MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations

Natural Language Processing Project Report

Completed in Fulfilment of the Requirements for the Natural Language Processing Project Guidelines at NIIT University

Submitted By: (Group Number: G14)

Ajinkya Bedekar (U101116FCS183) (D5)

Dhruva Agarwal (U101116FCS177) (D5)

Group Number:

G14

Group Members:

No.	Name	Contribution in Project
1.	Ajinkya Bedekar	Developing the main code for analysis of dataset
2.	Dhruva Agarwal	Developing the helper code to run the main file

Reference Paper Title:

MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations

Authors:

- Soujanya Poria
- Devamanyu Hazarika
- Navonil Majumder
- Gautam Naik
- Erik Cambria
- Rada Mihalcea

Introduction

Modality refers to a particular mode in which something exists or is experienced or expressed. So, the word multimodal means having different types of modes. Multi-party defines that more than two people participating in the conversations.

Emotion recognition defines the process of recognising human emotions, most typically using facial expressions and verbal expressions. Emotion recognition in Conversation refers to the practice of recognition using the latter expressions.

Multimodal Multi-Party Dataset (MELD) for Emotion Recognition in Conversation, as explained by the title of the research paper, is a dataset for recognising emotions in conversations. In these conversations, multiple people are participating and the dataset takes into account the text spoken by the speakers as well as the audio of the speakers.

MELD is an extension and enhancement of a previously existing dataset, namely, EmotionLines dataset. The dialogues in MELD are same as they are in EmotionLines. But the basic difference between the two datasets is that EmotionLines just considers text for Emotion Recognition, and MELD takes into account both audio and text to perform the task.

MELD contains more than 1400 dialogues and 13000 utterances from a famous TV series, Friends. Multiple speakers participated in the dialogues. Every utterance in a dialogue is annotated with seven emotions, and three sentiments. The emotions used for labelling are Anger, Disgust, Sadness, Joy, Neutral, Surprise, Fear, and the sentiments are Positive, Negative, Neutral.

Objectives

Multimodal data analysis uses information from multiple-parallel data channels for making any decision. As the field of Artificial Intelligence is growing rapidly, multimodal emotion recognition has become a major research interest. It is mostly due to its potential applications in a lot of challenging tasks, like dialogue generation, multimodal interaction, etc.

A conversational emotion recognition system is used to generate appropriate responses by analysing user emotions. There are a lot of previous works done on multimodal emotion recognition. But, only a very few of them are focussed on emotion recognition in conversations. However, they are limited to dyadic conversation (conversation between two people) understanding. So, these works are not useful for emotion recognition in multi-party conversations with more than two participants.

EmotionLines is a dataset which can be used as a resource for emotion recognition. But it can only be used for text, as it does not include data from visual and audio modalities. There is no multimodal multi-party conversations having more than two participants.

In the research, authors have extended, improved, and further developed EmotionLines dataset for multimodal scenario. There are a lot of challenges in emotion recognition in sequential turns. One of the challenges is context understanding. The emotion change and emotion flow in the sequence of turns in a dialogue make accurate context modelling a difficult task.

In the dataset developed by authors, multimodal data sources for each dialogue is available, it is hypothesized to improve the context modelling, thus improving the overall emotion recognition performance. The developed dataset can also be useful for developing multimodal affective dialogue system.

IEMOCAP, SEMAINE are multimodal conversational datasets containing emotion label for each utterance. But these datasets are dyadic in nature. Therefore, the importance of MELD is justified. Other publicly available datasets for multimodal emotion and sentiment recognition are MOSEI, MOSI, MOUD. But none of them is conversational.

Work Done

The raw data, stored in .mp4 format, is available at the link below and can be found in XXX.tar.gz files.
http://web.eecs.umich.edu/~mihalcea/downloads/MELD.Raw.tar.gz

Annotations are available at https://github.com/SenticNet/MELD/tree/master/data/MELD.

Description of Raw Data

- There are 3 folders (.tar.gz files) - train, dev and test; each of which corresponds to video clips from the utterances in the 3 .csv files.
- In any folder, each video clip in the raw data corresponds to one utterance in the corresponding .csv file. The video clips are named in the format: diaX1_uttX2.mp4, where X1 is the Dialogue_ID and X2 is the Utterance_ID as provided in the corresponding .csv file, denoting the particular utterance.

Labelling

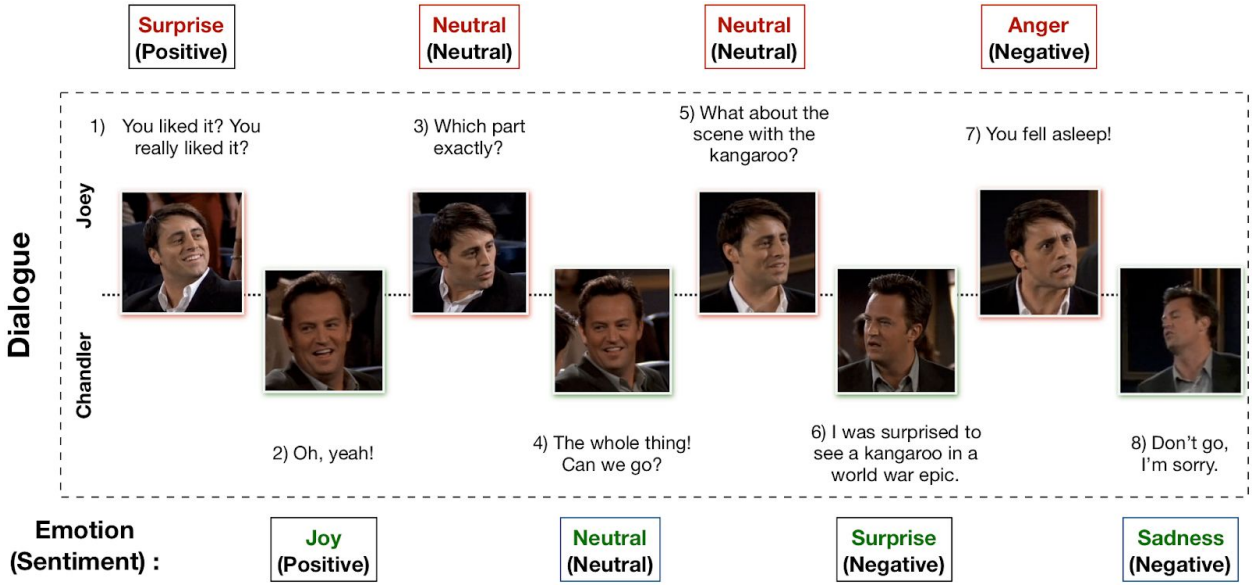
For experimentation, all the labels are represented as one-hot encodings, the indices for which are as follows:

