**Sentiment Analysis of Trending Topics**

**Project Report**

**By:**

**Abraham Ludlam**

**Student ID : 891610248**

**CPSC 531-01 (13548)**

**Spring, 2019**

**Professor: Dr. Chun-l Phillip Chen**

**Department of Computer Science**

**California State University, Fullerton**

**May 7th, 2019**

**Table of Contents**

Abstract ......................................................................................................................................... **3**

1.0 Introduction................................................................................................................................**4**

2.0 Literature Review ......................................................................................................................**5**

3.0 Methodology.............................................................................................................................**7**

**3.1** Software Used................................................................................................................**7**

**3.2**.Validation and Evaluation..............................................................................................**8**

**3.3** Limitations.....................................................................................................................**9**

4.0 Results.................................................................................................................................... **10**

**4.1** Model Comparisons....................................................................................................10

**4.2** Naive Bayes Final Model for Multiple Months.........................................................12

5.0 Conclusion…………………………………………………………………………………..**15**

Appendices...................................................................................................................................**17**

References.................................................................................................................................... **20**

**List of Figures**

**1.** Figure 1: Comparison of the accuracy for Naive Bayes, maximum entropy, and SVMs …... **11**

**2.** Table 1: Table of Final Naive Bayes accuracy and predictions …...….…….……………….. **12**

**2.** Figure 2: Naive Bayes Model predicted sentiment values for September - December 2018 ...**13**

**4.** Figure 3:WordCloud for September 2018 immigration tweets …………….…………...….... **14**

**5.** Figure 4: WordCloud for October 2018 immigration tweets ………….…………………….. **15**

**Abstract**

This report covers the results of a sentiment analysis project focused on viewing trending topics by using sentiment analysis techniques. Sentiment analysis has become a large topic for multiple companies; with the results being used for multiple purposes. Past sentiment analysis research will be covered to provide understanding of sentiment analysis and why this report is important.For data collection, the topics were taken from Twitter, with the main one being looked at being immigration. The sentiment analysis models that are used and compared are Naive Bayes classifier, maximum entropy classifier, and multi-class Support Vector Machines. The results gained from the comparison led to me further expanding upon the model that gave me the best results and was the easiest to work with. The results of the trending analysis is shown.

**1.0 Introduction**

As social media platforms have grown larger and larger over the years, so has the number of opinions and reactions to particular topics and media that appeared on these platforms. This increased growth has made it hard for an individual or company to measure the overall polarity of responses by analyzing manually with no special techniques. In this way, sentiment analysis has risen as a topic of research and discussion to better analyze and measure the overall trends of social media platforms. Specifically, the machine learning models for sentiment classification are the main way companies and individuals process the data obtained from these platforms.

For this project, the focus was on trending topics on Twitter using supervised machine learning sentiment classification techniques. The goal is to be able to show the change of sentiment polarity overtime for a topic. In doing so, that information can later be used by other parties for predicting how the sentiment of a topic will look like based on recent events or in by taking actions to change the sentiment polarity of the topic. Examples of these two things, respectively, would be gauging the sentiment of a politician after a debate, and improving the marketing of a product before release. The topic that was mainly looked at in depth was immigration, with a side look at vaccines and the movie “Bohemian Rhapsody”. Immigration has been a polarizing topic and this research lets me get a more indepth look at it. For model comparisons, Naive Bayes, maximum entropy, and support vector machines are looked at and used.

This project also expanded upon the work done by Unnikrishnan, Narayanan, and Joseph (2017) that used SVM and Naive Bayes for sentiment analysis of Twitter. It also tried to incorporate the grouping methods shown by Shaikh et al (2017) to better obtain and display the changes in sentiment polarity over time. Specifically, the contributions of Shaikh et al (2017) show how to group data gained by Twitter and then display the results of analysis with these groupings. Work done by Moise (2016) was big influence on this project, as she showed the trending sentiment polarity of vaccines and other medical topics from Twitter datasets. For the models, Medhat, Hassan, and Korashy (2014) description of Naive Bayes, maximum entropy, and Support Vector Machines influenced my choices for their inclusion for this project. .

