Sequential Emotion Recognition in Conversations

Mridul Khurana*
Virginia Tech
mridul@vt.edu

FNU Hardeep*
Virginia Tech
hardeep1@vt.edu

Harish Babu Manogaran*
Virginia Tech
harishbabu@vt.edu

Shri Sarvesh Venkatachala Moorthy*
Virginia Tech
sarvesh2k@vt.edu

Upasana Sivaramakrishnan*
Virginia Tech
upasana@vt.edu

Abstract

With the recent trends in the development of conversational Artificial Intelligence (AI), we notice that almost all commercial websites (mainly e-commerce) incorporate a "chatbot/support agent" to interact with potential customers. An analysis of the emotion and intent of the customer in a conversation is thus essential to study market growth. Understanding the flow of emotion in a conversation enables us to accurately predict if a customer may subscribe to or purchase a particular service. In this work, we investigate the performance of a recent method called EmotionFlow(1) on IEMOCAP(2), DailyDialog(3), and EmoryNLP(7) datasets to identify the spread of emotion in a given conversation. We will benchmark model performance on these datasets. The code can be found at https://github.com/mridulkhurana/emotionflow

1 Introduction

Despite the birth of AI around the 1950s, it was only during the later parts of the 20th century that the concept gained momentum among researchers, diversifying into fields such as Deep Learning, Computer Vision, Natural Language Processing (NLP), and Machine Learning. However, among all these fields, NLP stood out due to its uniqueness in computational and linguistic techniques via which computer systems comprehend and emulate human-computer interactions in the form of speech and text. It began with sentiment analysis, which is based on the modeling of perception and beliefs and assigning a positive, negative, or neutral stance to the text. But emotions analysis traverses beyond that, by distributing the analysis to the various aspects of the sentiment spectrum.

Emotion in Recognition in Conversations (ERC) has gained popularity for developing empathetic machines. It finds use in several applications, some of which are: mining opinions from publicly available conversational data on platforms such as Reddit, as a tool for psychological analysis in health care, and for understanding student frustration in the education sector, among others.

Recent ERC methods utilize graph-based neural networks to take the relationships between the utterances of speakers into account. There are two issues associated with this. Firstly, graph-based neural networks don't take sequential information into account, which is essential to infer emotion. Second, users' emotions would change due to the impact of others' emotions.

As a result, the current advances in Emotion Recognition in conversation take into account sequential emotions or contextual information to accurately identify the emotion that is associated with each utterance in the dialog[(1), (4), (5)]. The EmotionFlow(1) model is one such technique that considers the spread of emotion in the entire conversation. This model is implemented on the MELD Dataset(6).

* Equal Contribution

2 Related Work

Several studies have implemented various techniques for textual emotion recognition. Seal et al.(8) employed a keyword-based approach with an emphasis on phrasal verbs. The keyword-based approach was tested on the ISEAR dataset(9). On the other hand, Alotaibi(10) focused on an approach based on learning. He analyzed the usage of classifiers such as Logistic Regression, K-Nearest Neighbour, and Support Vector Machines (SVM) on the ISEAR dataset(9). He inferred that based on the classifier performance, deep learning techniques would contribute to model improvement.

Xu et al.(11) proposed an Emo2Vec method that vectorized emotion semantics. A multitask learning framework was utilized to train Emo2Vec on smaller (namely ISEAR and Olympic) and larger datasets. The results show its performance to be better than a CNN-based implementation. The Emo2Vec method found its use in stress detection and emotion analysis. Ragheb et al.(12) worked on a learning-based model with data that consists of the 6 types of emotion that were listed by Paul Ekman(21): disgust, anger, surprise, fear, sadness, and joy. Once the data is acquired, tokenized, and sent through the encoder, it goes through average stochastic gradient descent (ASGD) trained Bi-LSTM units. Dropouts are applied between LSTM units for preventing over-fitting, and a self-attention mechanism is used to prioritized emotion-influenced conversations. Upon testing the model, it showed an F1 score of 75.82%.

Further along the ladder of learning-based machine learning models, Suhasini and Srinivasu(13) made use of machine learning classifiers based on KNN and Naive Bayers for emotion detection. They used the tweets from Sentiment 140 Corpus. The accuracy of KNN and Naive Bayes was compared with each other, from which Naive Bayes triumphed over KNN (55.50%) with an accuracy of 72.06%. However, both the models faced a common issue of the inability in extracting low-level contextual information. Building upon the drawback, Hasan et al.(14) utilized SVM learning. An offline model(15) was created for classifying emotions, and the online system classified real-time tweets the way the offline model did it.

