Project for the course ”Text Mining and Sentiment Analysis” Happy or Not? Emotion Analysis of “Friends” based on Language Models

October 26, 2023

1 Introduction

The primary aim of this project was to investigate the capabilities of Language Models (LMs) in the realm of emotion detection within TV shows, with a specific focus on the widely popular TV sitcom ”Friends” as a case study. To accomplish this objective, a publicly available dataset was employed that contained emotion-annotated transcripts of dialogues spanning four seasons of the series. We then created two additional versions of this dataset, by extracting emotion labels using two distinct LMs, namely *DistilBERT-base-uncased* (DistilBERT for short) and *Emotion English DistilRoBERTa-base* (DistilRoBERTa for short). Throughout the course of this project, three principal tasks were undertaken for each of the three datasets (the two generated by the LMs and the original emotion-annotated textual corpus):

1. Creating an “emotional profile” for each of the main characters;
2. Investigating how the emotions of each character evolves over the course of the four seasons;
3. Analyzing the emotional influence of diverse characters on each other;

The findings offer intriguing insights into the emotional behavior of Friends’ characters. In addition, they allow for a direct comparison of the emotional extraction and classification capabilities of the two LMs, also against the baseline of the ”true emotions” represented by the emotional labels in the source dataset. This comparative analysis therefore contributes to addressing a challenging question: which of the two LMs under consideration (DistillBERT and DistilRoBERTa) is more accurate in extracting emotions and can potentially offer more reliable results when applied to text-based conversational data? The following sections will describe the general methodology and discuss the key findings.

2 Methodology: Data and Models

2.1 “Friends” dataset

”Friends” is a popular American TV sit com that aired during the 1990s. The choice of a sit com, among the various genres of TV shows, is motivated by several factors. Sitcoms are primarily characterized by use of straightforward and direct dialogues facilitating the classification of emotions to specific characters within the dialogue. In contrast, this task can become more intricate when dealing with more elaborate films, where several external factors may introduce confounding variables. For instance, in such films, external elements like music and set design might influence the portrayal of characters’ emotions, potentially rendering the process of emotion recognition more complex and nuanced.

The dataset used in this project was retrieved from ConvoKit [1]. From the extensive Friends corpus (which encompasses the labelled transcripts for all 10 seasons of the series), the project considered all the conversations that occurred throughout the first four seasons, comprising 97 episodes, 897 scenes, 12,606 utterances, and 263 characters (speakers). The original dataset is partitioned into three

[1] https://convokit.cornell.edu/documentation/friends.html

segments: training, validation, and testing. However, these segments were combined into a single dataset because the LMs employed in this project are pre-trained and do not require such segmentation. Every sentences in the original dataset are assigned an emotion classification, including the following categories: mad, joyful, peaceful, powerful, sad, scared, and neutral. The presence of these emotion labels within the dataset proved to be essential for our comparative analysis of the emotion detection capabilities of the LMs under consideration in this project.

2.2 LLMs and Emotion Detection

Different approaches and models exist for recognizing emotions from text. These approaches fall into two categories: lexicon-based approaches and machine learning models. Lexicon-based approaches rely on dictionaries of words associated with emotions and sentiments that rely on emotion-tagged word dictionaries to analyze text. They count emotion-associated words to determine overall sentiment. These methods are simple but may miss contextual nuances. While machine learning models, including deep learning, can be trained recognize emotional patterns in text. They capture context and nuances, enabling more accurate emotional analysis.

In emotion detection tasks, machine learning models are typically favored over lexicon-based approaches (Canales and Mart´ınez-Barco 2014). More recently, transformer models, especially BERT-based LMs, have gained significant attention in the field of emotion detection (Acheampong, Nunoo-Mensah, and Chen 2021). For this project, I employed two pre-trained variants of BERT (Bidirectional Encoder Representations from Transformers) model: "DistilBERT-base-uncased" [2] and "Emotion English DistilRoBERTa-base" [3],both sourced from Hugging Face.

DistilBERT-base-uncased (referred to as *DistilBERT*, or *Model 1* for short) is a compressed version of BERT model. DistilBERT upholds strong performance in natural language tasks while delivering a smaller model size and accelerated computational speed when compared to the original BERT model. It was fine-tuned using the Emotion Dataset, constructed from Twitter data, encompassing tweets that express six distinct categories of emotions: anger, joy, surprise, love, sadness, and fear.

