RELATED WORK

Earlier Sentiment Analysis was usually performed using machine learning and lexical based techniques. These techniques involved a lot of manual work in the form of preprocessing the data along with selecting and extracting relevant features. This process gets difficult as the number of data increases. Tang et al. [1], Medhat et al. [2], and Chaturvedi et al. [3] provide a detailed survey about the various machine learning techniques used for sentiment classification. Li et al. [4] discuss Support Vector Machines (SVM) for classification. They parsed the text to find the dependency of different words in order to calculate the modified distance between them. Lailiyah et al. [5] used Sentiwordnet and Indonesian lexical to classify the sentiments about the complaints of the Indonesian

978-1-7281-0886-5/19/$31.00 ©2019 IEEE
BCD 2019, May 29-31, 2019, Honolulu, Hawaii

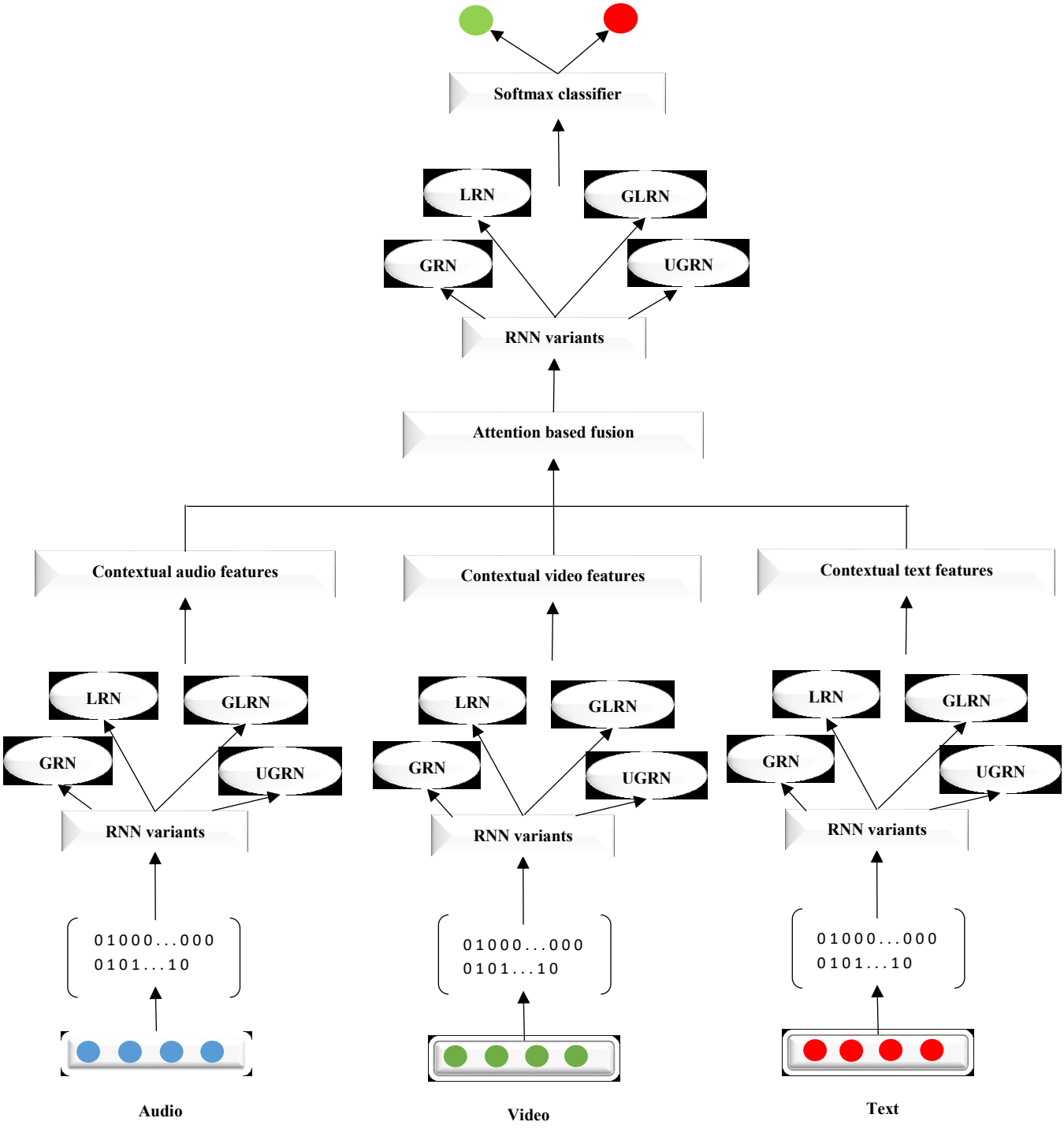

Fig. 1. Proposed Architecture for Multimodal Sentiment Analysis

public posted on the government website and Twitter. They also worked on reducing the ambiguity in some parts of the speech of the Indonesian language. Fan et al. [6] have compared the effect of different lexicons or dictionaries on the performance of sentiment analysis on the textual dataset. They applied word2vec approach for creating a lexicon which gave better performance than other sentiment lexicons and applied Naïve Bayes for sentiment classification. Sabra et al. [7] presented an approach for building a lexicon for the Arabic language using semi-supervised learning on the popular lexicon called WordNet. The lexicon consists of more than 70,000 terms with three different labels namely, neutral, positive and negative. They have used multiple techniques of classification to perform sentiment analysis using the Arabic lexicon. Alshari et al. [8] proposed a novel method for enriching the feature set by learning the score of non-opinionated words from

SentiWordNet and then performed. classification on a labeled movie review dataset. This method performed well than state-of-the-art techniques.

Deep Learning is a relatively new approach that has been employed to carry out sentiment analysis. Deep Learning has been found to perform better than machine learning or lexical-based approaches when an enormous amount of training data is available. Day et al. [9] compare the accuracy achieved by using Naïve Bayes, SVM, and Bi-directional LSTM on Google play store reviews dataset. The deep learning method significantly outperformed the efficiency achieved by the other machine learning based methods. They experimented with different activation functions like tanh, sigmoid, and ReLu where the highest accuracy is achieved by tanh. Ramadhani et al. [10] applied