Emotion Recognition in Dialogue Systems

MEHAK PIPLANI, DISHA KEDIGE CHANDRASHEKARACHAR, and RHUSHABH VAGHELA

ACM Reference Format:
Mehak Piplani, Disha Kedige Chandrashekarachar, and Rhushabh Vaghela. . Emotion Recognition in Dialogue Systems. 6 pages.

1 PROBLEM STATEMENT

Conversation in its natural form is multi-modal. In dialogues, we rely on others' facial expressions, vocal tonality, language, and gestures to anticipate their stance.
For emotion recognition, multi-modality becomes particularly important when the language is difficult to understand. In these cases, we resort to other modalities, such as prosodic and visual cues.
Our project aims to identify the verbal and non-verbal cues and their impact in determining the emotion of a person during a conversation.

2 MOTIVATION

Our research work is focused on studying AI's capability to understand human emotions in real-time as inspired by the study [1]. Understanding emotions will allow the systems to adapt according to the responses and behavioral patterns. Working with artificial agents such as Cortana, Alexa, Siri in our day-to-day life, we believe adding the emotional cognitive ability to an agent's responses would improve conversations. This feature can be utilized by real-time personal assistants such as Siri, Google Assistant to interact more naturally and responsively while communicating with the users via voice and text.

3 DATASET AND STATISTICAL ANALYSIS

We are utilizing the dataset: Multi-modal EmotionLines Dataset (MELD) [2]. It more than 1400 dialogues and 13000 utterances from the Friends TV series and is divided into three sets namely, Train, Development, and test with more detailed statistics mentioned in Table 1. An example from the dataset showcasing an emotion shift from surprise to sadness in a single conversation is visualized in Fig 1. The fig 1 also defines how the emotional dynamics depend on both the previous utterances and their associated emotions.
Multiple characters (aged 25 to 32 yrs) enacted in the series with 6 main characters as visualized in Fig 3b. Each utterance in dialogue has been labeled by emotion from Anger, Disgust, Sadness, Joy, Neutral, Surprise, and Fear as illustrated in Fig 2. A character-wise distribution of the emotions can be visualized in Fig 3a. MELD also has sentiment (positive, negative, and neutral) annotation for each utterance as described in Fig 3c.

Authors' address: Mehak Piplani; Disha Kedige Chandrashekarachar; Rhushabh Vaghela.

Manuscript submitted to ACM

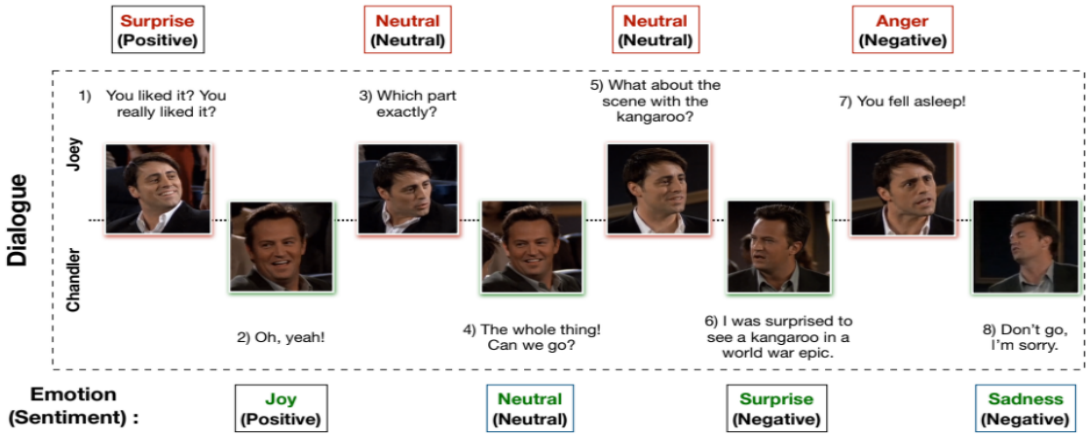


Fig. 1. An example from the dataset showing an emotion shift of speakers in a dialogue.

