Self-Driving Cars in the Public Eye of Twitter

Data Miners
Group 1
Ben Schiller
Mason Struthers
Brad Wyatt

EXECUTIVE SUMMARY

Tech and automotive companies, such as Google, Uber, Tesla, Audi, and Toyota, are getting closer to marketing self-driving cars directly to consumers. Before this can be done, they must solve the problem of understanding their potential customers' thoughts about their products. Our solution will give these companies a look inside their potential customers' heads, providing insights to help target them. Billions of dollars will be spent on marketing campaigns for self-driving cars, so it's vital that it will be spent effectively. We will address these companies' marketing challenges, give them insights to make their efforts effective, and, ideally, generate more sales for their self-driving car models in the future.

Our dataset originated from Amazon Mechanical Turk workers, who rated sentiments of tweets (related to driverless cars) from very negative, slightly negative, neutral, slightly positive, strongly positive, and non-relevant, as well as how confident they were of their rating. We narrowed the dataset down from 7156 to 6943 entries by removing non-relevant data, as well as only keeping the tweet, Sentiment_Confidence, and Sentiment attributes. Our goal is to build models that can accurately predict sentiment of tweets, and with those models we discovered keywords from negative, neutral, and positive sentiments that will be useful for our solution.

In R, we preprocessed the dataset through stemming words, removing punctuation, converting them all to lowercase, etc. We created a *userSelection* function where the user can input whether he/she want to analyze the data with a "Negative", "Positive", or "Neutral" outcome variable. We created several models: Logistic Regression, SVM Radial, SVM Linear, Naive Bayes, Decision Tree, and Neural Networks. We also removed specific sentiment levels in the dataset in order to try to obtain more relevant words from those models. Our metrics to pick the best models consisted of high accuracy, sensitivity, and precision. The results of the metrics and important words for the models in those combinations of sentiment levels are in the Appendix.

We were also able to gather what Twitter was saying about other car companies, such as Google, Uber, Audi. Google was mentioned the most by far out of potential competitors, being referred to once every three tweets (in the dataset), while others had a frequency of about 0.5-2% of tweets. Then, we used models to discover the variable importance of each company, for positive, negative, and neutral. Google was overwhelmingly positive, and Uber was very positive as well, but the others were not as much.

The conclusions can provide valuable and actionable insight to marketers as they prepare to market the first generation of self-driving cars. Twitter users who made negative tweets represent the people who don't want a self-driving car, likely due to their limited information, which can still be a huge market for companies. From our analysis, we determined that these people don't "trust" self-driving cars and feel that it is dangerous. To reverse this perception, companies can advertise about the safety of their car and the trust that current owners have placed in their vehicle. The second category of potential customers are the people who are excited about self-driving cars and more likely to purchase one when they are released. We gained insight on this market through our analysis of positive sentiment tweets. Those twitter users think self-driving cars are "cool", "awesome", and they "can't wait". Marketers can use these keywords to craft their messages to customers in this category. Our analysis was also able to provide us with valuable insight into the self-driving car's competitive landscape. We found that all companies other than Google have a big issue in a lack of name recognition. An advantage in public awareness could be gained through price point, being safer than other brands, celebrity endorsements, buzz from magazines, and ads making their model look cooler or more desirable than other brands vehicles.

BUSINESS PROBLEM

Our primary stakeholders are companies that are producing and marketing self-driving cars, as well as investors in these companies. Some of the key players include tech firms like Google, Uber, and Apple, as well as major automotive companies like Tesla, Audi, Nissan, Toyota, and General Motors. Even the US government is investing $4 billion in self-driving cars over the next 10 years.[1] The majority of these companies are planning on selling their self-driving car models directly to consumers, so their marketing campaigns will need to be effective if they're going to sell at a high rate. In the US alone, automotive companies spend $15 billion on advertisements[2], and with the addition of tech companies into the self-driving car space, there will likely be billions of dollars devoted to marketing these products alone, so it's essential this money is spent wisely. Self-driving cars are expected to enter the market in 2020, with an estimated ten million being on the road at this time.[3] While legality is still a major hurdle for this technology to overcome, the laws are looking to be turning in favor of self-driving cars, having already been legalized in states such as California, Michigan, Nevada, and Florida. All these facts show why marketing self-driving cars is a problem that these tech and automotive companies must start tackling now.

