1.125 Project Final Report (Draft, WIP)
Text Analytics - Sentiment Analysis of Tweets

Shweta Jindal
Aswini Prasad

Executive Summary:

The idea of the project is use data analytics to perform sentiment analysis on a given set of tweets, using various machine learning models such as classification trees, logistic regression, and randomForest. Using the predictive performance of the various models, we aim to make a recommendation to on which model to use!

The challenge of the project lies greatly in the first phase of the project - Data collection. Tweets are unorganized, and unstructured. It took a lot of time to take raw data (tweets) and convert it to the data format that can be be used in ML models.

Second challenge is to build the model. The text analytics is a very hard problem, since machines(computers) take text very literally. Computers cannot understand sarcasm, and ambiguity.

Finally, we will be proposing a couple of models that can we used to predict the sentiment of a tweet and compare them using the model's predictive power. Model with the best predictive power wins! This analysis has many use cases - specifically - to give an in-depth understanding of a company or a product to the senior leadership how a company or a product is being perceived by the customers.

Scope:

The goal of this project is to perform sentiment analysis based on the product reviews from the end customers/critics/experts on digital platforms like twitter and provide a summary of the overall impression of the product/service in the market. The main audience would be the senior leadership of the relevant product and technology firms, who are interested in getting a brief first-hand information about their recent launch as well as the potential buyers.

In general, these reviews are available in different platform but these data are large in volume and often contradictory and subjective too. Hence the objective is to use machine learning to make meaningful conclusion from these reviews. Accordingly, based on the product of interest, the reviews and comments will be collected from the social media, followed by sentiment analysis of these reviews based on the semantic structures of these sentences and tagging them as positive/negative/neutral.

Competitive Analysis:

Existing Tools/Codes	Pros	Cons
1. Predicting Stock market prices based on Tweets using psychology tool 'Profile of Mood States'	Accuracy as good as >85% in the tested cases	calmness index is not considered in predicting future prices Tweets are not filtered based on location and hence may not reflect rightly
2.http://twittersentiment.appspot.com - for twitter sentiment analysis based on machine learning	Nimble	Based on keyword and not sentence
3.http://brands.peoplebrowsr.com - Provides brand analyzer as part of a larger suite of tools.	Includes twitter as well as other social platforms	Paid version
4. Hootsuite – Enables marketing campaigns, schedule posts in advance, identify and grow audiences based on tracking of hashtags, mentions, Twitter lists	More focused on marketing for a brand among focused audience	Not strong in sentiment analysis, more for collaborative work
5. http://www.tweetfeel.com	Good with data visualization	Paid version Based on keyword and not sentence.

Challenges:

1. Data Collection: Tweets are inherently hard to use as it is, due to the following reasons:
    a. Loosely structured
    b. Textual
    c. Poor spellings, non-traditional
    d. Multilingual

Further, twitter APIs store data only for one week (duration), which is too less to build training models. Hence, we had to use various methods to dig more data. *(to add more specific details)*

2. **Text Analytics is very hard:**
    a) Computers find it hard to understand text, they tend to take everything literally
    b) Computer cannot deal with ambiguity - for instance: 'I put my bag in the car. It is big and blue in color' - computer doesn't know if the second 'it' refers to the bag or the car
    c) Computers find it to understand the context of the statement
    d) Computers cannot handle sarcasm, metaphors, and homonyms
3. **Sentiment Analysis**:

In sentiment analysis, the biggest challenge is to classify the tweets correctly as positive, negative, and neutral. To do so, we need to develop a robust training data set to calibrate the models.

a) Scape the website to find the tweets for a company or a product. In our analysis, we decided to use "APPLE" company, as it has many followers and many people who don't like the company or its products.
b) One the website is scrapped; we need to develop the training data frame. This involves allocating an average score to a particular tweet on a scale of -2 to 2. (-2): strongly negative, (-1) negative, (0) neutral, (1) positive, (2) strongly positive. There is no easy way to do develop the training model.
    i) There is an Amazon mechanical Turk - that breaks the task into small components and hires people to provide a score to a tweet for a minimal cost(2 cents/tweet). Same tweet is sent to a 6 people (let's say) and overall score of that tweet is the average score given by these 6 people.
    ii) We could have developed a survey system and asked our classmates to score our tweets. But due to limited time and large amount of data, we decided to score the tweets among us and use the average score.

Review of ML models:

The process of machine learning is similar to that of data mining. Both systems search through data to look for patterns. However, instead of extracting data for human comprehension -- as is the case in data mining applications -- machine learning uses that data to detect patterns in data and adjust program actions accordingly. Machine learning algorithms are often categorized as being supervised or unsupervised. Supervised algorithms can apply what has been learned in the past to new data. Unsupervised algorithms can draw inferences from datasets. In our project, we are doing sentiment analysis using machine learning models in a supervised fashion. Training data was used to train the model to detect the patterns of words used to convey the sentiment - whether positive, negative, and neutral.

In this project, ML literature was reviewed and we decided to pick two model namely - Classification Trees, and RandomForest. Sentiment analysis is a classification method where the outcome is either yes or no. Hence the regression trees were used in the classification mode. Similarly, randomForest is used in class method as well. Table below summarizes few points about CART and randomForest models.

