Reproduction of EmotionFlow with Gender Bias Detection

Project Team: Changhyun Lee, Yunjeon Lee, Jennifer Rodriguez-Trujillo
Project Mentor: William Merrill

Center for Data Science, New York University

Introduction

Sentiment analysis is a popular method. However, baseline sentiment analysis is mostly done by classifying and predicting based on a singular text with a matching sentiment. Instead, we reproduce the paper on **EmotionFlow**, which is an **Emotion Recognition in Conversations (ERC)** model to find the overall emotion of the entire dialogue participants. We extend this ERC model to detect and analyze the existence of gender bias.

Research Question

How can we capture the dominant emotion in a conversation between a group of people?

Subproblems addressed:

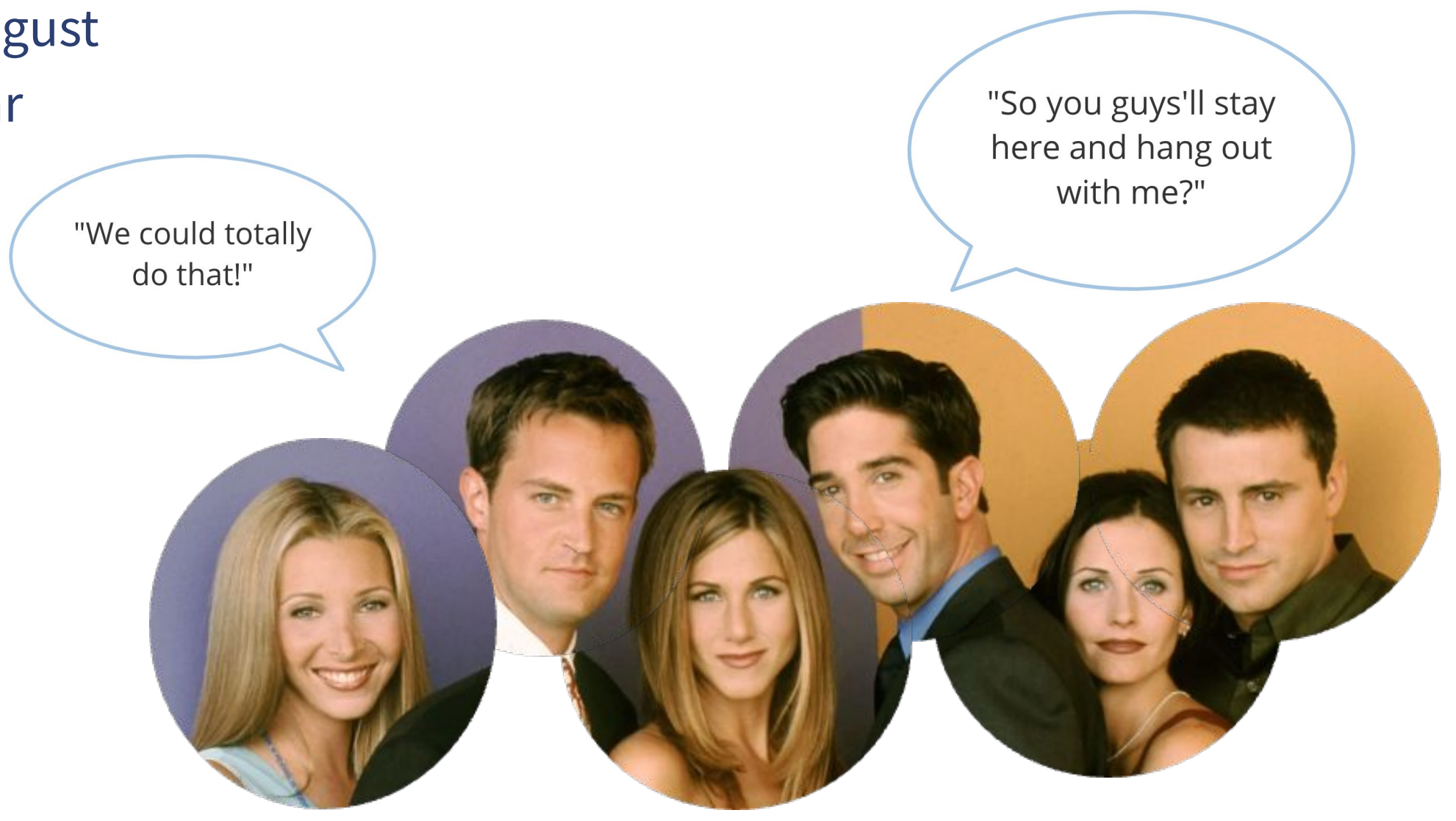
- Determine the methods to implement the *Friends* movie dataset
- Establish the baseline model that utilizes the pretrained **RoBERTa** as the utterance encoder.
- Addition of Question & Answering query in the input.
- Implementation of **Conditional Random Fields (CRF)** layer to capture sequential data on emotion.
- Optimizing both the pretrained model loss and the **CRF** layer loss.

