The Journal of Social Media in Society
Spring 2019, Vol. 8, No. 1, Page 211-230
thejsms.org

Power of Predictive Analytics: Using Emotion Classification of Twitter Data for Predicting the 2016 U.S. Presidential Elections

Satish M. Srinivasan[1], Raghvinder S. Sangwan[1], Colin J. Neill[1], and Tianhai Zu[2*]
[1]Department of Engineering, Penn State Great Valley, Malvern, PA, 19355
[2]Department of Operations, Business Analytics, and Information Systems (OBASIS), University of Cincinnati, Cincinnati, OH, 45221
*Corresponding Author: sus64@psu.edu, 610-427-9288

Predictive analytics using the Twitter feeds is becoming a popular field for research. A tweet holds a wealth of information on how an individual express and communicates their feelings and emotions within their social network. Large-scale collection, cleaning, and mining of tweets will not only help in capturing an individual's emotion but also the emotions of a larger group. However, capturing a large volume of tweets and identifying the emotions expressed in it is a challenging task. Different classification algorithms employed in the past for classifying emotions have resulted in low-to-moderate accuracies thus making it difficult to precisely predict the outcome of an event. In this study, we demonstrate the potentiality of a lexicon-based classifier, *NRC*, which can mine emotions and sentiments in tweets. Using the *NRC* classifier, we initially determined the emotions and the sentiments within the tweets and used that to predict the swing direction of the 19 US states towards the candidates of the 2016 US presidential election. Comparing the predictions from the *NRC* against with the actual outcome of the election, we observed a ~90% accuracy, a performance superior to the mainstream pollsters indicating the potential emotion and sentiment-based classification holds in predicting the outcome of significant social and political events.

Keywords: *machine learning, emotion classification, lexicon-based classifier, predictive analytics, social media, Twitter*

The advent of social media and microblogging sites have paved the path for individuals and communities to freely express their opinions, feelings, and thoughts on a variety of topics in the form of short and limited size texts. A commonly known social media site is Twitter through which short messages (*a.k.a. tweets*) can be posted by individuals. These tweets with a 140-character limitation hold a wealth of information on how individuals communicate their thoughts, emotions (happiness, anxiety, depression etc.) and feelings within their social network. Not only the emotions of individuals, but the emotions of larger groups (such as a certain country,

state, community, etc.) can also be identified by analyzing these tweets. Twitter houses billions of tweets which can serve as a rich ensemble of emotions, sentiments, and moods (Hasan, Rundensteiner, & Agu, 2014). For example, the tweet "*I felt quite happy and lighthearted; I put on the shoes and danced and jumped about in them*" expresses a happy mood and the tweet "*I left it but throughout the whole day I was really awful*" expresses sadness. Unlike conventional text, however, tweets are peculiar in nature due to their inherent structure and size making the determination of emotions of an individual or for larger group a challenging task. Additionally, since more than one emotion can be expressed in a tweet, emotion classification is considered more complex because a single text can be annotated with multiple different emotion classes.

In this research, the focus is on automatically detecting and classifying the emotions expressed within the tweets. The approach taken here will allow determining the emotions hidden in these short messages submitted around an event of interest and predict the outcome of that event. The goal of this study is twofold. First, we want to demonstrate that social media data, i.e. tweets, have the potentiality of predicting the outcome of an event if the emotions of an individual in those tweets can be properly determined, and second, we want to demonstrate the potentiality of a lexicon-based classifier, namely NRC, for emotion and sentiment classification.

LITERATURE REVIEW

Sentiment analysis and emotion classification has attracted much research during the last decade. One of the reasons for this increase can be attributed to the growing amount of opinion-rich text corpus being available due to the development of social media, giving researchers access to the opinions of the people. Another important reason for the increased interest in sentiment and emotion classification is the advances that have been made within the fields of natural language processing and machine learning. Peng, Lee, and Vaithyanathan (2002) have shown that an accuracy of 80% is achievable on a well-balanced dataset for the problem of classifying movie reviews as positive or negative. Several other studies have utilized the machine learning techniques on Twitter datasets to distinguish between positive and negative classes with accuracies ranging between 60% and 80% (Barbosa & Feng, 2010; Pak & Paroubek, 2010). Using Western-style emoticons

Go et al. (2009) have labeled and classified Twitter messages as positive and negative sentiment. Using different classification techniques including Naive Bayes, Maximum Entropy, and SVM they have reported an accuracy of 80% on their dataset collected from Twitter. Thelwall, Buckley, Platoglou, and Kappas (2010) have developed an application SentiStrength that utilizes machine learning approaches to extract the strength of the sentiments hidden in short informal text. They have reported that their applications can classify the positive sentiment with an accuracy of 60% and the negative sentiment with an accuracy of 72%.

In contrast to the sentiment analysis studies, Brynielsson et al. (2014) have looked in to another class of problems known as emotion classification. They collected tweets related to hurricane Sandy and tried to classify them into four distinct classes of emotion namely positive, fear, anger and others. Out of the two classifiers Support Vector Machine (SVM) and the Naïve Bayes (NB), they claim that the SVM classifier yielded the best classification accuracy which is close to 60%. Danisman and Alpkocak (2008) have proposed a Vector Space Model (VSM) based approach titled Feeler using which they were able to automatically classify the ISEAR (International Survey on Emotion Antecedents and Reactions) dataset into 5 emotion classes namely anger, disgust, fear, joy and sad. They have reported an overall accuracy of 67.4% using NB and an accuracy of 66.9% using SVM. The reported classification accuracies are based on the 10-fold cross validation technique on the stammered ISEAR dataset. Their observations also suggest that the VSM classifiers are as good as the NB and the SVM classifiers. Hasan et al. (2014) have proposed *EMOTEX* that employs different supervised classifiers to detect emotions in text messages. Using supervised classifiers NB, SVM, Decision trees, and KNN (*k*-Nearest Neighbour), they were able to demonstrate approximately 90% precision for a four-class model on the collected tweet dataset. In their studies they have incorporated many types of features that include the unigram, unigram emoticon, unigram punctuation, and the unigram negation. They have also reported a 90% classification accuracy on a larger tweet dataset using the supervised classifiers KNN and SVM. Choudhury, Gamon, Counts, and Horvitz (2013) have tried to classify the tweets that were posted by individuals with an onset of depression. Upon performing a 10-fold cross validation analysis on this dataset, they reported a classification accuracy of 70% using the supervised classifier SVM with

