Utilizing Quantitative Methods to Analyze Negative Sentiments in the United States.

By Lara Yaldo

Instructors: M. Al Hamando

Term: Spring 2017

Lawrence Technological University

31000 West Ten Mile Road,

Southfield, MI

Abstract

Sentiment analysis using quantitative methods and mathematical algorithms can classify humans' opinions and feelings. Those sentiments can be harassment, happiness, sadness, or even neutral. The purpose of this paper is to analyze and understand the negative sentiments usage on social media, specifically twitter, in the United States. The study is conducted using two data streams from twitter API. The first stream contained 22,732 tweets used to create the training model, and the second stream contained 557,870 tweets which were used as the test model for the training model created by the first stream. A rating system was built to analyze and rate the training model. Once the tweets in the training model were rated with a certain sentiment, five different classification algorithms were used against the rated tweets. The highest results for Cohen's Kappa was ~0.54; while the highest correctly classified percentage was at 72%. The mathematical algorithms utilize a Natural Language Processing method of creating a bag of words from the tweets. This paper studies the prediction of the sentiments used in the tweets from the test model in relation to the geographical location where the original tweet was created.

Keywords: sentiments, machine learning, quantitative methods, emotions, Natural Language Processing, rating system, Cohen's Kappa

Utilizing Quantitative Methods to Analyze Negative Sentiments in the United States.

Social Media, such as Facebook, Twitter, Instagram, etc., have created a new way of self-expression. Many Natural Language Processing systems use machine learning algorithms to define the sentiments of the writer using different methods including dictionary words. The purpose of this study is to analyze tweets generated from the United States using a rating system that analyzes expressions to define how many negative and abusive tweets have been used thought out the stream, without taking into consideration the United States population. The data used for this analysis will be taken from Twitter API streaming functionality. To be able to analyze the twitter data, a rating system is built to rate tweets with eleven different opinions. Those sentiments include five unpleasant opinions, five pleasant opinions and one neutral opinion.

Background & Literature Survey

Understanding self-expression on social media networks is very important since humans tend to express their opinions harshly to one another without thinking about the consequence and harm they could cause to other people. Some of those sentiments are defined as cyber bullying or abusive language. Bullying and attacking language are used widely starting from early ages of preschool to bullying in the work place. Many studies have been carried out to understand users' sentiments on social media and to be able to classify the sentiment using Natural Language Processing. Sentiments recognition have been one of the topics that researchers have been studying for the last decade or so. Retrieving data from social network sites, such as twitter's big data, demonstrates an example of how users use social networking sites to state their emotions to certain subjects, studied using Multinomial Naïve Bayes algorithm (Wang, Chen, Thirunarayan & Sheth, 2012).

Twitter raw data can be used as relevant information through Natural Language Processing algorithms and sentiment analysis to study certain subjects (Khanaferov & Wang, 2014). Researches have shown that sentiments, negative specifically, spreads faster through social networks that affects human psychology and physical health to negative perspective (AlSagri & Ykhlef, 2016). Not only does negative sentiments spreads in a great speed, those sentiments are also used in cyberbullying, which cannot be determined easily. Teenagers and young adults have used social media negatively to bully others and with the use of decision tree algorithm many studies could identify cyberbullying languages used in many site by teenagers (Reynolds, Kontostathis & Edwards, 2011). Another way used to determine and study cyberbullying is using natural language processing and pattern recognition strategy, in which researchers could identify cyberbullying languages (Dinakar, Jones, Havasi, Lieberman & Picard, 2012).

The use of clustering to identify negative sentiments are applied to identify online communities that can affects the negative thoughts. Some researchers find it very useful and necessary to get the users' information such as gender, habits over social media and general emotion mode (Roshanaei, Han & Mishra, 2015). Utilizing clustering strategy evaluates the speed of the emotion transfer, and at the same time using the recommended system strategy, positive sentiments can be applied through the same community to eliminate negative outcomes (AlSagri & Ykhlef, 2016). Researchers have also created and used many text analysis strategies and algorithms to be able to navigate through negative or positive sentiments in social media for "opinion mining;" however, long posts and blocks written by users are still a problem to analyze through mathematical computations (Chen & Zimbra, 2010).