In order to analyze and locate unstructured data, Rodriguez et al.(16) performed an analysis on emotion in select posts in social media that intend to spread hate. This was followed by Cao et al.(17), who highlighted challenges in textual emotion detection by exploiting deep learning and machine approaches to assess emotion in text. A survey on the emotion detection concept was conducted by Acheampon(18) to underline the research direction of text-based emotion detection system design. Nandwani(19) and Verma(20) prioritized on enhancing the process of emotion analysis in texts(16) and emphasized an emotion lexicon enriched with word intensities based on emotion.

3 Methodology

3.1 EmotionFlow

We adopt a technique to record the emotional sequence of events called EmotionFlow. It has two sub-parts: a CRF layer and an utterance encoder. The pre-trained language model RoBERTa(14), based on transformers, is used as our utterance encoder. The course of this approach is composed with the k turns of the latest utterances and their respective speakers, followed by an auxiliary inquiry about how the speaker feels at that instant. This is done for the model to learn the user-specific feature Fig. 1.

In order to understand the spread of emotions at the conversation level, we get the speaker's probability distribution of emotions from the utterance encoder. We calculate the emotion probability and classification loss L_1 from this. We then use this distribution as emission scores and as input to the CRF layer to obtain the negative log-likelihood L_2. The CRF layer aids in maximizing the ground truth's probability upon all the possible sequences of emotion(1). During the training phase, the utterance encoder and the CRF layer are jointly tuned using stochastic gradient descent, and the overall loss is the sum of losses from two sub-modules as shown in 1:

$$L_{total} = L_1 + L_2 \tag{1}$$

In order to decode the optimal emotion sequence at the prediction step, we run the Viterbi algorithm(8) over the CRF layer.

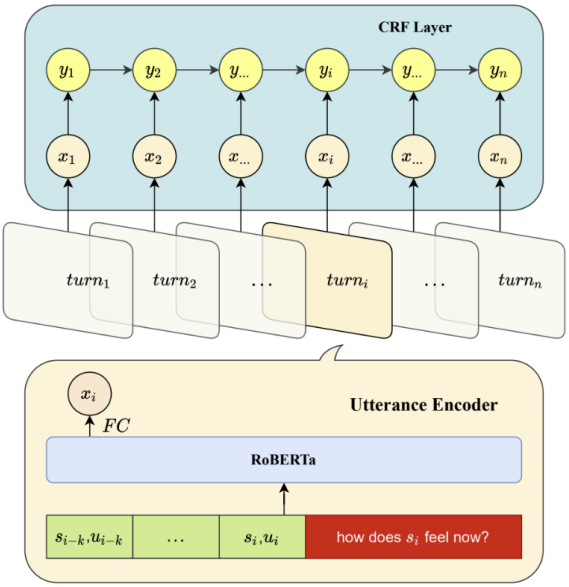


Figure 1: EmotionFlow Model Structure(1)

We investigate the performance of EmotionFlow on a variety of datasets namely IEMOCAP, DailyDialog, and EmoryNLP. We then perform a comparative analysis of the model's performance on the above-mentioned datasets against its performance on the MELD dataset by determining the F1-score. This will provide us insights into the robustness of the model, irrespective of the dataset used. We have also analyzed the performance of the model using two different variants of RoBERTa which are RoBERTa-base and RoBERTa-large.

We also further analyze the impact of speaker information on model performance. Currently, the model is taking into account the speaker prior to make a prediction about the speaker's current emotion. We aim to refine the model to remove this dependence.

4 Datasets

4.1 MELD

EmotionLines has been expanded upon and improved with the Multimodal EmotionLines Dataset (MELD). MELD has around 13,000 words from 1,433 conversations extracted from a popular TV show called "Friends". Each utterance includes audio, visual, and textual aspects and is tagged with emotion and sentiment descriptors. The dataset entails the following emotions i.e. anger, disgust, sadness, joy, neutrality, surprise, and fear, which have been assigned to each utterance in conversation.