the RBF kernel. Purver and Battersby (2012) have tried to detect six types of emotions namely happiness, sadness, anger, fear, surprise and disgust on a dataset that contains short messages from Twitter. They have constructed two training datasets, one that contains the tweets classified using emoticon and the other containing the tweets classified using hashtags. On these two datasets, they reported an overall 10-fold cross validation accuracy of less than 70% using SVM.

Roberts et al. (2012) have proposed *Empa Tweet*, an approach that can be used for annotating and detecting emotions on Twitter posts. In their research, they developed a synthetic corpus containing tweets for seven different emotion types namely Anger, Disgust, Fear, Joy, Love, Sadness and Surprise. Using 7 different binary SVM classifiers they tried to classify each tweet to determine if an emotion is present in the tweet or not. They reported their classification resulted in tweets with multiple emotion labels. Chaffar and Inkpen (2011) have tried to compare the performance of several different supervised classifiers including NB, Decision tree (J48), and SMO (an implementation of the SVM). A 10-fold cross validation analysis performed using these classifiers suggests that the SMO algorithm has the highest accuracy rate across all the datasets that were used as part of this study. Across all the datasets their feature set was represented using the Bag of Words (BOW). Aman and Szpakowicz (2007) have tried to compare the performance of the supervised classifiers namely the NB and the SVM on their constructed dataset. A stratified 10-fold cross validation analysis on their dataset containing six classes namely Happiness, Sadness, Disgust, Anger, Fear, and Surprise; resulted in an overall accuracy of 72.08% and 73.89% respectively suggesting the fact that the SVM classifier is slightly better than the NB classifier. Their feature set was a combination of the GI and the Word Net Affect. Ghazi, Inkpen, and Szpakowicz (2010) have tried to classify the emotion classes in both the Aman's and Alm's dataset using the SVM classifier. Using the BOW as the feature set and SVM as the classifier, they performed a 10-fold cross validation analysis and have reported an overall accuracy of 61.67% on the Aman's dataset and 57.41% on the Alm's dataset (Alm, 2008; Aman & Szpakowicz, 2007). Badshah et al. (2016) have proposed a divide-and-conquer approach to identify six emotions namely Happy, Surprise, Fear, Disgust, Angry and Sadness on a dataset in three different stages. Using the classifiers Decision Tree (DT), SVM, and Random Forest (RF) on a Surrey

Audio-Visual Expressed Emotion (SAVEE) dataset, they reported a maximum overall accuracy of 82.21%. According to them the RF was the best classifier in all the three stages. The features on the SAVEE dataset were derived using the Mel Frequency Cepstral Coefficients (MFCCs) technique. Lliou and Anagnostopoulos (2009) have compared the classification performance of Artificial Neural Networks (ANN) and RF on the emotional Berlin Database. To classify seven different classes Anger, Happiness, Anxiety/Fear, Sadness, Boredom, Disgust and Neutral, they have reported an overall accuracy of 83.17% and 77.19% using the ANN and the RF respectively. They have also reported a classification mean accuracy of 55% and 48% on the speaker independent framework thus suggesting the fact that the performance of the ANN classifier is superior than the RF classifier.

Challenges for This Study

Several challenges must be addressed in order to accurately classify the tweets in to different emotional and sentiment categories. First, unlike the conventional texts, tweets are peculiar in terms of their structure and size. Primarily, they are restricted to a length of 140 characters and secondly, due to this limitation the language used by people in tweets to express their emotions is very different when compared to the other digitized documents like blogs, articles and news (Ling & Baron, 2007). The language used on Twitter is often typically informal and the users tend to develop linguistically unique styles (Hu et al., 2013) and abbreviations, acronyms, emoticons, unusual orthographic elements, slang, and misspellings can be observed more frequently. Despite the character limitation, it is very common to find tweets with more than one emotion.

Second, a major challenge is posed by the availability of a very large number of features in the tweets. Each tweet, when presented as a vector of features, exponentially increases the size of the available features as the corpus would contain millions of features for a given topic. As a result, the feature vector for each tweet will be very large and sparse (Hasan et al., 2014).

Third, supervised classifiers need labeled data for training. Due to the large volume of Twitter messages, it would be time consuming and tedious to manually annotate them with emotion classes and later use it to identify the emotions expressed in an unlabeled data set. Researchers have previously tried to manually classify tweets however manually

annotating the texts may be ambiguous and does not guarantee 100% accuracy (Hasan et al., 2014).

Fourth, the inherent nature of the different types of emotions makes it very difficult to differentiate between them. According to the Circumplex model (Russell, 1980), there are 28 affect words or emotions. In the two-dimensional circular space, the 28 different emotion types differ from each other by a small angle. Few emotions are clustered so close that it becomes very hard to differentiate between them. When humans try to annotate short messages, there is a high probability of mislabeling the emotions that differ by a small angle. This in turn inhibits a classifier from learning the critical features that can enable it to differentiate between different emotion classes hidden in the tweets.

Though there are a limited number of labeled datasets available to train the classifiers, not all datasets are as efficient in providing a classifier with a critical set of features needed to differentiate between the various categories of emotions. When the supervised classifiers are cross-validated on a training dataset, the prediction accuracy in different folds are not so significant. On the other hand, unsupervised classifiers suffer from the fact that the emotion classes are clustered very close to each other thus making it very hard to accurately annotate the clusters with different emotion classes. Therefore, this study employs a lexicon-based classification technique which is preferable for emotion classification.