To better understand the public's perception about this industry, we are text mining twitter data to provide a sentiment analysis on self-driving cars. Marketing professionals at the tech and automotive companies discussed above can utilize this sentiment analysis to improve and better target their marketing campaigns, as it can only help to gain knowledge about their potential customers' thoughts on their future products. By addressing the concerns of potential customers, as well as capitalizing on their desires, these companies will be able to make highly effective marketing campaigns and greatly improve sales. As it relates to our model statistics, recall (minimizing false negatives) and precision (minimizing false positive rate) will be of high importance as it relates to the business problem. Any false classifications essentially mean many dollars, likely millions, wasted, so we want reduce these two numbers as much as possible to ensure that marketing dollars are spent as efficiently as possible.

DATA OVERVIEW

Our objective is to present an analysis of twitter users' sentiments regarding self-driving cars. We obtained the dataset from: http://www.crowdflower.com/data-for-everyone/ dated from June 8, 2015. The data comes from Amazon Mechanical Turk workers who read tweets that were related to driverless cars, and then classified them as very positive, slightly positive, neutral, slightly negative, very negative, or not relevant. The description was limited, so any other attributes in the spreadsheet besides the text of the tweets, Sentiment, and Sentiment_Confidence are not in our analysis. The analysis will be supervised learning, because sentences (tweets) were already labeled as sentiments from the AMT workers. Our objective is to find the important words through combining sentiment categories, and then training a variety of different models. Our analysis is predictive and not causal, since negative words do not cause a negative sentiment necessarily, but they can most likely predict it (same with neutral and positive). Our variables of interest are words that are important for prediction.

We modified the csv file to cut down on the necessities of the data (achieved using macros on Excel). Originally, there were 7156 entries. First, we removed tweets that were rated as "non-relevant". We also removed all columns except for Sentiment, Sentiment_Confidence, and text. The other columns were not of any use because the description of the dataset did not say anything about them. We also had

[1] http://www.nytimes.com/2016/01/15/business/us-proposes-spending-4-billion-on-self-driving-cars.html
[2] http://www.statista.com/topics/1601/automotive-advertising/
[3] http://www.businessinsider.com/report-10-million-self-driving-cars-will-be-on-the-road-by-2020-2015-5-6

no author of tweets. We changed sentiment from 1 to 5 instead to -2 and 2 for our own convenience (hence -2 and -1 were both "negative", 0 was "neutral", and 1 and 2 were both "positive"). The Rmd file was then analyzed using RStudio. See **Figure 1** for sample records of the dataset.

ANALYTICS SOLUTION

It is necessary to preprocess the data in order to analyze text from the dataset. We removed all tweets that had a sentiment confidence that was lower than 65%, to make our prediction more accurate. **Figure 2** displays the total amount of negative (rated -2 or -1), neutral, and positive (rated 1 or 2) tweets. The user must change the input of: *tweets <- userSelection(tweets, "Neutral")* in order to choose his desired sentiment for the outcome variable *desiredSentiment*.

In text analytics, it is necessary to preprocess the words in the dataset because it saves time and is efficient. For this project, we are instituting the bag of words approach, where each word is an independent variable. Using *gsub(),* we were able to remove twitter mentions (using @username), weblinks, and unrelated non-ASCII characters to avoid unnecessary variables. Because R is case-sensitive, we converted the tweets in the corpus to lowercase. We converted them to a PlainTextDocument, and then removed punctuation, English stopwords (including "i", "my", "we", etc), and stemmed the document. Next, we created a frequencies data frame, which only kept words that appeared in 0.5% or more of tweets using *removeSparseTerms*. We removed the following words because of their high frequency and their uselessness for interpretation: selfdriv, car, drive, driverless, take, will, driver, just, self, vehicl, still, can, autonom, isnt, anoth, like, dont, via. **Figure 3** shows a wordcloud of the overall frequency of words, after the preprocessing, with maximum words set to 75.

Next, we created a new variable in *tweetsSparse* called *desiredSentiment*, which equals *tweets$desiredSentiment*. We split the data where *trainSparse* consisted of 70% of the data, and *testSparse* consisted of 30%. We created the *prepare_testData* function to return testData based upon the model that is in the input. For each model, we set cross-validation to 5-fold, trained based upon the outcome factor variable *desiredSentiment* from the *trainSparse* data, plotted the most important words, and recorded the confusion matrix.

