SENTIMENTAL TEXT ANALYSIS ON MOVIE REVIEWS

Prepared for
Prof. Rachlin
DS 3500
Northeastern University

Prepared by
Jocelyn Ju
Ceara Zhang

February 27, 2023

ABSTRACT

This Natural Language Processing (NLP) library is a reusable framework with a tailored set of tools for natural language processing tasks. Built on top of libraries such as NLTK and matplotlib, this library offers a wide range of functionalities for processing, analyzing, and modeling textual data. In this report, the library is used to compare and analyze text and sentiment of consumer reviews for the movie *Twilight.*

The library can load and register up to 10 movie reviews as text files at a time. The data from each text file is then cleaned, and statistics including most common words, sentiment score, and word count are gathered. Users can visualize this information as the library supports three graphs: a Sankey diagram displaying the 20 most common words across all given reviews and the number of occurrences, a line plot overlaid on a bar graph showing correlations between word count and sentiment scores, and pie charts showing the positive, neutral and negative scores that make up each sentiment score.

Link to repository: https://github.khoury.northeastern.edu/jcju/nlp-twilight-reviews

INTRODUCTION

NLP stands for Natural Language Processing, which is a field of study in artificial intelligence and computational linguistics that focuses on enabling computers to understand, interpret and generate human language. It involves developing algorithms and models that can analyze and process large volumes of natural language data, such as text or speech, to perform language translation, sentiment analysis, and more. NLP is an increasingly important field of study, with applications in various industries. As the demand for NLP solutions continues to grow, the interest for such libraries and tools becomes more relevant.

This NLP library is designed specifically to provide textual statistics and sentiment analysis on text files. Users are able to register, process and analyze up to 10 textual data files at a time to compare and gain insight on. For demonstration purposes, this library has been used on a movie with a wide range of opinionated reviews, *Twilight*. *Twilight* is a classic American film released in 2008, and has a very split opinion based on consumer reviews, receiving a 49% on Rotten Tomatoes, 5.3/10 on IMDb, and a 56% on Metacritic. The reviews used have been gathered from Google reviews, which received 4.4 average stars of 1757 reviews. Each review was given by an independent party with their own Google account.

FRAMEWORK

Capabilities

The framework of this library is capable of loading up to 10 text files at a time, which preprocesses the data for each file by cleaning punctuation, capitalization, extra spaces, removing stop words and other irrelevant words. Then information such as most common words, sentiment scores, word count, and list of adjectives from each text file are collected to be used later on for visualizations.

Visualizations

Three visualizations are produced with the user defined text files to help the user visualize interesting textual trends in the data. The framework produces a Sankey diagram, a subplot of pie charts, and a bar plot overlaid by a line plot.

Sankey Diagram

The first graph is a Sankey diagram that displays the k most common adjectives (default 20) found across all text files being compared, mapping each text file to the words if the file contains the word. The thickness of the line also corresponds to the number of times the word occurs in the text, with a thicker line meaning the word appears more often. This diagram helps identify any trends in common words and which texts tend to contain more or less of them.

Pie Charts

The second graph is the sentiment subplot of an array of individual pie charts that correspond to each text file. Each pie chart represents how much of the positive, negative, and neutral scores make up the total sentiment score. Sentiment scores is a scaling system that reflects the emotional depth of emotions in a piece of text. Generally a lower sentiment score has a larger area for the negative score in the chart, and a higher sentiment score has a larger area for the positive score.

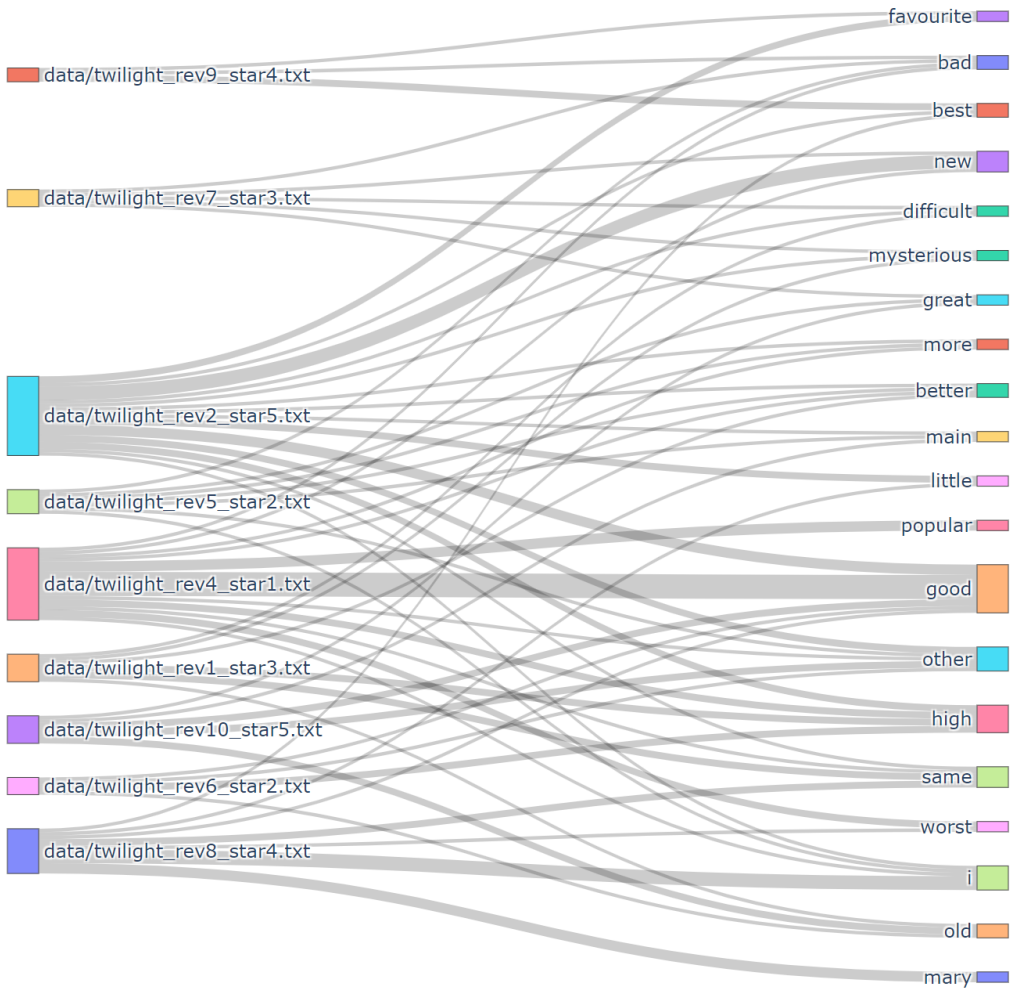
Bar Chart & Line Plot

The third visualization includes a bar chart with a line plot overlaid on top of it. The bars display the total word count for each text file, and the line indicates the sentiment score for each text file. Users can use this plot to understand if there is a correlation between the amount of words there are in a text file and the sentiment score.

CONCLUSION

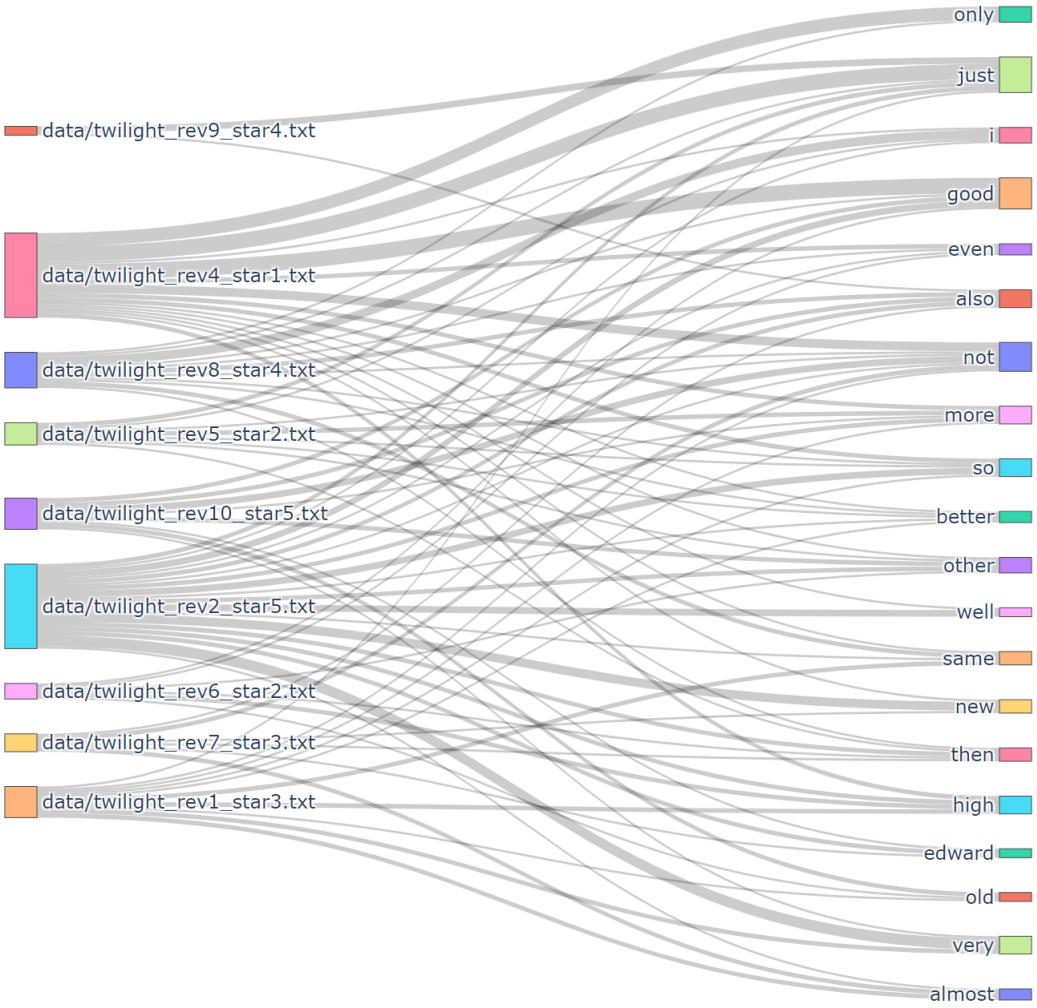
In this report, the NLP library was used to analyze sentiment of *Twilight* reviews. For demonstration purposes and to make sure the library was doing proper comparisons, the text files used have been named to include the number of actual stars the reviewer gave with their review. From the visualizations, there are many conclusions that can be drawn from the reviews.

The Sankey diagram displays some interesting trends, including a correlation between the number of branches to words and how extreme the movie review was, both positive or negative. This indicates that reviewers tend to write more content on a movie if they either strongly liked or disliked the movie. The other reviews with two, three, or four stars all tend to on average have less branches than the one or five star reviews. As to be expected, many of the common adjectives found in the reviews indicate positive and negative associations, such as “good”, “better”, “worst”, and “best.”



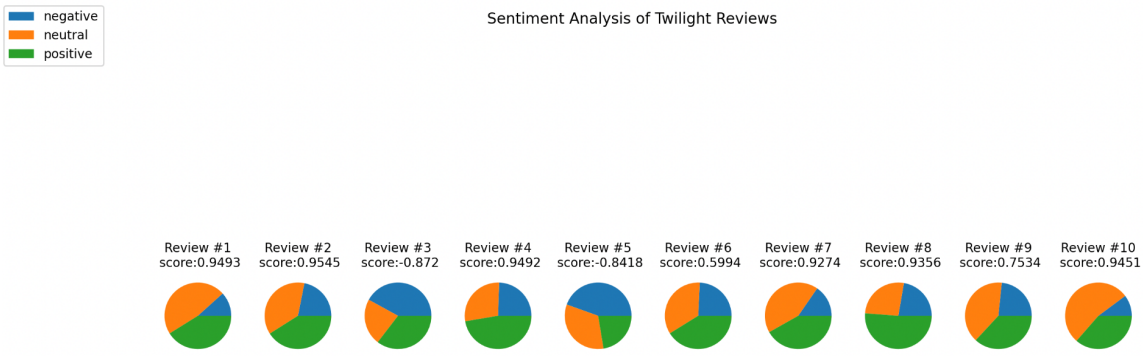
The five star review ‘data/twilight_rev2_star5.txt’ has several branches to many of the common words, and the one star review ‘data/twilight_rev4_star1.txt’ also has a lot of branches.

One of the one star reviews contained the most occurrences of the word “good” out of all the reviews. The match seems conflicting, however, if the Sankey diagram was done on adjectives and adverbs, the diagram will also reveal that the review also has a large number of occurrences of the word “not.” This finding indicates that sentiment analysis cannot only focus on adjectives as adverbs play an important role in determining the sentiment, as reviews can contain phrases such as “not good” or “not bad,” which have opposite meanings.