4.2 IEMOCAP

Interactive Emotional Dyadic Motion Capture (IEMOCAP) is a staged, multimodal, and multispeaker database. It has around 12 hours' worth of audiovisual content, which includes video, voice, facial motion capture, and textual content. It comprises of dyadic sessions in which actors act out improvised scenes or staged situations that have been chosen intentionally to stimulate emotional reactions. The IEMOCAP dataset has multiple annotators i.e. category labels such as "angry," "happy," "sad," and "neutral," and dimensional labels like "valence," "activation," and "dominance."

4.3 EmoryNLP

EmoryNLP is another dataset based on the Friends television series, but its array of emotions is described as "joyful, calm, powerful, terrified, crazy, sad, and neutral." There are no sentiment labels, however, the following groups of sentiment classes can be made: Positive: "powerful, joyous, and peaceful," negative: "scared, angry, and sad," and neutral: "neutral."

4.4 DailyDialog

DailyDialog is a dataset comprising two speakers' everyday dialogues, and the list of emotions is as follows: "anger, disgust, fear, joy, surprise, sorrow, and neutral." The conversations in the dataset span a total of 10 themes and follow typical conversational flow such as Question-Answer and Directives-Commissives bi-turn flows. Moreover, the dataset includes distinctive multi-turn conversation flow behavior that mirrors our real-world, emotional style of communication.

Tab. 1 shows the statistics for the datasets we have used in this project.

Table 1: Data Statistics

Datasets	Dialogues			Utterances			Classes
	Train	Dev	Test	Train	Dev	Test	
MELD	1038	114	280	9989	1109	2610	7
IEMOCAP	108	12	31	5163	647	1623	6
EmoryNLP	713	99	85	9989	1109	2610	7
DailyDialog	11118	1000	1000	87170	8069	7740	7

5 Experiments and Results

5.1 Experiments

In order to train our classification problem in hand, we utilized two RoBERTa-based pre-trained language models i.e. RoBERTa(14) base model and the large model as the backbone. The large model has 354 million parameters compared to only 123 million parameters in the base model. We did a comparative analysis of the EmotionFlow model performance on three datasets.

We also experimented with different accumulation steps, i.e., optimization for different batch intervals. We noticed that for smaller datasets (like MELD, EmoryNLP, and IEMOCAP), an accumulation step of 2 produced the best results. On the other hand, for larger datasets (like DialyDialog), an accumulation step of 8 produced better results.

In addition, we ran experiments to understand the model's dependence on the speaker's prior. We modified the model to make the speaker's prior information redundant.

Our experiments were performed on a single NVIDIA A100 Tensor core GPU, a highly specialised GPU for AI, data analytics, and high-performance computing applications.

5.2 Results

As discussed earlier, the EmotionFlow model proposed in the reference paper was trained on the MELD dataset. An F1 score of 65.05% was reported from the experiments conducted in the paper. In addition to reproducing the experiment successfully, we were also able to achieve results with a better F1 score (by 1%). We then refined the datasets we wanted to experiment on and developed model-compatible data loaders for the same. In a nutshell, we were able to get F1 scores in our training that were close to the current state-of-the-art results as seen in Tab. 2

In addition, we refined the model to test its dependence on speaker information. We trained the refined model with RoBERTa large backbone by removing the speaker information i.e. only using utterances. This speaker independence model came at the cost of decreased F1 score, we observed a drop of close to one percentage point in F1 scores for all the datasets. Tab. 3 shows the results of the model without the speaker information. The observed results were, however, comparable to the model trained with speaker information and is more robust in the real world where we might not have priors of the speaker.

Due to the time constraints for fine-tuning the RoBERTa-large model ($\sim$15-20 hours), we couldn't analyze the model's speaker information independence on the DailyDialog dataset but expect to see a similar trend of getting a better F1-score by 1-2%.