MATERIALS AND METHODS

A corpus was built using tweets retrieved from Twitter. All the tweets were related to either *Hillary Clinton* or *Donald Trump*, the candidates for the 2016 US precedential election. Retrieval of these tweets were facilitated using an automated script that leveraged the *Search API* of Twitter's *REST API*, and the in-built Twitter API package within the RStudio software. A developer account was set up on Twitter that provided access to various Tokens and the API key values that were necessary to successfully execute the automated script. Appropriate handles such as *@realDonaldTrump* and *@HillaryClinton* were identified and were provided to the automated script to selectively retrieve the tweets. Tweets were collected for the period of six weeks starting from

September 26 (Week 1 or W1) till November 6,2016 (Week 6 or W6). Tweets were collected for the following 15 states: Alabama, California, Florida, Idaho, Iowa, Massachusetts, Mississippi, New York, North Dakota, Ohio, Oregon, Virginia, Washington, Wisconsin, and Wyoming. These 15 states were strategically identified to have three different groups each with five states namely the Democratic, the Republican, and the Swing States or the Battlefield. Over the period of 6 weeks, we collected a total of 24,873,256 tweets. After data collection, extensive cleaning was performed on the tweets using the function *gsub* in the R statistical package *stringr*. Our cleaning process included all the steps outlined in Stanton (2013).

Both data collection and cleaning together took approximately 10 hours per Twitter handle across the 15 states. In order to manage this workload, the work was distributed evenly across several Google Cloud Computing engines. Data collection for each day took approximately 5 hours using two desktop computers each running 7 – 8 different Google Cloud Computing virtual machines. The entire data collection step was then validated using a two-fold mechanism. First, we used a custom-made Python script to compare the daily collected tweets against the streaming data provided by the "Streaming API" from Twitter in order to confirm the completeness of the tweet's content and attributes. Second, an additional *R* script implementing the same "Streaming API" was used, but instead of comparing the daily collected tweets, it compared on a weekly basis for each of the fifteen states in order to confirm the completeness of the collected tweets. Table 1 lists the total number of tweets that were collected over the period of six weeks across the 15 different states for both candidates, Hillary Clinton and Donald Trump.

Table 1
Number of tweets collected each week for Hillary Clinton and Donald Trump

Candidate	Sep26-Oct02 (W1)	Oct03-Oct09 (W2)	Oct10-Oct16 (W3)	Oct17-Oct23 (W4)	Oct24-Oct30 (W5)	Oct31-Nov6 (W6)	Grand Total
Hillary Clinton	2,843,307	1,718,258	1,977,067	1,940,672	1,842,366	2,015,748	12,337,418
Donald Trump	1,721,494	1,862,570	2,580,126	2,255,291	2,056,545	2,059,812	12,535,838
Total	4,564,801	3,580,828	4,557,193	4,195,963	3,898,911	4,075,560	24,873,256

Table 2 summarizes the total number of tweets that were collected across the different camps (Blue, Red, and Battlefields) over the period of six weeks.

Table 2
Number of tweets collected across different camps over the period of 6 weeks

Camp	Sep26-Oct02 (W1)	Oct03-Oct09 (W2)	Oct10t-Oct16 (W3)	Oct17-Oct23 (W4)	Oct24-Oct30 (W5)	Oct31-Nov6 (W6)	Grand Total
Battle	1,825,128	1,493,193	1,946,212	1,812,285	1,674,474	1,738,123	10,489,415
Blue	2,340,711	1,671,561	2,035,066	1,894,470	1,881,727	1,939,784	11,763,319
Red	398,962	416,074	575,915	489,208	342,710	397,653	2,620,522
Total	4,564,801	3,580,828	4,557,193	4,195,963	3,898,911	4,075,560	24,873,256

For the states of Pennsylvania, Michigan, Minnesota, and New Hampshire, we purchased the tweets from Twitter. A total of 8,74,228 tweets were purchased and we performed an extensive cleaning of the tweets based on the steps outlined by Stanton (2013). Table 3 lists the total number of tweets that were collected over the period of six weeks across the 4 states for both the candidates; Hillary Clinton and Donald Trump.

Table 3
Number of tweets collected for Clinton and Trump over the period of 6 weeks

State	Trump	Clinton	Total
Pennsylvania	266,035	167,424	433,459
New Hampshire	30,482	20,360	50,842
Minnesota	77,090	52,986	130,076
Michigan	159,207	100,644	259,851
Total	532,814	341,414	874,228

We used the *lexicon-based* classifier, *NRC*, for emotion classification of the tweets. The Lexicon-based classifiers search for axioms such as adjective, adverb, noun, etc. from a sentence and compare these words to their corresponding entries in a database of words that indicates their polarity, *i.e.* negative and positive sentiment (Rohini & Thomas, 2015). The database of words can be created either from a dictionary or from a corpus. In the dictionary-based approach a small list, also known as a seed, is initially prepared. Then

using the corpus *wordnet*, the synonyms and antonyms for a word are collected and this process continues recursively until there are no newer words to add. The major drawback of the dictionary-based approach is that sentiment words important to a particular domain (say, politics) may not be part of the list. The corpus-based approach helps overcome this drawback by including sentiment words relevant to the domain of study. However, unavailability of the domain-specific corpus is a major challenge in using this approach (Rohini & Thomas, 2015).