This research is conducted to understand the usage of sentiments in the United States and to underline abusive, negative, anger sentiments in social media. The scope of this study is emphasized on the sentiments of the tweets and the location the tweet was generated; therefore, certain aspects of each tweet will be ignored, such the number of followers and retweets. Also, some tweets did not have a geographical location listed on their account, which results in elimination of those tweets from this research. For the sentiments to be predicted accurately, a Natural Language Processing functionality will be used to rate the sentiment of each tweet, which will be followed by classification algorithms to create the models.

Data

The data used for this research include two streams from Twitter API, which is available from Twitter for development purposes. Both streams were location specific to the United States. The streams occurred during two different dates. The first stream was on Fri Feb 17, 2017, and the second stream was on Tue March 14, 2017. The first stream included ~97,792 tweets in which only 22,732 tweets were used for the rating system with a geographical location listed on the tweets. The purpose of the first stream is to create a training set for sentiments. The second stream included ~620,000 tweets which hit Twitter API stream limit. The API limit is defined on Twitter development site of "180 calls every 15 minutes." For every 180 tweets streamed, the API breaks for 15 minutes than streams tweets for another 15 minutes, or 180 tweets*. After cleaning the ~620,000 tweet, only 557,871 tweets will be used as test set. Figure 1 shows an

*Please see twitter API rules: https://dev.twitter.com/rest/public/rate-limiting

example of a streamed tweeted.


```
{"created_at":"Fri Feb 17 04:24:38 +0000
2017","id":832445629414592514,"id_str":"832445629414592514","text":"@realdonaldtrump 379-304=75. that's the biggest
margin.Shall I do the least or do you get the idea.Not all USA are cult -ivatible, donnieboy","source":"\u003ca
href=\"http:\/\/twitter.com\" rel=\"nofollow\"\u003eTwitter Web Client\u003c\/a
\u003e","truncated":false,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":25073
877,"in_reply_to_user_id_str":"25073877","in_reply_to_screen_name":"realDonaldTrump","user":
{"id":170296342,"id_str":"170296342","name":"LaNuevaMejicana","screen_name":"Uhuruhsiempre","location":"Southwester
n USA. NM. Arizona.","url":null,"description":"Abuse of power is the root of all evil and despair is the utlimate
sin.","protected":false,"verified":false,"followers_count":6753,"friends_count":7139,"listed_count":128,"favourites
_count":31055,"statuses_count":108016,"created_at":"Sat Jul 24 12:57:14 +0000
2010","utc_offset":-25200,"time_zone":"Mountain Time (US &
Canada)","geo_enabled":true,"lang":"en","contributors_enabled":false,"is_translator":false,"profile_background_colo
r":"000000","profile_background_image_url":"http:\/\/abs.twimg.com\/images\/themes\/theme18\/
bg.gif","profile_background_image_url_https":"https:\/\/abs.twimg.com\/images\/themes\/theme18\/
bg.gif","profile_background_tile":false,"profile_link_color":"3B94D9","profile_sidebar_border_color":"000000","prof
ile_sidebar_fill_color":"000000","profile_text_color":"000000","profile_use_background_image":false,"profile_image_
url":"http:\/\/pbs.twimg.com\/profile_images\/830428479128641536\/
YHb20jOH_normal.jpg","profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/830428479128641536\/
YHb20jOH_normal.jpg","profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners
\/170296342\/1486769544","default_profile":false,"default_profile_image":false,"following":null,"follow_request_sen
t":null,"notifications":null},"geo":null,"coordinates":null,"place":{"id":"813a485b26b8dae2","url":"https:\/\/
api.twitter.com\/1.1\/geo\/id
\/813a485b26b8dae2.json","place_type":"city","name":"Albuquerque","full_name":"Albuquerque,
NM","country_code":"US","country":"United States","bounding_box":{"type":"Polygon","coordinates":
[[[-106.791691,35.015891],[-106.791691,35.218114],[-106.473745,35.218114],[-106.473745,35.015891]]]},"attributes":
{}},"contributors":null,"is_quote_status":false,"retweet_count":0,"favorite_count":0,"entities":{"hashtags":
[],"urls":[],"user_mentions":[{"screen_name":"realDonaldTrump","name":"Donald J.
Trump","id":25073877,"id_str":"25073877","indices":[0,16]}],"symbols":
[]},"favorited":false,"retweeted":false,"filter_level":"low","lang":"en","timestamp_ms":"1487305478801"}
```


Figure 1: An example of a single tweet streamed using Twitter API

Although tweets streaming was based on the location, some of the tweets were streamed without city or state, and the others included tweets from Canada and Mexico. Those tweets were removed from the data file to be able to accurately locate each tweet in the United States. Also, few tweets did not include geo location, and as a result, the tweets were removed from the analysis study. Five specific attributes were parsed from each tweet. Table 1 shows the five attributes and their descriptions.

