Multi-Modal Methods for Sentiment Recognition

Natural Language Processing Course Project Report

D. Dehghani, A. Kashiri, S. Mesforoush, S.A. Mousavizadeh, F. Tohidian, S. Tabesh
Sharif University of Technology
Spring1401

ABSTRACT

Emotions could be recognized in a video by analyzing its text, frames, audio, or a combination of those called "multi-modal methods." In this project, we will propose a few algorithms to combine different features of a video in order to recognize the overall emotion of the scene. Our analysis reveals that this combination in video analysis can result in better emotion recognition models rather than using only one of its aspects.

KEYWORDS

Multimodal Sentiment Analysis, Transfer learning, Neural Network, Sentiment Recognition

1 INTRODUCTION

Context and Motivation. Analyzing and detecting emotions in images and text has been a growing trend in machine learning in recent years. A *video* consists of a number of consecutive images and dialogues or monologues consisting of voice and text. In order to recognize emotions in a video, we can either analyze one of these aspects or combine the result of two or all aspect's. In order to find which features effect the outcome most, we tried different approaches.

Research Problem. We considered scripts and frames as the constituents of a video and ignored the sound and tone of expressing speeches, since analyzing those needed additional knowledge which was out of this project's scope. Hence, for the purpose of recognizing the emotions of a video we observed the effect of both video's scripts and its scenes.

2 RELATED WORK

At the beginning of this project, we decided to find and read related articles about this area. Here are their brief explanation:

Is there an advantage for recognizing multi-modal emotional stimuli? [11]. This article indicated that emotion recognition is significantly better in response to multi-modal versus uni-modal stimuli. However, this article's results were based on human trials and it did not include any machine learning methods.

Context-dependent sentiment analysis in user-generated videos [12]. This article proposed a LSTM-based model that enables utterances to capture contextual information from their surroundings in the same video, thus aiding the classification process.

Multilogue-net: A context aware RNN for multi-modal emotion detection and sentiment analysis in conversation [14]. This article proposed a context-level inter-modal attention framework for simultaneously predicting the sentiment and expressed emotions of an utterance, evaluated using CMU-MOSEI [18] dataset.

Multimodal Sentiment Analysis on Video Streams using Lightweight Deep Neural Networks [16]. This article proposed a deep neural learning approach based on multiple modalities in which extracted features of an audiovisual data stream are fused in real time for sentiment classification.

Multi-task Learning for Multi-modal Emotion Recognition and Sentiment Analysis [3]. This article proposed a context-level inter-modal attention framework for simultaneously predicting the sentiment and expressed emotions of an utterance, by extracting an embedding from each multi-modal input (text, acoustic, and visual frames) from a GRU network and also from a CIM, which applies pairwise attention on the previous embeddings. The embeddings from all these processes will be concatenated. Afterward, a classifier on both emotions and sentiments will be trained. This framework has been evaluated on CMU-MOSEI [18] dataset.

3 DATASETS

There are several datasets for this particular problem, some of which are *MOSI* [17], *MOSEI* [18], and *MSCTD* [7]. In this section we will introduce these datasets and explore their main features and information.

3.1 MOSI

MOSI is a dataset of YouTube vlogs (videos of YouTubers expressing their opinion about general subjects) consisting of 2199 video clips from 98 speakers, with a total time of around 2 hours and 36 minutes. These videos vary in quality, distance from camera, lighting, background, etc. There may be more than one speaker in each video. Speakers talk in English and videos are manually transcribed by experts in multiple steps.
Sentiments in these videos are demonstrated with a linear scale: −3 (highly negative), −2 (negative), −1 (weakly negative), 0 (neutral), 1 (weakly positive), 2 (positive), and +3 (highly positive); there was also an "uncertain" choice for not being sure about video's sentiment. Sentiment labeling was performed by online workers with a high approval rate from the Amazon Mechanical Turk website. The distribution of sentiments over the entire dataset is shown in Figure 1.

3.2 MOSEI

MOSEI is the next generation of *MOSI* dataset, a collection of 23453 YouTube video clips from 1000 speakers, with a total time of around

Supervised by Ehsaneddin Asgari.

Project, NLP, Spring 1401
2018.