- Emotion - {'neutral': 0, 'surprise': 1, 'fear': 2, 'sadness': 3, 'joy': 4, 'disgust': 5, 'anger': 6}. Therefore, the label corresponding to the emotion 'joy' would be [0., 0., 0., 0., 1., 0., 0.]
- Sentiment - {'neutral': 0, 'positive': 1, 'negative': 2}. Therefore, the label corresponding to the sentiment 'positive' would be [0., 1., 0.]

Class Weights

For the baseline on emotion classification, the following class weights were used. The indexing is the same as mentioned above. Class Weights: [4.0, 15.0, 15.0, 3.0, 1.0, 6.0, 3.0].

Example Dialogue



Dataset Statistics

Statistics	Train	Dev	Test
# of modality	{a,v,t}	{a,v,t}	{a,v,t}
# of unique words	10,643	2,384	4,361
Avg. utterance length	8.03	7.99	8.28
Max. utterance length	69	37	45
Avg. # of emotions per dialogue	3.30	3.35	3.24
# of dialogues	1039	114	280
# of utterances	10016	1111	2620
# of speakers	260	47	100
# of emotion shift	4003	427	1003
Avg. duration of an utterance	3.59s	3.59s	3.58s

Dataset Distribution

	Train	Dev	Test
Anger	1109	153	345
Disgust	271	22	68
Fear	268	40	50
Joy	1743	163	402
Neutral	4710	470	1256
Sadness	683	111	208
Surprise	1205	150	281

Experiments and Results

The main focus of our group was to analyse the dataset developed by the authors from different perspectives. The analysis was done on three different parameters, i.e., classification report, confusion matrix, and precision-recall f-score.

The types of classifications that are tested are Emotion and Sentiment. The modalities tested are Text, Audio, Bimodal.

We used the pre-trained models files and pickle files and did the experiments for analysis using the functions available in Python module, namely, sklearn.

Run the baseline

Below steps need to be followed for running the baseline -

1. Download the features.
2. Copy these features into ./data/pickles/
3. To test the baseline model, run the file: group14.py as follows:
    - python group14.py -classify [Sentiment|Emotion] -modality [text|audio|bimodal]
    - example command to test text unimodal for sentiment classification: python group14.py -classify Sentiment -modality text
    - use python group14.py -h to get help text for the parameters.
4. For pre-trained models, download the model weights and place the pickle files inside ./data/models/.

The result obtained after running the python script for Emotion classification and Text modality is as below.

```
Confusion Matrix :
[[1119   38    0    5   74    0   20]
 [  77  110    0    0   67    0   27]
 [  27    2    0    0   11    0   10]
 [ 137    7    0    5   27    0   32]
 [ 137   16    0    0  220    0   29]
 [  39    4    0    0   10    0   15]
 [ 114   32    0    0   88    0  111]]
Classification Report :
C:\Program Files\Python37\lib\site-packages\sklearn\metrics\classifi
  'precision', 'predicted', average, warn_for)
              precision    recall  f1-score   support

           0     0.6782    0.8909    0.7701      1256
           1     0.5263    0.3915    0.4490       281
           2     0.0000    0.0000    0.0000        50
           3     0.5000    0.0240    0.0459       208
           4     0.4427    0.5473    0.4894       402
           5     0.0000    0.0000    0.0000        68
           6     0.4549    0.3217    0.3769       345

    accuracy                         0.5996      2610
   macro avg     0.3717    0.3108    0.3045      2610
weighted avg     0.5512    0.5996    0.5478      2610

Weighted FScore:
  (0.5511821869374117, 0.5996168582375478, 0.5478063952793075, None)
```

Conclusions and Future Work

In the research paper, MELD dataset is introduced, which is a multimodal multi-party conversational emotion recognition dataset. The authors described the process of building this dataset, and provided results obtained with strong baseline methods applied on this dataset. The dataset contains raw videos, audio segments, and transcripts for multimodal processing. The features used for baseline experiments are provided.

Future research on this dataset should be focussed on improving the contextual modelling. Another direction is to use visual information available in the raw videos. Enhancements can be made by extracting relevant visual features through processes utilizing audio-visual speaker diarization. There is room for further improvement using other more advanced fusion methods such as MARN.

This dataset will be useful as a training corpus for both conversational emotion recognition and multimodal empathetic response generation. Building upon this dataset, future research can explore the design of efficient multimodal fusion algorithms, novel ERC frameworks, as well as the extraction of new features from the audio, visual, and textual modalities.