Mohammad and Turney (2012) have compiled emotion annotations for about 14,182 words through crowdsourcing using Mechanical Turk. This lexicon, more commonly referred to as *NRC* emotion association lexicon or *EmoLex*, has annotations for eight different emotions including *anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, and *trust* and two sentiments *negative* and *positive*. This corpus was constructed based on two measures, namely *Strength of Association* (SOA) and *Pointwise Mutual Information* (PMI). To begin with *n*-grams (words in a sentence of varying length) were generated from the dataset containing emotion-labeled headline sentences. For each of the generated *n*-grams a PMI was computed which determines the association of an *n*-gram with a particular emotion class. At the same time a secondary PMI *i.e.* PMI' is computed for each *n*-grams that determines the association of an *n*-gram with other emotion classes. Finally, for each *n*-gram, the SOA is computed across each emotion class by taking the difference of PMI and PMI'. If an *n*-gram has a stronger tendency to occur in a sentence with a particular emotion class, than in a sentence that does not belong to that class, then that *n*-gram-emotion pair will have an SOA score greater than *zero*. Such *n*-grams are associated with that particular emotion class. These *n*-grams are considered as potential lexicons that can determine a particular emotion class in a sentence (Mohammad & Turney, 2011; Mohammad, 2012). The PMI values for *n*-grams that have a very low frequency of occurrence in the dataset are not robust. Such *n*-grams should be removed from the dataset (Mohammad & Turney, 2011; Mohammad, 2012). One drawback of the lexicon-based classifier is that the classification of a sentence containing words not present in its lexicon is not possible.

RESULTS AND DISCUSSIONS

Initially, we classified all the 25,747,484 tweets in to 8 emotional classes and 2 sentiment classes using the NRC classifier implemented in R. Using the results from the sentiment analysis, we tried to predict the swing direction of each state. Across each state, we computed the net positive score for each candidate, which is the difference between the fraction of the positive sentiment tweet to the total number of tweets, and the fraction of the negative sentiment tweet to the total number of tweets. Here, we postulate that a state would swing in favor of a candidate if that candidate has received the highest number of net positive tweets in that state. Table 4 provides a comparison of the net positive score for each candidate across the 19 states.

Table 4

Net positive scores for each candidate and the predictions of the swing direction

States	Net positive score		Predicted Margin	Predicted Result	Actual Margin	Actual Result
	Clinton	Trump				
Alabama	-7.191	1.593	R8.79%	Likely Republican	R27.72%	Trump
California	4.768	-0.597	D5.37%	Likely Democrat	D30.11%	Clinton
Florida	0.107	1.684	R1.58%	Likely Republican	R1.20%	Trump
Idaho	-9.040	-1.397	R7.64%	Likely Republican	R31.77%	Trump
Iowa	2.589	17.462	R14.87%	Solid Republican	R9.41%	Trump
Massachusetts	25.088	13.326	D11.76%	Solid Democrat	D27.20%	Clinton
Mississippi	-6.461	-1.132	R5.33%	Likely Republican	R17.83%	Trump
New York	22.486	14.835	D7.65%	Likely Democrat	D22.49%	Clinton
North Dakota	-3.873	3.290	R7.16%	Likely Republican	R35.73%	Trump
Ohio	-2.173	-0.849	R1.32%	Likely Republican	R8.13%	Trump
Oregon	9.813	-10.817	D20.63%	Solid Democrat	D10.98%	Clinton
Virginia	20.228	15.050	D5.18%	Likely Democrat	D5.32%	Clinton

Washington	11.717	-11.958	D23.68%	Solid Democrat	D15.71%	Clinton
Wisconsin	**7.256**	**3.483**	**D3.77%**	**Likely Democrat**	**R0.77%**	**Trump**
Wyoming	-8.221	-3.814	R4.41%	Likely Republican	R45.77%	Trump
Pennsylvania	-28.590	10.450	R39.04%	Solid Republican	R0.72%	Trump
Michigan	9.400	11.090	R1.68%	Likely Republican	R0.23%	Trump
Minnesota	12.950	10.620	D2.33%	Likely Democrat	D1.52%	Clinton
New Hampshire	**8.480**	**11.190**	**R2.71%**	**Likely Republican**	**D0.37%**	**Clinton**

From the computed net positive score, we were able to correctly predict the swing directions of 17 out of 19 states. Table 4 provides a head-to-head comparison of our predictions using the *NRC* classifier against the outcome of the election. Our prediction was incorrect for the states of Wisconsin and New Hampshire (shown in bold in Table 4).

Table 4 also shows two measures, namely the *actual margin* and the *predicted margin*. An *actual margin* is the difference between the number of favorable votes received by Clinton and Trump (a margin greater than zero goes in favor of Clinton indicated by D and a number, and a margin less than zero goes in favor of Trump indicated by R and a number). Similarly, the *predicted margin* is the difference between the percentage of the net positive tweets obtained by the Democratic candidate and the Republican candidate.

Table 5 compiles the list of predictions by the different pollsters across the 19 states (Katz, 2016).

Table 5
Pollster's prediction on the outcomes of the US presidential election 2016

States	Pollster Predictions								
	NYT	538	HP	PW	PEC	DK	Cook	Roth.1	Sabato
Alabama	>99% Rep.	>99% Rep.	>99% Rep.	>99% Rep.	>99% Rep.	>99% Rep.	Solid Rep.	Solid Rep.	Solid Rep.
California	>99% Dem.	>99% Dem.	>99% Dem.	>99% Dem.	>99% Dem.	>99% Dem.	Solid Dem.	Solid Dem.	Solid Dem.
Florida	**67% Dem.**	**55% Dem.**	**88% Dem.**	**77% Dem.**	**69% Dem.**	**86% Dem.**	Tossup	**Lean Dem.**	**Lean Dem.**
Idaho	>99%	99%	>99%	>99%	>99%	>99%	Solid	Solid	Solid