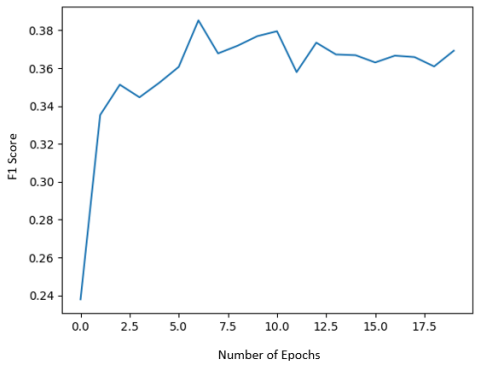
Table 2: Comparing accuracies for our model EmotionFlow vs the current state of the art on different datasets keeping the speaker information intact

Dataset	Weighted F1 Score		
	EmotionFlow		Current State-of-the-Art
	RoBERTA-base	RoBERTA-large	
MELD	64.51%	66.02%	67.25% (SPCL-CL(22))
EmoryNLP	38.53%	39.11%	40.94% (SPCL-CL(22))
IEMOCAP	60.73%	62.41%	71.77% (Bi-LSTM(23))
DialyDialog	55.58%	-	66.98% (Kl-Net(24))

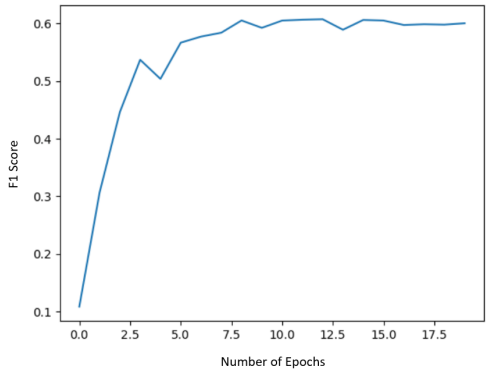
Table 3: Results of the model without speaker information

Dataset	F1 Score
	RoBERTa-large
MELD	65.35%
EmoryNLP	38.43%
IEMOCAP	62.96%

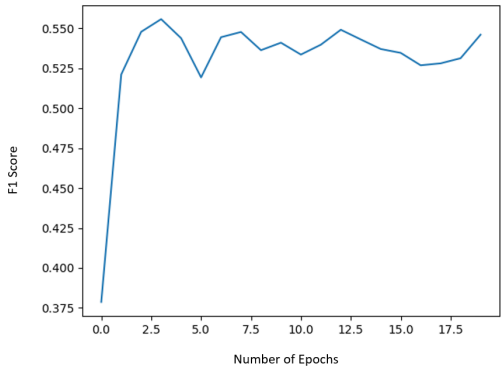
Fig. 2 provides the weighted F1-score achieved over different epochs on the test set. This takes into account the speaker's information with each utterance.



(a) EmoryNLP



(b) IEMOCAP



(c) DailyDialog

Figure 2: Weighted F1-scores for different datasets on the test data

6 Conclusion

In our project, we were able to refine the EmotionFlow model for real-time ERC. We developed dataloaders for testing on our chosen datasets: MELD, EmoryNLP, IEMOCAP, and DailyDialog. Upon testing our refined model on these datasets, we achieved comparable performance with state-of-the-art methods. Adding to this, we were also able to surpass the performance on the MELD dataset as compared to the model tested in the reference paper. This indicates that our model is now robust and can be easily extrapolated to various industry standard datasets. We also evaluated the model's performance after removing the speaker information to mimic real-life scenarios where we would not have any priors for the speaker.

Acknowledgement

We would like to thank our guide Dr. Ismini Lourentzou for her constant support and invaluable insight throughout the project. We have gained tremendous knowledge and experience, and we are grateful to her. We would also like to thank Advanced Research Computing at Virginia Tech for providing the required computational resources for our project.

References

[1] X. Song, L. Zang, R. Zhang, S. Hu, and L. Huang, "Emotionflow: Capture the Dialogue Level Emotion Transitions," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 8542-8546, doi: 10.1109/ICASSP43922.2022.9746464.

[2] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," Journal of Language Resources and Evaluation, vol. 42, no. 4, pp. 335-359, December 2008

[3] Li, Yanran, et al. "Dailydialog: A manually labeled multi-turn dialogue dataset." arXiv preprint arXiv:1710.03957 (2017)

[4] Joosung Lee and Wooin Lee. 2022. CoMPM: Context Modeling with Speaker's Pre-trained Memory Tracking for Emotion Recognition in Conversation. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5669–5679, Seattle, United States. Association for Computational Linguistics.

[5] M. R. Makiuchi, K. Uto and K. Shinoda, "Multimodal Emotion Recognition with High-Level Speech and Text Features," 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2021, pp. 350-357, doi: 10.1109/ASRU51503.2021.9688036.