Emotion English DistilRoBERTa-base (*DistilRoBERTa* or *Model 2* for short) is a fine-tuned checkpoint of the DistilRoBERTa-base model (where RoBERTa stands for "Robustly optimized BERT pretraining approach) that was trained on six diverse datasets such that texts from Twitter, Reddit, student self-reports, and utterances from TV dialogues. This Language Model encompasses emotions that align with the "basic" emotional states as proposed by Paul Ekman, including anger, joy, surprise, disgust, sadness, and fear, along with a category for "neutral" emotions. Paul Ekman's theory posits that these emotions are universally recognized and experienced across various cultures, representing innate human emotional responses.

The table below delineates the categorical models of emotions employed in each of the three datasets (DSs) analysed in the following sections. These datasets include the above mentioned subset of the emotion-annotated textual corpus of "Friends," which is referred to as *True Label DS*, along with the new datasets that utilize emotional features extracted by the two LMs mentioned earlier. These new datasets are respectively denoted as *DistillBERT DS* and *DistilRoBERTa DS*.

True Label DS	Distilbert DS	DistilRoBERTa DS
mad	anger	anger
joyful	joy	joy
sad	sadness	sadness
peaceful	surprise	surprise
powerful	love	disgust
scared	fear	fear
neutral		neutral

Table 1: Label comparison

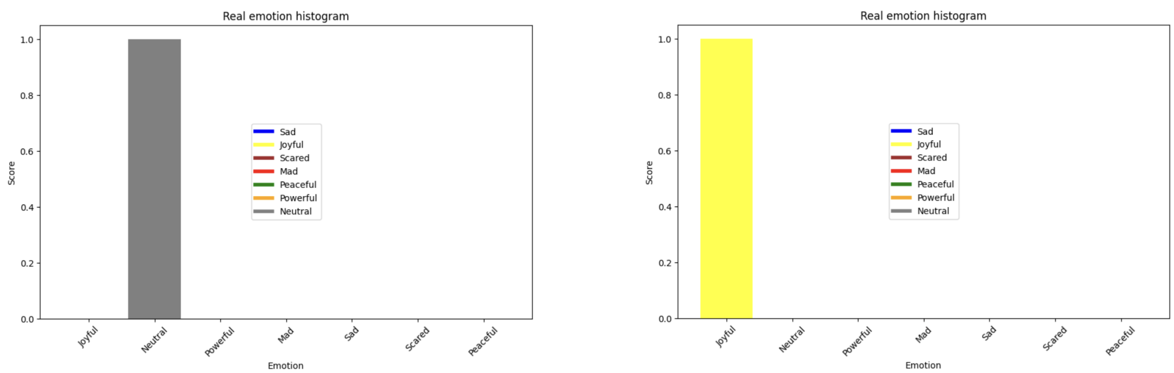
[2] https://huggingface.co/bhadresh-savani/distilbert-base-uncased-emotion?text=I+hate+football
[3] https://huggingface.co/j-hartmann/emotion-english-distilroberta-base?text=I+hate+football

3 Preliminary experiment

As a preliminary experiment to compare the emotional classification capability of the two models, mutually and against the True Label DS, we compared their outcome for two sentences: “Sorry I’m late, I was stuck at work. There was this big dinosaur.. thing.. anyway” (Sentence 1) and “”Oh my God! Ohh! Look at this one! It’s so beautiful!” (Sentence 2). For every sentence, the two LMs provided the percentage of each emotion present, while the True Label DS indicated a single emotion label only.

As evident from Fig. 1a, for Sentence 1 the True Label dataset provided Neutral as the only emotion. The DistilBERT results for the same sentence, shown in Fig. 2b, indicated that the main emotion was Anger (70%) followed by Sadness (approx. 30%). Still, it is worth noticing that DistilBERT does not include a “neutral” emotion in its classification scheme. For DistillRoBerta , Fig. 3a shows that the main emotion in Sentence 1 was Sadness (40%) followed by Neutral (approx.35%) and Surprise” (15%). The findings suggest that for Sentence 1 DistilRoBERTa model was more consistent with the emotion classification indicated in the true label DS than the DistilBERT model.