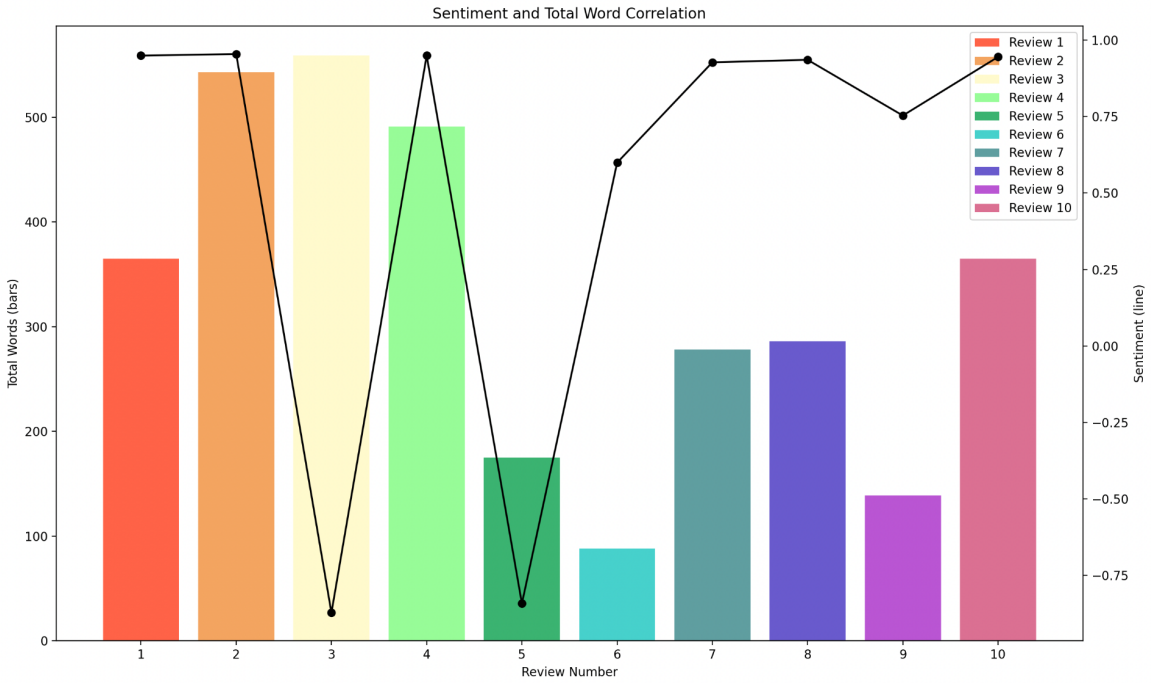
For ‘data/twilight_rev4_star1.txt, there is a thick line for ‘good’, but also ‘not.’

The pie charts of each review's sentiment scores give users a sense of what proportion of positive, neutral, and negative scores go into each sentiment score. Overall, the (sentiment) score would be negative if the negative score made up about 30% or more of the chart. Neutral scores tend to make up a lot of each pie chart.



Breakdown of sentiment scores for each review.

The bar and line plot further supports the trend that reviewers will tend to write a longer review when they have a very positive or negative opinion (sentiment) of the movie. As shown in the 2nd and 3rd bar, the reviews both had a very large word count, but one had a very low sentiment score value and other a very high value. For other reviews, there did not seem to be any other strong correlations.



There is a trend between very polar sentiment scores and high word count.

Overall, the use of this NLP library can be very extensive, as users can compare text files in any sort of context, not limited to only movie reviews, to gain more insight into the user defined textual data.

Works Cited

"Change the Legend Position in Matplotlib." *GeeksforGeeks*, GeeksforGeeks, 26 May 2021, https://www.geeksforgeeks.org/change-the-legend-position-in-matplotlib/.

"Creating Multiple Subplots Using Plt.subplots#." *Creating Multiple Subplots Using Plt.subplots - Matplotlib 3.7.0 Documentation*, https://matplotlib.org/stable/gallery/subplots_axes_and_figures/subplots_demo.html.

"Get Value from Dictionary by Key with GET() in Python." *Get Value from Dictionary by Key with Get() in Python*, https://note.nkmk.me/en/python-dict-get/#:~:text=In%20Python%2C%20you%20can%20get,key%20like.

Google Translate, Google, https://translate.google.ca/.

"How to Change from Lowercase to Uppercase in Excel: 13 Steps." *WikiHow*, WikiHow, https://www.wikihow.com/Change-from-Lowercase-to-Uppercase-in-Excel.

"How to Plot Collections Counter Histogram Using Matplotlib." *Tutorials Point*, https://www.tutorialspoint.com/how-to-plot-collections-counter-histogram-using-matplotlib.

newbie1newbie1 13922 gold badges22 silver badges55 bronze badges, et al. "Plotting a Pie Chart out of a Dictionary." *Stack Overflow*, 1 May 1967, https://stackoverflow.com/questions/62593913/plotting-a-pie-chart-out-of-a-dictionary.

OParker, et al. "Trying to Plot a Line Plot on a Bar Plot Using Matplotlib." *Stack Overflow*, 1 Aug. 1962, https://stackoverflow.com/questions/32474434/trying-to-plot-a-line-plot-on-a-bar-plot-using-matplotlib.

"Python String Lower()." *Programiz*, https://www.programiz.com/python-programming/methods/string/lower#:~:text=The%20lower()%20met.

"Tight Layout Guide#." *Tight Layout Guide - Matplotlib 3.7.0 Documentation*, https://matplotlib.org/stable/tutorials/intermediate/tight_layout_guide.html.

Tom, et al. "What Is Sentiment Analysis? an Ultimate Guide for 2023: Brand24." *Brand24 Blog*, 14 Feb. 2023, https://brand24.com/blog/sentiment-analysis/.

"Unable to Add Group as Favorite in Outlook PWA." *TECHCOMMUNITY.MICROSOFT.COM*, 10 June 2021, https://techcommunity.microsoft.com/t5/microsoft-365-groups/unable-to-add-group-as-favorite-in-outlook-pwa/m-p/1166312.