***Nathan Gilman, 34785552***

**CIS4930 Individual Coding Assignment**

**Spring 2023**

1. **Problem Statement**

*In an increasingly digital world, people comment and interact on social media very frequently. While providing an unparalleled sense of connectedness, online communication lacks important factors such as tone, facial expressions, and body language. Thus, it can be difficult to determine the true sentiment of a message.   
  
To address this problem, we can try to utilize machine learning in a way that accurately labels the sentiment of a textual message. This will be done through a combination of text preparation, text feature extraction and developing a model that can accurately classify the extracted features.*

1. **Data Preparation**

*Like many data analysis projects, I began with importing important libraries. The libraries I used in this project for data preparation and feature extraction were pandas, nltk, and itertools. I chose these libraries because of the following reasons: pandas is an popular data science library that offers a efficient and easy to use structure for handling large datasets, nltk is a popular natural language processing library which is the core of this project, and itertools helped me efficiently iterate through the extremely large dataset we were provided.*

*After loading the data from CSVs, I began by ensuring there were no discrepancies within the data I was going to use. From there, I converted characters to lowercase and used regular expressions to remove user mentions, hyperlinks, and any special characters. I then used other Text Prepocessing steps outlined in the course materials. I tokenized the messages and subsequently removed stop words and stemmed the remaining words in one pass. Dealing with the stop words and stemming the remaining words was an extremely time consuming process and iterated over several millions of words.*

1. **Model Development**
   * Model Training
   * *In this problem, I needed to compare the applications of 4 types of classification algorithms: Logistic Regression, SVM, Naïve Bayes, Random Forest. Because of that, I implemented them all. Additionally, I needed to compare the performance of the models on the different linguistic features. Due to the time constraints, I first determined the best linguistic feature and then used that for the models.*

* + Model Evaluation

*My models, frankly, were inefficient and not great. I had the most success with Logistic Regression on bag-of-words. TF\*IDF feature extraction with Logistic Regression yielded slightly similar results, with a lower accuracy. My implementation of word2vec had an overall accuracy of around 70 but was abysmal at predicting true positives. The confusion matrices and classification metrics of the three features are below.  
Chart, treemap chart

Description automatically generatedChart, treemap chart

Description automatically generatedChart

Description automatically generated*precision recall f1-score support

0 0.83 0.94 0.88 7542

1 0.68 0.41 0.51 2458

accuracy 0.81 10000

macro avg 0.75 0.67 0.69 10000

weighted avg 0.79 0.81 0.79 10000

precision recall f1-score support

0 0.83 0.94 0.88 7604

1 0.66 0.39 0.49 2396

accuracy 0.81 10000

macro avg 0.74 0.66 0.68 10000

weighted avg 0.79 0.81 0.79 10000

precision recall f1-score support

0 0.76 1.00 0.87 7628

1 0.17 0.00 0.00 2372

accuracy 0.76 10000

macro avg 0.46 0.50 0.43 10000

weighted avg 0.62 0.76 0.66 10000

* + *Using the bag-of-words with the other 3 classification models did not yield me any significant gain over the Logistic Regression.*

1. **Discussion**
   * *I feel that my model does not perform well enough. While the accuracy of the bag-of-words logistic regression classifier is 81%, all other feature extraction algorithms did not yield good results. There are many false negatives and false positives that I am unsure of how to deal with. I also believe that I could have utilized other libraries to perform a more efficient and correct implementation for this assignment, however I attempted to pursue a path that was within my comprehension.*
   * *The primary challenge during this project was inefficiency. I was attempting to quickly create my model through iterative development, but was being help back through the time consuming process of attempting to figure out how to properly clean the data. This was mainly removing stop words and stemming the remaining words. After creating several versions of my code and refactoring heavily, I was able to create code that was bearable. I would like to improve my code further, however I feel I need to read further on efficient data management strategies.*
   * *This was a challenging, yet rewarding, assignment. I learned many things and would not have exposed myself various data management strategies had I not taken this course. I did not anticipate the data preparation taking a majority of my time but it did.*
2. **Appendix**

*The link to my github repository is* [*https://github.com/npgilman/SentimentAnalysisHomework*](https://github.com/npgilman/SentimentAnalysisHomework)*.*