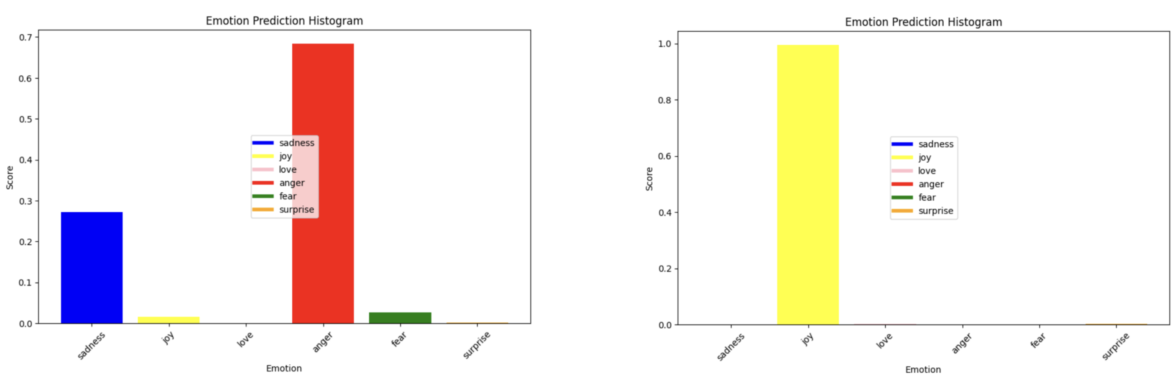
In Sentence 2, the True Label dataset (Fig. 1b) provided Joyful as the only emotion, and also DistilBERT extracted the same result, with Joy at 100% (Fig. 2b.) In contrast, Fig. 3b shows that for DistilRoBERTa the main emotion was Surprise (80%) followed by Joy (1%). Therefore, the findings suggest that for Sentence 2 the DistilBERT model was more consistent with the classification in the True Label dataset than DistilRoBERTa model.



(a) True Label- Sentence 1

(b) True Label - Sentence 2

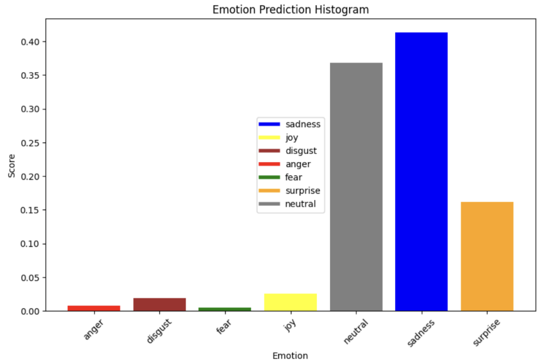
Figure 1



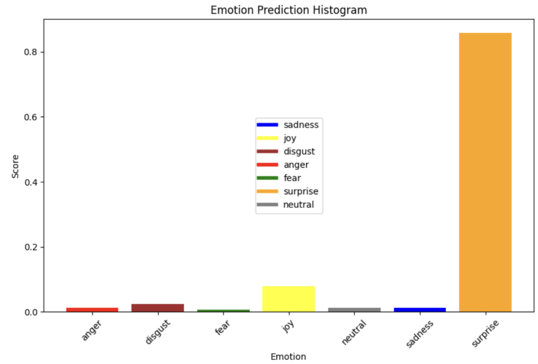
(a) DistilBERT - Sentence 1

(b) DistilBERT - Sentence 2

Figure 2



(a) DistilRoBERTa - Sentence 1



(b) DistilRoBERTa- Sentence 2

Figure 3

4 Main characters' emotional profiles

Although the LMs calculated the percentage of *each* emotion expressed in a sentence, to determine the distribution of emotions in the resulting datasets and the emotional profile of a character, only the *predominant emotion* was considered for each sentence, i.e., the emotion with the highest percentage in that sentence. The distribution of an emotion was calculated as the number of occurrences in all sentences of the dataset where such emotion was the predominant one.

The emotional profile of a character was then defined by a set of values, where each value indicates the total occurrences of an emotion as the predominant emotion, normalized relative to the number of sentences spoken by the character.

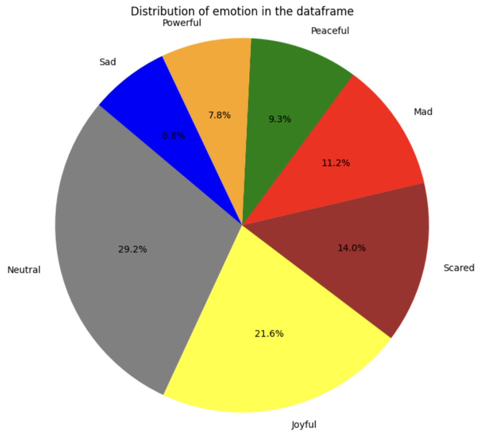
The following figures allow us to compare the overall distribution of emotions against the emotional profiles of the six main characters (e.g., Rachel Green, Monica Geller, Phoebe Buffay, Joey Tribbiani, Chandler Bing, and Ross Geller) as derived from the three datasets.

Regarding the overall distribution, the emotions resulting from the True Label DS indicated a more balanced distribution among emotions, with a prevalence of Neutral (29%), followed by Joy (22%), and Scared (14%) as shown in Fig.4a. Model 1 exhibited a predominance of Joy (48%), followed by Anger (38%) - see Fig. 5a. Model 2 suggested a prevalence of Neutral (39%), followed by Surprise (24%), Anger (11%), and Joy (10%) - see Fig. 6a.