Attributes Name	Description
Screen Name	Displays the tweeter user screen name used for their account.
Date	Shows the date when the user tweeted.
Text	Displays the text of the tweet.
City	City Name in the USA.
State	State name in the USA where the geo points to.

Table 1: Five attributes collected from each tweet.

Since the attributes are part of the tweets, the parsing method was easily done using CRAN R and Python programing. The only attribute missing from the list is the Rating attribute. This attribute is generated later once the Rating System. An example of how the data table would look like is shown in table 2.

ScreenName	Date	Text	City	State
cassydieberryy	Fri Feb 17 04:40:06 2017	i miss noaboa11 wtf	Tulsa	OK
casillas_shawn	Fri Feb 17 04:38:42 2017	If you feed them they will come great Thursday night Refuge Youth youthpastorlife https t co rdHVUzWbdg	tulsa	OK
AndreaRusherR5	Fri Feb 17 04:56:35 2017	Patz DT Why won t you ever let me leave just fucking let me out Everytime that I m out of reach she says turn around	YumaCity	AZ
Erik_Streetwise	Fri Feb 17 04:52:11 2017	She so Fucking mean to me come on what did I do https t co VqKHiPAbtA	Tulare	CA
eintdave	Fri Feb 17 04:39:26 2017	NormEisen I apologize to the Czech Republic for having such a bitter moron like norm as ambASSador	Upland	CA
Ferraribоots14	Fri Feb 17 04:28:57 2017	I will always retweet this https t co M0Vn0di6Yc	Vallejo	CA
xAnthonytejedax	Fri Feb 17 04:47:01 2017	ElChefMJ littlecaesars true but I got half pepperoni half cheese	Ventura	CA

Table 2: Example of Parsed Tweets

Method

A rating system was created to define the sentiment for each tweet. Once the tweets had sentiments rating, five different classification algorithms are used to create a train model. The machine learning algorithm that would have the most accurate results will be used to predict the sentiment for the testing model. For using GUI and command line job submissions, Weka, data mining and machine learning software, was used to run the analysis for the training model and test model (Frank, Hall & Witten, 2016).

Using Rating System

The rating system is built using Windows Form and C# programing language. The purpose of the rating system is to analyze the tweets' text and identify the sentiments behind the tweet. Utilizing Part-of-speech tagging process (Brill, 1992), which used to be part of Natural Language Processing, each tweet's text is divided into few parts according to the part of speech. The rating system uses eleven sentiment classifications, which are neutral, happy, abusive, negative, sad, love, depressed, good positive, open and anger. To use the rating system, the program can be executed through visual studio using binary code, or through the released executable. Once the program is started, user will be required to open the data file. Then, the

sentiments classification will start once the rating button is clicked. Figure 2 shows a sample of how the rating system is designed from the GUI stand point.