In order to find more potentially relevant words as well as build better models, we experimented through forming different combinations of sentiment categories. For example, for one of the combinations we removed "slightly positive" from the dataset, set "strongly positive" to TRUE and "strongly negative", "slightly negative" and "neutral" to FALSE, and then we found the top variables for the model (**Figure 13**). Every possible modification is *not* in the code because there are so many, however there is a comment in the code (in *userSelection* function) that shows where it happens and how to do it. In regards to the formatting of this report, for each combination, the underlined numbers in parenthesis are TRUE while those that are not are FALSE (the example would be Pos (-2, -1, 0, 1)). Some of the combinations were left out due to a poor accuracy, recall, or precision. For each combination, we experimented with logistic regression, support vector machines (radial and linear), naive bayes, decision trees, and neural networks. Our objective was to achieve the best balance of high accuracy, recall, and precision out of the models for those combinations from each sentiment. Please see **Figure 4** for the model we chose for each combination, with the included information of accuracy, recall, and precision. We were able to find the most important variables (words) for each model, and they are all plotted in **Figures 7 to 16**.

A further explanation behind what words we will recommend companies to utilize will be under the **Recommendations** section. For every confusion matrix of each model for each combination, please

refer to the **EveryModelCombinations** document attached to this project zip file. For the summary of the models, please refer to the **SummaryCombinations** document also attached to the project zip file.

FURTHER INSIGHTS

1) Using *mean(tweetsSparse$companyName)* we were able to find the average amount of times that a company is mentioned in a tweet in the dataset. Refer to **Figure 5** for the results.
2) We want to find how important these companies are in terms of how much negativity versus positivity is tweeted about them. Hence, the Negative and Positive models will remove neutral sentiment tweets for a better judgment of these companies. *svm.imp$importance* is used to find variable importance for all variables in the model. Refer to **Figure 6** for the results.
3) When looking at the results of several of the models, we noticed that "cant" and "wait" were high on all three of the positive models (**Figure 12, Figure 13, and Figure 14**). It would be contradicting for "cant" to be important on positive models by itself, considering it's considered a negative word. Using the command *findAssocs(frequencies, "cant", .3)* we discovered that "cant" and "wait" had a 0.42 correlation, confirming our hypothesis that those two words frequently happen together.
4) Both Neutral models from **Figure 15** and **Figure 16** have poor accuracy, precision, and/or recall, so it isn't that worthy of analysis. However, the words do relate to states that legalized driverless cars (California and Michigan), testing, and car companies.

RECOMMENDATIONS

The conclusions drawn from this analysis provide valuable and actionable insight to marketers as they prepare to market the first generation of self-driving cars. An analysis of the data gathered from negative and positive sentiment tweets and the current competitive landscape according to twitter provides valuable business intelligence to organizations all over the world. As 2020 approaches, marketers need to start forming their plans and this analysis helps them to better understand and adapt to their market.

A major goal of every automotive marketing department is to convince customers to buy the car. We think of these potential customers in two groups: people who don't want a self-driving car and people who do. Twitter users who made a negative tweet represent the people who don't want a self-driving car with the information they have been given. It is a key goal for marketers to understand why these people won't consider purchasing a self-driving car so marketers can work to change their minds. From our analysis we determined that these people don't "trust" self-driving cars and feel that it is a dangerous product. Companies can change their perception by placing advertisements emphasizing the safety of their car and the trust that current owners have placed in their vehicle. Pop culture magazines can feature stories of owners of driverless cars, and have an "angle" of the owners feeling safer. And, it would be important to try to publicize research and studies that improve the image of the safety of driverless cars. Through our analysis of the negative tweets, the word "idea" was recurrent. This shows that potential customers in this group don't view a self-driving car as a feasible product and still consider it an idea, in this case a bad idea. If advertisers worked to show that their cars have the capabilities to function on American roads these potential customers will be more likely to purchase the car.

The second category of potential customers are the people who are excited about self-driving cars and more likely to purchase one when they are released. We gained insight on this market through our analysis of positive sentiment tweets. Twitter users think self-driving cars are "cool", "awesome", and they "can't wait". Marketers can use these keywords to craft their messages to customers in this category.

By creating materials showing that the car is more cool and awesome than their competitors they can generate additional interest in their product. This could be done by making a video of a self-driving car doing incredible stunts to escape gun wielding pursuers or by having a self-driving car do a stunt no human piloted car has ever achieved. A marketing campaign with the goal of letting potential customers feel what owning a self-driving car would be like could help to shorten the wait that so many twitter users are frustrated by. While the technology for a fully autonomous car hasn't been fully developed yet a remote control car could be used to give riders the same sensation. Tons of commercials could be made following this campaign, showing anywhere from an elderly person's reaction to the car to a toddlers.