feedforward neural networks with three hidden layers along with sigmoid and ReLu activation functions to test the neural network with different learning rates on the tweets dataset crawled using Twitter API, labeled as positive and negative. They used a relatively new method of Stochastic Gradient Descent to perform backpropagation in the neural network. Chachra et al. [11] discuss an approach that applies Convolutional Neural Networks (CNNs) for the classification and detection of various sentiments on Twitter tweets. They found that a simpler network known as shallow network performs better than deeper CNNs. Sun et al. [12] proposed a three-layer CNN-LSTM network on the Tibetan microblog dataset along with a two-layer LSTM fusion network. This method revealed heavy semantic relationships between the words and hence was able to achieve better results. The method suffered from overfitting because of the shortage of tweets in the Tibetan language as compared to other common languages. Preethi et al. [13] proposed an RNN-based deep learning sentiment analysis framework on the movie and restaurant review dataset which analyzes the different reviews of the users to recommend the places near to them.

Apart from studying the unimodal data, a steady drift has been seen for exploring the multimedia data. Poria et al. [14] applied the LSTM based network to extract features from the utterances of a video that can be used for representing and defining the context in multimodal sentiment analysis. The proposed method has performed very well and has shown significant performance improvement over the baseline. In another work, Poria et al. [15] identified the sentiment from the videos where they proposed an LSTM Network based on attention mechanism to extract contextual features from the opinionated utterances on the CMI-MOSI dataset. They also proposed a novel fusion based mechanism to amplify the quality of classification. Chen et al. [16] proposed an approach of using the LSTM based temporal attention method along with Gated multimodal embeddings. They have introduced a method of using reinforcement learning for performing fusion at word-level in textual data on the MOSI dataset. Yu et al. [17] discuss another novel attention method called Temporally Selective Attention Model (TSAM) on the MOSI Dataset to selectively pick the video sequences resulting in better performance. They used speaker-distribution loss which has been designed to recognize social states. Wang et al. [18] discuss a method called Select-Additive Learning (SAL) to improve the generalizability of the trained neural networks for performing multimodal sentiment analysis on MOSI dataset. This approach improved the unimodal as well as the multimodal accuracy.

