Influence Factor Based Opinion Mining of Twitter Data Using Supervised Learning

Malhar Anjaria, Ram Mohana Reddy Guddeti
Department of Information Technology
National Institute of Technology Karnataka,
Surathkal, Mangalore - 575025, India
malhar.anjaria@gmail.com, profgrmreddy@nitk.ac.in

***Abstract* — Social Networking portals have been widely used for expressing opinions in the public domain through internet based text messages and images. Among popular social networking portals, Twitter has been the point of attraction to several researchers in important areas like prediction of democratic electoral events, consumer brands, movie box office, stock market, popularity of celebrities etc. Sentiment analysis over Twitter offers a fast and efficient way of monitoring the public sentiment. In this paper, we introduce the novel approach of exploiting the user influence factor in order to predict the outcome of an election result. We also propose a hybrid approach of extracting opinion using direct and indirect features of Twitter data based on Support Vector Machines (SVM), Naive Bayes, Maximum Entropy and Artificial Neural Networks based supervised classifiers. We combined Principal Component Analysis (PCA) with SVM in an attempt to perform dimensionality reduction. This paper shows two different case studies of entirely different social scenarios, US Presidential Elections 2012 and Karnataka Assembly Elections 2013. We conclude the conditions under which Twitter may fail or succeed in predicting the outcome of elections. Experimental results demonstrate that Support Vector Machines outperform all other classifiers with maximum successful prediction accuracy of 88% in case of US Presidential Elections held in November 2012 and maximum prediction accuracy of 58% in case of Karnataka State Assembly Elections held in May 2013.**

Keywords* — *big data analysis electoral prediction; micro-blogs; opinion mining; sentiment analysis; social intelligence; social network analysis; supervised machine learning; twitter; twitter analytics

I. INTRODUCTION

Big data is the widely used term for a collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications. The challenges include analysis, pattern recognition, visualization etc. Given the challenges big data also provides a chance to understand the data patterns and helps with prediction of events using this huge chunk of data. Big data analysis is performed in various domains like cloud computing, network simulation and prediction, user behavioral study etc. With advent of web 2.0, where users are the central focus for any organization, big data analysis of user data comes very handy in predicting the correct strategies for success of any product. Study of the user data of social networks is one of the current trends of the times.

1. http://www.facebook.com
2. http://www.tumblr.com
3. http://www.twitter.com

Social Networks have become one of the admired communication medium used over internet. Millions of text messages are appearing daily on popular web-sites that provide micro-blogging services such as Facebook[1], Tumblr[2], Twitter[3]. Authors of messages publish their opinions on variety of topics and discuss several current issues. Because of restriction-less message format and also due to an easy accessibility of micro-blogging platforms, Internet users tend to shift from traditional communication tools such as traditional blogs or mailing lists to micro-blogging services. Avalanche of messages on social networks, make it very attractive medium of user data analysis through posts about consumer brands and services they use, or expression about political and religious views [1]. Figure-1 depicts the basic idea of social network analysis as a collaborative study of social networks, big data and the statistical analysis.

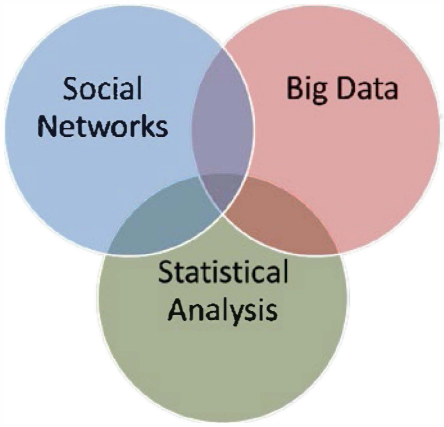

Figure 1: Visualizing Social Network Analysis

Twitter is a novel micro-blogging platform launched in 2006 with more than 25 million unique monthly visitors. On Twitter, any user can publish a short message referred to as tweet with a maximum length of 140 characters, which is visible on the public display. The public timeline conveying the tweets of all users worldwide is an extensive real-time information stream of more than one million messages per hour. Especially during the process of democratic elections, political issues are clearly on the minds of several users. In addition, politicians are communicating with the electorate and trying to mobilize their supporters.

Natural Language opinions are often expressed in restrained and multifarious ways, presenting issues which are difficult to solve by basic text processing methodologies.

978-1-4799-3635-9/14/$31.00 ©2014 IEEE

However, approaches like utilizing n-grams, Part-of-speech tagging have been employed effectively in [2] and [3] for finding the twitter sentiment using the machine learning techniques and other methodologies.

Currently established methods use literal meaning of the word to categorize into positive or negative irrespective of the target while performing sentiment analysis. It is important to understand the target and the relative terms of the sentiment analysis as well. For example *"Jobs is unpredictable"*, will be considered as a positive sentiment when "Jobs" is a movie, which does not necessarily hold true if "jobs" is the name of a person. We try to address this issue in this paper by using extended target based features explained in section-3.

Incorporating sentiment analysis with social network data such as twitter for prediction, as explained in section-2, has been widely studied. We attempt to use the social network features such as mentions and retweets to understand the behavioral analysis for the party participating in democratic electoral polls.