Our analysis was also able to provide us with valuable insight into the self-driving car's competitive landscape. We found that for all companies other than Google their biggest issue is lack of name recognition. Google is mentioned in 36.5% of all the tweets in our data set, the next most mentioned company is Uber at 1.7% of total tweets. This clearly shows companies need to be investing substantially more into public relations to make the public aware of their plan to release a self-driving car. An advantage in public awareness could be gained through price point, being safer than other brands, celebrity endorsements, buzz from magazines, and ads making their model look cooler or more desirable than other brands vehicles.

In regards to further improvement of our analysis, there are still a lot of unrelated words that could be removed. And, we noticed some words consisting only of non-ASCII characters that were misformatted that was still included in our models. We tried N-gram, but it would detect words that were important that were very neutral. We also could not get the names of the people who tweeted, which was the fault of the dataset itself. One other limitation to our business goals is that Twitter does not necessarily represent the population at-large. Twitter users tend to represent a younger, more tech-saavy demographic. People may also be tweeting emotional tweets for the goal of attention and retweeting, not because they actually excessively hate or love the idea of driverless cars.

APPENDIX

Figure 1. Sample 5 records of Dataset

Tweet	Sentiment_ Confidence	Sentiment (-2 to 2)
GAHHH GOOGLE SELF DRIVING CAR! #socool @ Kahuna HQ http://t.co/qEKOQxw7UD	0.6355	2
Wall-E world is getting closer >> With a Push From Google, California Legalizes Driverless Cars - NYT http://t.co/R0hdOy07	1	0
Why all these people taking pics of the rain driving talking about "be careful out there" u need tobe careful ur self driving/raining pixing	0.7155	Not relevant
Defiantly going to get my self-driving car license. :)	1	1
that self driving car is terrifying, especially if it runs linux.	0.6306	-2

Figure 2. Amount of Tweets for Each Sentiment Bar Chart (where -2 and -1 are negative, 0 is neutral, and 1 and 2 are positive)

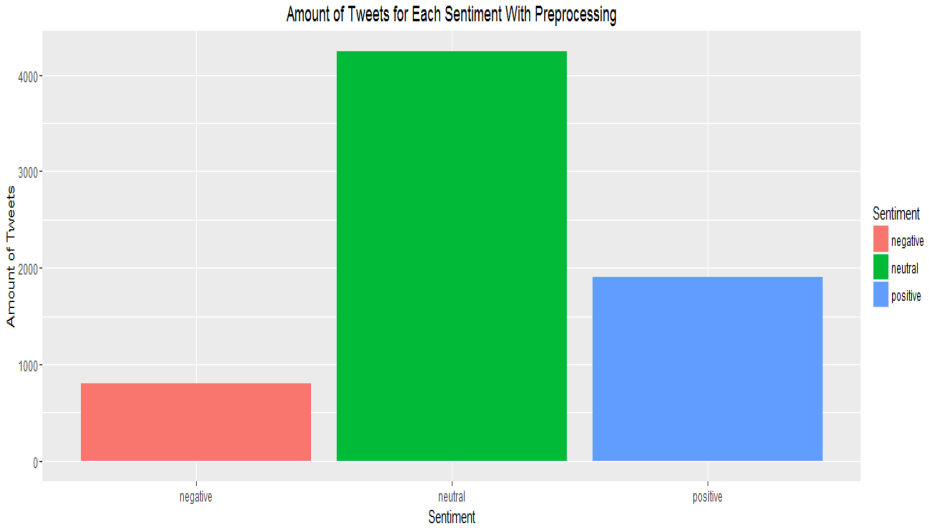

Figure 3. Word cloud, overall after preprocess and removing neutral words:

Figure 4. Top prediction models for each desired sentiment

Sentiment&Levels	**ModelChosen**	**Accuracy**	**Recall**	**Precision**
Negative (-2, -1, 0, 1, 2)	SVM Radial	0.9007	1	0.9007
Negative (-2, 2)	SVM Radial	0.8851	0.9740	0.9036
Negative (-2, 1, 2)	SVM Radial	0.9337	1	0.9337
Negative (-2, 0, 1, 2)	SVM Radial	0.9871	1	0.9871
Negative (-1, 0, 1, 2)	SVM Linear	0.9151	0.99497	0.91899
Positive (-2, -1, 0, 1, 2)	SVM Radial	0.8146	0.9565	0.8325
Positive (-2, -1, 0, 2)	SVM Radial	0.9555	0.9967	0.9583
Positive (-2, -1, 0, 1)	Naive Bayes	0.8261	1	0.8261
Neutral (-2, -1, 0, 1, 2)	Neural Networks	0.7015	0.4969	0.5255
Neutral (-2, 0, 2)	SVM Radial	0.9419	0.16216	0.75