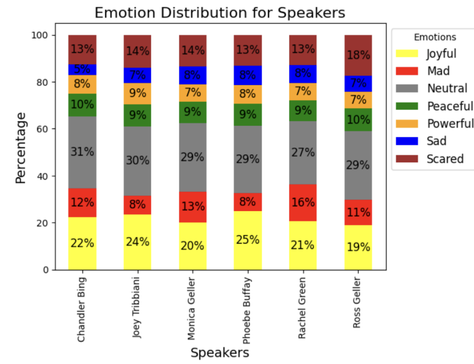
The three figures also allow us to compare the profile of each character as it emerged from each dataset. Looking at each single figure it is also possible to compare the overall distribution of emotions with that of each individual character, and observe that the results within the same approach were consistent. In addition, this comparative analysis of the distributions is useful to understand why each character predominantly exhibited one specific emotion, depending on the dataset. For example, using the True Label DS, the findings indicated that Phoebe had a percentage of Joy that is not only the highest one among the six main characters, but also higher than the percentage of Joy calculated for the entire dataset. A similar phenomenon emerged from the emotional features extracted using the two LMs (Figg. 4b, 5b, 6b) As an additional example, using Model 1 (DistilBERT), Monica resulted to be the one with the highest level of anger, while using Model 2 (DistilRoBERTa) she was the one with the lowest level of anger. The opposite occurred for Chandler: using Model 2 he was the one with the highest level of anger while using Model 1 he was the third.

The results also indicate that there was a no predominant emotion that distinctly characterized the individuals based on their gender. This suggests that gender might not be a significant factor in determining the characteristics of an emotional profile.

Finally, the status of "being a couple eventually," which means experiencing romantic feelings at some point during the series (as occurred with characters like Ross and Rachel, or Monica and Chandler), does not seem to imply the presence of specific similarities in the emotional profiles of the romantic partners.