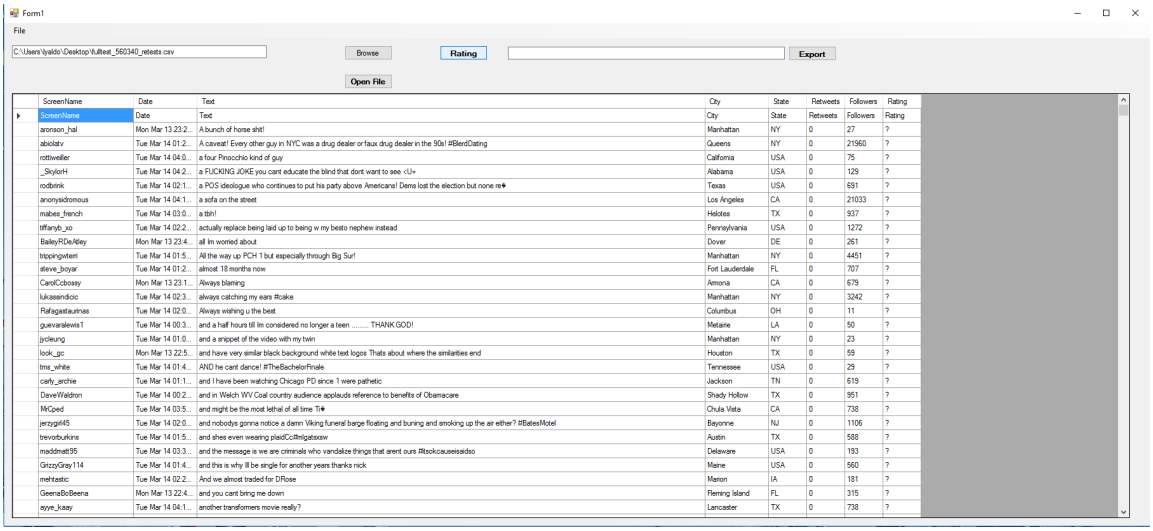

Figure 2: a sample picture of the Rating System using Windows Form and C#

Once the rating is completed, the data can be saved to another csv file with the rating information. After the data is populated, then the user would click the Rating button on the GUI, which will display the sentiments depending on the data analysis part. The export button will allow the users to export the data after the rating to a csv file.

Analysis Methods

Five machine learning algorithms will be used to be able to predict accurate results for the test set. Some of the mathematical methods used to predict the sentiments are K-nearest Neighbor, Bayes Network, 1R, Decision Stump, and Locally Weighted Learning. The K-nearest Neighbor uses the closest neighbor to predict the sentiments (Aha, Kibler & Albert, 1991). For this study, K is set to the number 1, which means to look only for one closest neighbor. Another classifier used is Bayesian Network, which uses probability distribution for every case and define cases that closer to the original one to represent its parent (Bouckaert, 2004). 1R is also used to run the training model, which uses the attribute with the least error to make an accurate prediction (Holte, 1993). Another classifier algorithm is decision stump, which classifies data depending on the quantitative measurement

used to predict data (Kudo & Matsumoto, 2004). An additional classification method used for sentiment analysis is Locally Weighted Learner, a lazy learner, that creates a weight for each node and create a calculation according to the weight (Atkeson, Moore, & Schaal, 1997).

Results

The rating system results for the training model shows that ~8,700 from the 22,732 total tweets are classified to the abusive and harassment sentiment. While ~9700 tweets are neutral. Table 3 shows the number of tweets that are predicted for each sentiment.

Sentiment	Number of Tweets corresponding to the sentiment using the Rating System.
Neutral	9753
Abusive	8730
Negative	1987
Happy	773
Love	535
Sad	347
Depressed	203
Good	149
Open	144
Anger	63
Positive	48

Table 3: Results of the sentiments from the rating system.

The training model is created by using five different machine learning algorithms. The algorithm with the most accuracy results in sentiment rating will be used to predict the sentiments for the 557,871 tweets. Bayes Network has the highest accuracy for the sentiments analysis with a result of 72% correctly classified sentiments. It predicted that 16,396 tweets are rated correctly from the total of 22,732 tweets. The second highest classification result was the K-nearest neighbor with an accuracy percentage of 71% and a total of 16,333 tweets correctly rated out of 22,732 total. On the other hand, 1R classification has the lowest correctly classified

results with a 47% accuracy. That included 10,765 tweets that were correctly classified. Figure 3 shows the correctly classified results for all five machine learning algorithms used.