Figure 5. Average Mentions of Competitors in Tweets

Google	**Uber**	**Toyota**	**Audi**	**Nissan**	**Apple**	**Tesla**
0.365	0.017	0.010	0.007	0.006	0.006	0.013

Figure 6. Variable Importance for Best Models in Negative, Positive, Neutral

Company	**Neg (-2, -1, 1, 2)**	**Pos (-2, -1, 1, 2)**	**Neutral (-2, -1, 0, 1, 2)**
Google	4.341765	95.65824	93.794848
Uber	37.048397	62.95160	75.241897
Toyota	N/A	N/A	66.756953
Audi	46.305517	53.69448	68.191251
Nissan	48.950408	51.04959	60.960588
Apple	48.950408	51.04959	64.833026
Tesla	54.622585	45.37741	58.738718

Figure 7. Negative (-2, -1, 0, 1, 2)

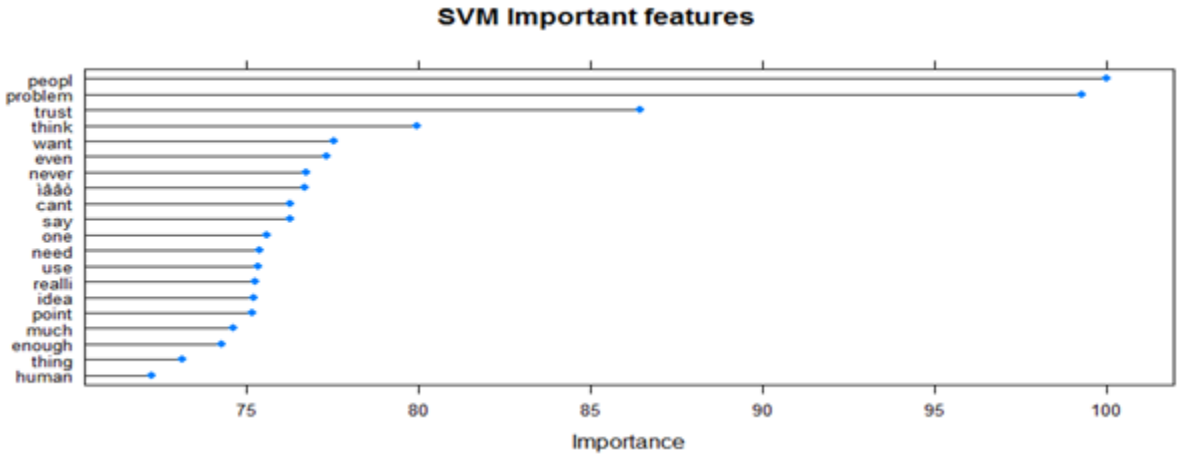

Figure 8. Negative (-2, 2)

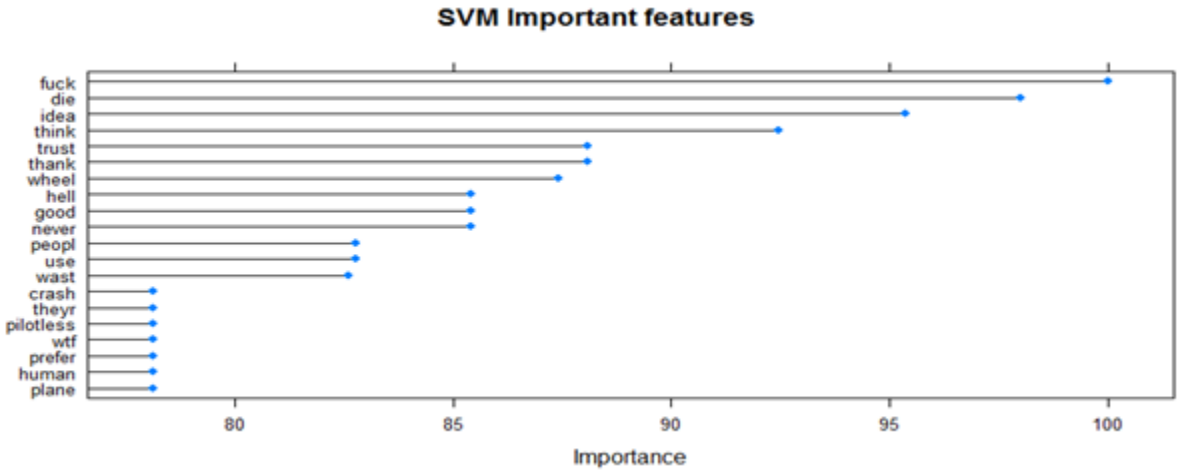

Figure 9. Negative (-2, 1, 2)