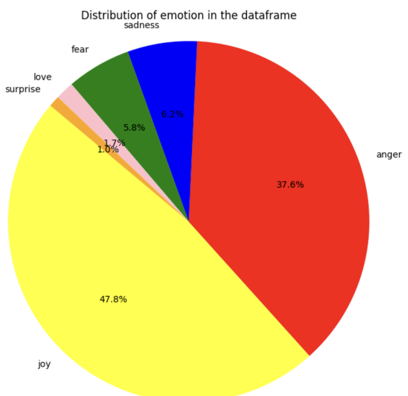


(a) Overall distribution of emotions - True Label

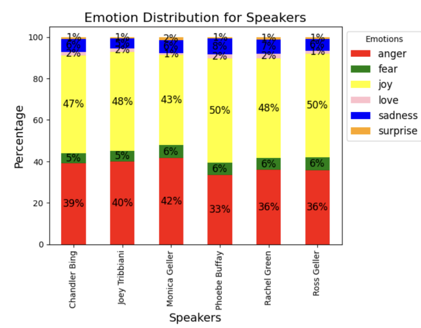


(b) Characters's profile - True Label

Figure 4

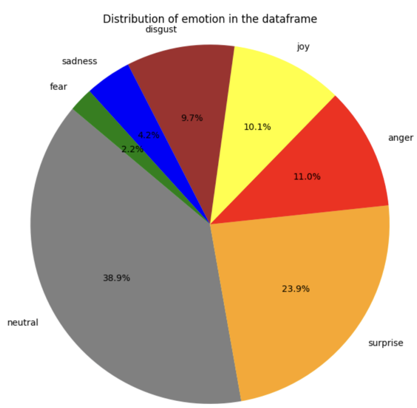


(a) Overall distribution of emotions - DistilBERT

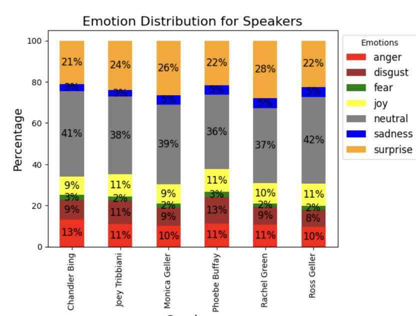


(b) Characters' profile - DistilBERT

Figure 5



(a) Overall distribution of emotions - DistilRoBERTa

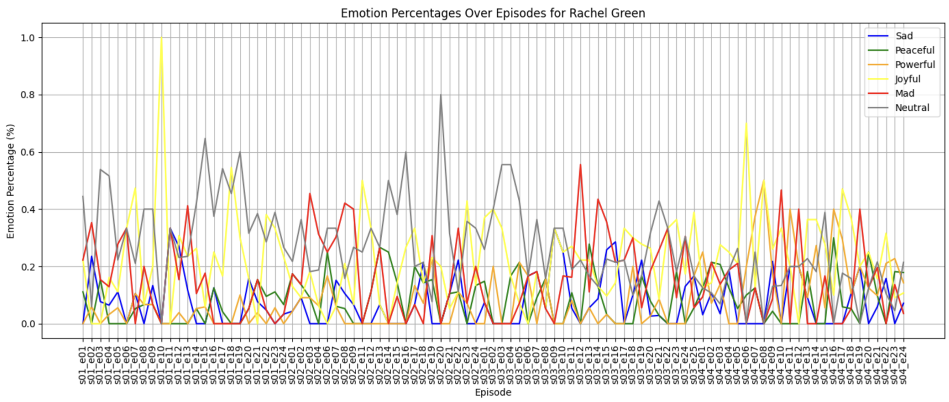


(b) Characters's profile - DistilRoBERTa

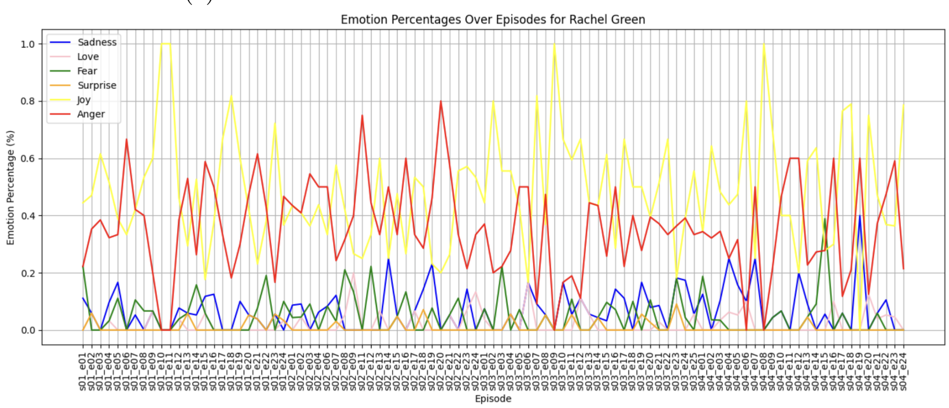
Figure 6

5 Characters' emotional evolution

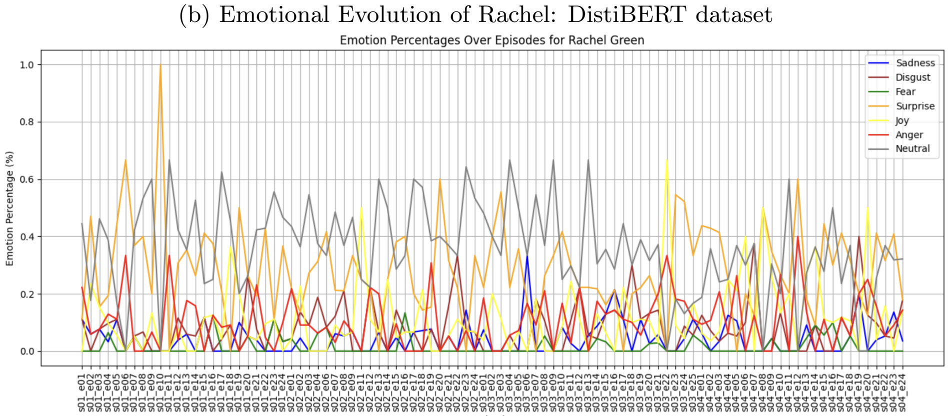
The analysis covered in the previous section does not provide us with insights into potential changes in the characters' emotional states as the comedic storyline progresses. In order to delve deeper into this matter, time series plots were generated to elucidate the "emotional evolution" of each character across all scenes in each episode, spanning the four seasons of the series. As an example, Fig. 7 illustrates Rachel's emotional evolution across all scenes, as it emerged in each dataset.