III. Proposed Methodology

In the upcoming subsections, we have discussed the proposed methodology for unimodal and multimodal sentiment analysis. We have also considered the different variants of RNN for multimodal sentiment analysis. Our work is based on the existing work of [14] and [15]. Fig. 1 explains the proposed architecture.

A. For unimodal Analysis

We have initially trained text, audio, and video modalities. The training data has a shape of (62, 63,100) and the testing data has a shape of (31, 63,100). The training and testing labels are encoded using one hot representation. Each modality is passed through the different variants of RNN to get the contextual unimodal activations. Finally, the unimodal activations of each modality are saved for further analysis.

B. For Multimodal Analysis

The contextual unimodal activations of each modality obtained from Section *A* above, are loaded for multimodal sentiment analysis. Now, each of the unimodal activations is fused using the Attention Networks. The fused modalities are again passed through the different variants of RNN to get the final sentiment polarity.

C. Network Architecture

Before explaining the different variants of RNN, we have discussed Bi-directional RNN (BRNN) and Attention Networks.

a) Bi-directional RNN (BRNN)

BRNN is a type of RNN which overcomes the major issue of unidirectional RNN, which could capture the information from previous time steps only. On the other hand, BRNN has two types of connections, one moving in the right (forward) direction and another moving in the left (backward) direction. Hence, they are able to understand the context of the problem by efficiently capturing the information from previous and future time steps. The following articles discuss the BRNN for sentiment analysis: [19], [20], [21]

b) Attention Network

Generally, when we need to generate an output of the fixed length, we pass the entire input to the RNN based network. This might result in some irrelevant information. Hence, attention networks are used to "attend" the more important inputs depending on the task. It decides what part of the text should get more focus by generating the weighted combination of the input, while ignoring the weight of other inputs. The following articles discusses Attention network in the area of sentiment analysis: [22], [23], [24], [25].

c) Variants of RNN

We have applied the following variants of RNN which are based on LSTM and GRU models.

- **GRU based RNN (GRNN):**

We have used the GRU block cell to create a basic GRU cell with output size 100 and dropout rate 0.2. The output

received is passed to BRNN with the sequence length of 62 and 31 for training and testing data respectively. Finally, the dense layer is applied with ReLu activation function and the output is passed into the softmax classifier for the final sentiment classification.

- **LSTM based RNN (LRNN):**

We have used the basic LSTM recurrent unit cell based on the implementation by Zaremba et al. [26] with the output size of 100 and dropout rate 0.2. The output received is passed to BRNN with the sequence length of 62 and 31 for training and testing data respectively. Finally, the dense layer is applied with ReLu activation function and the output is passed into the softmax classifier for the final sentiment classification.

- **Group LSTM based RNN (GLRNN):**

We have used Group LSTM cell based on the implementation by Ginsburg et al. [27] with output size 100 and dropout rate 0.2. A Group LSTM cell consists of one LSTM sub-cell per group, where each sub-cell operates on an evenly sized sub-vector of the output. The output received is passed to BRNN with the sequence length of 62 and 31 for training and testing data respectively. Finally, the dense layer is applied with ReLu activation function and the output is passed into the softmax classifier for the final sentiment classification.

- **Update Gate based RNN (UGRNN):**

Finally, we have used an Update Gate RNN cell based on the implementation of Etworks et al. [28] with an output size 100 and dropout rate 0.2. This cell is a combination of LSTM and GRU units, wherein there is only one gate, to determine whether the unit should be integrating or computing instantaneously. This is the recurrent idea of the feedforward highway network. The output received is passed to BRNN with the sequence length of 62 and 31 for training and testing data respectively. Finally, the dense layer is applied with ReLu activation function and the output is passed into the softmax classifier for the final sentiment classification.