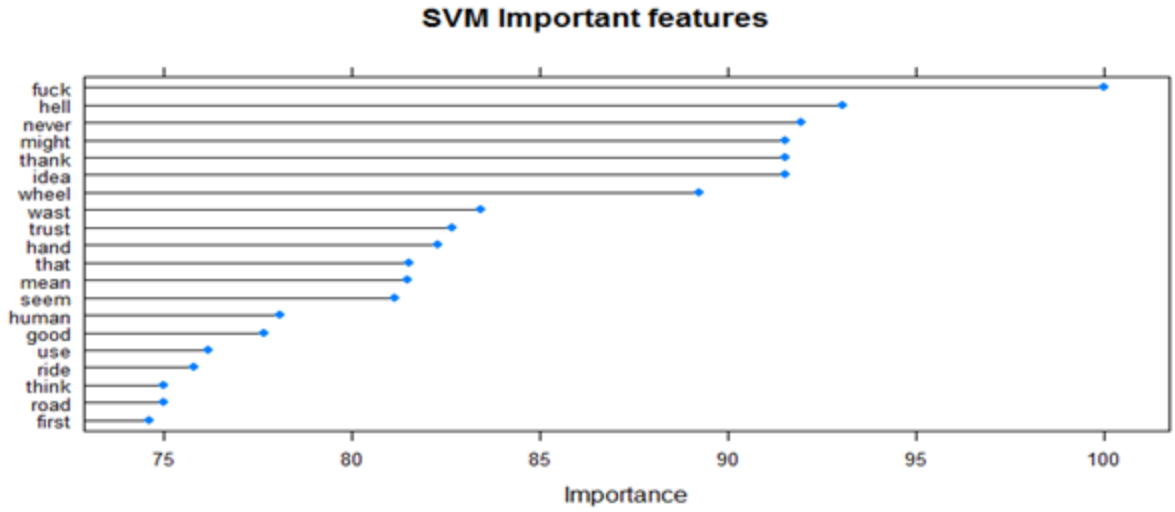

Figure 10. Negative (-2, 0, 1, 2)

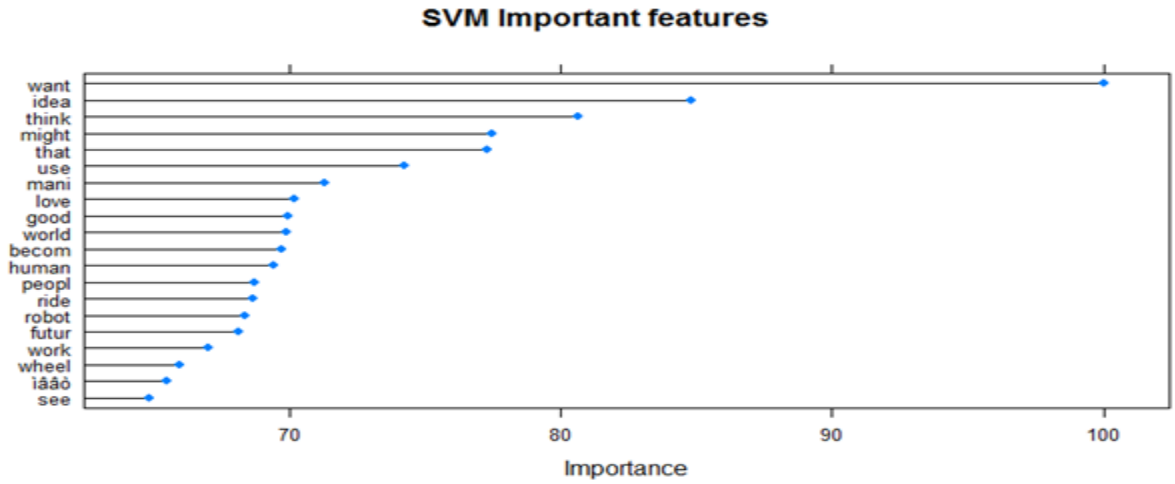

Figure 11. Negative (-1, 0, 1, 2)