(a) Emotional Evolution of Rachel: True label dataset



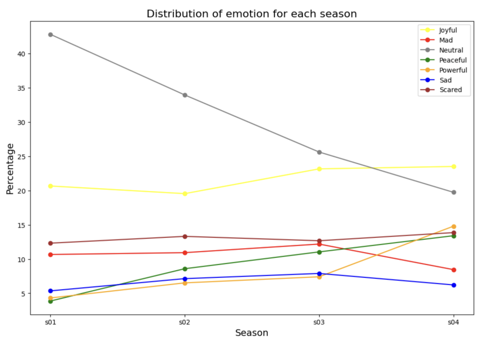
(b) Emotional Evolution of Rachel: DistiBERT dataset



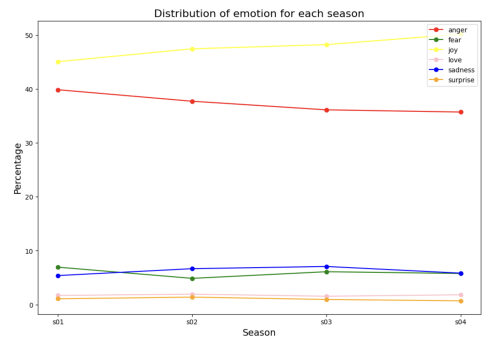
(c) Emotional Evolution of Rachel: DistilRoBERTa dataset

Figure 7

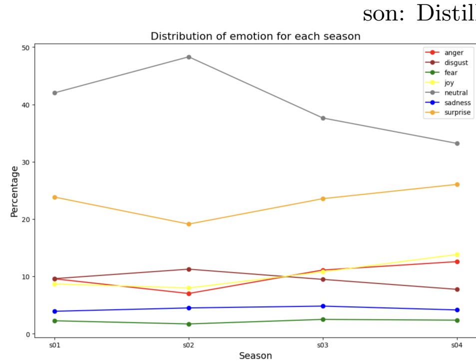
To effectively analyze the above plots and determine whether the character's emotional profile undergoes any significant changes, it is useful to consider the results also in light of additional data illustrating how the distribution of emotions across the *entire dataset* evolved over time *season by season* (Fig. 8). For instance, the plot depicted in Fig. 7a reveals a decline in the prevalence of the "neutrality" emotion in Rachel during seasons 3 and 4, with other emotions taking precedence. As evidenced in Fig. 8a, this change in emotional distribution is consistently observed across the entire dataset. Through this temporal analysis, it becomes apparent that, irrespective of the dataset under examination, there are no discernible changes in Rachel's emotional behavior. This observation suggests the existence of a fundamental consistency and stability in the character's underlying emotional state.



(a) Emotional Evolution of Rachel by season: True Label dataset



(b) Emotional Evolution of Rachel by season: DistilBERT dataset

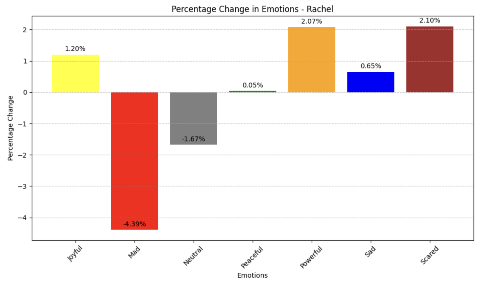


(c) Emotional Evolution of Rachel by season: DistilRoBERTa dataset

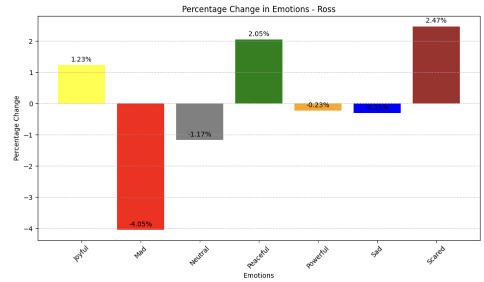
Figure 8

6 How different characters have impact on each other

To explore the emotional dynamics between various characters, and how the "social presence" of other characters affects one individual's emotional state, a comparative analysis was performed concerning the emotional expressions involving pairs of main characters in two separate scenarios: scenes that included both characters and scenes where only one character was present while the other one was absent. A character was considered as "present" in a scene if he or she spoke at least once during that scene. The primary emphasis in this analysis was placed on the characters Rachel and Ross, owing to their iconic love story and Ross's enduring infatuation with Rachel, which began at the outset of the series. The overall percentage of each emotion was calculated in all the scenes in which Rachel (resp. Ross) was present but Ross (resp. Rachel) was not, and in all the scenes in which they were both present. The difference between the different scenarios was then analysed. The results using the True Label dataset indicated that both Rachel and Ross exhibites less anger when the other was not present (Fig.9) In contrast, the data derived from DistilBERT suggest that Rachel, when without Ross, tended to be more angry, while Ross, when without Rachel, appeared to be less angry (Fig. 10). The results from DistilRoBERTa (Fig. 11)indicate that Rachel was less angry when without Ross, consistently with the findings from the true label dataset). However, the values emerging from the same model also show that Ross was more angry when Rachel was not present, which is not aligned with the results from the True Label dataset.