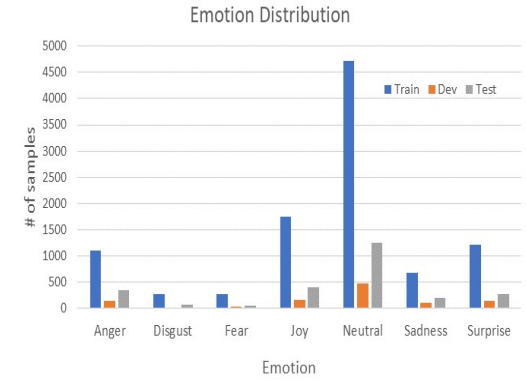
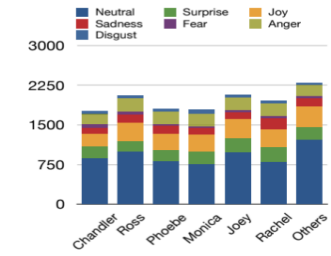


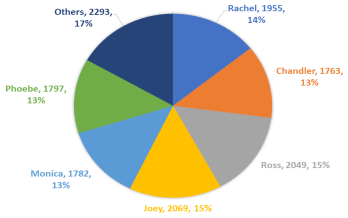
Fig. 2. Emotion distribution for three sets of the dataset: train, development, and test.

Statistics	Train	Dev	Test
# of unique words	10,643	2,384	4,361
Avg.utterance length	8.0	7.0	8.2
Max utterance length	69	37	45
# of dialogues	1039	114	280
# of utterances	9989	1109	2610
# of speakers	260	47	100
# of emotion shift	4003	427	1003
Avg. duration of an utterance	3.59s	3.59s	3.58s

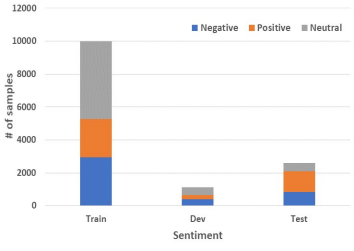
Table 1. Dataset statistics



(a) Emotion distribution character wise



(b) Character distribution



(c) Sentiment distribution

Fig. 3. Dataset statistics

4 FEATURE EXTRACTION

The study [3] was followed to extract features for each utterance in MELD.

4.1 Acoustic Features

The popular toolkit openSMILE [4] has been utilized to extract acoustic features. The toolkit extracts 6373-dimensional features constituting several low-level descriptors and various statistical functionals of varied vocal and prosodic features. As the audio representation is high dimensional, an L2-based feature selection was employed such as SVMs, to get a dense representation of the overall audio segment.

4.2 Lexical Features

The utterance transcripts were tokenized and embeddings were extracted using the pre-trained Glove model [5] to create 300-dimensional glove vectors.

4.3 Visual Features

For the visual features, each video was divided into frames. Each frame was passed through a ResNet model to extract its embedding. Later, average-pooling was used to merge the embedding for each frame corresponding to one training sample.

5 METHODOLOGY

5.1 Data Pre-processing

To capture an efficient relationship between the features and their labels in this skewed dataset, techniques like up-sampling and down-sampling were utilized. The training samples belonging to class "neutral" and "Joy" were down-sampled to maintain a constant ratio of examples between all the classes. The training samples for class "sadness" were up-sampled for the same reason. Various experiments and learning's from the article [6] led to belief that keeping the training samples belonging to class "disgust" and "fear" were affecting the efficiency of the model and hence only 5 labels are being considered for further experiments and observations.

5.2 Models

The motive of our experiments was to understand which modality captures the context efficiently and how these modalities interact with each other. To get an idea of the affect of each modality on the emotion label, we developed uni-modal models using Multi-Layer Perceptron (MLP) networks consisting of two or three linear layers.
Further, to experiment how each of these modalities comprehend each other, we tried creating bi-modal networks for these modalities pairwise via the techniques of early fusion and late fusion. For early fusion, features were concatenated and modelled through a MLP network, and for late fusion, a weighted average of the output probabilities from the individual modality models was calculated and the class label was predicted.

5.3 Experimental Setup

ADAM optimizer [7] and cross-entropy loss [8] were employed during the training. The number of epochs experimented was in range of 20 to 300 with Early Stopping enabled and learning rate was varied from 1e-05 to 1e-04.

6 RESULTS

The process of emotion detection can be mapped to a multi-class classification problem. Hence, evaluation metrics like weighted average F1 score and class wise F1-scores are used to measure the performance of our model. A comparison between the Uni-modality models has been showcased in Table 2 and for Bi-modal models in Table 3.