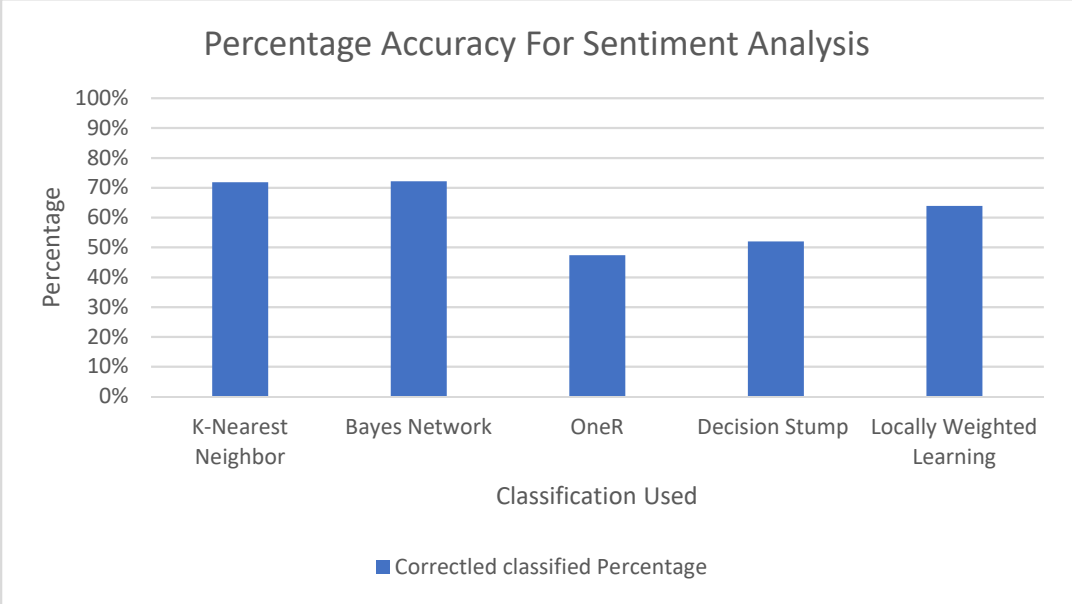

Figure 3: Correctly Classified Tweets Results.

In addition to the correctly classified results, both Bayesian Network and the K-Nearest Neighbor classifications has the result of 0.5452 for Cohen's Kappa, which means that there is a slight agreement when running the data against itself using the methodology of leave-one-out. On the other hand, the Cohen's Kappa for 1R has the lowest agreement between the five algorithms of 0.1166. Figure 4 shows the results of Cohen' Kappa.

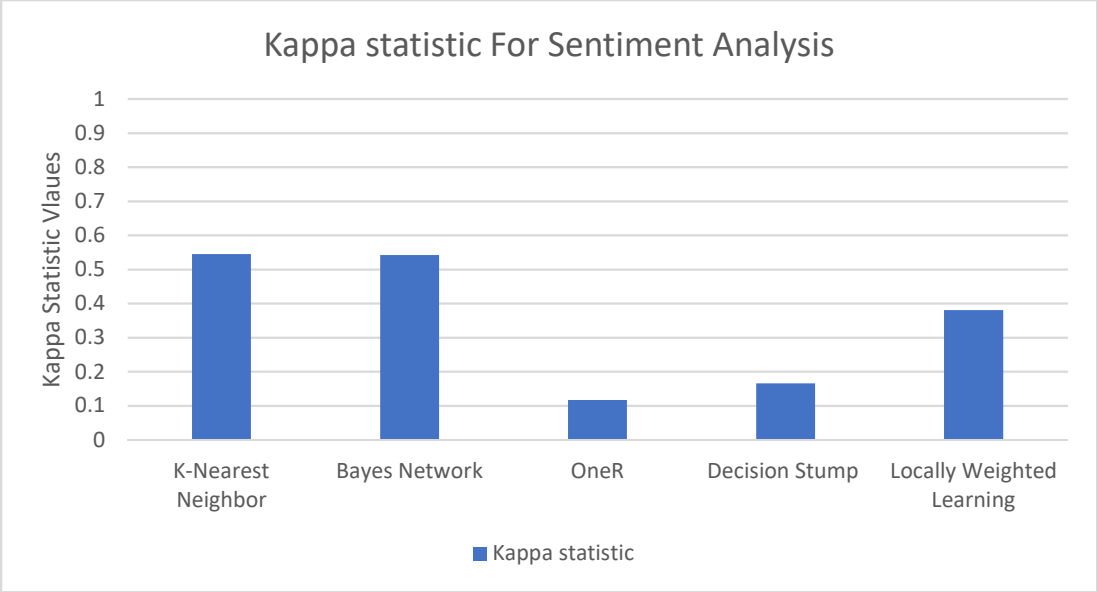

Figure 4: Kappa Statistics Results.