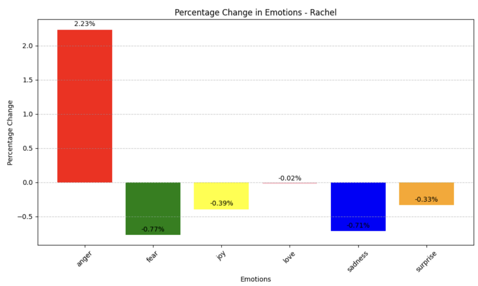


(a) Impact of Ross on Rachel - True Label

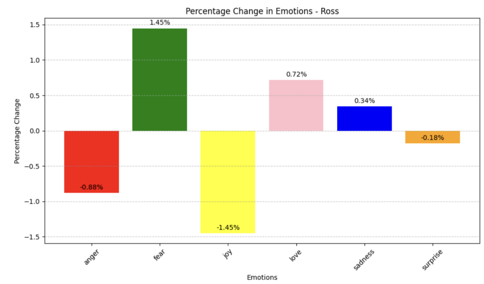


(b) Impact of Rachel on Ross - True Label

Figure 9

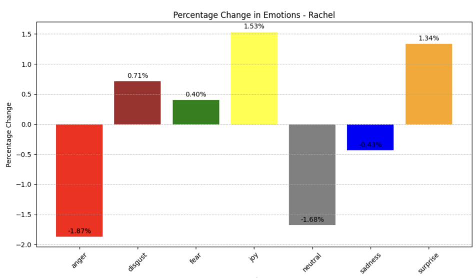


(a) Impact of Ross on Rachel: DistiBERT

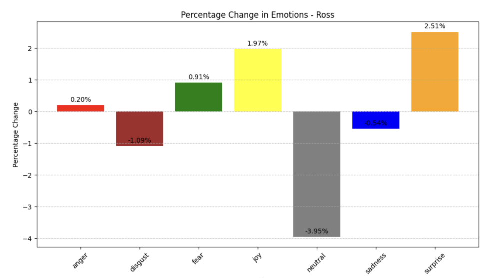


(b) Impact of Rachel on Ross - DistiBERT

Figure 10



(a) Impact of Ross on Rachel - Distil-RoBERTa



(b) Impact of Rachel on Ross - Distil-RoBERTa

Figure 11

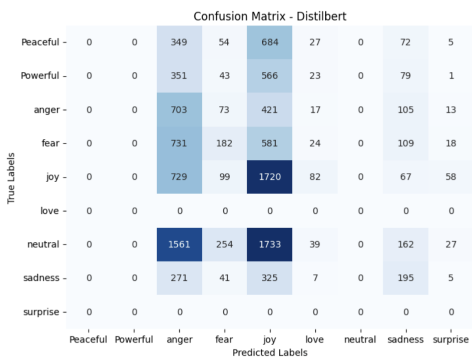
7 Models' Performance

Considering the disparities in the results from the previous sections, it is crucial to systematically investigate the performance of DistilBERT and DistilRoBERTa in terms of their respective emotional classifications aligned with the actual emotional labels provided in the source dataset. To accomplish this, we generated a confusion matrix for each model (Fig. 12). We also calculated several key etrics, including Precision, Recall, F1-score, Support, Accuracy, Macro-average, and Weighted average (Fig. 13).

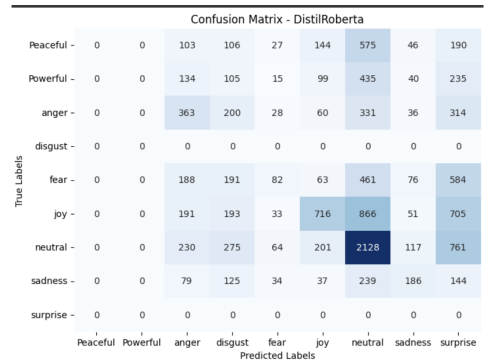
For the purpose of conducting this analysis, a modification was made to the emotional labels in the True Label dataset. Specifically, the label "mad" was changed to "anger" since these two emotions are closely related and can be treated as equivalent for the scope of our investigation.

