Harish Ram, Sisi Zhang, Zhuohan Yu

Group 5 - Final Project Proposal

Background
The problem that we selected was analyzing movie reviews and interpreting the text in each review to determine whether the review contained an overall positive or negative sentiment. By doing this, we can demonstrate specific words and phrases reviewers typically utilize to indicate how they feel about a movie.

Data Source
The data we are going to use comes from 'The Stanford Sentiment Treebank' which is a movie review dataset where each row indicates a unique movie review with a corresponding binary label. This dataset contains 67,349 rows in the training set, 872 rows in the validation set and 1,821 rows in the testing set. The text in our dataset is not standardized meaning the length, ambiguity and formatting of each movie review is different from one another; thus, this dataset contains a diverse sample set to yield better results.

Modeling Concepts
For this project we will, primarily, be using Recurrent Neural Networks along with Rule Based Modelling to model and determine the accuracy of our results. In terms of the coding packages and libraries used for this project, NLTK and Spacy will be used for cleaning and preprocessing the data; while, RNN will be used for the core model generation. Due to the fact that the dataset contains a binary label, classification is the main NLP task that we will focus on to train and test our model and use the F1 score as the metric of our model.

Schedule

11.5 (1 week)	Finish proposal and research
11.19 (2 weeks)	Preprocessing (tokenization, create features)
11.26 (1 week)	Modeling (try different models) and improve score
12.3 (1 week)	Report write up
12.9 (1 week)	Final project presentation and presentation

Data Source URL:
https://nlp.stanford.edu/sentiment/index.html

Github URL:
https://github.com/HarishRam10/NLP_FinalProject.git