	Rep.	Rep.	Rep.	Rep.	Rep.	Rep.	Rep.	Rep.	Rep.
Iowa	62% Rep.	70% Rep.	89% Rep.	79% Rep.	74% Rep.	99% Rep.	Lean Rep.	Lean Rep.	Lean Rep.
Massachusetts	>99% Dem.	>99% Dem.	>99% Dem.	>99% Dem.	>99% Dem.	>99% Dem.	Solid Dem.	Solid Dem.	Solid Dem.
Mississippi	86% Rep.	98% Rep.	>99% Rep.	>99% Rep.	>99% Rep.	>99% Rep.	Solid Rep.	Solid Rep.	Solid Rep.
New York	>99% Dem.	>99% Dem.	>99% Dem.	>99% Dem.	>99% Dem.	>99% Dem.	Solid Dem.	Solid Dem.	Solid Dem.
North Dakota	>99% Rep.	98% Rep.	>99% Rep.	>99% Rep.	>99% Rep.	>99% Rep.	Solid Rep.	Solid Rep.	Solid Rep.
Ohio	54% Rep.	65% Rep.	73% Rep.	67% Rep.	63% Rep.	88% Rep.	Lean Rep.	Tossup	Lean Rep.
Oregon	98% Dem.	94% Dem.	>99% Dem.	>99% Dem.	>99% Dem.	>99% Dem.	Solid Dem.	Solid Dem.	Solid Dem.
Virginia	96% Dem.	86% Dem.	99% Dem.	98% Dem.	98% Dem.	>99% Dem.	Likely Dem.	Likely Dem.	Likely Dem.
Washington	>99% Dem.	98% Dem.	>99% Dem.	>99% Dem.	>99% Dem.	>99% Dem.	Solid Dem.	Solid Dem.	Solid Dem.
Wisconsin	**93% Dem.**	**84% Dem.**	**99% Dem.**	**98% Dem.**	**98% Dem.**	**>99% Dem.**	**Lean Dem.**	**Lean Dem.**	**Likely Dem.**
Wyoming	>99% Rep.	99% Rep.	>99% Rep.	>99% Rep.	>99% Rep.	>99% Rep.	Solid Rep.	Solid Rep.	Solid Rep.
Pennsylvania	**89% Dem.**	**77% Dem.**	**99% Dem.**	**93% Dem.**	**79% Dem.**	**>99% Dem.**	**Lean Dem.**	**Lean Dem.**	**Lean Dem.**
Michigan	**94% Dem.**	**79% Dem.**	**99% Dem.**	**95% Dem.**	**79% Dem.**	**>99% Dem.**	**Lean Dem.**	**Lean Dem.**	**Lean Dem.**
Minnesota	94% Dem.	85% Dem.	>99% Dem.	99% Dem.	98% Dem.	>99% Dem.	Likely Dem.	Likely Dem.	Likely Dem.
New Hampshire	79% Dem.	70% Dem.	92% Dem.	84% Dem.	63% Dem.	99% Dem.	Lean Dem.	Lean Dem.	Lean Dem.

NYT – The New York Times Upshot; 538 – FiveThirtyEight; HP – Huffingtonpost; PW - PredictWise PEC – Princeton Election Consortium; DK - Dailykos; Cook – The Cook Political Report; Roth.I – Rothenberg Gonzales; Sabato – Sabato's Crystal Ball

As seen from Table 5, almost all the pollsters mis-predicted the swing directions of the states of Florida, Wisconsin, Pennsylvania and Michigan. A head-to-head comparison of the predictions from the *NRC* (Table 4) and the pollsters (Table 5) suggest that the predictions by *NRC* are superior to that of the pollsters. All the pollsters were able to correctly predict the outcome in 15 out of 19 states compared to 17 out of 19 by *NRC*. When *NRC* mis-predicted the swing direction of the states of Wisconsin and New Hampshire, the actual margin (of victory) was less than 1% (that is, the states were closely contested).

In a democratic country where the leaders are elected by the voting process, the voter turnout is highly influenced by who they trust. Therefore, we were more interested in determining the *trust* factor for both presidential candidates. Using *NRC*, we explored the landscape of trust for both candidates over the period of six weeks across the 19 different states. There was a marginal increase in trust across all the camps for Clinton, in and after, the third week which was about the same time when she won the second presidential debate. In particular, the people in the red states showed more trust in her compared to the people in the battlefield and blue states. In the mid of week 4, specifically October 19, Clinton narrowly won the final presidential debate, but Trump did make some really good points which meant that the final debate was his best performance. Since Clinton's performance was not so exceptional compared to that of Trump the trust factor for Clinton dropped in week 5 and continued to do so thereafter (see Figure 1).

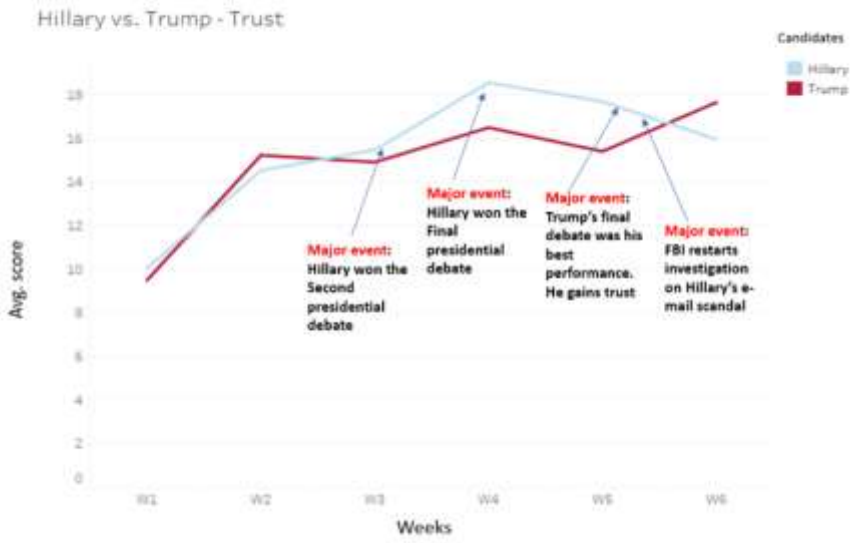

Figure 1. Clinton vs. Trump: Trust over the period of 6 weeks according to NRC

This pattern was observed consistently across all states. Overall people were not angry, disgusted, or sad with her but *anticipation* towards her had increased significantly, almost doubling in the red camp and trust continued to decline. Figure 2 shows this trend starting at week 4 (W4) all the way into week 6 (W6), the final week before the election.

