Rideshare Rodeo: An examination of sentiment towards Uber using Machine Learning Techniques

Authors: Jeff Bertucci, Abraham Freeman, and Antonio Llorens

**Introduction**

When it comes down to it, a company is only as good as what its name represents to its audience. A brand has the power to foster customer or client expectations and to generate trust. In older times, companies may not have had this down to a science yet, but they knew the power that they held when people were able to not only remember the brand of product they were buying, but also see their product as the best among all other brands of the same product. It is no surprise that a company that is popular with its clientele will have the greatest success in selling their products. As such, companies began to put out advertisements to shape how the public saw them. These ads not only tried to sell a product, but instead an ideal of what purchasing that product meant about the vendor and buyer.

Today, advertisements are all around. They are on televisions, magazines, and billboards, but that does not seem to be enough. In this over-saturation of people trying to sell other people something, buyers have become desensitized. There is also the issue of media moving into the cloud, creating less opportunity to market directly to customers. However, paradoxically, there are more opportunities to market than ever before. There are influencers on social media that command their own legions of followers that trust their guidance and opinions. And, in the case of social media, there is enormous power that comes with a fanbase.

In the age of social media, twitter seems to be a mainstay. In 280 characters or less, everyone can make their own opinions known to everyone. They can tweet about the day they are having, which celebrities they are into, or about the products and services they encounter in their daily lives. Companies have fostered this last method by creating their own social media accounts that can problem solve, joke around, and make them seem as human and relatable as a friend from another town. A relationship is then formed between the company and its userbase in which the users say everything that can be said about the company, good or bad, and the company then pushes back with make-goods and supportive comments.

The “twitter-sphere” that surrounds a company contains everything that anyone is saying about the company at any given time. Simply by including the company in one’s tweet with a simple “@” sign, a person now has bought in to this universe. A company can leverage this sphere of influence by spreading their own message to their fans and following how their userbase is describing them. These messages about a company are a goldmine when it comes to discovering how the public feels about certain decisions, or company news and allowing for the company to pivot and test how that pivot was interpreted in the public eye.

The problem this tactic presents is one of volume. Hundreds of thousands of tweets are generated every week, and it is important for a social media manager to be constantly viewing the stream of data. Generally, there aren’t enough hours in the day to determine the sentiments of each tweet that is written about a public facing company. By the time someone did do something like that, the sentiments will be old hat, and the public will have moved on to something new. It is imperative that a company be able to rapidly look at tweets and categorize them by what they mean for the company. This is where computers can shine.

While humans may not be able to rapidly read and categorize massive amounts of tweets at a time, a computer can not only read in minutes what it would take a human days to do, it can also assign values to these tweets in comparison to a corpus of works that have come before. Tweets can be ranked in a multitude of categories that allow for in depth description of how the public feels towards a company at any given time. This power is immense when it come to marketing as companies have access to second-by-second actionable data.