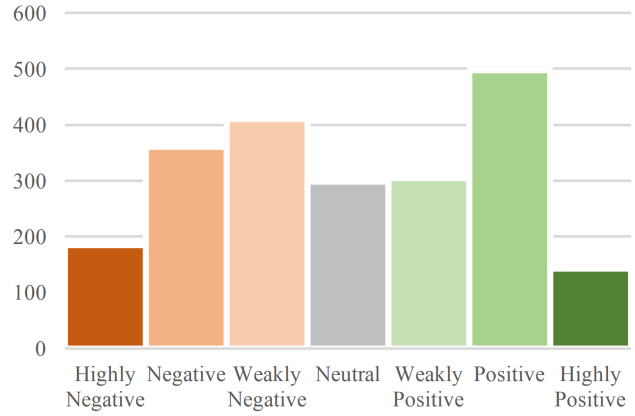

Figure 1: Distribution of sentiments over the MOSI dataset, taken from [17]

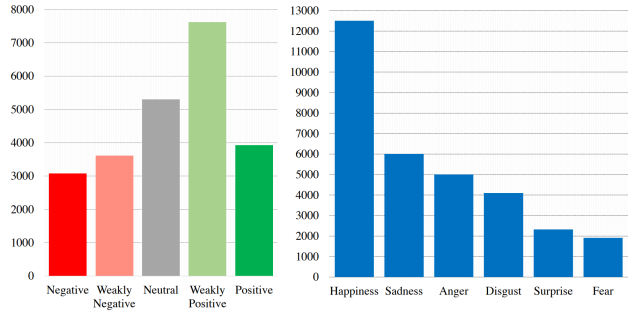

Figure 2: Distribution of sentiments and emotions over the MOSEI dataset, taken from [18]

65 hours and 54 minutes. It has the same sentiment labels as *MOSI*, in addition to emotional labels (happiness, sadness, anger, fear, disgust, surprise) labeled by experts. Each video has one speaker and transcribed by its uploader. The distribution of sentiments and emotions over the entire dataset is shown in Figure 2. Due to large number of videos contained in this dataset, HPS are needed to process it.

3.3 MSCTD

Eventually, we decided to use the *MSCTD* dataset. It is a new dataset created in the year 2022 and includes over 17k multi-modal bilingual conversations, consisting of more than 142k English-Chinese and 30k English-German utterance pairs, where each utterance pair corresponds with the associated visual context indicating where it happens. Each visual context of this dataset is a sequence of series or movies images and its dialogues are from OpenViDial dataset [9]. In addition, each utterance is annotated with one sentiment label (i.e., positive/neutral/negative). The distribution of sentiments over the entire dataset is shown in Figure 3.
This dataset's annotation includes two steps: automatic annotation and then human annotation, for checking and correcting automatic labeling. The size details of the English-German section of

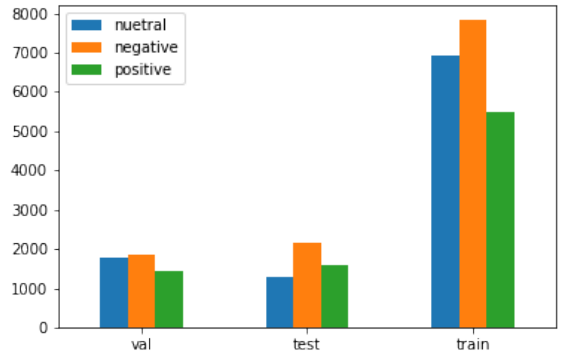

Figure 3: Distribution of sentiments over the MSCTD dataset

the dataset, which we used in our project, can be seen in table 1. The advantage of this dataset for us was that we wanted to extract numerous features from each scene. In the previous datasets, we mainly had the face of just one speaker, but here we have the complete scene, which allows us to extract additional features, such as different faces, poses, scene details, etc.

Sentiment	Train	Evaluation	Test
Positive	5483	1454	1606
Neutral	6921	1764	1298
Negative	7836	1845	2163

Table 1: Number of each utterance in each section in *MSCTD* dataset

3.4 Other Datasets

Moreover, there are other datasets that can be used in future works in order to make this project's results more precise. Some of them are:

(1) *IEMOCAP [4]*: Consisted of almost 12 hours of monologues from 10 actors with markers on their faces, heads, and hands, eliciting different emotions during speaking. It needed verification for giving this database which could take a week.
(2) *EmoReact [10]*: A collected multi-modal basic and complex emotion dataset of children between the ages of four and fourteen years old. It also required up to one-week of verification.
(3) *PATS [1, 2, 5]*: Contains transcribed pose data with aligned audio and transcriptions of 25 speakers of talk shows, lecturers, etc. Using the pose and audio of this dataset can improve our project's outcomes and, therefore, could be considered in future works.