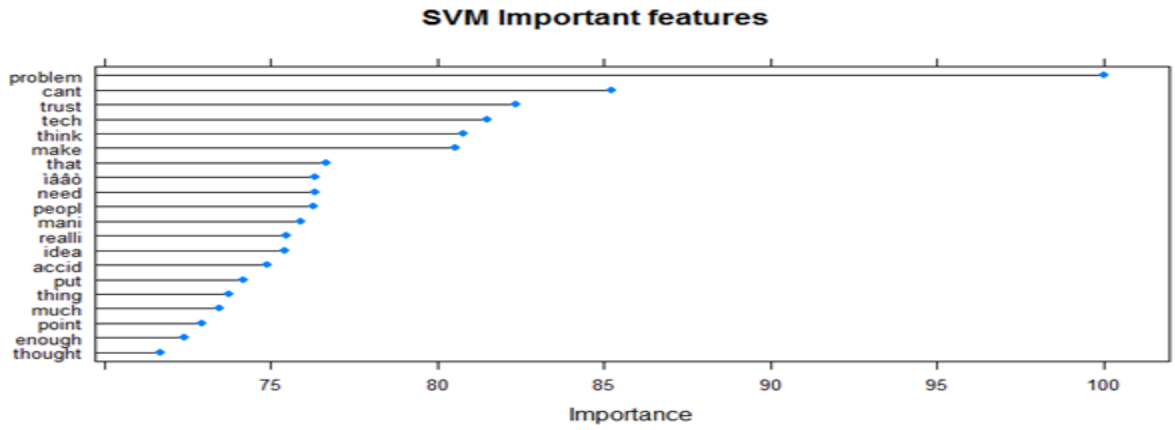

Figure 12. Positive (-2, -1, 0, 1, 2)

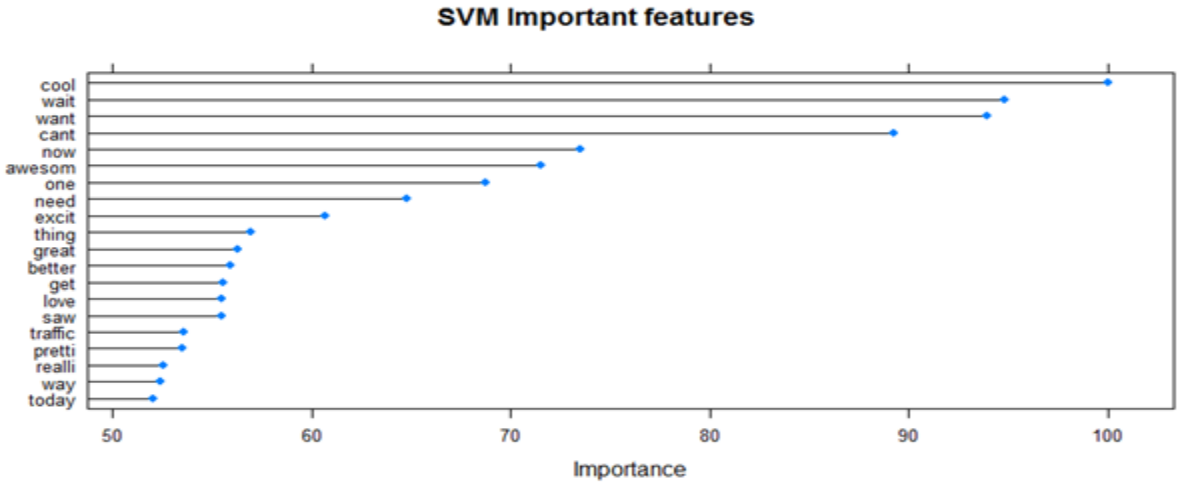

Figure 13. Positive (-2, -1, 0, 2)

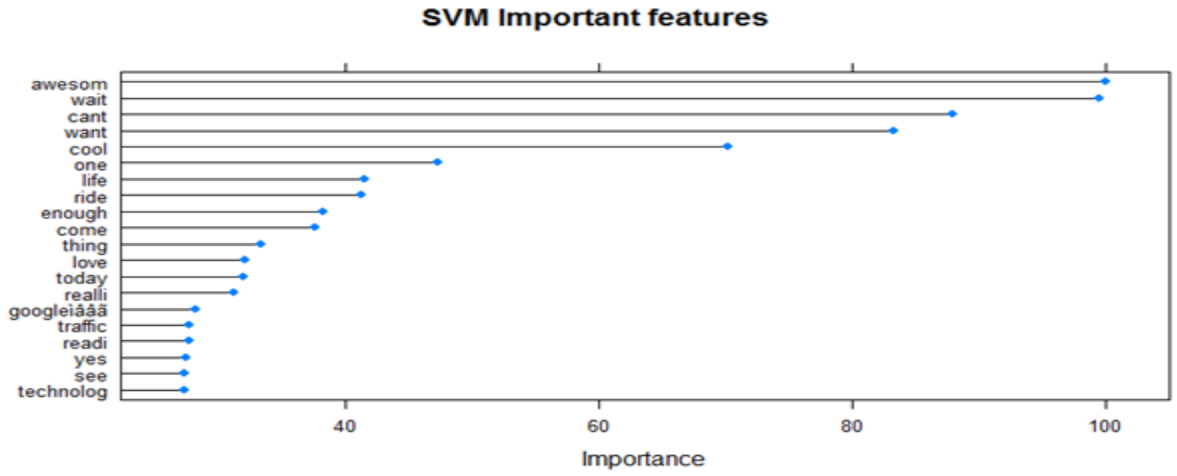

Figure 14. Positive (-2, -1, 0, 1)