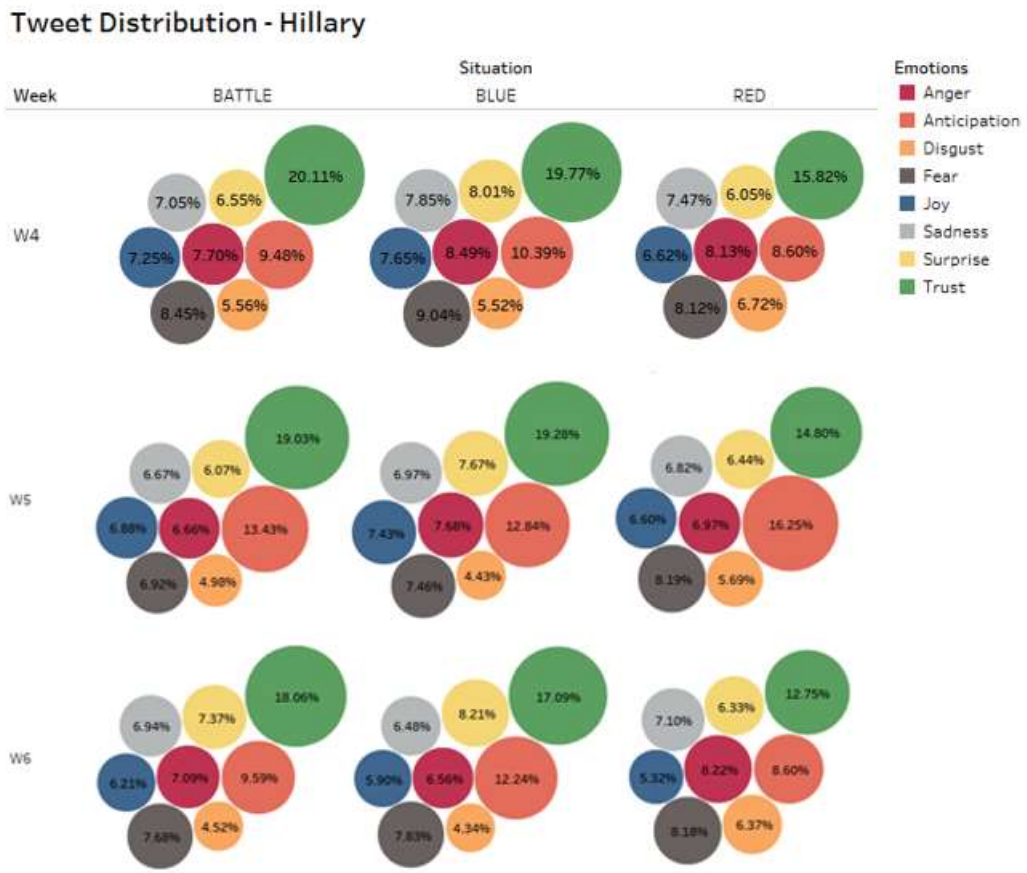

Figure 2. Classification of tweets into 8 different emotions for Clinton by NRC

People were rather disappointed with Clinton after the final debate, not at all a good prospect for Clinton. In week 5, the e-mail server controversy against her was reopened. Across all the camps people lost trust in her as she entered in to final week of the election (see Figure 1 and 2). The fear for her also started to move up in the final week in the battle and blue states (see Figure 2). Looking in to just the positive and negative sentiments, the reopening of the e-mail server controversy did not bother the people in the blue camp much, but the controversy did hit hard (positive sentiment fall, and negative sentiment raise) on the people in the battlefield and red states.

On the other hand, Trump's campaign experienced a rise in trust factor during the first two weeks (week 1 and week 2). This can be attributed to the fact that Trump's running mate, Mike Pence, ended up narrowly winning over Tim Kaine (Clinton's running

mate). It can be speculated that after the vice presidential debate on October 4, 2016, Trump became more popular, probably because of the remarks made by his running mate Mike Pence. On October 7 when Trump's video bragging about his sexual exploits leaked, there was a significant change in the trust factor for Trump that was captured by the *NRC*. From figure 1, it is evident that both the candidates experienced a roller-coaster of variation in the trust factor but eventually it was Trump who gained more trust among the voters during the final week (week 6) before the election.

Exploring the landscape of the emotions *joy* and *disgust* over the period of six weeks (see Figures 3 and 4) paints a very similar picture as discussed above. From Figure 4, it is evident that the people were more disgusted with Trump throughout the six weeks probably because of his radical insulting remarks and controversial proposals. According to *NRC,* the public's disgust reached its peak sometime in the third week (week 3) when the video of Trump making sexist remarks was released. Also, in the second week (week 2) during the vice presidential debate, his running mate made some insulting comments against the Latino which caught a lot of attention of the public causing uproar.

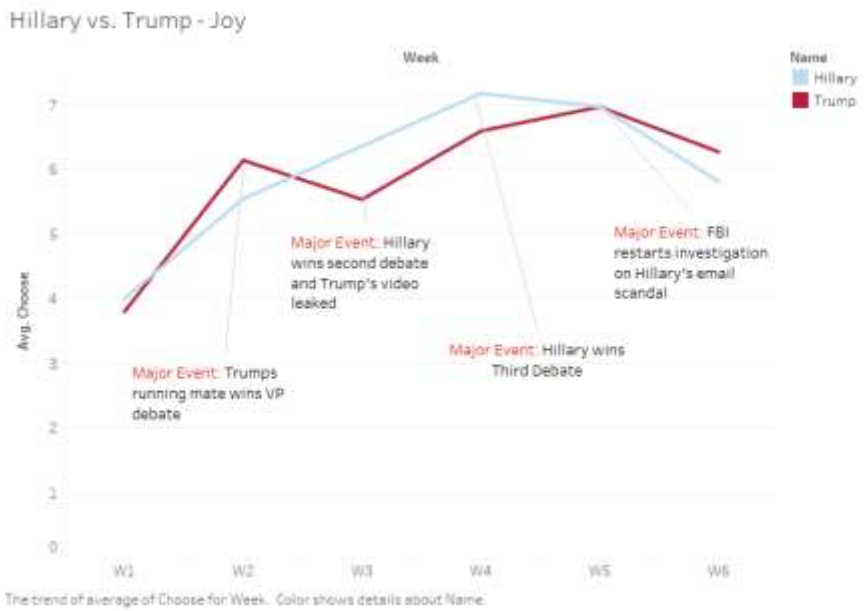

Figure 3. Clinton vs. Trump: NRC analysis of the emotion joy over the period of 6 weeks