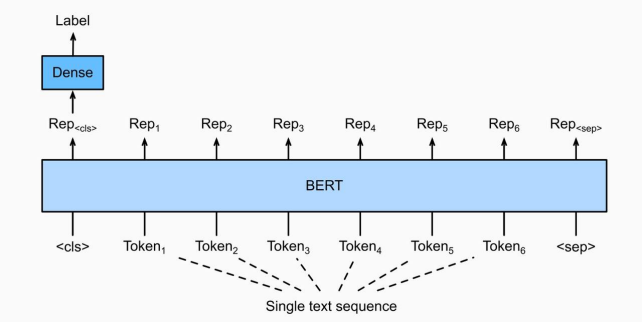

Figure 4: Roberta For Sequence Classification

4 METHODOLOGY

It should be noted that we could not use the MSCTD dataset completely, since some scenes did not have a person in them, (for face or pose recognition), or their face and/or pose could not be recognized. Due to these problems, first, we filtered this dataset, which led to omitting about 20 percent of MSCTD, and then it was used for our model.

In order to create a multi-modal network, we decided to extract two embeddings from pre-trained networks, then concatenate them and add a linear layer, which leads to a classifier with three possible outcomes for existing sentiments (i.e., positive/ neutral/ negative). We compared the result of this classifier with single-modal networks with a number of criteria.

4.1 Text Classifier

We used the "RobertaForSequenceClassification" classifier based on RoBERTa [8], a robustly optimized method for pretraining NLP systems that improves on BERT, the self-supervised method released by Google, which gives us the sentiment of a text. Sequence classification uses the final embedding for *CLS* token of each sentence. Its details can be seen in Figure 4. *MSCTD* dataset's scripts are classified by this pre-trained network.

4.2 Image Classifier

Face recognition. In order to extract a face from an image, four steps must be done:

(1) *Face detection*: We used the *MTCNN* [19] implemented with facenet pytorch (since it was much faster than implemented with TensorFlow) in order that the fundamental face points, such as eyes, nose, mouth, etc. are determined.
(2) *Face alignment*: All faces should be aligned, using the angle of eyes with a horizontal line. As a result, all face points will be in similar conditions after this rotation.
MTCNN can detect and align the face in three stages: In the first stage it uses a shallow CNN to quickly produce candidate windows. In the second stage it refines the proposed candidate windows through a more complex CNN. And lastly, in the third stage it uses a third CNN, more complex than the others, to further refine the result and output facial landmark positions.
(3) *Landmark detector*: With some help from the *Dlib* library [6] in order to find landmarks around the face, we masked the face and removed extra points out of the face. By doing this, emotion recognition from the face will be more precise. After localizing the face in the image, detecting these landmarks will be possible. With this pre-trained facial landmark detector, 68 (x, y) coordinates of facial structures will be estimated and those of them which are around the face and specify its boundary will be used.
(4) *Emotion recognition*: The sentiment of a face will be recognized using a model called *Multi-task EfficientNet-B2* [13], by a single efficient neural network. This network is pre-trained on face identification and fine-tuned for facial expression recognition on static images. At the end of this network, there is a sub-network for classifying the emotion, based on its input which is emotion embedding. As we decided to keep this embedding in order to combine it with other embeddings, this sub-network is omitted.

After all these steps, the last layer's features of the dense model will be used as an embedding for a face. By this method, the embedding of each face can be extracted. Moreover, getting the average of all face embeddings is also possible.

Pose detection. As an improvement, we decided to include the pose of humans in our model, using keypoint extraction [15]. By this method, 17 (x, y) keypoints' coordinates of a person, such as elbows and ears, in a frame will be detected and concatenated. This method is based on a few deconvolutional layers added to a backbone network, which is ResNet-50.

Scene recognition. We used a model which its backbone is ResNet-50 to extract an embedding for each scene and added this embedding to the previous ones, although scene details did not improve our accuracy in recognizing emotion from a frame. As a result, we found out that a scene on its own does not have any additional information about sentiment. Perhaps a scene in which its faces have been masked includes extra data, but we did not try this method and we trained our model with scenes with faces.

5 EXPERIMENTAL EVALUATION

The final results of our trained multi-modal sentiment recognition model in comparison with a pre-trained model on the text and a sentiment recognition model based on face, trained on train dataset, on the test dataset can be seen in Table 2. We used *macro* mode of each criteria.

Model/Criteria	F1	Accuracy	Precision	Recall
Text	0.408	0.424	**0.608**	0.466
Face	0.299	0.454	0.379	0.363
Text, Face, Pose	0.557	0.579	0.558	**0.556**
Text, Face, Pose, Scene	**0.559**	**0.581**	0.564	**0.556**

Table 2: Final results of multi-modal and single-modal models

As it can be seen, the multi-modal model had better outcomes rather than the uni-modal ones. Only when models are being compared with precision, sentiment recognition based on text performs better. Moreover, recognizing sentiment solely using face has never been superior in comparison to others.

