Project Documentation: Friends Scripts Sentiment Analysis

Overview

We aim to explore the sentiments expressed in the Friends TV show scripts. Our focus involves processing the screenplay script text data and conducting an in-depth analysis using text retrieval, sentiment analysis and topic modeling.

GitHub Directory

Final submission

- **Project code**: CS410_Team_Project_Workspace.ipynb
- **Final report**: CS410 Team Project Final Report.pdf
- **Presentation recording**: https://www.youtube.com/watch?v=GOjaOA9hOOk
- **Tableau file**: Tableau.twbx
- **Combined raw** data: combinedData_raw.csv

Project proposal & progress report

- **Project proposal**: CS410 Team Project Proposal.pdf
- **Progress report**: CS410 Team Project Progress Report.pdf

Set up

The environment we used for this project is Google Colab. We imported different libraries to complete tasks of data cleaning, tokenization, and lemmatization, and conducting sentiment analysis and topic modeling analysis. We also leveraged Tableau to visualize the results from sentiment analysis, which provides more flexibility in analyzing results in various dimensions.

To run the analysis, the user can locate the Colab code and Tableau files on the GitHub pages. The user can run each step in Colab to repeat our analysis. To save time from re-running the steps of combining script data for each episode, we also uploaded the combined script data to GitHub for reference.

Implementation

In sentiment analysis, we found out the 3-level sentiments (positive, neutral, negative) and multi-level emotions (anger, disgust, fear, joy, neutral, sadness, surprise) for each line in the show.

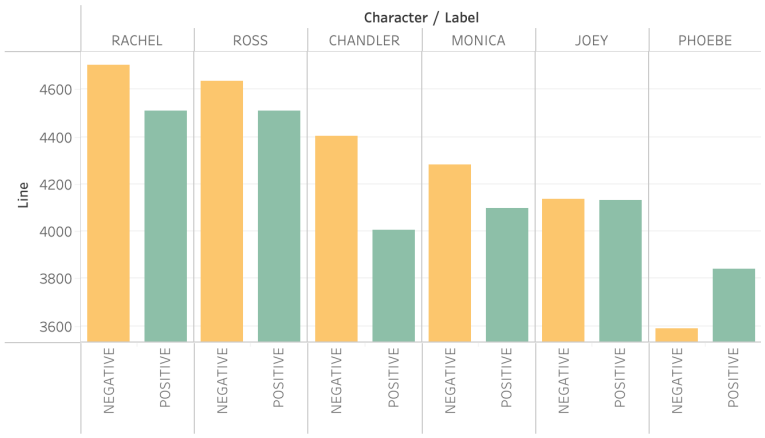
- We leveraged the transformers pipelines from Hugging Face. Transformers are deep learning models that are based on Transformer architecture and use self-attention mechanisms to process string input. Transformers have good performance in capturing the context and dependencies between words.
- For 3-level sentiment analysis, we used the default "sentiment-analysis" pipeline and "DistilBERT base uncased finetuned SST-2" model. This model is a fine-tune checkpoint of DistilBERT-base-uncased, fine-tuned on SST-2. DistilBERT is a distilled version of BERT, according to this paper (https://arxiv.org/abs/1910.01108), it is smaller, faster, cheaper and lighter than the language representation model BERT.
- For multi-level emotion analysis, we used the "text-classification" pipeline and "Emotion English DistilRoBERTa-base" model. The model is a fine-tuned checkpoint of DistilRoBERTa-base.

For topic model analysis, we used many different libraries, including Gensim for topic model, NLTK for stopwords removal, and Spacy for lemmatization. The code preprocesses the text data by removing stop words, creating bigrams, and lemmatizing dialogue data. Then, it builds a dictionary and a term frequency corpus as inputs for the Latent Dirichlet Allocation model to identify topics in all the dialogues.

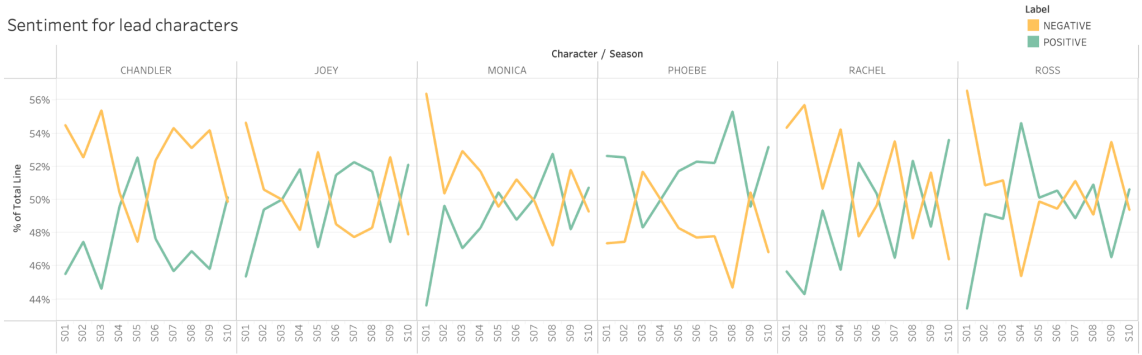
Results

For sentiment analysis, we analyzed the results in the following aspects:

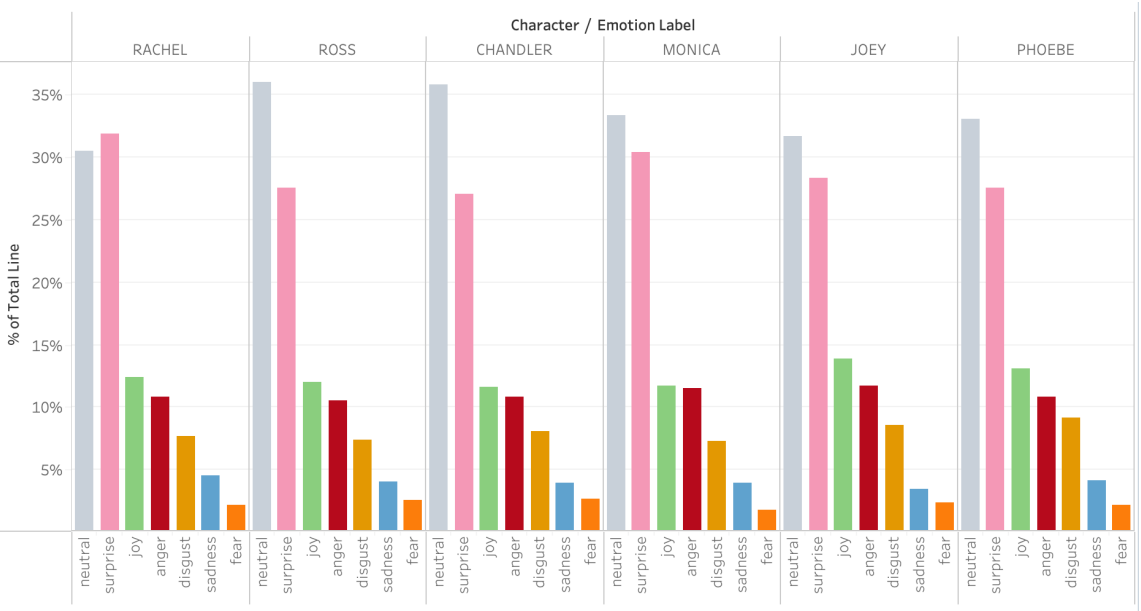
- Overall sentiment of the lead characters
    - Lead characters having the most lines (Rachael, Ross, Chandler and Monica) are more inclined towards negative sentiments than positive ones.
    - Phoebe emerged as the happiest or the most optimistic character among the lead cast.
    - As anticipated, Chandler is positioned as the least happy or the most pessimistic character among the main characters.



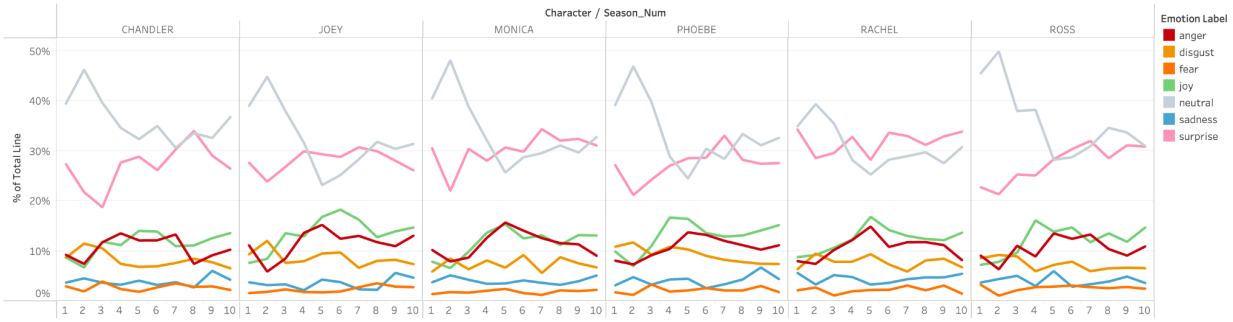
- Trend of lead characters' sentiments over time
    - Chandler appears to be happier in Season 5, coinciding with the onset of his relationship with Monica.
    - Rachael and Ross exhibit noticeably reversed sentiments starting from Season 4 (after they were on a break).
    - In Season 1, all characters began with more extreme sentiments. However, as the show progresses, the lead characters eventually find themselves in a positive or happy mood.



