PROPOSAL:

Graham Harris, Judy Zheng, Christine Ma, Jong Heon Han

ROLES:

The team consists of four members. Graham Harris will be the writer, editor, and assistant researcher. Judy Zhang will lead the project research, implement the dataset for the computer to read, and develop methods to tokenize the data inputted into the model. Christine Ma will also work on tokenization methods and general code maintenance. Jong Heon Han will be the data analyst and create an evaluation method to calculate the correlation between the answer key and test data.

PROBLEM STATEMENT:

There are many difficulties when attempting sentiment analysis of a given corpus because human language is so complex and has many different interpretations. Since movies reviews contain concise paragraphs of text with usually strong sentiment, they are a great starting point for analysis. Given a dataset of reviews with an assigned binary classified sentiment of positive and negative, represented in the data set as probabilistic range between 1 and 0 respectively, our team aims to test different tokenization methods on a reputable model in order to analyze the differences of how various tokens are weighted in their sentiment value. Several different tokenization methods will be compared to discern the best system.

EVALUATION PLAN:

Our team plans to use a variety of tokenization methods on a constant model to evaluate sentiment across the data set. On the NLP Progress website's page dedicated to sentiment analysis, there is a given list of accuracies for certain models on data sets pertaining to movie reviews. To keep the analysis consistent across trials, one model will be used, while various tokenizations of the input set will be developed as independent variables. We will then perform experiments by embedding these tokenizations into the model.

Since the dataset of 50,000 IMDb movie reviews has already been annotated as for sentiment (positive or negative), it will serve as the answer key and be compared to the output of the model. The model we decided on is a SVM model which as a regression, will output the sentiment analysis as a range, but for the sake of simplicity, the range will then be condensed into a grading of positive and negative

where [0.0, 0.6) is considered negative and [0.6, 1.0] is considered positive. We will then compare and contrast the accuracy of the system to the annotated set to determine which system performs better on the bases of calculating recall, precision, and f-measure.

Data Set: https://ai.stanford.edu/~amaas/data/sentiment/

50,000 IMDb movie reviews annotated as positive/negative, including a training set

Tokenization Methods:

Splitting by space, Adjectives as tokens, Ignoring punctuation, TF-IDF weights, Stemming Included

Model: Support Vector Machines (SVM)

https://scikit-learn.org/stable/modules/svm.html

https://scikit-learn.org/stable/modules/linear_model.html

STRATEGY FOR SOLVING THE PROBLEM:

Each member will work on their assigned roles and communicate their findings. The team will start by reading various articles and papers on state-of-the-art sentiment analysis algorithms that solve the problem. By comparing the results of each tokenization though the model, we will draw conclusions regarding how sentiment has a role in language and how the development sentiment analysis by a computer aims to mimic understanding similar human evaluations.

COLLABORATION PLAN:

Our team will be able to collaborate by communicating our findings as we discover them to the rest of the team. As Graham and Judy have roles regarding research of the topic, they will communicate essential concepts and experiment recommendations for each tokenization process. Judy and Christine will work together on the initial stages of coding and implementing various tokenizations for the model. Then Jeong will evaluate the output and communicate the data to the rest of the group, as well as any figures and data representations. Then, everyone collectively will work together to draw conclusions as well as focus on improving and fine-tuning the experiment. While each member will be responsible for understanding the final report, Graham will synthesize the group's conclusions in a presentable and clear manner.

ARTICLES RELATED TO RESEARCH:

1.https://www.aclweb.org/anthology/W17-5410.pdf
Title: Breaking Sentiment Analysis of Movie Reviews
Author(s): By Ieva Staliūnaitė and Ben Bonfil (Utrecht University)
The article outlines how certain words and parts of speech have various meanings by context of the surrounding words. Essentially, the team concluded that since movie reviews allow for pragmatic and stylistic manipulations, it is difficult for systems to properly recognize sentiment. This research will be used to guide some of our tokenization methods, because our team will try to identify some difficulties and ambiguities that might be able to be improved upon.

2. https://arxiv.org/abs/1704.01444
Title: Learning to Generate Reviews and Discovering Sentiment
Author(s): Alec Radford, Rafal Jozefowicz, Ilya Sutskever
The work in this paper features sensitivity of learned representations of data models on distributions that they are trained on. Although our team's process will not be learning from our data set, we will be training it. This research team's sentiment analysis model was precise, interpretable, and manipulatable, which is a standard that our team will strive to achieve.

3. https://www.aclweb.org/anthology/Q17-1021.pdf
Title: Overcoming Language Variation in Sentiment Analysis with Social Attention
Author(s): Yi Yang and Jacob Eisenstein
This paper analyzes how different heterogeneous language variations make sentiment analysis more robust. The research team also used language variations from socially linked individuals to better perform sentiment analysis on Twitter users. This paper is of interest to our team language variation is a difficulty that POS-tagging and tokenization algorithms face. Our team can also pursue a tokenization method similar to theirs, where we group together reviews with similar language patterns.

4. https://arxiv.org/pdf/1905.05583.pdf
Title: How to Fine-Tune BERT for Text Classification?
Author(s): Chi Sun, Xipeng Qiu, Yige Xu, Xuanjing Huang
The goal of this research was to rigorously test different fine-tuning methods of BERT on text classification. Sentiment analysis tasks were run on the IMDb set (just like our research) and other sets,

for testing the BERT system. This paper also proposed other possible data sets for experimentation, such as Yelp food reviews.

5. https://arxiv.org/abs/1602.02373

Title: Supervised and Semi-Supervised Text Categorization using LSTM for Region Embeddings

Author(s): Rie Johnson, Tong Zhang

This team used region embedding for text characterization. Their most interesting finding was that after combining two types of region embedding techniques, the best results were obtained on untrained data, which suggests that there were complementary text characterizations in their model. This is something our team will remain aware of - it may be beneficial to try to combine multiple tokenization methods for optimal results (if the methods are complementary).