IV. Experimental results

A. Dataset

We have used the CMU-MOSI dataset [29] which contains opinionated utterances on various topics such as movies, books, and products by 89 people. A total of 2199 utterances are contained in the 93 videos crawled from YouTube. The videos are highly segmented. Each segment of the video features an utterance annotated by its textual and audio data. Each utterance has been categorized as either positive or negative. The dataset has been split into training and testing set consisting of 62 videos and 31 videos respectively. Each video has been padded to 63 utterances. All the four RNN variants have been trained using categorical cross entropy loss on the output of each utterance in a video.

$$loss = -\frac{1}{(\sum_{i=1}^{M} L_i)} \sum_{i=1}^{M} \sum_{j=1}^{L_i} \sum_{c=1}^{C} y_{i,c}^{j} \log_2 (\hat{y}_{i,c}^{j}) \quad (1)$$

Where, M = total number of videos, Li= number of utterances for i^{th} video, C = number of classes $y_{i,c}^{j}$ *and* $\hat{y}_{i,c}^{j}$ are the expected and predicted output respectively for the j^{th} occurrence of the i^{th} video. We have used ADAM optimizer [30] for training with learning rate as 0.0001, *beta1* (exponential decay rate for 1st moment estimates) as 0.9 and *beta2* (exponential decay rate for 2nd moment estimates) as 0.999.

For training the RNN variants we used a batch size of 20 with the number of epochs as 200 in case of unimodal sentiment analysis, whereas for multimodal sentiment analysis we have used a batch size of 20 and number of epochs as 1000.

B. Performance of different models

This section first compares the performance of the different variants of RNN (GRNN, LRNN, GLRNN, and UGRNN) for unimodal and multimodal sentiment analysis of CMU-MOSI dataset by comparing the accuracy achieved by each on them as shown in Table I. Further, we have also calculated the loss obtained by the different variants in Table II.

TABLE I. COMPARISON OF PERFORMANCE BETWEEN THE RNN VARIANTS IN TERMS OF ACCURACY (%)

Modality	GRNN	LRNN	GLRNN	UGRNN
Text (T)	**80.85**	79.92	80.58	79.92
Audio (A)	**62.10**	56.78	**62.10**	56.11
Video (V)	59.44	54.65	57.04	**59.70**
Multimodal (T+A+V)	77.65	**78.05**	**78.05**	75.66

TABLE II. LEAST LOSS ACHIEVED BY RNN VARIANTS

Modality	GRNN	LRNN	GLRNN	UGRNN
Text	0.014294	0.014319	0.014162	0.014274
Audio	0.021647	0.021769	0.021433	0.022086
Video	0.023051	0.022811	0.021991	0.024126
Multimodal (T + A + V)	0.016726	0.015695	0.016876	0.017345

As seen in Table I, GRNN performs best for textual modality, GRNN and GLRNN perform best for audio modality, and UGRNN gives the best results for video modality. After fusing the individual modality, the best results for multimodal sentiment analysis are shown by LRNN and GLRNN networks. The results shows than RNN models can be popularly applied for natural language processing applications as they can model sequential information. Moreover, LSTM and GRU are able to capture the contextual features in the data. Table II shows the least loss obtained by the RNN variants.

Table III compares our approach with some prominent work which are based on CMU-MOSI dataset for unimodal and multimodal sentiment classification. The highest accuracy is marked in bold.

TABLE III. RESULT COMPARISON ON THE BASIS ACCURACY (%) FOR UNIMODAL AND MULTIMODAL SENTIMENT CLASSIFICATION

Models	Text	Audio	Video	Multimodal (T+A+V)
Chen et al. [16]	71.3	55.4	52.3	75.7
Wang et al. [18]	73.2	61.8	**63.6**	73
Yu et al. [17]	74.5	60.9	61.8	75.1
Poria et al. [14]	78.1	60.3	55.8	80.30
Porial et al. [15]	79.1	60.1	55.5	**81.3**
Proposed approach	**80.85**	**62.10**	59.70	78.05

From Table III, we can see that our approach has obtained the highest accuracy on text and audio modalities, whereas second best accuracy for video modalities. This may be due to

the fact that 3D CNN used for video feature extraction is not able to extract the features properly. For multimodal sentiment analysis, we obtain an accuracy of 78.05%.