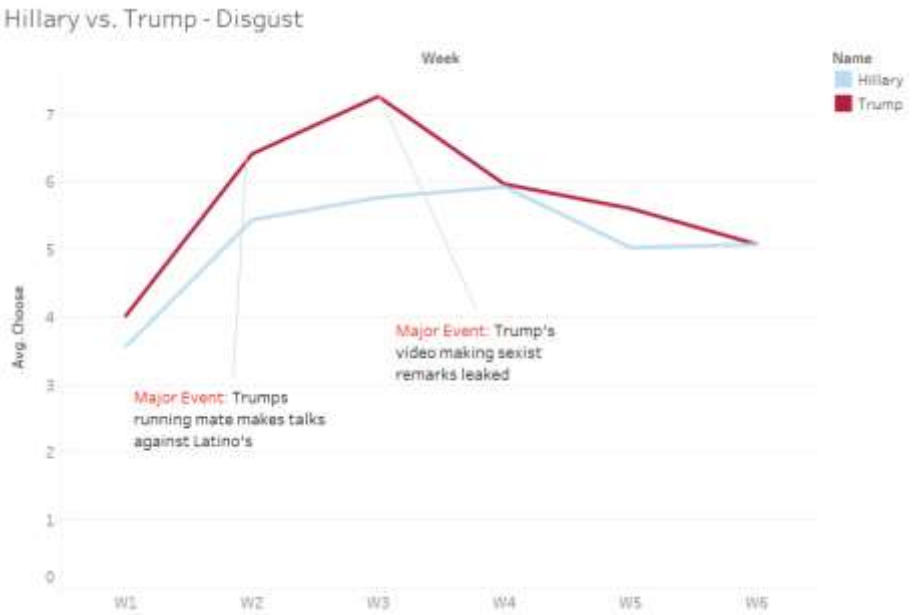

Figure 4. Clinton vs. Trump: NRC analysis of the emotion disgust over the period of 6 weeks

The emotion classification by NRC for each week corroborated well with the various political events that took place in that particular week. Therefore, it was easy to understand the swing in the emotions of the people for each week by utilizing the *NRC* classifier. According to the *NRC* classifier, the people in all the camps were much happier with Clinton than with Trump until week 4 but in the later weeks people's emotion toward Clinton changed. We also observed that after the fourth week (week 4) people in all the camps were unhappy (sad) with both the candidates. As shown in Figure 5, we also noted that the people were happier with Trump than with Clinton in the final week (week 6) of the election which is very much consistent with all the other observations we made by analyzing the different emotion types.

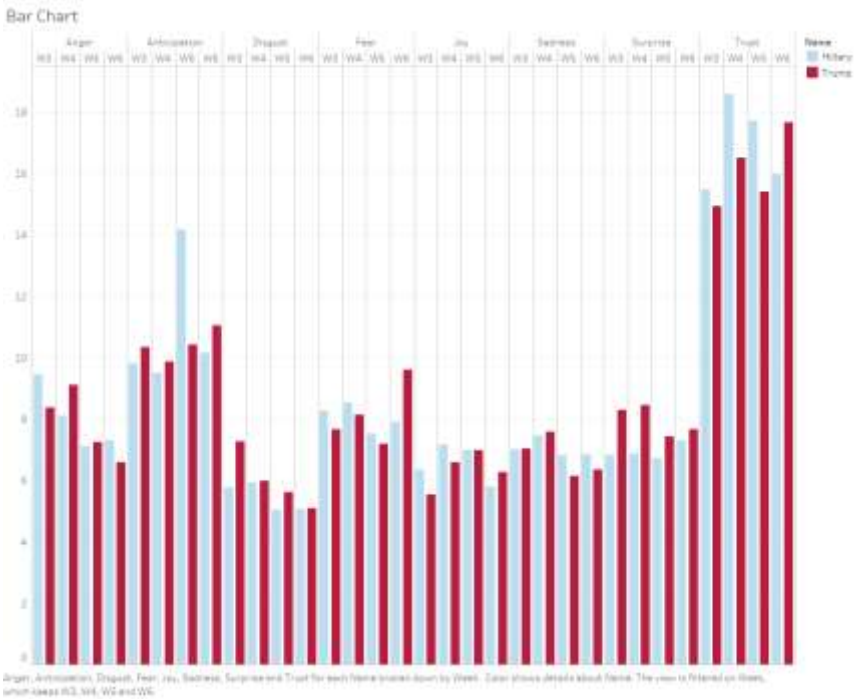

Figure 5. Bar chart comparing all the 8 emotions for both Clinton and Trump over the period of 6 weeks as determined by the NRC

CONCLUSION

This study investigated the potentiality of a lexicon-based classifier, namely *NRC*, for emotion classification. Using the *NRC* classifier, we classified 25.74 million tweets related to the event of 2016 US presidential election into 8 different emotions, and into positive and negative sentiments. Based on this classification, we were able to correctly determine the swing directions for 17 out of 19 states, approximately 90% accuracy. In comparison to the predictions from 9 different pollsters, our predictions were more accurate, especially for the states of Florida, Pennsylvania and Michigan that were critical to this election. The pollsters correctly predicted the swing direction of 15 out of 19 states, approximately 79% accuracy.

