Final Report : Reproduction of EmotionFlow with Gender Bias Detection

Changhyun Lee, Yunjeon Lee, Jennifer Rodriguez-Trujillo
New York University, Center for Data Science
{cl4017, yl7143, jr5951}@nyu.edu

Abstract

This paper intends to propose a system that takes into account the sequence of human emotions in a conversation to better determine the resulting emotion of the last speaker. This conversation utilizes a Multimodal EmotionLines Dataset (MELD) (Poria et al., 2018). Through this data set, the experiment uses a Conditional Random Field (CRF) to encapsulate the tone of the speakers as a whole. Compared to other papers, this model has the ability to account for the effect of emotions at a social context level. EmotionFlow captures semantics through the composition of an utterance encoder and a CRF layer, where the utterance encoder is a pre-trained model RoBERTa. Once the user's scores are converted to probabilities, they are used as input values for the CRF layer(Song). To model the semantic language model, the creators use RoBERTa alongside cross-entropy loss as the objective function in the training phase. To further prove and confirm the effectiveness of Emotion-Flow, they conduct an ablation study on the MELD test set from the series Friends. Two main components that help with the effectiveness of the approach are the CRF layer and the QA-style input construction. Compared to existing sentiment analysis models, EmotionFlow takes into "consideration of the spread of participant's emotions during a conversation"(Song).

1 Summary of Process

We first explore the MELD data set from the drama Friends, which is used to train and predict sentence-level emotions through a pre-trained BERT-type language model. The CRF layer combines those emotions and calculates the overall conversation-level emotion predictions. We evaluate the model with an f1-score and explore different parameters to fine-tune the model. Finally, our extension tests and analyzes the existence of gender bias from the model.

2 Data Sets

In order to achieve the goal of predicting conversation-level emotions, we must also use conversational data.

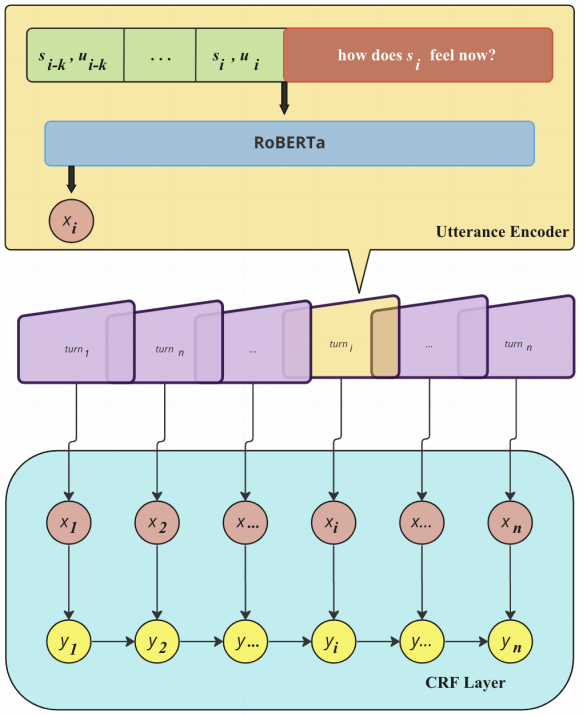

Figure 1: Diagram demonstrates the connection between the different components of EmotionFlow.

For our replication thesis (Song et al., 2022) we are using the MELD (Poria et al., 2018) data set collected from the TV show Friends. The data set consists of 1432 conversations, 13708 utterances, and 274 speakers in total. These files have Sr No., Utterance, Speaker, Emotion, Sentiment, Dialogue_ID, Utterance_ID, Season, Episode, StartTime, and EndTime as attributes. From this set of columns, only Utterance_ID, Utterance, Speaker, Emotion, and Dialogue_ID are used for the training. The data was split into 3 sets: train, dev, and test. In addition to the data set described above, we are using the entirety of the friends_transcript.json file. This file is a down-sampled version of the original data that extracts all speaker names from the full data set.

3 Model Implementation

In terms of implementing the sentiment analysis model pipeline, we completed the recreation of the utterance encoder structure using a pre-trained RoBERTa-base

- **L1 loss** from the encoder

$$\mathcal{L}_1(\theta_1) = -\sum_{i=1}^{M}\sum_{t=1}^{N_i} \log p_{i,t} * y_{i,t}$$

- **L2** loss from the **CRF**

$$\mathcal{L}_2(\theta_2) = -\log(\mathrm{P})$$

- **CRF layer** is optimized by maximizing the probability of the ground truth emotion sequence **P**, over all emotion sequence

$$\mathrm{P} = \frac{1}{Z(\boldsymbol{x})} \exp\left(p_{1,y_1} + \sum_{t=2}^{N}\left[g\left(y_{t-1}, y_t\right) + p_{t,y_t}\right]\right)$$

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$$

Figure 2: Composition of EmotionFlow's Formula.

and its baseline training function. This class is used to predict the emotion of each utterance from a speaker before it is then fed into the CRF layer of the model. We initialized the RoBERTa model with pre-configurations (default padding value, number of classes, cross-entropy loss function, speaker embeddings, and dropout). Then, the model's forward function was implemented to find sentence and speaker batch sizes and maximum turns of the speakers in the conversation. These were used as input to the specified encoder (i.e. RoBERTa). The encoder output is used to find the initial emission score necessary to run the CRF layer. The CRF layer takes the RoBERTa predicted emotion from each utterance and produces a final emotion for a given dialogue. The CRF was taken from the paper's EmotionFlow model, per William Merrill's permission. The diagram in Figure 1 walks through the steps described above.

The baseline train function performs the optimization on the train data set and saves the best model based on the f1-score over the input number of epochs. The overall model's loss function that linearly combines RoBERTa encoder loss and CRF loss is shown in Figure 2.

4 Experiments

We experiment with the EmotionFlow by using the same Friends Multi-modal EmotionLines Data Set (MELD) used in the original paper, which consists of seven emotion labels(Poria). Furthermore, this replication has two sub-modules: RoBERTa and a CRF layer to capture the sequential information of emotions. We calculate f1-scores for each epoch to evaluate the model's performance.

In Figure 3, we can observe the progress of the f1-score for both the train and dev data. Initially, we commenced running the model at 100 epochs. We hypothesize that, like most models, the accuracy for both training and dev sets will also increase as the number of epochs increases. Furthermore, considering the size of the data and the variation of the text data(some sentences being more extensive than others and the tone of the dialogue having the potential to vary), we require the number of epochs to be substantially larger than 5. However, we were not able to run with many epochs due to memory and time limitations. Instead, we decided to run the model with 5 epochs, from which we were able to reach an f1-score of approximately 0.64, similar to the original thesis.

5 Model Results

We test the model with varying values of the parameters: number of epochs, learning rate, and pre-trained learning rate. The performance is observed from the test data set using an f1-score. Each parameter is tuned with three different values. We keep the rest of the model parameters the same with 5 epochs, 0.3 dropout, 1 batch size, and a RoBERTa-base model.

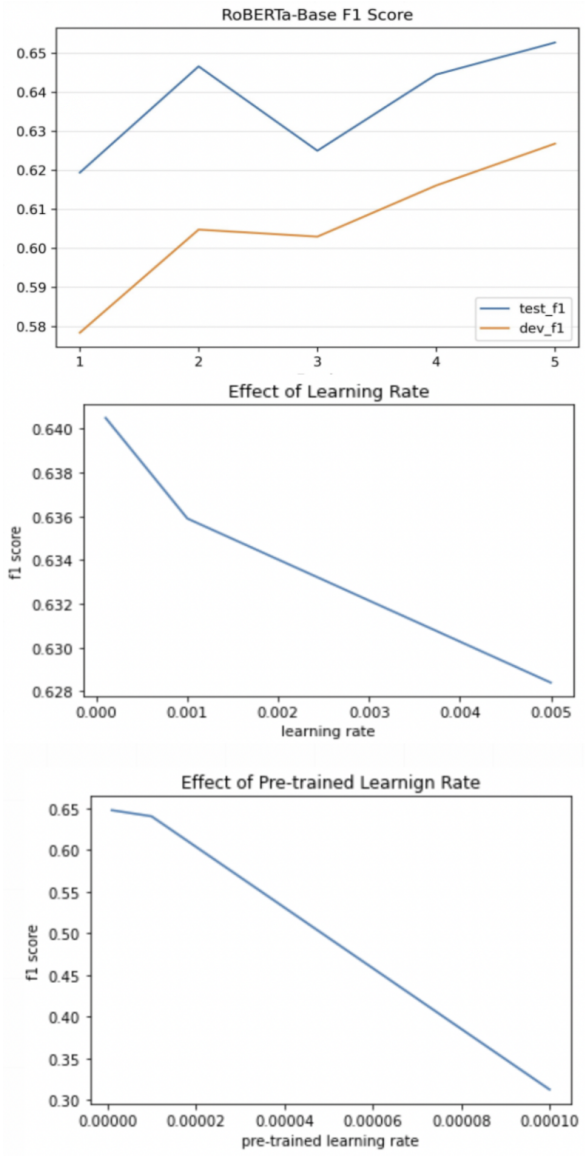

Figure 3: Graphs above demonstrate the models' accuracies with different epochs, learning rates, and pre-trained learning rates.

The first parameter was the epoch of the model. In general, both the dev and test set reflected better f1-score

after running on more epochs. Due to "Out of Memory" issues, we were limited in the number of epochs to run. In Figure 3, image 1, we observe the overall increasing trend of the performance, despite the drop in the test set's epoch 3.

The second parameter of interest was the learning rate on the CRF model. In Figure 3, image 2, we can once again see that the lower the learning rate, the higher the f1-score. On the contrary, the higher the learning rate, the lower the f1-score. The lower learning rate is more likely to converge to the global minimum loss than the higher learning rate, resulting in better performance. The values we tested are 0.0001, 0.001, and 0.005.

The third parameter we used for tuning was the pre-trained learning rate for RoBERTa. Figure 2, image 3, demonstrates a similar pattern as the CRF learning rate. The higher pre-training learning rate resulted in lower performance while the lower pre-trained learning rate resulted in a higher f1-score. The values we tested are 0.000001, 0,00001, and 0.0001.

6 Extension

Compared to other existing models, the EmotionFlow model predicts emotions based on dialogue data from the Friends series. To further extend the existing paper, we explored the existence of gender bias within the model. The original data set is divided into train, dev, and test. The train contains 260 unique "speakers"; the dev data consists of 47 unique speakers, and the test data consists of 100 unique speakers. The entire data set consists of 304 unique speakers.

6.0.1 Extension: Data Set

We take the subset of approximately 1000 utterances from the original data. By limiting the size of the utterances to approximately 1000, we focus on finding bias rather than increasing accuracy. From this subset, we extract the unique set of characters within the speaker column. We then take five approaches to create five different data sets.

The first, second, and third data set only consist of the American female, American male, and American gender-neutral names, respectively. The fourth data set maps existing male names to new female names and existing female names to new gender-neutral names. The fifth data set maps existing female names to new male names and existing male names to new gender-neutral names. All of the new name replacements come from the data set we created using 300 American female names, 300 American male names, and 300 American gender-neutral names.

To minimize the types of bias that may exist within the model, we control for the ethnic subgroups of names and ensure that each name is specifically categorized as female, male, or neutral. By fixing the factor of ethnicity, we make sure that gender bias is the only possible factor that can affect the model.

6.0.2 Extension: Methods

To assess the outcomes, we use a confusion matrix to compare the model's correctness in predicting emotions from five manipulated data sets. We also compare the total number of emotion predictions from the same data sets. As we set all factors the same except for the replaced gender names, the unbiased model should return similar results (i.e., f1-score and accuracy) across the five data sets.

7 Extension Results

We explore the results from running the five data sets and determine the difference in the prediction accuracy that may signify gender bias in the model.

The overall accuracy of the EmotionFlow using five data sets did not show a significant difference, and the f1-scores stayed at approximately 0.60 according to Table 1. This is expected since the change in names is not designed to affect the accuracy of the model. However, we explored further detailed emotion prediction results.

The confusion matrix for seven emotion classes allows us to directly see how the model predicts the emotion labels compared to the true labels. In general, fear and disgust have a small number of occurrences, which means no important findings can be derived. Neutral emotion is the majority of the labels, and its accuracy is similar between different data sets. The number of correct labels for each emotion from five data sets is shown in Table 2.