Looking at the mean absolute error for all five algorithms, the K-Nearest Neighbor has the lowest value comparing to the Decision stump which has the highest value. Figure 5 shows the results for the mean absolute error for all five algorithms.

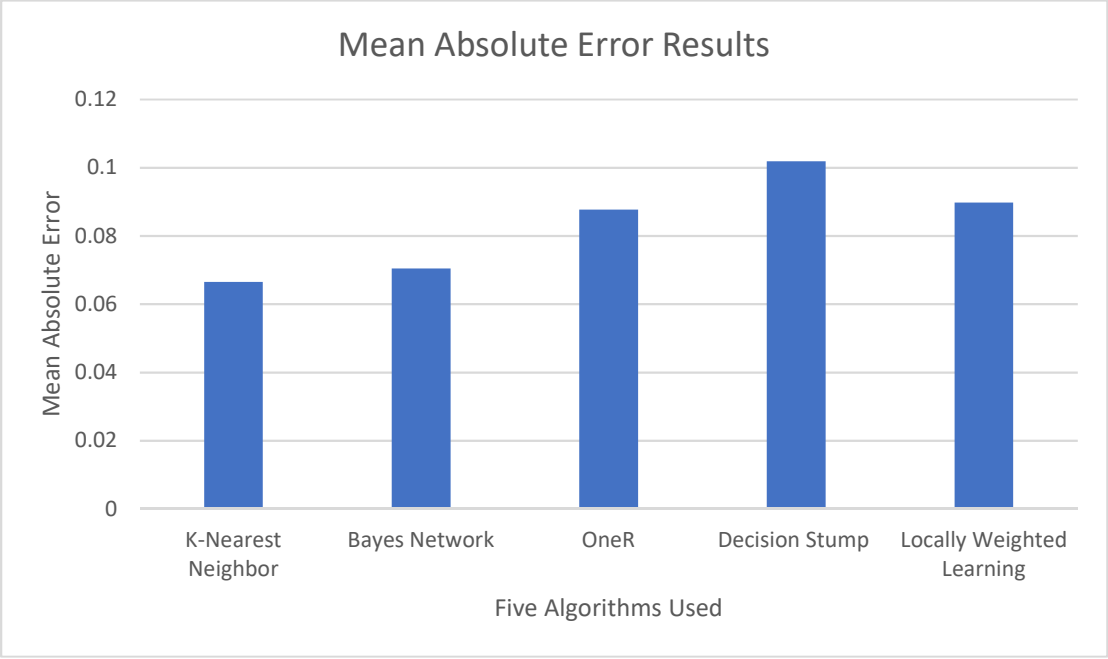

Figure 5: Mean Absolute Error Results.

Table 4 shows the actual and predicted results using Bayesian Network used for training.

actual	predicted	ScreenName	Text	City/State	Date Tweeted
7:depressed	1:neutral	Shabba6Ranks	Lightskin women should only be eye candythem fuckers terrible	Miami, FL	Fri Feb 17 04:41:38 2017
5:sad	1:neutral	bamabev79	neta America loves you So sorry you had to deal with 8 yrs of a POTUS who did not reflect that We re in much bett	Enterprise, AL	Fri Feb 17 04:26:51 2017
7:depressed	6:love	jamiemgiller	katiehawk my heart sank reading this news earlier today I m so sorry for your loss and hope you relish in beautiful memories 1 2	Miami Beach, FL	Fri Feb 17 04:55:21 2017
4:negative	5:sad	MikeBellATL	Just learned Cooper at Carnevino in Vegas lost his Father So sorry my brother Thoughts and prayers love you Man	Atlanta, GA	Fri Feb 17 04:36:34 2017
2:happy	2:happy	AngelaCribben17	YodaForces I m glad you are enjoying all the posts gt lt 3	Chicago, IL	Fri Feb 17 04:40:29 2017
3:abusive	3:abusive	aaron_seabooty	BustosBella People who fuck you over then show no remorse are the WORST	Casa Grande,AZ	Fri Feb 17 04:54:18 2017
3:abusive	3:abusive	Lokis_FanGirl	GeorgeTakei This better be the last reality TV star pussy grabbing blathering idiot of a Russian puppet we ever elect Cuz I m not d	Pittsburgh,PA	Fri Feb 17 04:34:37 2017
3:abusive	3:abusive	itsquayvo	whoever heavy footed ass keep running down this hallway please go to bed thank you	Troy,AL	Fri Feb 17 04:44:10 2017

Table 4: Results From Bayesian Network Training Model

Once the prediction model was used on the 557,871 tweets, most of the prediction was neutral. Table 5 displays an example of the predicted sentiment with the corresponding twitter account screen name.

ScreenName	Date	Text	City	State	Rating
Javiercousteau	Mon Mar 13 22:40:43 2017	Always trynna make me eat. let me get my work done and die first sheesh.	Washington	DC	3:abusive
tina_press	Mon Mar 13 22:40:43 2017	Happy Birthday Keegs! Have a great day	Cottage Grove	MN	2:happy
ultabeautyjobs	Mon Mar 13 22:40:43 2017	Interested in a in IL? This could be a great fit	Bolingbrook	IL	2:happy
ACCOC	Mon Mar 13 22:40:43 2017	Thank you Community Development Director Ben Metcalf for leading our housing affordability discussion!ï¿½	Sacramento	CA	2:happy
Cherwood4	Mon Mar 13 22:40:43 2017	You have destroyed what little trust the people had of you. You just dont go around accusing people of wiretapping!!	McAlester	OK	2:happy
paranoidliotta	Mon Mar 13 22:40:43 2017	Hey Fuckhead? Why dont you advertise your cat euthanasia services so people can thankyou in person?	Los Angeles	CA	1:neutral
agrullon16	Mon Mar 13 22:40:43 2017	accurate	South Hackensack	NJ	1:neutral

Table 5: Predicted Sentiments for the test data

The quantitative method assigns a number to each sentiment as showing above to be able to predict accurately. Although, sentiment analysis for sarcasm have yet to be identified correctly. This research has shown negative sentiments is used widely in the United States, and many people express their happiness, sadness and even aggressiveness using vulgar language. Table 6 shows the results from the test for the prediction of the Bayes Network.

Type of Sentiment	**Number of Tweets**
Abusive	62594
Neutral	375788
Open	4785
Negative	30442
Positive	4831
Happy	34218
Sad	9830
Depressed	4326
Anger	8576
Love	17467
Good	5013

Table 6: Results of Bayes Network on 557,870 tweets

More than ~62,000 uses abusive and harsh language against other people, which mean that 11% of the streamed tweets are considered abusive sentiment. Although, 11% is not considered high, but in social media, it might affect millions. The results show that abusive sentiments are more used over twitter in the East Coast states. The US map in figure 3 shows the outcome results of the sentiment prediction for the United States.