The emotion classification by *NRC* for each week leading up to the elections corroborated well with the various political events that took place during that period making it easier to understand the swing in the emotions of the people. According to the *NRC* classifier, the people in all the camps were much happier with Clinton than with Trump until week 4 but in the later weeks people's emotion toward Clinton changed. Also, after the fourth week people in all the camps were unhappy (sad) with both the candidates

but people were relatively happier with Trump than with Clinton during the final week of the election. This was consistent with all the other observations from analyzing the different emotion types captured by *NRC* which demonstrates the superior performance the *NRC* classifier exhibited in predicting the results of the 2016 US presidential election.

This study clearly demonstrates that both the emotion and sentiment analysis have the potentiality in understanding and gauging the emotional state of an individual, and the society as a whole. It also shows the potentiality of computer-based algorithms, such as the NRC classifier, in predicting the outcomes of significant events when compared against the predictions made by the pollsters that are purely based on analysis of the data collected through surveys and opinion polls. This study also highlights the value and power of the Twitter data, and the wealth of information hidden in such data for predictive analytics.

The advances in big data infrastructure has paved the path to capture, store and process large volumes of social media data from different sources making it possible to design and implement automated real-time predictive analytics systems. We intend to leverage these capabilities in the future to make improvements to our current model.

References

Alm, C. O. (2008). Affect in Text and Speech. *PhD Dissertation.* University of Illinois at Urbana-Champaign.

Aman, S., & Szpakowicz, S. (2007). Identifying Expressions of Emotion in Text. *TSD 2007*, *LNAI 4629*, 196-205.

Badshah, A. M., Ahmad, J., Lee, M. Y., & Baik, S. W. (2016). Divide-and-Conquer based Ensemble to Spot Emotions in Speech using MFCC and Random Forest. *Proceedings of the 2nd International Integrated Conference & Concert on Convergence,* 1-8.

Barbosa, L., & Feng, J. (2010). Robust sentiment detection on Twitter from biased and noisy data. *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), 36-44.*

Brynielsson, J., Johansson, F., Jonsson, C., Westling, A. (2014). *Emotion Classification of social media posts for estimating people's reactions to communicated alert messages during crises. Security Informatics, 3(7).*

Chaffar, S., & Inkpen, D. (2011). *Using a Heterogeneous Dataset for Emotion Analysis in Text. Advances in Artificial Intelligence – 24th Canadian Conference on Artificial Intelligence.*

Choudhury, M. D., Gamon, M., Counts, S., & Horvitz, E. (2013). *Predicting depression via social media. International AAAI Conference on Weblogs and Social Media (ICWSM'13).*

Danisman, T., & Alpkocak, A. (2008). Feeler: Emotion Classification of Text Using Vector Space Model. *AISB Convention Communication, Interaction and Social Intelligence, 53-59.*

Ghazi, D., Inkpen, D., & Szpakowicz, S. (2010). Hierarchical versus Flat Classification of Emotions in Text. *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, 140-146.

Go, A., Bhayani, R., & Huang, L. (2009). *Twitter sentiment classification using distant supervision. CS224N Project Report, 1-12.*

Hasan, M., Rundensteiner, E., & Agu, E. (2014). EMOTEX: Detecting Emotions in Twitter Messages. *Academy of Science and Engineering.*

Hu, X., Tang, J., Gao, H., & Liu, H. (2013). Unsupervised sentiment analysis with emotional signals. *Proceedings of the 22nd international conference on World Wide Web, WWW'13. ACM.*

Katz, J. (2016, November 8). Who Will Be President? *New York Times*. Retrieved from https://www.nytimes.com/interactive/2016/upshot/presidential-polls-forecast.html

Ling, R., & Baron, N. S. (2007). Text Messaging and IM: Linguistic Comparison of American College Data. *Journal of Language and Social Psychology*, 26(3), 291-298.

LLiou, T., & Anagnostopoulos, C.N. (2009). Comparison of Different Classifiers for Emotion Recognition. *13th Panhellenic* IEEE *Conference on Informatics*, Retrieved from http://ieeexplore.ieee.org/document/5298878/

Maleki, R. E., Rezaei, A., & Bidgoli, B. M. (2009). Comparison of classification methods based on the type of attributes and sample size. *Journal of Convergence Information Technology*. 4(3). 94-102

Mohammad, S., & Turney, P. (2011). Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. *Proceedings of the NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text.*

Mohammad, S. (2012). Emotional Tweets. *Proceedings of the First Joint Conference on Lexical and Computational Semantics.*

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010), 1320-1326.*

Peng, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs us? Sentiment classification using machine learning techniques. *Proceedings of the Seventh Conference on Empirical Methods in Natural Language Processing (EMNLP-02)*, 79-86.

Purver, M., & Battersby, S. (2012). Experimenting with distant supervision for emotion classification. *Proceedings of the 13th EACL.* Association for Computational Linguistics, 482-491.

Roberts, K., Roach, M. A., Johnson, J., Guthrie, J., & Harabagiu, S. M. (2012). EmpaTweet: Annotating and Detecting Emotions on Twitter. *LREC,* 3806-3813.

Rohini, V., & Thomas, M. (2015). Comparison of Lexicon based and Naïve Bayes Classifier in Sentiment Analysis. *International Journal for Scientific Research & Development*, 3(4).

Russell, J. A. (1980). A Circumplex model of affect. *Journal of Personality and Social Psychology*, 39, 1161-1178.

Stanton, J. (2013). An Introduction to the Data Science. Retried from https://www.scribd.com/document/194116122/Data-Science-Book-v-3

Thelwall, M., Buckley, K., Platoglou, G., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology, 61(*12), 2544-2558.

Funding and Acknowledgements

The authors declare no funding sources or conflicts of interest.