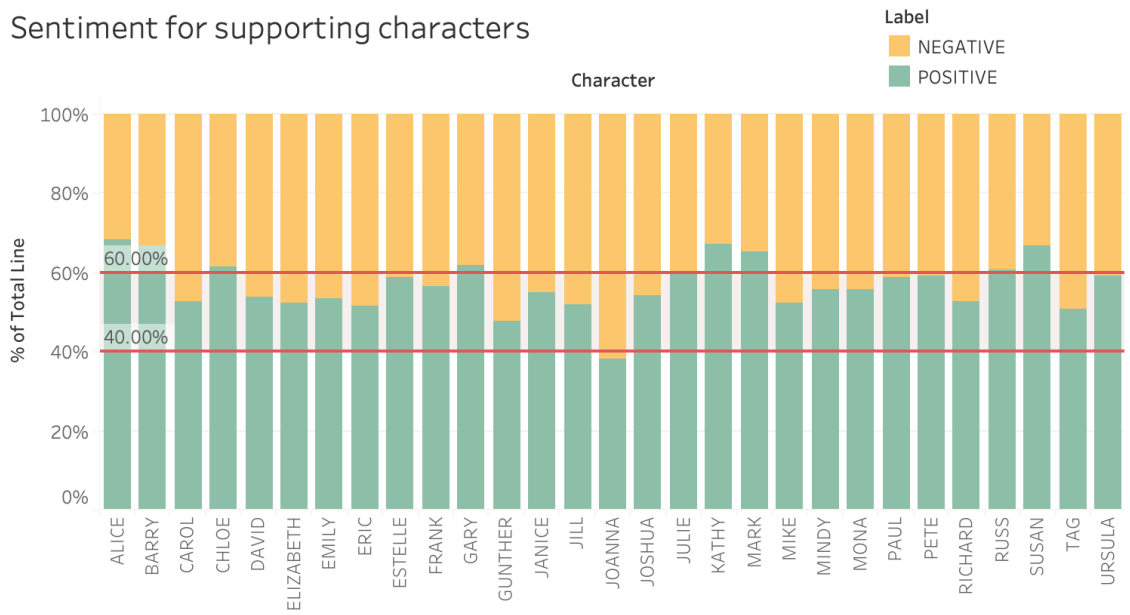
- Overall emotion of the lead characters
    - After looking at the % of lines per emotion per character, it's clear that:
        - Ross and Chandler tend to have neutral emotions more frequently than the other lead characters.
        - Rachael tends to be the most emotional among lead characters.



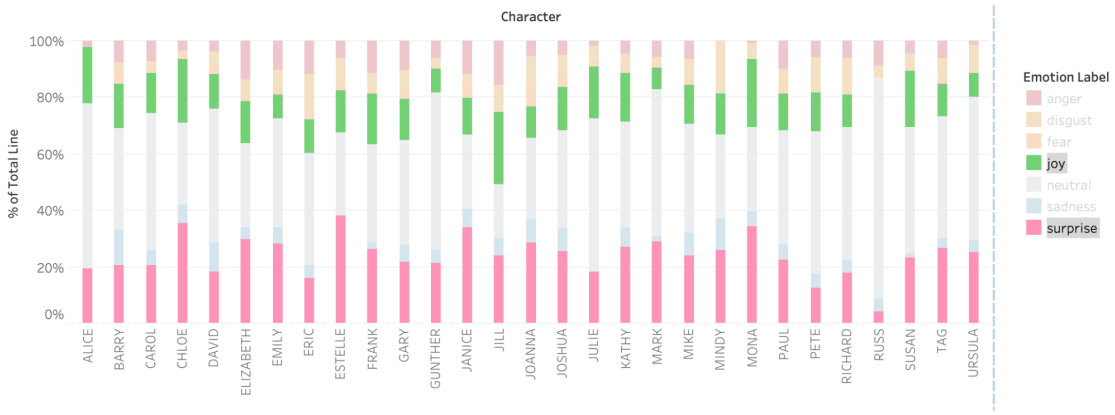
- Trend of lead characters' emotions over time
    - In Season 2, all lead characters exhibit relatively neutral emotions, suggesting a somewhat uneventful season.
    - Season 5, in contrast, proves to be intriguing. Rachel and Monica display nearly as much anger as joy, while Phoebe experiences a significant amount of joy. This season marks Phoebe's pregnancy with the triplets, and Ross is preparing for his marriage to Emily.
    - Following Chandler and Monica's marriage in Season 7, Chandler consistently experiences more joy than anger throughout the remainder of the show.



- Overall sentiment of supporting characters
    - Some supporting characters consistently exhibit a generally positive demeanor, among them are Alice, Barry, Kathy, Mark, and Susan
    - In contrast, Joanna and Gunther emerge as the least happy supporting characters.



- Overall emotion of supporting characters
    - Chloe, Elizabeth, Estelle, Janice, Jill seem to be the most emotional supporting characters.



For topic modeling analysis, we built the MDA model with 10 different topics where each topic is a combination of keywords and each keyword has a different weight. Topics are represented as the top N words with the highest probability of belonging to that particular topic. We used CV coherence score to measure the performance of the LDA model. The coherence score measures how similar these words are to each other. Our model has a coherence score of 0.31, which is not ideal. This could be due to the nature of our modeling topic, the *Friends* TV show. There are many terms with high frequency in the show because the dialogue data is colloquial, which could explain why the topic words are very similar to each other.

To visualize the results of the LDA topic model, we created a word cloud: