Math 156: Final Project Proposal

Students:
Mansa Krishna (UID: 305085380)
Advika Ruia (UID: 605130258)
Andrew Chao (UID: 805109717)

Final Project Outline

Sentiment analysis is a natural language processing (NLP) technique that enables us to determine the sentiment behind a piece of writing or text. Typically, such models extract meaning from textual data and assign a numerical score to it, allowing us to classify the piece of writing as positive, negative, or neutral. In the modern age, sentiment analysis is used by a wide variety of businesses to help them understand the overall sentiment of their customers through reviews on online platforms, thus allowing businesses to make accurate and quick decisions catered towards customer sentiment.

One such industry that benefits from sentiment analysis would be the on-demand video streaming service industry. Streaming service platforms such as Netflix or Disney+ would be better able to recommend filmography to their users if they are acquainted with the overall sentiments of their users. For instance, if viewers enjoy film A (positive sentiment), it follows that they might enjoy a similar film B. Moreover, through sentiment analysis of various filmography, streaming services would also know which on-demand videos audiences would want to see and stream only those.

As users of various streaming services, we were interested in conducting a sentiment analysis of various films that would enable us to recommend/choose films to stream! Through our final project, we plan on creating a sentiment analysis model that will allow us to detect polarity in a piece of writing (i.e., text classification problem). More specifically, we will aim to classify movie reviews as positive, negative, or neutral.

Links: (Additional links may be added)

Data Sets:

https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews
http://ai.stanford.edu/~amaas/data/sentiment/
https://www.kaggle.com/shashank1558/preprocessed-twitter-tweets?select=processedPositive.csv

References:

https://monkeylearn.com/blog/sentiment-analysis-deep-learning/ ht
https://journals.flvc.org/FLAIRS/article/download/128562/130025/
https://data-flair.training/blogs/data-science-r-sentiment-analysis-project/

Machine Learning Techniques and Methods:

NLP - used to determine sentiment (various implementations) of a given input string, and associate it with emotions, either positive or negative. This is the input for our clustering recommendation.

Clustering Recommendations (Limited Scope) - gathers data to associate with similar data from similar users. Given data on similar users, we attempt to match the preference of a given user to similar users, and then recommend movies that the given cluster favors.

Code Structure/Process:

The process is straightforward, although the structure may change.

Preprocess Data:

- Using movie reviews, we parse each review as a string and a score.
- Score may consist of single normalized score, or a normalized sentiment vector
- Input string is the review, absent punctuation and nonstandard characters.
    - Input string may be truncated or randomly sampled depending on length.

NLP Layer:

- We train a model to recognize sentiment given the sentiment vector and input.
- This process is independent of the second ML layer.
- No error backpropagation from Clustering layer.

Clustering Layer:

- We input the NLP sentiment vector to identify if the user favors a particular movie.
- This should be matched with the most similar users from the pool of users.

Output Layer:

- The user is associated with a given cluster of similar users.
- A movie is selected if it is associated with that cluster.
- The movie is selected from a preprocessed dictionary, and then outputted.

Experiments and Validation:

Upon training the model, we wish to consider different types of inputs - our standard input will consist of legitimate reviews of movies, but there may be specific types of text that our model will fail to detect. This includes natural language that may be coded beyond the scope of a simple algorithm and would specifically require a complex model to parse. Interconnected reviews, Sarcastic or facetious reviews, and language that may fall beyond the reach of a simple model may also fail to be detected using our model.

The datasets will also require some parsing, we may want to perform simple statistical analysis on datasets to ensure that the data is not malformed in some way - ex: having a strong sampling of negative reviews on a single entry may result in reviews for a particular movie being perceived as negative independent of the movie's merits or the reviewers feelings. We want well balanced datasets.

Possible Challenges:

Difficulties in:

ML Processes:

- How do you parse a segment of text that might discuss multiple Movies?
- How do you distinguish movies within a given context?
- If a review is sectioned, how do you parse it?

Feasible Implementation:

- Is it practical to implement relative to other methods?
- Can you show that the model won't overfit existing data?
- When would you use this?

Data Collection:

- Reliance on existing datasets with outdated language?
- What if input data is corrupted/malformed?
- What if data is malicious? Copypastas, memes, trolls, or ambivalent reviews?
- How do you gather relevant data if no such data exists?

Time Constraints:

- Scope of the project may be too large?
- Implementation of NLP/Cluster may take significantly longer if labeling is required?