**2.0 Literature Review**

As mentioned above, this project builds upon the research done by others. Knowledge was able to be gained and used to be taken advantage for in this project such as the fact that Unnikrishnan et al. (2017) mention about an advantage of SVM compared to Naive Bayes is that insertion of attributes can be dependent on others, allowing more accurate results. For Shaikh et al (2017) work, the specific thing that was looked at was how to group data gained by Twitter and then display the results of analysis with these groupings. Moise’s ( 2016), allowed her to see the trend of populations on how they reacted the news of certain outbreaks and vaccinations. Specifically, it showed how overall Twitter sentiment polarity changed depending on the recent events and how it was less about people changing their sentiment, and more of people of the other sentiment polarity speaking up more. For Naive Bayes, Medhat et al. (2014) states that Naive Bayes model is about weighing the probability of the features of the data intersecting with the features of the class labels and then assigning a class label based on the class label that had the highest probability. For maximum entropy, Medhat et al. (2014) states that vectors based of the labels of the data are made with weights and that these weights from the vectors for each classification label are used to create probabilities of which class the data point most likely belongs to. For support vector machines (SVMs), Kaur, Mangat, and Nidhi (2017) state that SVMs work by creating a hyperplane of the respective labeled classes, and then tries to find the best linear separator that defines the classes from each other. New words and their features are assigned to the class based off the sector the data point lands in on the hyperplane.

For comparisons of these models against each other that has already been done, the main focus is usually SVMs models against Naive Bayes based models. For Naive Bayes, advantages of using the model, as Vanaja and Belwal (2018) state, is that the implementation of Naive Bayes model is easier to implement and runs quicker than almost all other algorithms. For SVMs, feature selection research done by Jianqiang.and Xiaolin (2017), showed SVMs most benefit in increased accuracy from the implementation of the feature selection methods tested. A disadvantage of SVM brought up by Medhat et al. (2014) is that SVMs can have a higher computational cost at initial runtime than both the Naive Bayes and maximum entropy models. For research of Naive Bayes and SVMs against each other, Rana and Singh (2016), where for the purpose of classifying movie reviews from IMBD, they found linear SVMs to be more accurate that Naive Bayes by about 5%, where the average accuracy provided by the linear SVMs was about 78.75%. The results of Vanaja and Belwal (2018) work showed that Naive Bayes was more accurate, had better precision, and better recall than SVMs for correctly labeling sentiment values to customer reviews. An example where SVM was clearly better than Naive Bayes is the airline twitter sentiment analysis done by Rane and Kumar (2018), where all of SVMs’ precision, recall, and f-measure were above 80%, where as Naive Bayes had around 64% for those three statistics. One reason that can explain this discrepancy is the dataset that Rane and Kumar (2018), where the majority of the tweets were negative, which Naive Bayes already has some problems analyzing without proper feature selection techniques.For maximum entropy models, ne of the only benefits is mentioned by Medhat et al. (2014), in that maximum entropy model is able to achieve similarly accurate results to other classifiers with a smaller training dataset.

**3.0 Methodology**

**3.1 Software Used**

The very first thing I did was collect the data from Twitter. The main program and test builds were built in Python, mainly Python 3. The original plan was to strictly use Twitter’s API with Python with the library “python-twitter”, but the limitations of the API made me restrict its usage for the final implementation of the best model. This is due to the fact that the Twitter API is able to grab any tweet as long as you have the ID, but can only grab Tweets with search terms within the last seven days. As such, I moved to a Twitter-scraper program, specifically one the uses the web scraper extension for Google’s Chrome and a “twitter-scraper” JSON file. With this, I was able to gather over 15,000 tweets regarding immigration over a four month period of time from September 2018 to December 2018. The limitations of the twitter-scraper program that was used was that it couldn’t grab more than a thousand tweets without a risk of crashing. To get around this, I simply searched and gathered tweets within 7-8 days, and compiled these together for each month. From there, I used the Python library “Textblob” to take the CSV file outputted by the scraper and assign a base sentiment value for each Tweet. For feature selection, “nltk” was used to define stopwords, and the library “re” was used to remove extraneous features from the tweets such as special characters and URLs. For interacting with the csv files, storing the data in data frames for the program, and saving the results, the “pandas” library was used.

For the models themselves, different libraries were used. For Naive Bayes, the Python library of “nltk” was used to take advantage of its “NaiveBayesClassifier” class and other supporting functions. “Nltk” was also used later for its “MaxentClassifier”, which was used for the maximum entropy model. To save these classifiers for future use, the library “pickles” was used. For SVMs implementation, the “sklearn” library was used for its various SVM models and functions. In particular, to make a multi-class SVM, the “onevstheRest” function was used to support the training of the SVM that was used, which was a radial basis function SVM. To help with feature selection for SVMs, “wordnet” was used to help normalize the data, and both“sklearn” and “scipy.hstack” were used to help vectorize the features. “Wordcloud” library was used to display word clouds of the most popular words of the four months that were examined for this project. For displaying the results gained from these processes, matplotlib was used for bar graphs and line graphs. Basic user interface was made with “tkinter” for the final model that was used.

**3.2 Validation and evaluation.**

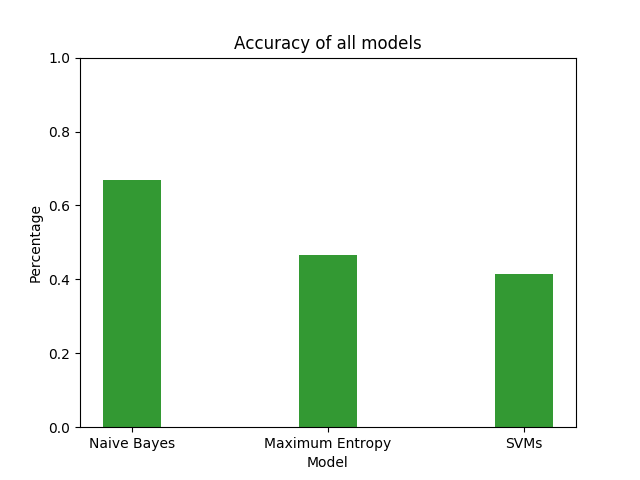
For validation, a hold out method is applied, where one months worth of data is tested and the other three months worth of data is used to train. In this case, all of September tweets about immigration are used for testing, about 12.5%, total, and the rest of the immigration data, about 87.5%, is used to train the dataset. These September tweets totaled at 2799 and were composed of about 47% positive tweets, 30% negative tweets, and 23% neutral tweets. For the testing dataset, the composition of values is 1336 positive tweets, 828 negative tweets, and 634 neutral tweets, so the desired prediction results would be about 47.7%, 29.5%, and 22.6% respectively. For evaluation, the overall accuracy and errors are looked at, and the best model is chosen from the one that had the highest accuracy and lowest amount of errors. For the model that showed the most promise, it was further developed to work with the other months, and switched to nth-crossfold validation. This is due to the fact that the datasets being used with it from here on out can be split into parts by time and the datasets per time period are of varying sizes. For example, for immigration, the other three months were used as a testing dataset for a new classifier that was trained from the other three months. So October’s dataset classifier was trained with data from September, November, and December, compared to the previous comparisons where September’s dataset classifier was trained with data from October, November, and December.

**3.3 Limitations of the Models**

Unfortunately, the Maximum Entropy model had to be scaled down, since during training the model would run out of memory or encounter an overflow error. When it ran out of memory, it would crash the program, and if it overflowed, it would overfit the data and only predict a single sentiment polarity. As such, training data is now only 72 tweets for the maximum entropy classifier, which was about the maximum it could be without encountering either of the two errors mentioned above.. While this likely hurts the final results of the maximum entropy model, it was better than having no results at all. For specific implementation of Naive Bayes, prediction for the testing set could sometimes cause a memory error, so the accuracy of the testing data set is taken 500 at a time, with the total accuracy being the sum of all them divided by the number of times this process had to happen. For iterations of each classifier, Naive Bayes doesn’t iterate at all, maximum entropy iterates 10 times, and the multi-class radial basis function SVM iterates 100,000 times. The reason there is a set limit for the SVM model is for time complexity sake, and mainly due to the fact that the first implementation with no set iteration limit went on for three days without outputting an answer. When setting the iteration limit at 100,000 with a linear kernel for the SVM, the model tended to overfit the training data and exclusively predict a single sentiment value for the test data. As such, I switched over to using a radial basis function kernel for the SVM model as it seemed to correct the over fitting experienced with the linear kernel.

**4.0 Results**

**4.1 Model Comparisons**

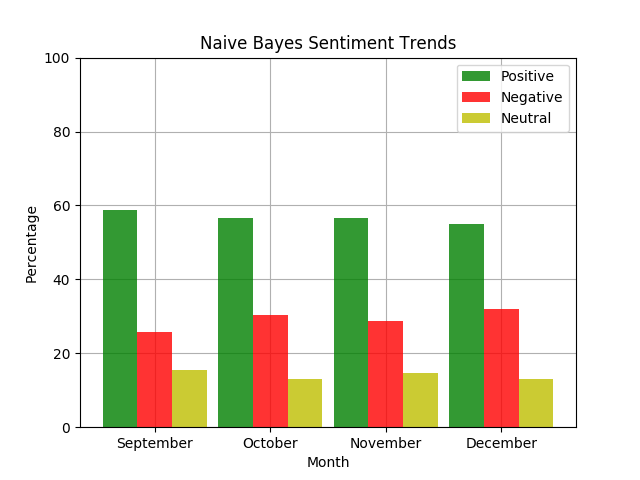
The final results showed that for the initial dataset of immigration tweets from Twitter, that Naive Bayes was the best at about 66% accuracy. In comparison, the maximum entropy classifier had around 46% accuracy, and the SVM classifier had about 42% accuracy. Maximum entropy model’s deficiency can be explained by the lack of training data it was given when it was compared to the other two. As for SVM model coming in dead last, this could be due to the fact that there was still not enough iterations for the SVM model to create an accurate hyperplane. I traded time for accuracy in the case of SVMs and it shows in the results. As for Naive Bayes, it tended to overfit for the positive values, with a predicted 58% positive values for the test dataset comprised of only 47%. Surprisingly, the SVM model had the closest predicted total sentiment values to the original model, with 56% positive, 26% for the negative, and 18% of the neutral. That is, however, of course due to the fact that it was widely inaccurate in its predictions and got more wrong than the others.This shows that overall sentiment value is basically a useless statistic in terms of strength of a classifier, as it ignores the individual results that each modifier does. That being said, as this project was focused on trend analysis, it is important for the models to predict somewhat close to the original models, otherwise it makes it difficult to accurately view the changing of sentiment polarity overtime. Charts made from this data are shown in the appendices at the end of this paper.

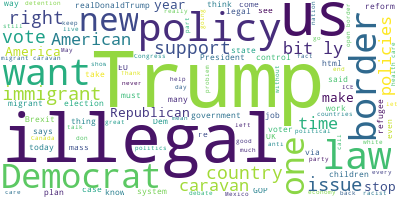
**4.2 Final Naive Bayes Model for Multiple Months**

As Naive Bayes showed itself to be the most accurate, efficient, and least prone to error of the bunch, it was chosen to continue the work with the rest of the datasets. This meant the validation techniques would turn into four-cross fold validation, with each month being trained by the other months in full dataset. The results of this process are shown in the figure and table below. The initial prediction ended being a good indicator of what the rest of the predicted values would look like, with the average accuracy of the Naive Bayes model used for this project being about 66.35%, around .5% from the initial predicted value. While no large shifts can be seen in the data gathered from the model, small indicators show that the positive sentiment polarity started to slowly shift downwards and negative polarity slightly spiked up a bit around October and December.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Month | Total Accuracy | Predicted Positive | Predicted Negative | Predicted Neutral |
| September | 66.87% | 58.79% | 25.70% | 15.51% |
| October | 65.51% | 56.66% | 30.38% | 13.07% |
| November | 67.22% | 56.71% | 28.68% | 14.61% |
| December | 65.78% | 55.10% | 31.91% | 13.00% |
|  |  |  |  |  |
| Average | 66.35% | ~56.82% | ~29.17% | ~14.05% |



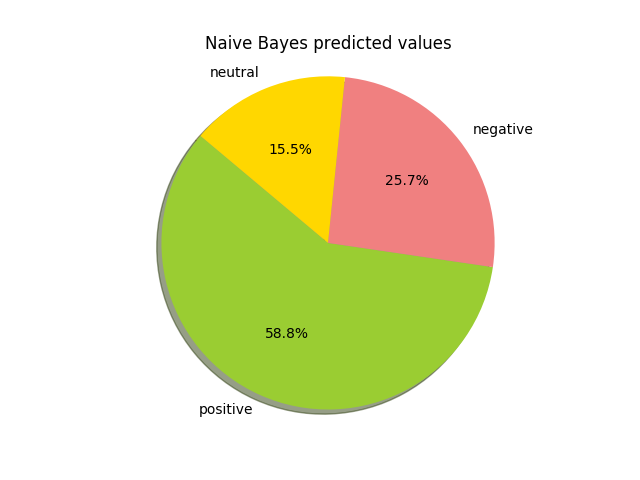
To look more closely at these shifts, I incorporated word clouds that look at the most used words for the Tweets of these four months. Two of these are shown below, with figure 3 being the word cloud of September, and figure 4 being the word cloud of October, as these show the changes between what is being talked about quite decently. With the September one, we can see that Donald Trump is probably the most talked about person on Twitter when it comes to immigration, but it also seems that a new policy or law was passed and was talked about just as much or more. By just googling “September 2018, immigration new policy”, we are able to see that these tweets are probably talking about the changes in policy issued by the United States Citizenship and Immigration services that allowed the group to deny applications and requests without notifying the applicants. By just showing this one collection of words, a person is able to quickly grasp the whys’ for the reason people are talking and tweeting. This is further supported with the one for October, where the new policies and laws are talked about less, and more discussion is about illegal immigrants and voting. Below the massive “illegal” you can see the word “caravan” which suggests that this was around the time Trump started to talk about the caravan coming up through Mexico to the border. Meanwhile on the left hand side, vote has become more prominent as elections are held the next month in November 2018 and there's a surge of discussion about Democrats, while there’s no discussion at all about Republicans. This could explain the 5% spike in negative sentiment between September and October, where the discussion becomes more volatile around the fact there’s a caravan heading to the border of the United States and that elections are coming up to try to change things in some way shape or form. The other two belonging to November and December can both be found in the appendices at the end of this paper.