[6] Chen, S.Y., Hsu, C.C., Kuo, C.C. and Ku, L.W., "EmotionLines: An Emotion Corpus of Multi-Party Conversations," arXiv preprint arXiv:1802.08379 (2018).

[7] Sayyed M. Zahiri, Sayyed M. Zahiri, "Emotion Detection on TV Show Transcripts with Sequence-based Convolutional Neural Networks," arXiv preprint arXiv:1708.04299 (2017).

[8] D. Seal, U. K. Roy, and R. Basak, "Sentence-level emotion detection from text based on semantic rules," Information and Communication Technology for Sustainable Development, Springer, Singapore, pp. 423–430, 2020.

[9] A. A. Alnuaim, M. Zakariah, P. K. Shukla et al., "Human-computer interaction for recognizing speech emotions using multilayer perceptron classifier," Journal of Healthcare Engineering, vol. 2022, Article ID 6005446, 12 pages, 2022.

[10] S. M. Mohammad and F. Bravo-Marquez, "WASSA-2017 Shared Task on Emotion Intensity," 2017, http://arxiv.org/abs/1708.03700.

[11] P. Xu, A. Madotto, C. S. Wu, J. H. Park, and P. Fung, "Emo2vec: learning generalized emotion representation by multi-task training," 2018, http://arxiv.org/abs/1809.04505.

[12] W. Ragheb, J. Azé, S. Bringay, and M. Servajean, "Attention-based modeling for emotion detection and classification in textual conversations," 2019, http://arxiv.org/abs/1906.07020.

[13] M. Suhasini and B. Srinivasu, "Emotion detection framework for twitter data using supervised classifiers," Data Engineering and Communication Technology, Springer, Singapore, pp. 565–576, 2020.

[14] M. Hasan, E. Rundensteiner, and E. Agu, "Automatic emotion detection in text streams by analyzing twitter data," International Journal of Data Science and Analytics, vol. 7, no. 1, pp. 35–51, 2019.

[15] A. S. Rajawat, P. Bedi, S. B. Goyal et al., "Fog big data analysis for IoT sensor application using fusion deep learning," Mathematical Problems in Engineering, vol. 2021, Article ID 6876688, 16 pages, 2021.

[16] A. Rodriguez, Y. L. Chen, and C. Argueta, "FADOHS: framework for detection and integration of unstructured data of hate speech on facebook using sentiment and emotion analysis," IEEE Access, vol. 10, pp. 22400–22419, 2022.

[17] L. Cao, S. Peng, P. Yin, Y. Zhou, A. Yang, and X. Li, "A survey of emotion analysis in text based on deep learning," in Proceedings of the 2020 IEEE 8th International Conference on Smart City and Informatization (iSCI), pp. 81–88, IEEE, Guangzhou, China, December 2020.

[18] F. A. Acheampong, C. Wenyu, and H. Nunoo-Mensah, "Textbased emotion detection: a," Engineering Reports, vol. 2, no. 7, Article ID e12189, 2020.

[19] A. S. Navarrete, C. Martinez-Araneda, C. Vidal-Castro, and C. Rubio-Manzano, "A novel approach to the creation of a labelling lexicon for improving emotion analysis in text," The Electronic Library, vol. 39, 2021.

[20] K. Sailunaz and R. Alhajj, "Emotion and sentiment analysis from Twitter text," Journal of Computational Science, vol. 36, Article ID 101003, 2019.

[21] P. Ekman, "Basic emotions," Handbook of cognition and emotion, vol. 98, no. 45-60, p. 16, 1999.

[22] Xiaohui, Song and Huang, Longtao and Xue, Hui and Hu, Songlin. (2022). Supervised Prototypical Contrastive Learning for Emotion Recognition in Conversation.

[23] Li, Zaijing and Tang, Fengxiao and Ming, Zhao and Zhu, Yusen. (2022). EmoCaps: Emotion Capsule based Model for Conversational Emotion Recognition.

[24] Yunhe Xie, Kailai Yang, Chengjie Sun, Bingquan Liu, and Zhenzhou Ji. 2021. Knowledge-Interactive Network with Sentiment Polarity Intensity-Aware Multi-Task Learning for Emotion Recognition in Conversations. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2879–2889, Punta Cana, Dominican Republic. Association for Computational Linguistics.