V. Conclusion

Multimodal Sentiment Analysis is a relatively new field of research. Here we only combined three forms of data whereas many other types of data can also be combined for a better analysis. We have used an attention network for fusing the unimodal activations along with four variants of RNN namely, GRNN, LRNN, GLRNN, and UGRNN RNN and GLSTM for obtaining the contextual features. This work can further be evolved to be used for medical purposes such as detection of autism in a child. Further, this model can also be used for predicting the sentiments, emotions, and genre of a movie by its trailer. We will also work to improve the performance of our multimodal sentiment analysis.

This work mainly incorporates text, video and speech data. We can work on using other types of data generated such as contextual, crowd-source, and relationship information. The broader context of the this work comprehends Multimedia Information Processing, Multimedia Embedding, Web Mining, Machine Learning, Deep Neural Networks, and Artificial Intelligence. This work can be further explored for other applications such as credibility analysis. It can also be modified so as to make it better at preserving meaning and understanding the context. This will help us in deploying the model in real-world scenarios.

References

[1] H. Tang, S. Tan, and X. Cheng, "A survey on sentiment detection of reviews," *Expert Syst. Appl.*, vol. 36, no. 7, pp. 10760–10773, 2009.

[2] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, 2014.

[3] N. M. S Chaturvedi, V Mishra, "Sentiment Analysis using Machine Learning for Business Intelligence," in *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, 2018, vol. 14, no. 3, pp. 2162–2166.

[4] J. Li and L. Qiu, "A sentiment analysis method of short texts in microblog," *Proc. - 2017 IEEE Int. Conf. Comput. Sci. Eng. IEEE/IFIP Int. Conf. Embed. Ubiquitous Comput. CSE EUC 2017*, vol. 1, pp. 776–779, 2017.

[5] M. Lailiyah, S. Sumpeno, and I. K. E. Purnama, "Sentiment Analysis of Public Complaints Using Lexical Resources Between Indonesian Sentiment Lexicon and Sentiwordnet," in *Intelligent Technology and Its Applications (ISITIA), 2017 International Seminar on. IEEE, 2017*, 2017, pp. 307–312.

[6] X. Fan, X. Li, F. Du, X. Li, and M. Wei, "Apply word vectors for sentiment analysis of APP reviews," in *2016 3rd International Conference on Systems and Informatics, ICSAI 2016*, 2017, no. Icsai, pp. 1062–1066.

[7] K. S. Sabra, R. N. Zantout, M. A. El Abed, and L. Hamandi, "Sentiment Analysis : Arabic Sentiment Lexicons," in *Sensors Networks Smart and Emerging Technologies (SENSET)*, 2017, pp. 6–9.

[8] E. M. Alshari, "Effective Method for Sentiment Lexical Dictionary Enrichment based on Word2Vec for Sentiment Analysis," in *2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP)*, 2018, pp. 1–5.

[9] M. Y. Day and Y. Da Lin, "Deep learning for sentiment analysis on google play consumer review," in *Proceedings - 2017 IEEE International Conference on Information Reuse and Integration, IRI 2017*, 2017, pp. 382–388.

[10] A. M. Ramadhani and H. S. Goo, "Twitter sentiment analysis using deep learning methods," *2017 7th Int. Annu. Eng. Semin.*, pp. 1–4, 2017.

[11] A. Chachra, P. Mehndiratta, and M. Gupta, "Sentiment analysis of text using deep convolution neural networks," *2017 Tenth Int. Conf. Contemp. Comput.*, no. August, pp. 1–6, 2017.

[12] B. Sun, F. Tian, and L. Liang, "Tibetan Micro-Blog Sentiment Analysis Based on Mixed Deep Learning," in *2018 International Conference on Audio, Language and Image Processing (ICALIP)*, 2018, pp. 109–112.

[13] G. Preethi, P. V. Krishna, M. S. Obaidat, V. Saritha, and S. Yenduri, "Application of Deep Learning to Sentiment Analysis for recommender system on cloud," in *IEEE CITS 2017 - 2017 International Conference on Computer, Information and Telecommunication Systems*, 2017, pp. 93–97.

[14] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-Dependent Sentiment Analysis in User-Generated Videos," *Proc. 55th Annu. Meet. Assoc. Comput. Linguist. (Volume 1 Long Pap.*, no. January, pp. 873–883, 2017.

[15] S. Poria, E. Cambria, D. Hazarika, N. Mazumder, A. Zadeh, and L. P. Morency, "Multi-level multiple attentions for contextual multimodal sentiment analysis," in *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2017, vol. 2017–Novem, pp. 1033–1038.

[16] M. Chen, "Multimodal Sentiment Analysis with Word-Level Fusion and Reinforcement Learning," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction. ACM, 2017*, 2017, pp. 163–171.

[17] H. Yu, L. Gui, M. Madaio, A. Ogan, and J. Cassell, "Temporally Selective Attention Model for Social and Affective State Recognition in Multimedia Content," in *Proceedings of the 2017 ACM on Multimedia Conference. ACM, 2017*, 2017, pp. 1743–1751.

[18] H. Wang, A. Meghawat, L. Morency, and E. P. Xing, "Select-Additive Learning: Improving Generalization in Multimodal Sentiment Analysis," in *arXiv preprint arXiv:1609.05244 (2016)*.

[19] T. Chen, R. Xu, Y. He, and X. Wang, "Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN," *Expert Syst. Appl.*, vol. 72, pp. 221–230, 2017.

[20] K. Baktha and B. K. Tripathy, "Investigation of recurrent neural networks in the field of sentiment analysis," in *Proceedings of the 2017 IEEE International Conference on Communication and Signal Processing, ICCSP 2017*, 2017, pp. 2047–2050.

[21] Q. Liu, Y. Zhang, and J. Liu, "Learning Domain Representation for Multi-Domain Sentiment Classification," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol 1)*, 2018.

[22] M. Jiang, J. Wang, M. Lan, and Y. Wu, "An Effective Gated and Attention-Based Neural Network Model for Fine-Grained Financial Target-Dependent Sentiment Analysis," *Int. Conf. Knowl. Sci. Eng. Manag.*, vol. 214, pp. 42–54, 2014.

[23] K. Song, T. Yao, Q. Ling, and T. Mei, "Boosting image sentiment analysis with visual attention," *Neurocomputing*, vol. 312, pp. 218–228, 2018.

[24] T. Liu, S. Yu, B. Xu, and H. Yin, "Recurrent networks with attention and convolutional networks for sentence representation and classification," *Appl. Intell.*, vol. 48, no. 10, pp. 3797–3806, 2018.

[25] M. Yang, Q. Qu, X. Chen, C. Guo, Y. Shen, and K. Lei, "Feature-enhanced attention network for target-dependent sentiment classification," *Neurocomputing*, vol. 307, pp. 91–97, 2018.

[26] W. Zaremba, "Recurrent Neural Network Regularization," in *arXiv preprint arXiv:1409.2329 (2014)*, 2015, no. 2013, pp. 1–8.

[27] B. Ginsburg, "Factorization Tricks For LSTM Networks," in *arXiv preprint arXiv:1703.10722 (2017)*, 2017, pp. 1–6.

[28] N. E. N. Etworks and J. Collins, "Capacity and Trainability in Recurrent Neural Networks," in *arXiv preprint arXiv:1611.09913 (2016)*, 2017, pp. 1–17.

[29] A. Zadeh, R. Zellers, E. Pincus, and L. Morency, "MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos," *IEEE Intell. Syst.*, 2016.

[30] D. P. Kingma and J. L. Ba, "Adam: A Method for Stochastic Optimization," in *arXiv preprint arXiv:1703.10722 (2017)*, 2015, pp. 1–15.