Data Collection

The initial paper uses a MELD dataset collected from the TV show *Friends*. This data is conversational data that consists of 1432 conversations, 13708 utterances, and 274 speakers in total. This dataset is splitted into train, dev, and test dataset. These files have Sr No., Utterance, Speaker, Emotion, Sentiment, Dialogue ID, Utterance ID, Season, Episode, StartTime, and EndTime as attributes. Among these, only Utterance ID, Utterance, Speaker, Emotion, and Dialogue ID are used for the training the model.

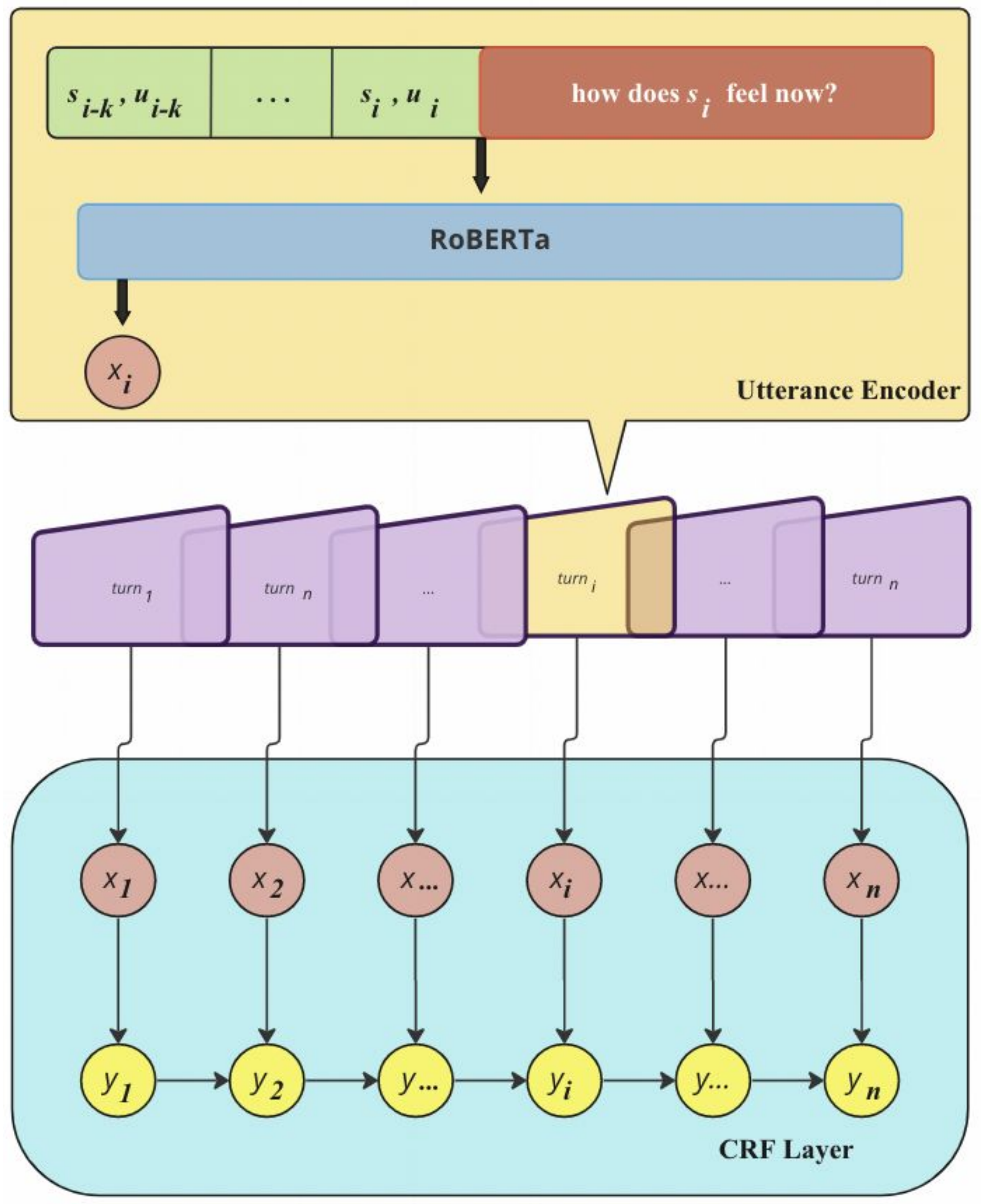
The “Emotion” columns has following labels:

- neutral
- joy
- surprise
- sadness
- anger
- disgust
- fear


"We could totally do that!"
"So you guys'll stay here and hang out with me?"

Model Design

How is the model composed?



The preprocessed input texts are changed into QA type query with ‘<speaker> now feels <mask>’. The pretrained RoBERTa encodes the input and extracts the utterance-level sentiments.
Furthermore:

- **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

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

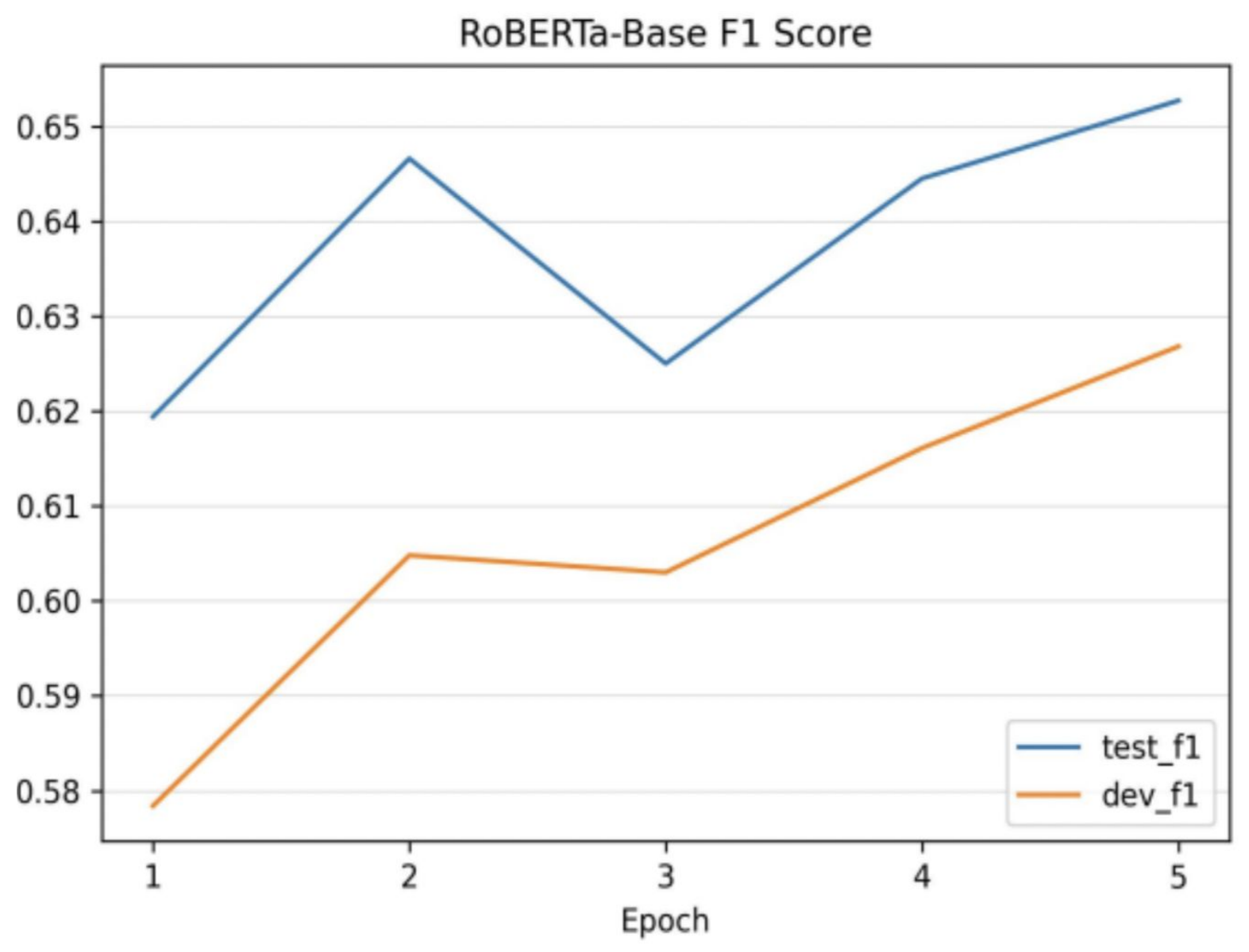
- **CRF layer** is optimized by maximizing the probability of the ground truth emotion sequence **P,** over all emotion sequence