**About the Data**

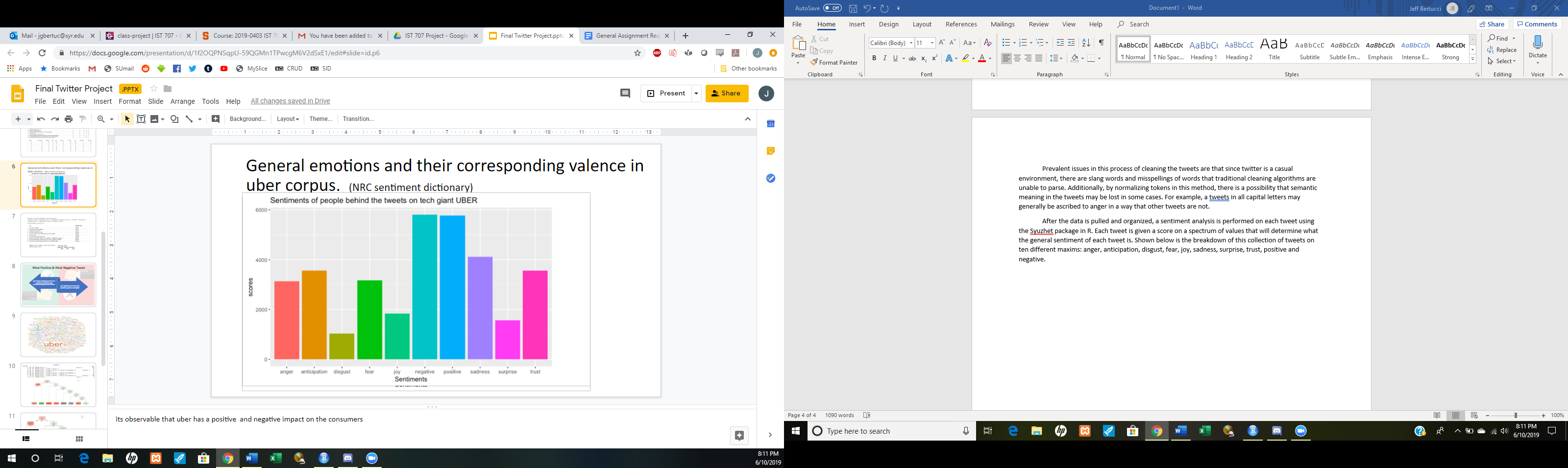
The data set used is a collection of 10,000 tweets that were generated between the dates of 5/26/19 – 6/2/19. This is the maximum range that the Twitter API would allow for. The search word used for the TwittR package was “@Uber”. At the time of the data pull, the full text of the tweet is included for each of the 10,000 cases. This includes numbers, symbols, letters, and html hyperlinks. Emoji symbols are not included when the tweet is pulled. Tweets can be as short as five characters, “@Uber”, and up to 280 characters. A benefit of the search approach of pulling the tweets is that there aren’t any null values in the dataset.

In order to clean the dataset, any symbols are removed as well as the hyperlinks in their entirety. Numbers are also removed, and the entire dataset is normalized into lower case so that each unique word can be a unique variable. Standard English stop words are also removed from the collection of tweets. Lastly, words are truncated into root stems that common English word endings are removed from. Extremely common words are removed from the dataset as well as words that are extremely rare. A word cloud is shown below describing the more popular words and their frequencies.

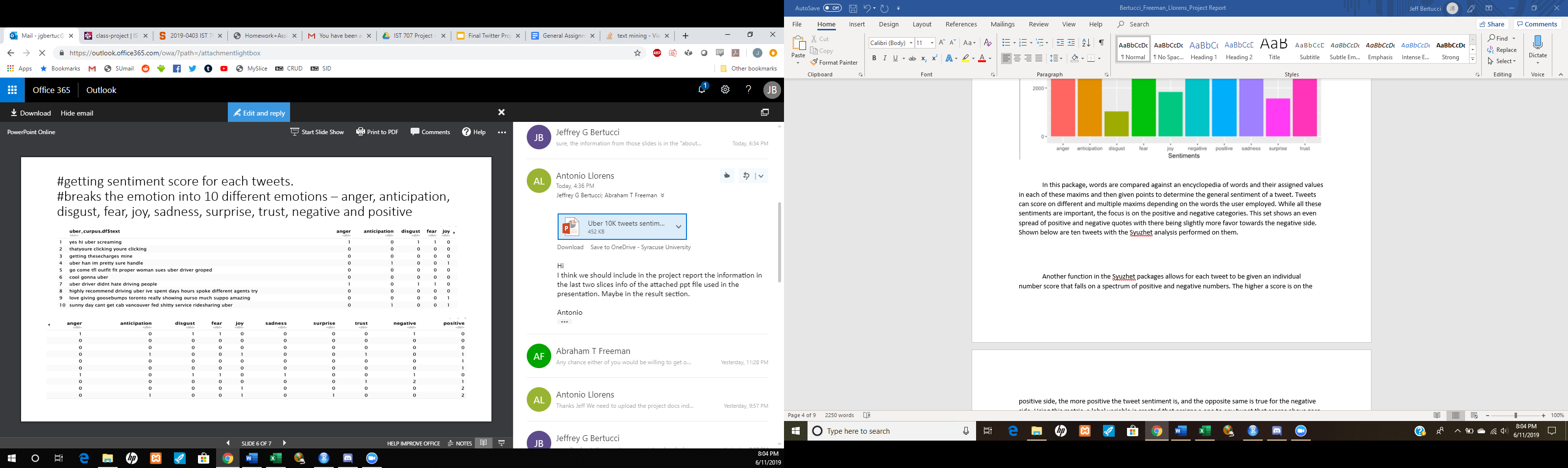


Prevalent issues in this process of cleaning the tweets are that since twitter is a casual environment, there are slang words and misspellings of words that traditional cleaning algorithms are unable to parse. Additionally, by normalizing tokens in this method, there is a possibility that semantic meaning in the tweets may be lost in some cases. For example, a tweets in all capital letters may generally be ascribed to anger in a way that other tweets are not.