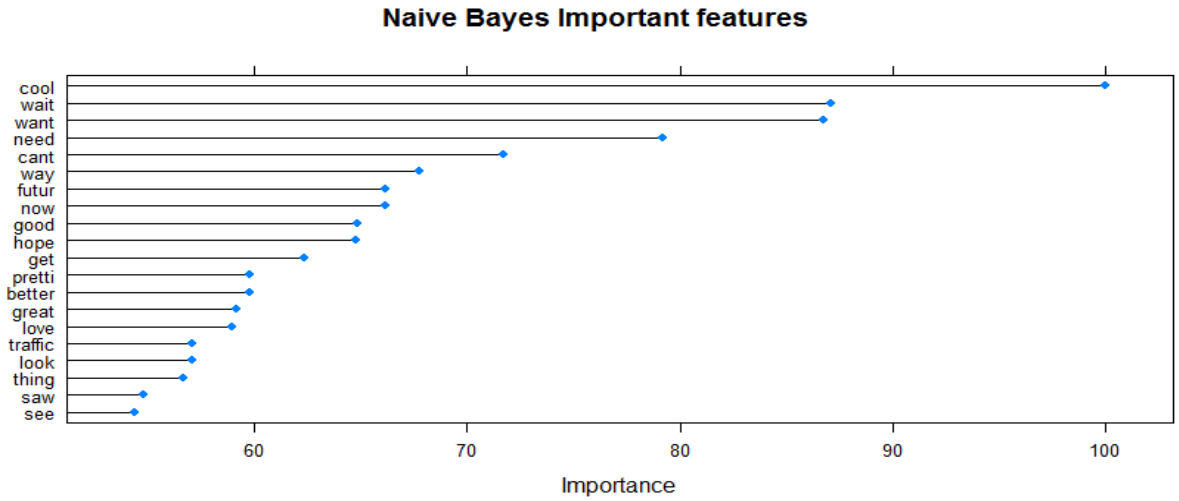

Figure 15. Neutral (-2, -1, 0, 1, 2)

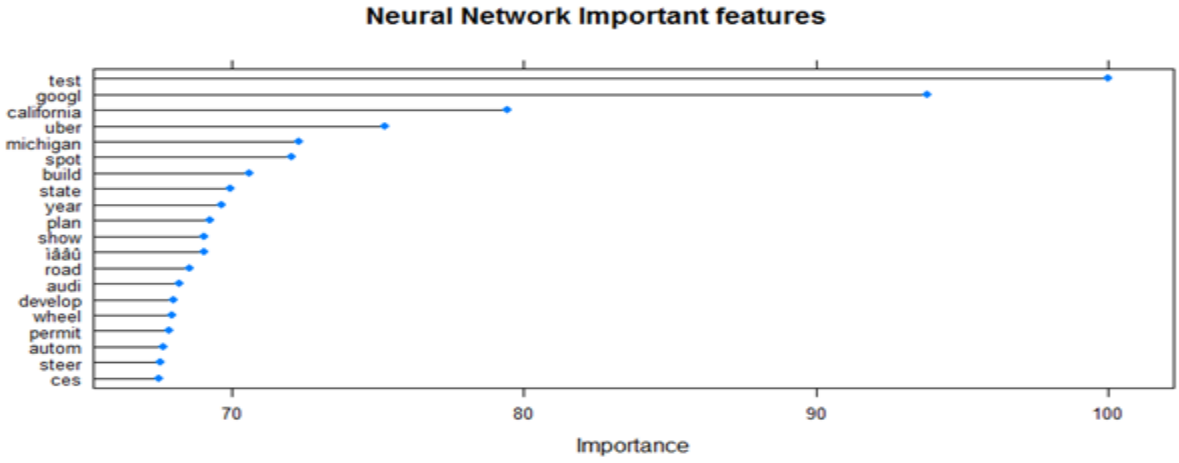

Figure 16. Neutral (-2, 0, 2)