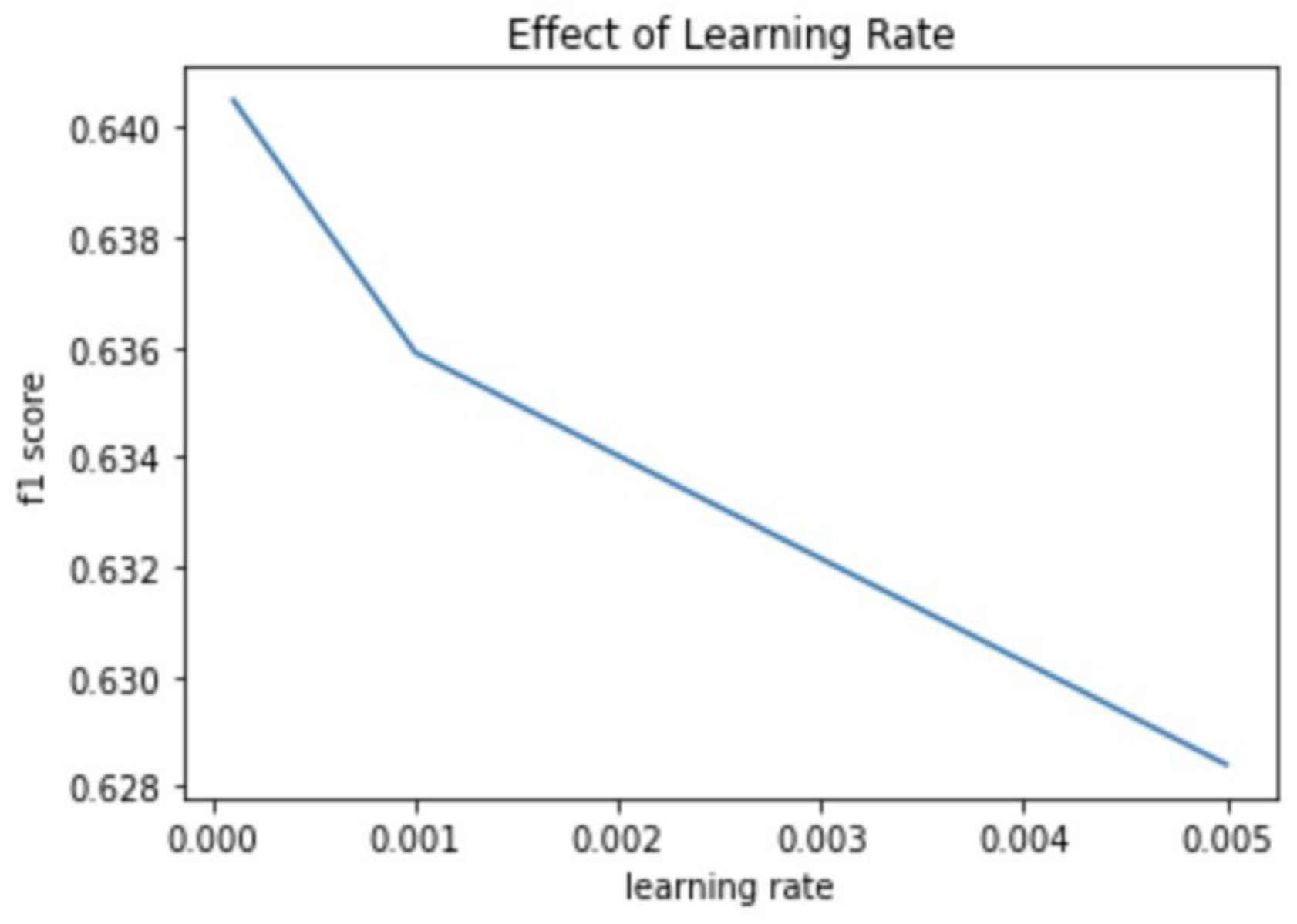
$$\mathrm{P} = \frac{1}{Z(\boldsymbol{x})} \exp\left(p_{1,y_1} + \sum_{t=2}^{N}\left[g\left(y_{t-1}, y_t\right) + p_{t,y_t}\right]\right)$$