After the data is pulled and organized, a sentiment analysis is performed on each tweet using the Syuzhet package in R. Each tweet is given a score on a spectrum of values that will determine what the general sentiment of each tweet is. Shown below is the breakdown of this collection of tweets on ten different maxims: anger, anticipation, disgust, fear, joy, sadness, surprise, trust, positive and negative.



In this package, words are compared against an encyclopedia of words and their assigned values in each of these maxims and then given points to determine the general sentiment of a tweet. Tweets can score on different and multiple maxims depending on the words the user employed. While all these sentiments are important, the focus is on the positive and negative categories. This set shows an even spread of positive and negative quotes with there being slightly more favor towards the negative side. Shown below are ten tweets with the Syuzhet analysis performed on them.



Another function in the Syuzhet packages allows for each tweet to be given an individual number score that falls on a spectrum of positive and negative numbers. The higher a score is on the positive side, the more positive the tweet sentiment is, and the inverse is true for the negative side. Using this metric, a label variable is created that assigns a one to any tweet that scores above zero, a negative one to any tweet below zero, and a zero for any tweet at exactly zero. The resulting spread is shown below:

A screenshot of a cell phone

Description automatically generated

In this distribution, there is a definitive skew towards negative tweets, meaning that results may also be skewed to favor negative identification.

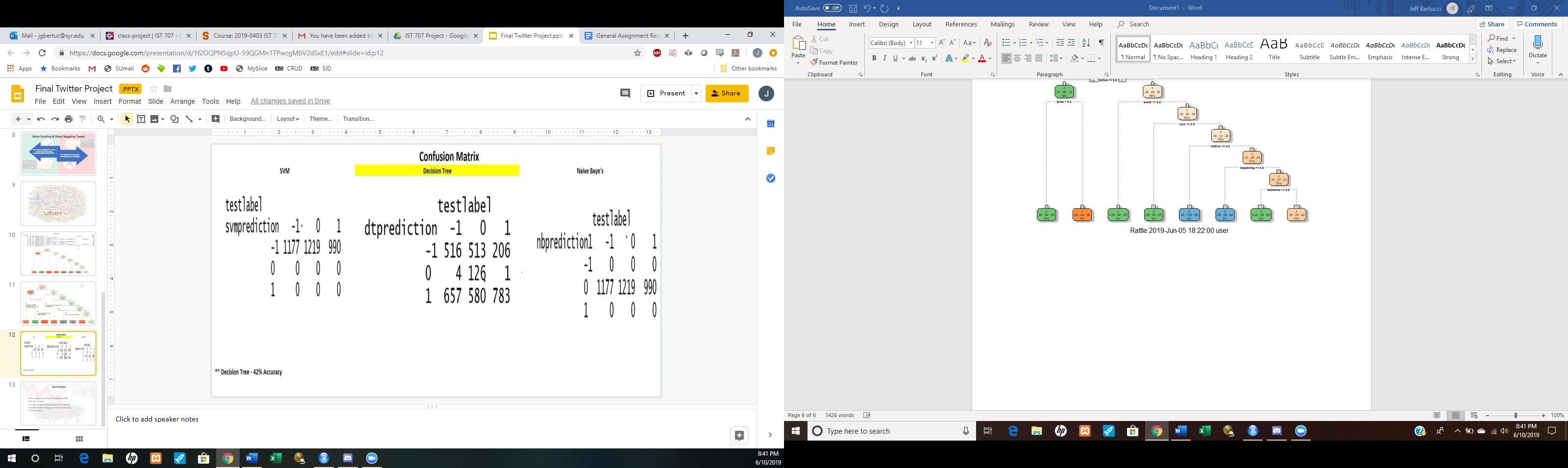
The dataset is then transformed into a document term matrix (dtm) that converts each word that appears in the corpus into a variable. A number is assigned to each tweet counting how many times each word appears in the tweet. The data is very sparse considering that each tweet can only have a maximum of 280 characters. includes any punctuation and symbols the tweet uses as well. In all, there are approximately 9,860 variable words that are considered in order to predict the label variable previously generated by Syuzhet. There are 10,000 tweets in the set. After the dtm is generated, the tweets are separated into training and testing datasets. The training set uses approximately two thirds of the tweets and the testing set uses the remaining third. The labels that were previously generated are kept for the training set and are removed and saved separately from the test set.

**Models**

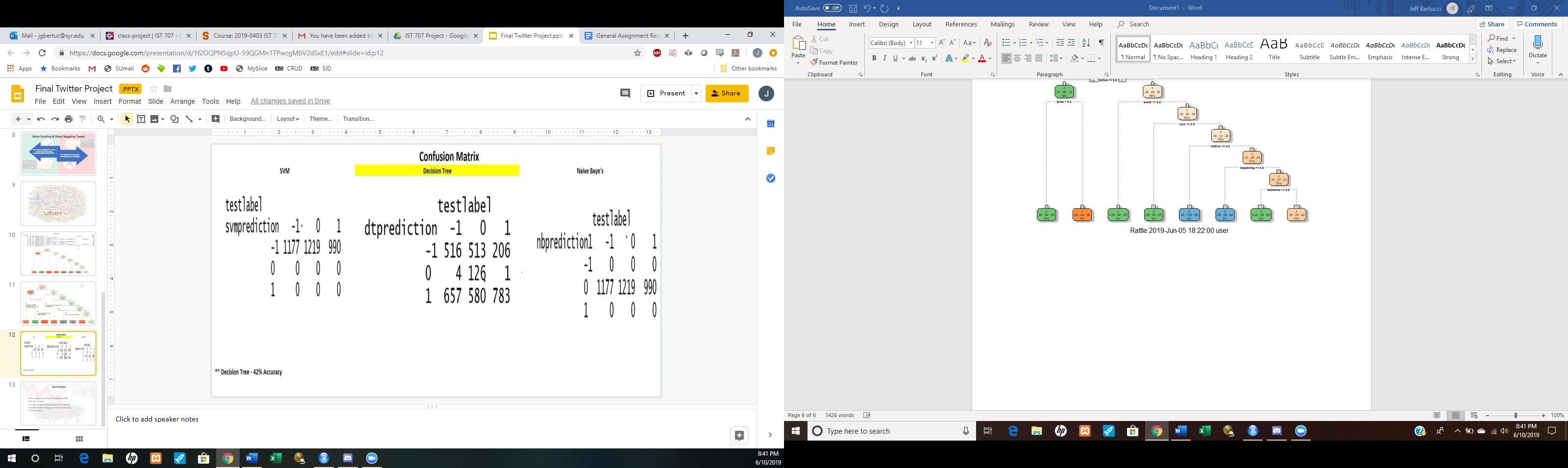
Model 1: Decision tree model Accuracy = 42%