Modality	Anger	Joy	Surprise	Sadness	Neutral	Weighted
Acoustic	0.48	0.14	0.26	0.28	0.37	0.343
Lexical	0.44	0.52	0.36	0.61	0.57	0.50
Visual	0.39	0.39	0.21	0.31	0.25	0.31

Table 2. Uni-Modal results

Modality	Fusion	Anger	Joy	Surprise	Sadness	Neutral	Weighted
Acoustic + Lexical	Early	0.50	0.35	0.28	0.47	0.47	0.43
	Late	0.49	0.51	0.37	0.61	0.58	0.52
Visual + Lexical	Early	0.32	0.44	0.19	0.32	0.27	0.31
	Late	0.22	0.10	0.18	0.41	0.14	0.20
Visual + Acoustic	Early	0.31	0.46	0.19	0.34	0.23	0.31
	Late	0.35	0.19	0.15	0.17	0.25	0.23

Table 3. Bi-Modal results

7 OBSERVATIONS

- The emotion classes disgust and fear were very complex to train due to lack of training samples.
- Pre-processing techniques like Upsampling and Downsampling improved the F1-score.
- While testing the acoustic model, we observed that joy was misclassified to anger for a lot of samples.
- The results of the Acoustic-Lexical model show that when the textual and acoustic features were trained together, the model performed better on ambiguous samples.
- Batch Normalization improved the acoustic model's results drastically whereas, the lexical model's results didn't show any difference.
- Early Fusion model of Acoustic-Lexical modality pair outperformed all the other models.

8 LESSONS LEARNT

- Emotion for a dialogue predominantly depends on four factors interlocutor state, interlocutor intent, the preceding and future emotions, and the context of the conversation [9].
- The context captured by the glove embeddings lead to a higher and class balanced F1-score.
- Additive fusion techniques like early fusion and late fusion rely on figuring out relative emphasis on different modalities. However, in reality, every modality may not be reliable due to sensor noise or background noise.
- Inter-speaker influence plays an important role in determining emotion as people tend to change their responses and behavior based on the response from the person they are in conversation with.

9 FUTURE MILESTONES

We plan to carry out experiments with functionals extracted from the openSMILE toolkit as acoustic features, Bert embeddings as lexical features, and various pooling techniques for concatenating frame embeddings for visual features. To extract the interlocutor state, the preceding, and future emotions to model our task better, the current state-of-the-art method: DialogueRNN [10] will be explored. The study mentions that it captures context by tracking individual speaker states throughout a conversation to effectively emotion emotional states. Fusion techniques like Multilogue-Net [9] and Tensor fusion networks [11] will also be explored to model the interactions between all the three modalities to predict the emotions. We also plan to inspect gender bias in the dataset and take steps to handle it. Last, we also plan to check for bias due to enacted emotions and see if our models perform efficiently in a conversation possessing spontaneous emotions.

10 CONTRIBUTION

- Mehak
  - Responsible for analyzing the dataset and showcasing the statistical analysis in an interpretable way.
  - Worked on pre-processing of the dataset for extraction of acoustic features.
  - Worked on creating a subset of data with a balanced number of training samples for each class.
  - Worked on creating the uni-modal models for the acoustic modality.
  - Worked on bi-modal models: early fusion and late fusion models for the acoustic and lexical pair.
- Disha
  - Worked on pre-processing of the dataset for extraction of textual features.
  - Worked on creating a subset of data with a balanced number of testing samples for each class.
  - Worked on creating the uni-modal models for textual modality.
  - Worked on bi-modal models: early fusion and late fusion models for the lexical and visual pair.
- Rhushabh
  - Worked on pre-processing of the dataset for extraction of visual features.
  - Worked on creating the uni-modal models for visual modality.
  - Worked on bi-modal models: early fusion and late fusion models for the acoustic and visual pair.

REFERENCES

[1] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(02):1359–1367, Apr. 2020.

[2] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. pages 527–536, 01 2019.

[3] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 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)*, pages 873–883, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[4] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1459–1462, New York, NY, USA, 2010. Association for Computing Machinery.

[5] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[6] L. Wikarsa and S. N. Thahir. A text mining application of emotion classifications of twitter's users using naïve bayes method. In *2015 1st International Conference on Wireless and Telematics (ICWT)*, pages 1–6, 2015.

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[8] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in neural information processing systems*, pages 8778–8788, 2018.

[9] Aman Shenoy and Ashish Sardana. Multilogue-net: A context-aware RNN for multi-modal emotion detection and sentiment analysis in conversation. In *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, pages 19–28, Seattle, USA, July 2020. Association for Computational Linguistics.

[10] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. Dialoguernn: An attentive rnn for emotion detection in conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6818–6825, Jul. 2019.

[11] Saurav Sahay, Shachi H Kumar, Rui Xia, Jonathan Huang, and Lama Nachman. Multimodal relational tensor network for sentiment and emotion classification. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pages 20–27, Melbourne, Australia, July 2018. Association for Computational Linguistics.