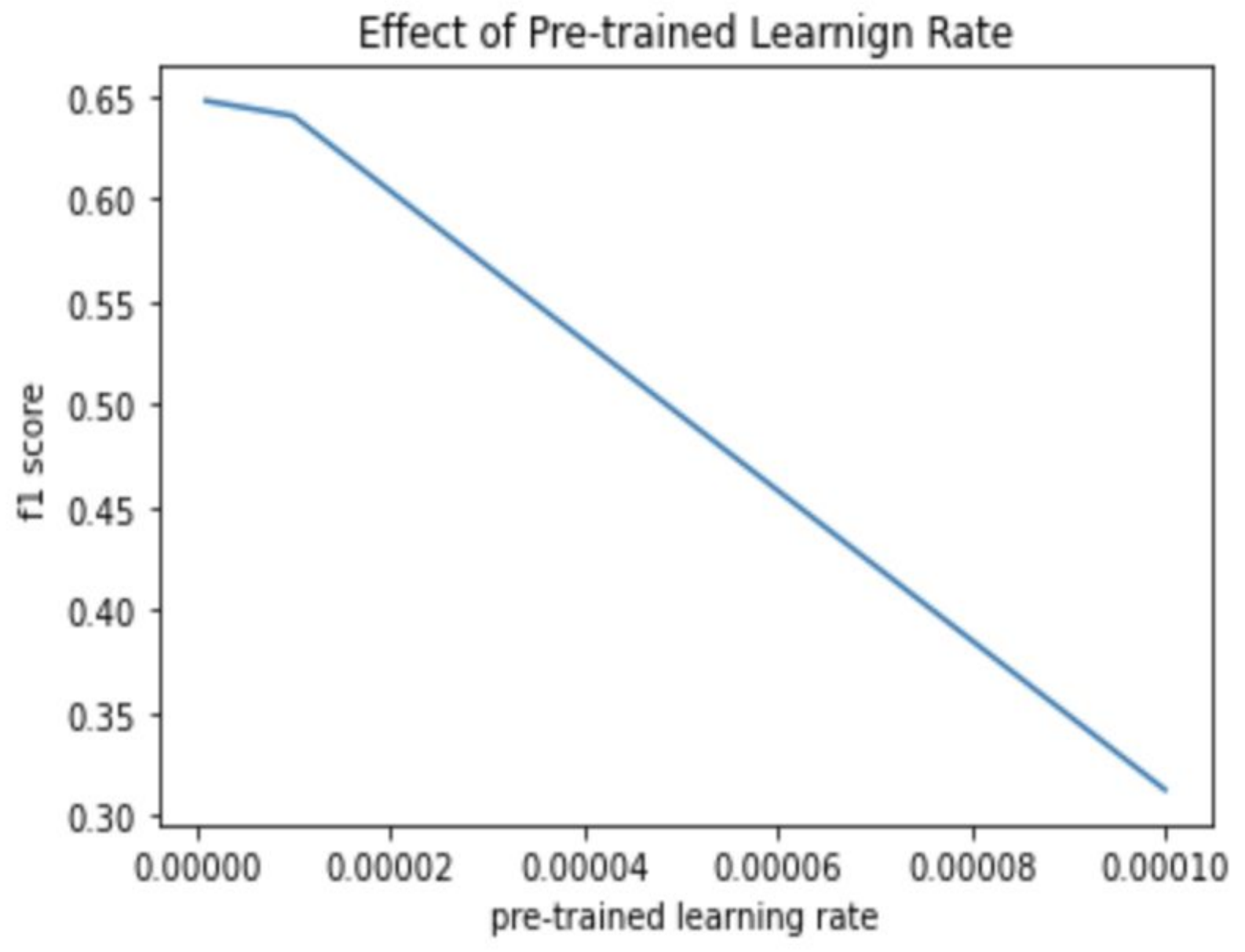
$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$$