Model	Used For	Pros	Cons
Classification Trees (CART)	Predicting a categorical outcome (quality rating 1-5, Buy/Sell/Hold or a continuous outcome (regression trees) (salary, price, etc.)	Can handle datasets without a linear relationship Easy to explain and interpret	May not work well with small datasets
Random Forest	Same as CART	Can improve accuracy over CART	Many parameters to adjust Not as easy to explain as CART

How exactly the text analytics work?

In simple terms, the technique applied for the text analytics to work is called **'Bag of words' approach**. The idea is to take text and convert it into the matrix format, often known as CORPUS. The structure of corpus looks like a 2D matrix, where the rows represent the data (tweet#1, tweet#2, and so on), columns represent all the distinct words that appeared in the data. Subsequent rows and columns in the matrix is a combination of 1s and 0s, representing the frequency of any particular word that appeared in a tweet.
As an example: We picked two distinct tweets, and as an example showed three words and matrix shows the frequency of these words. In our project, we have a matrix 1000X400

	best	customer	beautiful
I have to say, Apple has by far the best customer care service I have ever received! @Apple @AppStore	1	1	0
iOS 7 is so fricking	0	0	1

smooth & beautiful!! #ThanxApple @Apple			

Step by step process followed in text analytics:

1. Data cleaning:
    a. Convert the raw tweets to lowercase
    b. Remove punctuation; as an example - @apple, #apple, apple! - converted to apple
    c. Remove stop words - such as I, he, she, the, etc. Some common words that don't have predictive power are also removed, for eg. apple
    d. Stemming - to retain the root word.
2. Corpus Matrix
    a) Remove words that are very sparse
    b) Convert the data frame into document matrix format as explained above.
3. Divide the dataset into training and test dataset. We decided to go with 70% (training(/30%(test) data splits
4. Data is ready to train the models
5. Build ML models - CART, and randomForest
6. Predict the accuracy of the models using test data - done using confusion matrix technique. The outcome of the accuracy is compared to the baseline model.
7. Compare the predictive power of the models and use that model for analyzing the new tweets to classify the sentiment as positive, negative, and neutral.

Predicting new tweets:

1) To prepare the data for predicting new tweets, we used the above steps to clean the data.
2) Once the data to be predicted is in the proper format, we used our built ML models to test the sentiment on the tweet data. We use three separate attributes - Positive, Negative, and neutral, this was used to create a matrix to assign a particular sentiment to the tweet data
3) Visualization in the form of word cloud, and pie chart was done to give a visual impression of how the company or the product is doing.

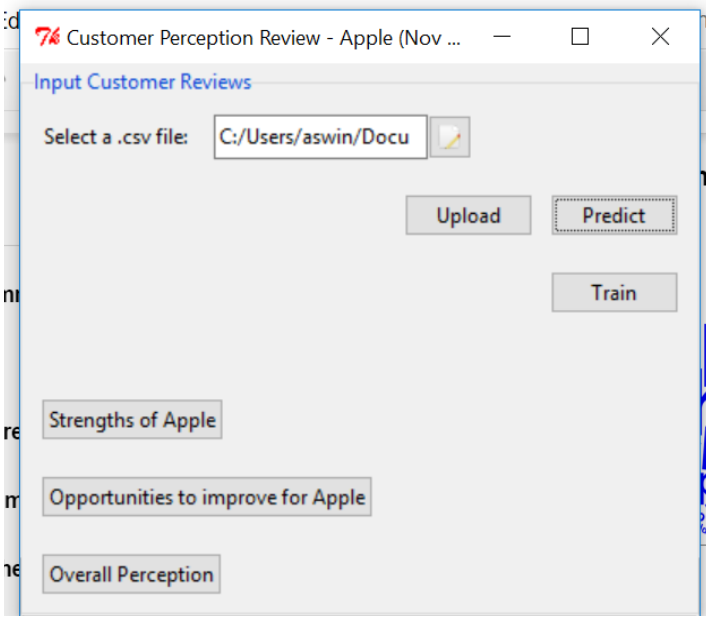
1.125 Final Term Project

Code written in JavaScript, ML models in R, Data visualization techniques - word cloud

Code can be found here:
https://drive.google.com/drive/folders/0BxTCSZrPjZvTLXhLNFNjQXlDaGc

Data Visualization:

User Interface:



Conclusion:

The goal of the project was to learn machine learning methods used to develop a new tool to perform text analytics in R. We developed a ML model to perform sentiment analysis on twitter data for a company - APPLE. This involved understanding how the text analytics is done, how to take raw data and convert it to the useable form, how to train CART/randomForest models, and finally how to use validation data for the model to give the required outcome. We believe that we were able to achieve all the goals put forward at the beginning of the project. Furthermore, we were able to deep dive and push the scope of the project by developing data visualization methods such as word cloud, pie chart to give a visual representation of the sentiments. We learnt how to integrate R and JavaScript, and learnt how to call JavaScript functions in R. There was a steep learning curve involved in the completion of this project, ranging from learning how to use R, Node.js, how to integrate JavaScript in R and many more skills.