In addition, including scenes can also lead to improving our model, however, it cannot enhance it extremely and it contributes mildly. Understanding emotion in a scene is usually dependent on previous scenes, therefore we decided to examine a transformer-based model in order to use attention to find related scenes. In this model, each embedding of a frame, consisting of face, pose, and text embeddings was an input for a sequence of transformer encoders. Then, the outcome was the input for a fully-connected layer, which resulted in our logits for our labels. We used different loss functions such as Cross-Entropy and negative likelihood loss to compute our loss in training. Although, when we examined this method we observed that the results are not as good as we expected which might be the result of a complex network or the need for other training methods.

6 CONCLUSION

In this project, we examined a multi-modal model for sentiment recognition in a sequence of frames, in order to evaluate some factors' contribution to this task, i.e. text, face, pose, and scene details. Based on the results, we realized that in sentiment recognition a combination of these factors can lead to better outcomes with higher accuracy and recall.

This method's results can help label data in other cases. For instance, sentiment recognition in other languages with a small number of data is challenging, but via this model, many data could be classified and prepared for sentiment recognition training.

7 FURTHER WORK

These multi-modal methods can also be implemented in other languages using transcripts and subtitles from those languages.

There have been major successful methods in this field and some of them have been introduced in section 2 (Related Work) articles. Adopting their pre-train models and fine-tuning them on our datasets would lead to enhancing the results.

There are different approaches for fixing transformer-based models. Using Masked Language Models we can mask the embedding of a few scenes and use the model to get their embeddings. We can also try other transformer-based networks with various loss functions to capture the main problem in the training.

REFERENCES

[1] Chaitanya Ahuja, Dong Won Lee, Ryo Ishii, and Louis-Philippe Morency. 2020. No Gestures Left Behind: Learning Relationships between Spoken Language and Freeform Gestures. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 1884–1895.

[2] Chaitanya Ahuja, Dong Won Lee, Yukiko I. Nakano, and Louis-Philippe Morency. [n. d.]. Style Transfer for Co-Speech Gesture Animation: A Multi-Speaker Conditional-Mixture Approach.

[3] Md Shad Akhtar, Dushyant Singh Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multi-task learning for multi-modal emotion recognition and sentiment analysis. *arXiv preprint arXiv:1905.05812* (2019).

[4] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, 4 (2008), 335–359.

[5] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. 2019. Learning individual styles of conversational gesture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3497–3506.

[6] Vahid Kazemi and Josephine Sullivan. 2014. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1867–1874.

[7] Yunlong Liang, Fandong Meng, Jinan Xu, Yufeng Chen, and Jie Zhou. 2022. Msctd: A multimodal sentiment chat translation dataset. *arXiv preprint arXiv:2202.13645* (2022).

[8] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[9] Yuxian Meng, Shuhe Wang, Qinghong Han, Xiaofei Sun, Fei Wu, Rui Yan, and Jiwei Li. 2020. Openvidial: A large-scale, open-domain dialogue dataset with visual contexts. *arXiv preprint arXiv:2012.15015* (2020).

[10] Behnaz Nojavanasghari, Tadas Baltrušaitis, Charles E Hughes, and Louis-Philippe Morency. 2016. Emoreact: a multimodal approach and dataset for recognizing emotional responses in children. In *Proceedings of the 18th acm international conference on multimodal interaction*. 137–144.

[11] Silke Paulmann and Marc D Pell. 2011. Is there an advantage for recognizing multi-modal emotional stimuli? *Motivation and Emotion* 35, 2 (2011), 192–201.

[12] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*. 873–883.

[13] Andrey V Savchenko, Lyudmila V Savchenko, and Ilya Makarov. 2022. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Transactions on Affective Computing* (2022).

[14] Aman Shenoy and Ashish Sardana. 2020. Multilogue-net: A context aware rnn for multi-modal emotion detection and sentiment analysis in conversation. *arXiv preprint arXiv:2002.08267* (2020).

[15] Bin Xiao, Haiping Wu, and Yichen Wei. 2018. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*. 466–481.

[16] Atitaya Yakaew, Matthew N Dailey, and Teeradaj Racharak. 2021. Multimodal Sentiment Analysis on Video Streams using Lightweight Deep Neural Networks.. In *ICPRAM*. 442–451.

[17] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259* (2016).

[18] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2236–2246.

[19] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters* 23, 10 (2016), 1499–1503.