From the Confusion Matrix reported in Figure 12a and considering what is reported in 1, Anger, Joy, and Sad are the emotions that the True Label dataset has in common with the DistilBERT dataset. It is noticeable that Joy is the most correctly predicted label. This can be coherent with the fact that it is one of the most frequent emotions in both datasets (Fig. 4a - 5a). At the same time, it's also apparent that almost a third of the utterances labeled as Joy in the True Label dataset are predicted by the DistilBERT model as anger. Another discrepancy is the prediction of Sadness, which has been predicted more frequently as Joy and Anger than as Sadness. Additionally, it's interesting to analyze the prediction of the emotion that the two datasets don't have in common. For instance, Neutral is predicted more frequently as Anger and Joy, as Peaceful and Powerful.

On the other hand, considering what's reported in 12b, Neutral is the most correctly predicted label, while Joy is more frequently predicted as Neutral. At the same time, Fear is more frequently predicted as Neutral, Surprise, Anger, and Disgust, and then Fear. Analyzing the results of the two sets of metrics, it can be noticed that in the second table, some Precision and Recall values have slightly improved compared to the first table. For example, for the labels "anger," "fear," and "joy," there has been an increase in precision and recall, indicating an improvement in the models' ability to correctly identify instances of these classes. The overall accuracy of the first model, 22%, indicates that Model 1 had more difficulty correctly labeling most instances compared to the second model with an overall accuracy of 28%.



(a) Confusion Matrix - DistilBERT model



(b) Confusion Matrix - DistilRoBERTa model

Figure 12

	precision	recall	f1-score	support
Peaceful	0.00	0.00	0.00	1191
Powerful	0.00	0.00	0.00	1063
anger	0.15	0.53	0.23	1332
fear	0.24	0.11	0.15	1645
joy	0.29	0.62	0.39	2755
love	0.00	0.00	0.00	0
neutral	0.00	0.00	0.00	3776
sadness	0.25	0.23	0.24	844
surprise	0.00	0.00	0.00	0
accuracy			0.22	12606
macro avg	0.10	0.17	0.11	12606
weighted avg	0.13	0.22	0.15	12606

(a) Various Metrics - DistilBERT model

	precision	recall	f1-score	support
Peaceful	0.00	0.00	0.00	1191
Powerful	0.00	0.00	0.00	1063
anger	0.28	0.27	0.28	1332
disgust	0.00	0.00	0.00	0
fear	0.29	0.05	0.09	1645
joy	0.54	0.26	0.35	2755
neutral	0.42	0.56	0.48	3776
sadness	0.34	0.22	0.27	844
surprise	0.00	0.00	0.00	0
accuracy			0.28	12606
macro avg	0.21	0.15	0.16	12606
weighted avg	0.34	0.28	0.28	12606

(b) Various Metrics - DistilRoBERTa model

Figure 13

8 Conclusions and final remarks

This project may contribute to advance our understanding of emotion detection within TV show transcripts using Language Models. We condider two specific LMs (DistilBERT-base-uncased and Emotion English DistilRoBERTa-base) and used "Friends" as case study. Still, the code created have the potential to be readily adapted for application to other sitcoms, and the methodology employed could be exploited to explore the use of other LMs.

The analysis performed in this project provides insights into the emotional behavior of Friends characters and its progression throughout the seasons, and into the impact of "social presence" - meaning the presence of other characters in the scene - on the emotional state of individual characters.

It is important to note that the findings derived from the use of the two LMs and from the emotion-annotated dataset highlight some similarities but also exhibited disparities. For example, from the comparative analysis of both the labeled data of the source dataset and the data emerging from the two models, the differences in the emotional profiles of all main characters are low. Furthermore, all approaches highlight that "social presence" influences the characters' emotions. However, the three approaches provide different indications about which emotions, and to what extent, are most influenced by social presence. Other notable differences are in the elicited emotional profiles of the different characters. The disparities observed in the results are not entirely unexpected, given the distinct characteristics of the LMs used, their diverse pre-training datasets, and variations in the categories of emotions employed to label sentences in the True Label dataset versus the two models.

The project has also shed a light on a challenging question: among the two LMs under consideration (DistillBERT and DistilRoBERTa), which one is more proficient in predicting the "true" emotions expressed by the characters? The divergences in the findings underscore the complexity inherent in addressing the above question. They emphasize the intricacies and challenges associated with accurately detecting and interpreting emotions within textual data extracted from TV shows, particularly when relying on LLM-based emotion detection methods. Further investigation and refinement of the emotion detection models would be necessary to improve alignment and consistency in results, particularly within the narrative-rich and context-dependent environments commonly found in sitcoms.

References

- https://aclanthology.org/W14-6905.pdf
- https://www.researchgate.net/publication/324042746_Emotion_Detection_from_Text_and_Speech_-_A_Survey
- https://www.researchgate.net/publication/348740926_Transformer_Models_for_Text-based_Emotion_Detection_A_Review_of_BERT-based_Approaches