Team NoDjo
ISYE 6740
March 31, 2020
Prashant Kubsad, Krishna Kumar Ganesh, Nevi Shah

Building a Brand Perception Model Using Twitter Data

Introduction:

In this day and age, social media has brought a whole new meaning to marketing. Gone are the days when marketing consisted only of designing singular ads, commercials and emails. Marketing today is fueled by intense market research on current and potential consumers, driving more personalized experiences and products with great market fit. Marketing today utilizes a breadth of technologies in order to appeal to and understand the market of their company.

With more than 3.96 billion people utilizing social media, analyzing social networking data is a fast and easy way for companies to get free and raw data about their consumers. Some of the most important insights drawn by this sort of analysis includes a comprehensive understanding of their audiences such as to who they are, what their interests are, where they come from and how they react to certain trends, products and ideas. Social media has been a consistent source of real time data giving accurate depictions of target consumers and allows companies to predict the adoption of new products and services.

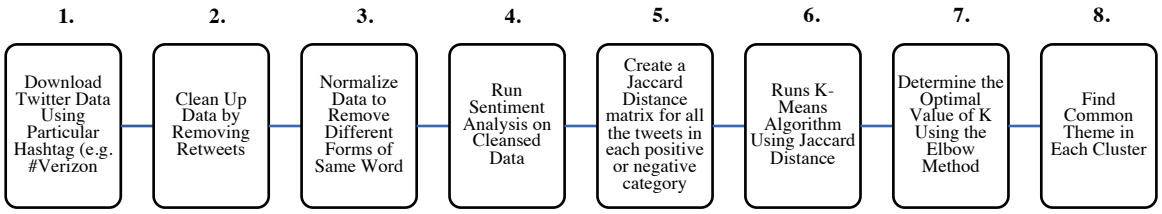
Twitter in particular has been an instrumental contribution to every company's understanding of their consumer's attitudes. Consumers are always quick to turn to Twitter to discuss dissatisfaction of their experiences, specifically as it relates to areas like customer support experience, in-store experience and online experience. Not only does this help companies to understand and document these experiences when it is negative but also enables them to document positive anecdotal evidence with respect to the brand. This naturally informative way of sharing customer insight is the crux of business development and market research and it is completely free to companies. Twitter has provided a series of APIs allowing companies to pull Tweet data as a raw dataset enabling them to see information like geographic location, Tweet text, number of followers and more. Data scientists are able to extract this data but are then required to cleanse and normalize the data on their own. While Twitter provides the data for analysis, it is the responsibility of each company to build and analyze this data for themselves. A big challenge that companies face is trying to find a quick and automated way to digest and derive actionable insights on live Twitter data without having to read each tweet manually document it. Things bring us to our problem statement of:

How might we derive the specific areas of both consumer satisfaction and dissatisfaction of a given company by utilizing Twitter data?

Methodology:

In order to answer this question, we will first start off by performing Sentiment Analysis via Natural Language Processing. Specifically, we will apply a Naïve Bayes classifier to determine the sentiment of all tweets. This process will include using a bag of words model to determine whether or not a tweet is in favor of the company, not in favor of the company or neutral. Because we are using free text data it is difficult to immediately classify tweets by a certain attribute, thus using an NLP/Naïve Bayes approach will be most appropriate. While performing Sentiment Analysis, we will optimize this model by (a) manipulating our bag of words, (b) adjusting the training sample size for positive sentiments and negative sentiments and (c) changing the threshold that determines a positive versus negative tweet.

Once we have initially performed the Sentiment Analysis on the data set, we will then perform a series of K-Means clustering within both the positive and negative sentiment groups. K-Means relies on factors to determine distance between data points, however, in this use case we don't have factors but free text. To combat this issue, we will be utilizing the *Jaccard Distance* to calculate the distance between two data points. In other words, we will be measuring the similarity/dissimilarity between two tweets. In this section of the analysis, we can adjust the K-value (number of clusters) for optimization. Below is an overview of the process we will use to build our models:



Data Source:
In order to pull the proper Twitter data, we will be using Twitter's Developer API and possibly using a supplemental tool we found that allows us to pull data from given a certain hashtag. The data we will pull from Twitter will be with respect to a specific company or a specific product by use of the hashtag. For example, this could be '#iphoneXR' or simply '#Verizon. We will set a range of dates to pull this data from to limit our data set to a certain time period and avoid getting an unmanageable amount of data.

We will be writing Python programs and utilizing Python libraries to clean and normalize the data. Some of the activities included in data cleansing/normalization are:

- Removing any retweets to avoid duplicates
- Setting all letters (uppercase and lowercase) to lowercase
- Eliminating any unwanted HTML text, special characters, white space and emojis
- Tokenizing text in the tweet by utilizing NLTK
- Removing words that do not contribute value to the model (e.g. ~~My~~ son loves ~~the~~ company)
- Converting multiple tenses of the same word into the same word (e.g. stop-words, stop words, stopwords)

Evaluation and Final Results:
The goal of this project is to first be able to segment tweets based on the sentiment. Then within the sentiment we will be able to find clusters within each sentiment group that describe some of the broader positive and negative opinions. Our hypothesis is that for a given hashtag we should be able to establish that a certain percentage of tweets is positive, and a certain percentage of tweets are negative. As an example, if we use #Apple to pull our dataset and find 80 negative tweets after Sentiment Analysis, we can further classify these 80 tweets into three categories: customer service, network coverage and call wait times based on our clustering. We are assuming that the elbow method will give us the correct number of clusters. We will run the accuracy of this model by manually labeling a training set to determine actual themes and then create a confusion matrix to observe our results.

References:

- https://www.pluralsight.com/guides/building-a-twitter-sentiment-analysis-in-python
- https://developer.twitter.com/en/docs/twitter-api/tweets/search/api-reference/get-tweets-search-all
- https://www.vicinitas.io/free-tools/download-search-tweets
- https://ieeexplore.ieee.org/document/5340335