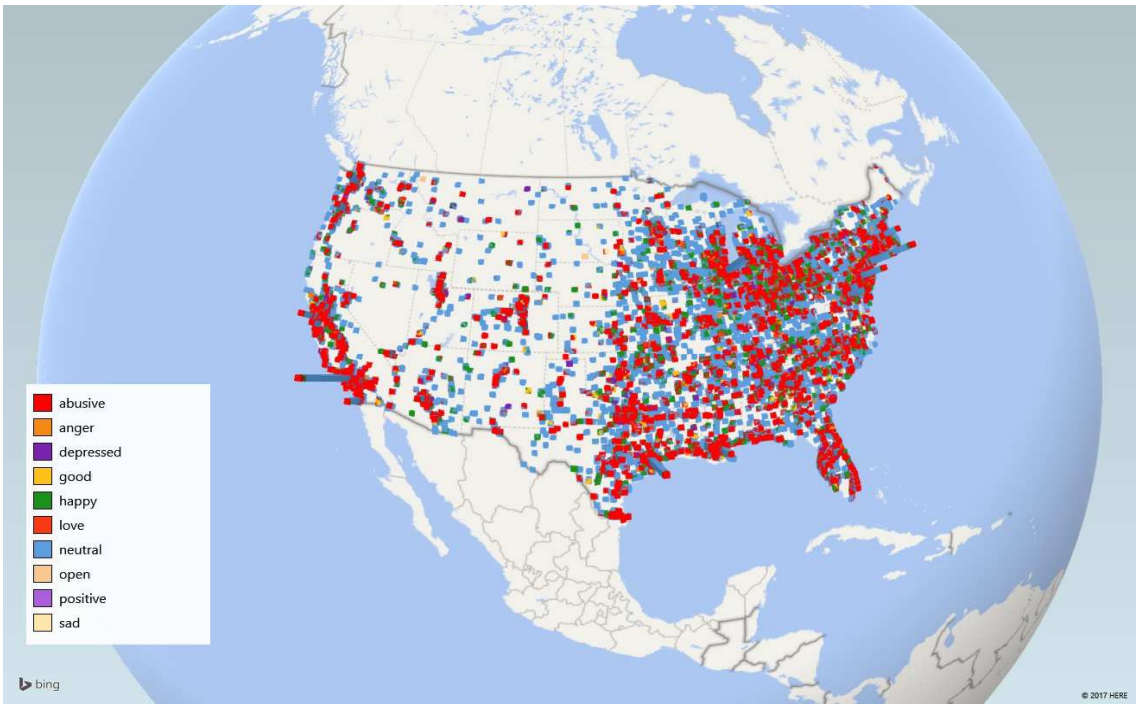

Figure 6: US Map for Sentiment Prediction.

Conclusion

The research shows that the use of harassments, negative sentiments and cyberbullying in social media are distributed widely around the United States. Although some locations use more abusive sentiments than others, the harassment sentiments are still used in large in the United States. Even though most of the tweets from the ~550,000 have a sizable number of neutral and positive sentiments, the negative sentiments spread faster and have a greater effect on people. In the future study, the sentiments analysis would be more specific to each state individually to identify the states with the most abusive language with respect to their population.

References:

Wang, W., Chen, L., Thirunarayan, K., & Sheth, A. P. (2012, September). Harnessing twitter" big data" for automatic emotion identification. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom)* (pp. 587-592). IEEE.

Khanaferov, D., Luc, C., & Wang, T. (2014, June). Social network data mining using natural language processing and density based clustering. In *Semantic Computing (ICSC), 2014 IEEE International Conference on* (pp. 250-251). IEEE.

AlSagri, H. S., & Ykhlef, M. (2016, April). A framework for analyzing and detracting negative emotional contagion in online social networks. In *2016 7th International Conference on Information and Communication Systems (ICICS)* (pp. 115-120). IEEE.

Roshanaei, Mahnaz, Richard Han, and Shivakant Mishra. "Features for mood prediction in social media." *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2015.

Chen, H., & Zimbra, D. (2010). AI and opinion mining. *IEEE Intelligent Systems*, *25*(3), 74-80.

Reynolds, K., Kontostathis, A., & Edwards, L. (2011, December). Using machine learning to detect cyberbullying. In *Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on*(Vol. 2, pp. 241-244). IEEE.

Dinakar, K., Jones, B., Havasi, C., Lieberman, H., & Picard, R. (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*,*2*(3), 18.

Csurka, G., Dance, C., Fan, L., Willamowski, J., & Bray, C. (2004, May). Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV* (Vol. 1, No. 1-22, pp. 1-2).

Vayssières, M. P., Plant, R. E., & Allen-Diaz, B. H. (2000). Classification trees: An alternative non-parametric approach for predicting species distributions. *Journal of vegetation science*, *11*(5), 679-694.

Kahle, D., & Wickham, H. (2013). ggmap: Spatial Visualization with ggplot2.*R Journal*, *5*(1).

Frank, Eibe, Hall, Mark A., & Witten, Ian H. (2016). The WEKA Workbench. Online Appendix for "*Data Mining: Practical Machine Learning Tools and Techniques*", Morgan Kaufmann, Fourth Edition, 2016.

Brill, E. (1992, February). A simple rule-based part of speech tagger. In *Proceedings of the workshop on Speech and Natural Language* (pp. 112-116). Association for Computational Linguistics.

Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. *Machine learning*, *6*(1), 37-66.

Bouckaert, R. R. (2004). *Bayesian network classifiers in weka*. Hamilton: Department of Computer Science, University of Waikato.

Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine learning*, *11*(1), 63-90.

Kudo, T., & Matsumoto, Y. (2004, June). A Boosting Algorithm for Classification of Semi-Structured Text. In *EMNLP* (Vol. 4, pp. 301-308).

Atkeson, C. G., Moore, A. W., & Schaal, S. (1997). Locally weighted learning for control. In *Lazy learning* (pp. 75-113). Springer Netherlands.