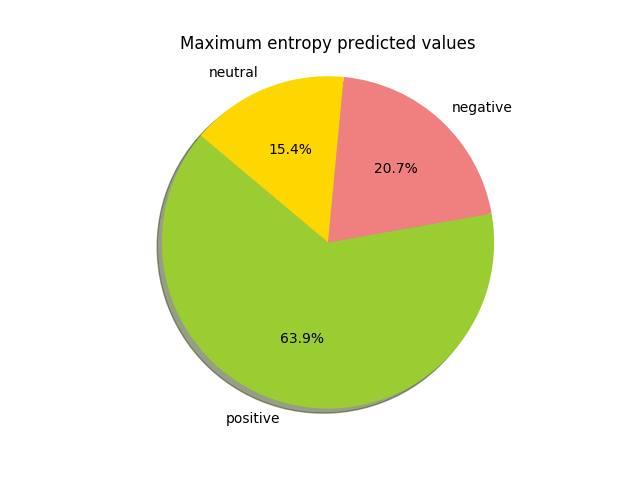


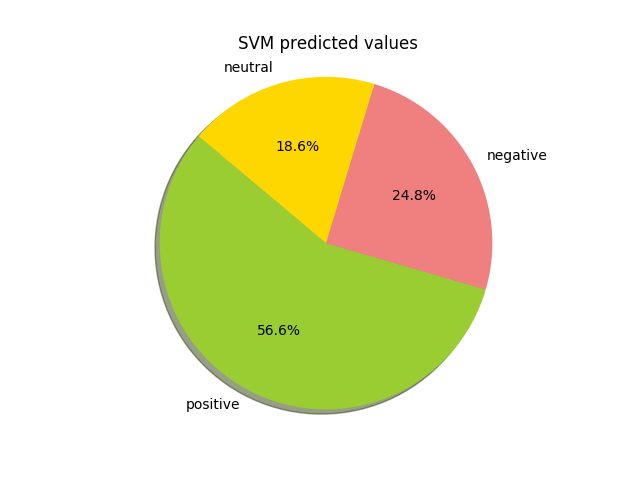
**5.0 Conclusion**

As what was just shown, Naive Bayes seemed like the best model for the datasets monitored of changing sentiment polarity overtime regarding immigration. That being said, the limitations encountered with the two other models without a doubt hindered them, but even Naive Bayes model had limitations with large testing sets, and its issue was the easiest to get around. So even if the models could show more accurate results in the long run, their limitations and overall complexity would make obtaining the goal quite a bit longer. As for trending sentiment polarity, while there were no monitored big shifts in the polarity of the sentiment values, small shifts and spikes can be seen and data can be extrapolated from the text of the tweets themselves to give context. In more volatile datasets where sentiment shifts quickly, the methods done for this project should be equally applicable and be able to extrapolate correlations between the words written and the sentiment values of Twitter.

Beyond just that, this research can be extended to companies in allowing them to see how their product is viewed and why it is being discussed in the manner it is. This provides invaluable insight for businesses, allowing them to improve their product’s image in the most effective way that can be found out. Also the models themselves can be improved upon, allowing more memory and time for execution for more accurate results and analysis to be obtained. There is also a growth of knowledge to be obtained from the mixing of sentiment analysis research and the word clouds of the topics, showing how things are discussed and why. These clouds allow people to immediately glance at a bunch of words and understand the conversations being made of that time. In that way, this research can be used to analyze tweets and other posts from other websites, expanding the data that can be analyzed and gaining more insights to the overall sentiment polarity of the yesteryears. For stronger analytical methods, neural networks can be considered to expand upon the linear classification field along with SVMs. The goal of this research overall was less about the sentiment values of a specific month or week, but to see if that sentiment changes over time and the reasons for the change. Because of this, this research is relevant beyond just its comparisons of sentiment classification techniques, and is able to be relevant for multiple fields due to its overarching implications in understanding how the datasets came about. In this way, I hope the work done here is continued by others to create better correlations between sentiment and words, and allow people to gain more insight to the ever changing polarity of sentiment for every topic and discussion.

**Appendices**









**References**

Jianqiang, Z.& Xiaolin, G. (2017). "Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis," in *IEEE Access*, vol. 5, pp. 2870-2879.

Kaur, H., Mangat, V., & Nidhi. (2017). "A survey of sentiment analysis techniques," *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pp. 921-925.

Medhat, W., Hassan, A., & Korashy H. (2014). “Sentiment analysis algorithms and applications: A survey,”. Ain Shams Engineering Journal, Volume 5, Issue 4, Pages 1093-1113.

Moise, I. (2016) The technical hashtag in Twitter data: A hadoop experience. *2016 IEEE International Conference on Big Data (Big Data)*, pp. 3519-3528.

Rana, S. & Singh, A. (2016). "Comparative analysis of sentiment orientation using SVM and Naive Bayes techniques," *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, pp. 106-111.

Rane, A.& Kumar, A. (2018). "Sentiment Classification System of Twitter Data for US Airline Service Analysis," *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, pp. 769-773.

Shaikh, F., Rangrez, F., Khan, A., & Shaikh, U. (2017). Social media analytics based on big data. *2017 International Conference on Intelligent Computing and Control (I2C2)*, Coimbatore, 2017, pp. 1-6.

Unnikrishnan, A., Narayanan, U. & Joseph, S. (2017). Performance analysis of various supervised algorithms on big data, *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, pp. 2293-2298.

Vanaja, S. & Belwal, M. (2018)."Aspect-Level Sentiment Analysis on E-Commerce Data," *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)*, pp. 1275-1279.