To: Dr. David Allee
From: Divyanshu Khare
Subject: Sentiment Analysis – Python (EEE 498)
Date: 30th April 2018

Executive Summary:

The confusion of negatives and positives has many criticalities in today's world. The typical understanding is from what we see in movie reviews, within social media, by restaurant critics and even text messages. Multiple tools can be drafted to guide users regarding their sentiments and about the statements produced by them to better understand the under tone of a text in a standardized manner. The scope of this application can move over to sectors like healthcare. For example, consider a scenario where a patient is meeting their psychiatrist, the sentence spoken by their patient is entered into a tool which gives psychiatrists specifics as to whether the statement is positive or negative or maybe even complex thoughts and emotions about a subject in the future. The tool can basically be used to formalize and standardize an important medicine behind psychiatry. Our current and greatest problem is the reliance of our systems on human intuition and understanding which only grows as the technological sphere advances. Our machines will soon need to comprehend basic human intuitions and understand the stability of our thoughts conveyed through Facebook, twitter, Instagram and even news media. Additionally, the use of sentiment analysis as a whole opens a pathway for generalizing connotations, emotions and feelings behind words that cannot be recognized easily. Furthermore, this methodology can also be used as a great business tool. For example, you know that you have frequent negative mentions of your room service quality, but your competitors tend to have high mentions of their food and beverage topics, you can consider this a particularly important topic to investigate further. On the flip side, if you have some trending topics that are positive for your property, but not as popular for your competitors, consider highlighting those in your sales and hotel marketing materials [3]. In both scenarios, the giant problem is use of human intuition. Human intuition is far too variant to be considered as standard in the area of investment, healthcare and even politics as seen in the recent past. For these reasons, I have chosen this important technique that adds standardization and generalization to current areas where human intuition is prevalent.

There have been multiple projects varying from basic sentiment analysis to complex ones involving high level machine learning algorithms like tensor flow to determine every emotion and reasoning for given datasets. I started this project with some research on how to create the train set for my algorithm and soon figured out that Sentiment Analysis isn't a new concept. There are thousands of labeled datasets out there, labels varying from simple positive and negative to more complex systems that determine how positive or negative is a given text [1]. Another interesting example involving this topic is the inclusion of hierarchical classification to determine the neutrality of a string using external APIs [2].

My code is somewhat basic in the big data world, it starts with importing multiple libraries from sklearn, pandas, glob and errno. The most important, sklearn was used to manipulate data and run through 4 machine learning algorithms, **Naive Bayes Model, Logistic Regression Model, Support Vector Machine Model and K-Nearest Neighbors Model**. The

dataset I chose contains reviews of multiple hundreds of movies separated into two folders. These reviews are around a paragraph long stored in .txt files within two folders, "**pos**" and "**neg**". The first part of my code finds and runs through the data in both folders following the given path (needs to be changed by grader) and stores it into separate lists (called "posdata" and "negdata"). A loop then runs through the text data and transfers it to a single list called "data" while assigning "1" (for positive text) or "0" (for negative text) to each transferred item on the list. The assigned value was respectively transferred to another list called "data_labels" and all of the data is sorted out using pandas to create a data frame that contains **reviews** and **sentiments** (1 or 0) column (Figure 1).

	review	sentiment
0	For a movie that gets no respect there sure ar...	1
1	Bizarre horror movie filled with famous faces ...	1
2	A solid, if unremarkable film. Matthau, as Ein...	1
3	It's a strange feeling to sit alone in a theat...	1
4	You probably all already know this by now, but...	1
5	I saw the movie with two grown children. Altho...	1
6	You're using the IMDb. You've given...	1
7	This was a good film with a powerful message o...	1
8	Made after QUARTET was, TRIO continued the qua...	1
9	For a mature man, to admit that he shed a tear...	1
10	Aileen Gonsalves, my girlfriend, is in this fi...	1
11	Jonathan Demme's directorial debut for Roger C...	1
12	When I rented this movie to watch it, I knew t...	1
13	It's hard to say sometimes why exactly a film ...	1
14	Yes, this gets the full ten stars. It's plain ...	1
15	Hello. This movie is.....well.....okay. Ju...	1
16	This is a film that was very well done. I had ...	1
17	A typical romp through Cheech and Chong's real...	1
18	OK heres what I say: The movie was...	1
19	This is one of the first films I can remember,...	1
20	This film is worth seeing alone for Jared Harr...	1
21	I was 10 years old when this show was on TV. B...	1
22	Here's another movie that should be loaded int...	1
23	To all the reviewers on this page, I would hav...	1
24	A favourite of mine,this movie tells of two fe...	1
25	As Most Off You Might off Seen Star Wars: Retu...	1
26	I thought this movie was LOL funny. It's a fun...	1
27	Hood of the Living Dead had a lot to live up t...	1
28	Pierce Brosnan the newest but no longer James ...	1
29	This is one of those Film's/pilot that if you ...	1
...	...	...
24970	Everybody who wants to be an editor should wat...	0
24971	Made one year before ILSA, SHE-WOLF OF THE SS....	0

Figure 1

```
Naive Bayes
Accuracy Score: 84.82%

Logistic Regression
Accuracy Score: 86.8%

Support Vector Machine
Accuracy Score: 85.08%

K Nearest Neighbors (NN = 3)
Accuracy Score: 59.12%
```

Figure 2

Next, a feature extraction function was used (CountVectorizer( )) to remove all of NLTK's stop words (her, hers, their, his, Itself, them etc.) along with special characters to create a matrix of features like a feature vector. This data then goes through splitting into train and test data to perform ML analysis and display an analysis census of the total amount of positive/negative tokens in the full dataset. Finally a verification tool was created that checks the sentiment of two given words, simply "good" and "bad". Since logistic regression algorithm had the highest accuracy, I used LR to determine the sentiment of a user input sentence that the grader can type and verify themselves (Figure 3).

```
Search Results for token : ['good']
      Token  Positive  Negative
10549  good    6264.0    6034.0

Search Results for token : ['bad']
     Token  Positive  Negative
1971   bad    1526.0    6074.0

Test a custom review message
Enter review to be analysed:
python is cool
The review is predicted Positive
```

Figure 3

1. [1] Cordero, Lesley. “Posts By Stack.” *Twilio Cloud Communications Blog*, Twilio, www.twilio.com/blog/2017/12/sentiment-analysis-scikit-learn.html.

2. [2] “Sentiment Analysis with Python NLTK Text Classification.” *Python NLTK Sentiment Analysis with Text Classification Demo*, text-processing.com/demo/sentiment/.

3. [3] “Three Practical Uses for Sentiment Analysis.” *Revinate*, 7 Mar. 2018, www.revinate.com/blog/2015/10/three-practical-uses-sentiment-analysis/.