For prediction, we decode using the paper’s version of Viterbi algorithm over **CRF layer** to find the best emotion sequence

Experiments & Results







We experimented the model with several parameters including epoch, learning rate, and pre-trained learning rate. For the epoch, as the epoch increases the f1 score also increases as we can see in the first graph.
With regard to learning rates, learning rate is for model CRF, and pre-trained learning rate is for RoBERTa in our model. For both of learning rate and pre-trained learning rate, the f1 score becomes higher when the learning rates are lower. Especially for the pre-trained learning rate, when the values is 1e-4, the f1 score was around 0.31 and from 1e-5 the f1 score goes around 0.64.

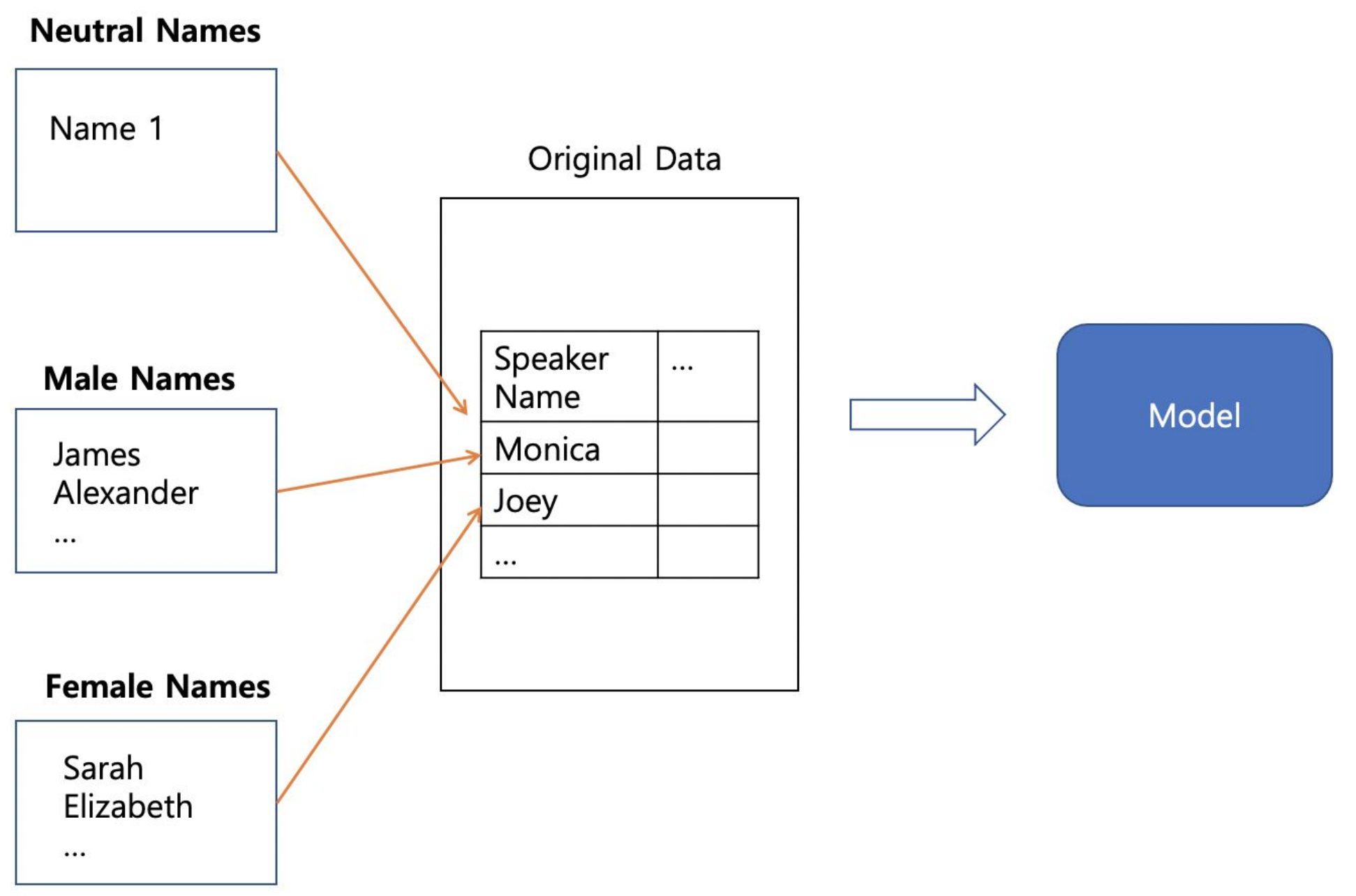
Extension

As an extension, we figure out whether the model has gender bias or not. The model utilizes the speakers’ names for training, meaning the name of the speaker can be considered for the model to decide the emotion (result).

We will make 3 groups of speaker names: neutral names, male names, and female names. The differences between these names and the original speaker names are gender and we are controlling race. Using these groups of names, we will replace the original speaker names and run the model.

To assessment the model, we will use a confusion matrix. Based on the confusion matrix, we can decide whether the model does or does not contain bias towards gender.

How do we find gender bias?



Conclusion

Through the process and implementation of our supervised model, we have come across resolutions and recommendations on how we can better adapt to the fast-paced realm of the web. After analyzing the data collection, data wrangling, and feature selection steps, we have observed ways to further improve upon and extend the scope of the project.

- **Unforeseen Power-Transfer & Updates to Twitter’s platform:**
  - Rise in concerns of the stability of a data provider (concept drift)
  - Brings our attention to the importance of building durable, future-proof Machine Learning (ML) models
- **Language (non-English):**
  - Domains are not solely from the U.S, since they may also stem from other countries
  - Improve our ML models by training on non-English content
- **Issues:**
  - Data encompasses primarily news-sites and social media platforms
  - ML models should further be applied to investigate the spread of content through multimodality, non-text features (Videos, Images, Apps etc.)

References

1. Poria, S. et al. (2018) Meld: A multi-modal multi-party dataset for emotion recognition in conversations, arXiv.org. Available at: https://arxiv.org/abs/1810.02508v4.
2. Song, X. et al. (no date) Emotionflow: Capture the dialogue level emotion transitions, IEEE Xplore. Available at: {https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9746464&tag=1