Furthermore, if we look at the total number of predictions of the model using different data sets, the model generally predicts more "anger" when there are more female characters. The model also predicts the emotion "surprise" more for the male gender, regardless of accuracy. The rest of the emotion prediction numbers show the same pattern as the number of correct labels, given in Table 3.

Data	f1-score
male_to_female_1000.csv	0.6063
female_to_male_1000.csv	0.6000
all_female.csv	0.5983
all_male.csv	0.5890
all_neutral.csv	0.5930

Table 1: f1-score results for extension

	N	M	F	F to M	M to F
Neutral	955	986	1002	991	1011
Surprise	183	194	182	202	184
Fear	2	1	1	2	4
Sadness	49	43	33	48	45
Joy	267	246	257	229	231
Disgust	1	1	1	1	3
Anger	128	111	143	134	145

Table 2: correct prediction for each emotion. (Legend: N = Neutral, M = Male, and F= Female)

	N	M	F	F to M	M to F
Neutral	1278	1333	1362	1328	1368
Surprise	413	465	382	468	400
Fear	13	26	5	11	20
Sadness	96	90	56	97	100
Joy	527	450	488	407	419
Disgust	2	4	2	3	6
Anger	265	226	299	280	283

Table 3: the total number of predictions for each emotion from the model (Legend: N = Neutral, M = Male, and F= Female)

7.0.1 Extension: Analysis

In terms of the emotion "angry," the data set with all female names have a higher f1-score accuracy. In two out of the five data sets that contained the majority of female names, "Anger" was correctly predicted more than the other data sets with male or neutral names. This finding veers towards the biased model because this kind of performance behavior should not exist in the unbiased model.

A second surprising finding within the confusion matrix is that, in general, males presented higher accuracy about the emotion "surprise." In two out of the five data sets that contained the majority of male names, "surprise" was labeled correctly most often. This suggests the model behaves differently just by changing the speaker names, which is a sign of model bias.

Finally, despite the similar f1-scores of five name-manipulated data sets, we observed changes in the numbers of correct labels and total labels for "anger" and "surprise." Because we controlled for the possibility of other biases in the model by using a highly selective set of names, we believe the model has a slight gender bias toward these two emotions. It is important to note that the model is designed to train only from the context (a.k.a. utterances), not from the speaker names. Nonetheless, the model did not output similar confusion matrices depending on different gender names, so we conclude the existence of gender bias.

8 Member Contributions

Given the complexity of the model, it was impossible to have different components of the model to be built in an unordered manner. This led us to all simultaneously work synchronously. Throughout the construction of EmotionFlow, all members were involved to prevent malfunctions that hindered the model from running in general (Note: the model fails to run when all components are not connected at once).

9 Issues and Future Considerations

Some of the challenges we faced were with the batch size and the number of epochs. Although we used GPU, the number of available resources did not suffice. If resources were not an issue, we would have opted to increase the number of epochs until the model ceases to improve or demonstrates signs of over-fitting.

Had time and the number of resources not been limited, we would have taken this paper a step further by considering other dialogue languages in order to investigate whether the model's accuracy outcomes only pertain to the English language. It is certainly a possibility that these outcomes only exist within the English dialogue, creating a bias in itself.

The topic of gender is complex and the ability to identify bias in a model is challenging. Given that multiple factors play a role in creating bias, we would have also reached out to responsible data science experts and social scientists.

10 Conclusion

We implemented the data processing and RoBERTa model for train and prediction of each sentence, and we referred to CRF from the original thesis. Through the implementation and parameter experiments, we reached approximately 0.64 for the f1 score. Considering that the f1-score in the original thesis was 0.6547, we successfully achieve a similar f1-score. With regard to the extension, the model presented a slight bias towards gender in certain emotions. Generally, there was no bias within Neutral, Fear, Sadness, Joy, and Disgust. However, the model tends to predict more "Anger" for females, and more "Surprise" for males, which leads us to make a conclusion that the model has a gender bias.

References

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *Computation and Language*, arXiv:1503.06733. Version 4.

Xiaohui Song, Liangjun Zang, Rong Zhang, Songlin Hu, and Longtao Huang. 2022. Emotionflow: Capture the dialogue level emotion transitions. *IEEE Xplore*.