A close up of a map

Description automatically generated

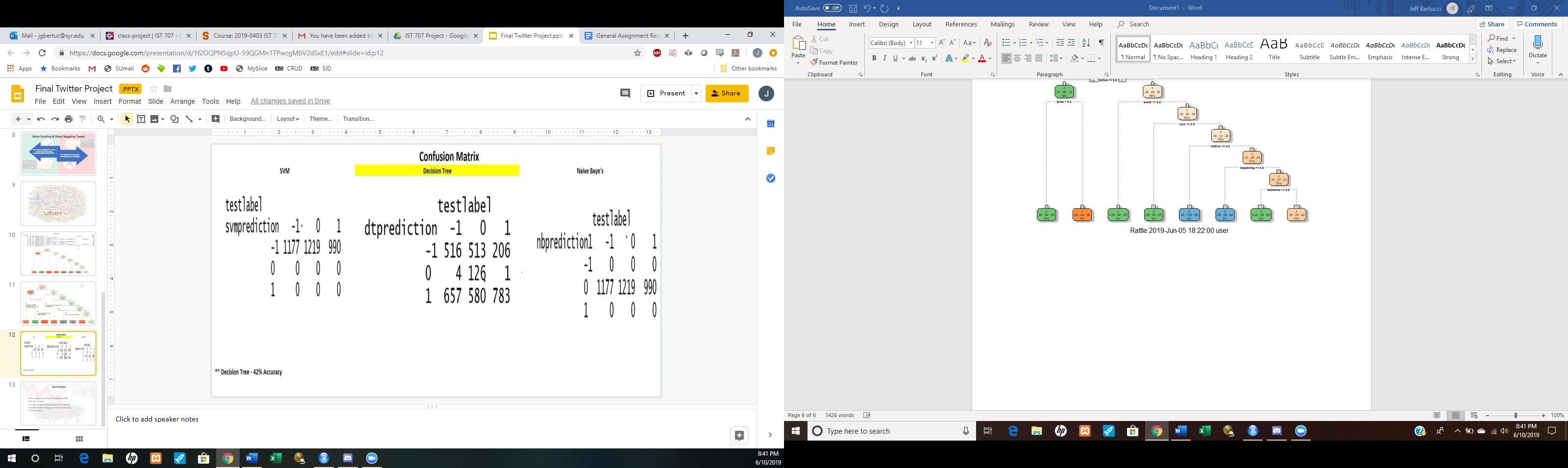


Model 2: Naïve Bayes Accuracy: 36%



Model 3: SVM Accuracy=35%

parameters: kernel = Polynomial, Cost = 10,000



**Results**

The data was most conducive for the decision tree model as it was able to cut away the words that had the most impact on a tweet’s identification outcome down to a handful of words that needed to appear exactly once for the model to make a decision. The Naïve Bayes model was severely limited by the style of text that it was working with; yielding a result where it could not make a guess one way or another and instead chose to go with the neutral option. The SVM model had a similar issue with modeling the data and inevitably chose to put all its guesses into the option that had the highest frequency in the training data. This same decision occurred with cost values up to 10,000 and still did not have any fitting.

Focusing on the decision tree model, a handful of words were chosen as the map for success, with most of the words leading to an identification of negative tweets with the else condition yielding either positive or neutral identifications. The words “billion” and “grow” are the most interesting identifiers for positive and negative tweets as they allow for the analyst to glean what some of this sentiment is regarding. These key words in 43% of the observations allowed for correct identification. Under these circumstances, random guessing would yield a success rate of approximately 33%. As such, this model shows that there is an improvement when using the tool.

**Conclusions**

It is important to note that this data is very sparse. When one crafts a tweet, word choice is important and there is very little room for bloviating. It is imperative that meaning be conveyed in as few words as possible in order to make a good tweet. Words are not doubled up on unless they are extremely important to the user. As such, categorizing tweets in this way is difficult as two tweets have the capacity to be incredibly different. The writing styles are not of a mere few authors, but of 10,000 voices all speaking at the same time and about a wide variety of things.

However, this data was still able to point in a direction at a specific event and gauge what it meant to many of Uber’s Twitter-sphere. During the week that this data was pulled, Uber had gone public with their new IPO showing a billion-dollar loss, and they were taking an incredible amount of heat from this loss. The term “billion” showed up in the decision tree not by accident, but because it was a word that many users were buzzing about. It is also why the words “motive”, “loss”, and “growth” appear in this model. The more positive tweets are focusing on growth while the negative ones are harping on the losses presented.

This is one case of PR, but it is a very good case study into how these words can be leveraged. If Uber can focus their brand on the words that generate positive tweets during a crisis and at the same time build bridges to those words that are resonating poorly, they can craft their brand in a way the public can accept rather than reject. The case truly shows that there is no such thing as bad news so long as a company can filter the bad news into a way that fits the narrative they seek to tell with their brand. Implementations for this method can be used to monitor what the userbase is talking about in one day or week or even year and use that to create their talking points for the next cycle. It is a way of staying ahead of the game by predicting reactions. There is no way for a company to please everybody all the time, but that is not what their goal should be. Instead, if they can tend directly with the customers that show a vested interest in what they are doing, positively and negatively, the company can continue to succeed.

Future steps for this data specifically would be to further refine the model with more in depth pruning of the decision tree. By discovering better keywords, it will be easier for Uber to figure out what their consumer base’s status is up to the minute. The more data collected, the better the model will be. This leads into the next point, while this model needs to be refined, next steps also include working with the next set of data in order to keep on pace with the hundreds of thousands of tweets that are constantly flowing in through the system. If this power can be refined and harnessed by companies, they will be one step closer to understanding their clientele’s needs, wants and fears.