In this paper, we try to validate prediction results for the US Presidential Elections-2012 and Karnataka State Elections-2013 using twitter data and the results are compared with different classification methods. Further, we propose a novel approach of utilizing twitter features along with sentiment analysis. We also provide the gender based analysis of US Presidential Elections held in November 2012. To the best of our knowledge this is the first attempt to use target dependent features to parse sentiment analysis for electoral prediction. We establish the model to find influence factor for a query term using aggregated retweets count as well.

Rest of the paper is organized as follows: Section 2 describes the Background and Related Work; Section 3 deals with Methodology and Framework; Section 4 gives the Results and Analysis; final Conclusions are detailed in Section 5.

II. Background And Related Work

Several Researchers carried out research work in Social Network Analysis and sentiment analysis. Sentiment analysis is a text processing technique to derive an opinion or mood intention based on the terms used in a real language sentence. A vast number of researchers have concentrated on generating statistical inference from social network data using sentiment analysis models.

Bo Pang and Lilliam Lee [2] provided an insightful discussion on sentiment analysis. They considered the ratio of positive words to total words to estimate the opinion. This system has achieved good results for the direct targets using machine learning algorithms, but the extended targets or related terms have not been considered.

Meeyong Cha et. al. [4] (2010) studied the influence factor and established that more number of followers doesn't necessarily mean more influence over twitter. However, Antoine Boutet et. al [5], made prediction with party characteristics with the use of user behavior analysis and influence factor using the UK general elections of 2012 case study. This model assumes the influence factor as follower and following ratio; which is a clear contradiction to the results established by, Meeyong Cha et al.

Johan Bollen [6] used POMS score to establish the sentiment values classified in moods category. This method does not use any machine-learning algorithm for training of the data as positive or negative. The system measures this sentiment using a syntactic, term-based approach, in order to detect as much mood signal as possible from very brief Twitter messages.

Marko Skoric et. al [7] proposed word and term frequencies for tweets corresponding to key-terms related to the subject, such as democratic leader names and democratic organization names which played an important role in predicting Singapore Elections 2011. This system depends more on the relationship of the users who publish the tweets for the key-words related to case-study. Romero et. al. [8] instituted that the people would use the Twitter hashtags, to get noted over trends and the use of the hashtags remains consistent across time for politically debated topics.

Support Vector Machine (SVM), Naïve Bayes (NB) and Maximum Entropy (MaxEnt) classifiers are well discussed in many literatures such as Pang and Lee [2][9][10], where as Artificial Neural Networks (ANN) have been discussed very limited number of times. Rodrigo Moraes et al. [11] discussed the comparative features of ANN and SVM in detail for the document level sentiment classification. They used TF-IDF (Term Frequency – Inverse Document Frequency) to extract feature values for unbalance datasets. Long-Sheng Chen et. al [12] implemented the feed-forward BPN network and uses Sentiment Orientation to compute the values at each neuron. The model depends upon the sentiment orientation of terms used in the documents.

Cozma and Chen [9] studied the US 2010 midterm elections. They found that the currently chaired politicians and challengers used Twitter in different ways. Currently chaired politicians were more concentrated upon recent events, whereas challengers preferred the strategy of attacking the incumbents. Following the same elections, Pew Research Center researchers [10] found that tweets from election participants urged users more for active participation and publish the votes over twitter. Several such studies display that Twitter plays a vital role in the political communication environment in many countries. It offers an extremely rich source of information for those attracted towards studying public opinion and political behavior.

Min Song and Min Chul Kim [13], attempted to mine the twitter data in real time Korean Elections 2012. The system offers the prediction based on term occurrence and mentions based social network analogy. The system reflects well for term calculation based trend change in user opinion. They also take user mentions into account to predict the trend behavior. This trend behavior can help to find out the outcome of a topic in near time analysis.

As Twitter becomes more popular, sentiment analysis on Twitter data has attracted more attention. Go et al., (2009); Parikh and Movassate [14]; Barbosa and Feng [15]; Davidiv et al. [16], followed the machine learning approach for sentiment

analysis of tweets. Davidiv et al., (2010) proposed multiple sentiment types to classify tweets using hashtags and smileys as labels. Further, Barbosa and Feng, (2010) proposed a two-step approach to classify the sentiments of tweets using Support Vector Machine (SVM) classifiers with abstract features.

Majority of the researches carried out in this domain, focuses on the TF-IDF or Sentiment Orientation with original features and ignores the related extended features that can possibly add value to the prediction of the outcome. Researchers also use the batch mode training method for classifiers with large datasets, which requires very heavy amount of main memory usage and processing power. We use extended features extracted while processing and also use incremental learning for training phase of the classifiers.

III. METHODOLOGY

We intend to evolve the standard election prediction model with our prediction model based on tweet influence factor using the number of re-tweets, each party garners.

Figure-2 depicts the proposed system based on utilizing salient features of different classifiers. We have used Naïve Bayes, SVM, MaxEnt, ANN classifiers with features extracted from Twitter data using feature extraction methods such as Unigram, Bigram and Hybrid (Unigram + Bigrams) for sentiment analysis. The process is carried out for all the participating parties based on keywords.

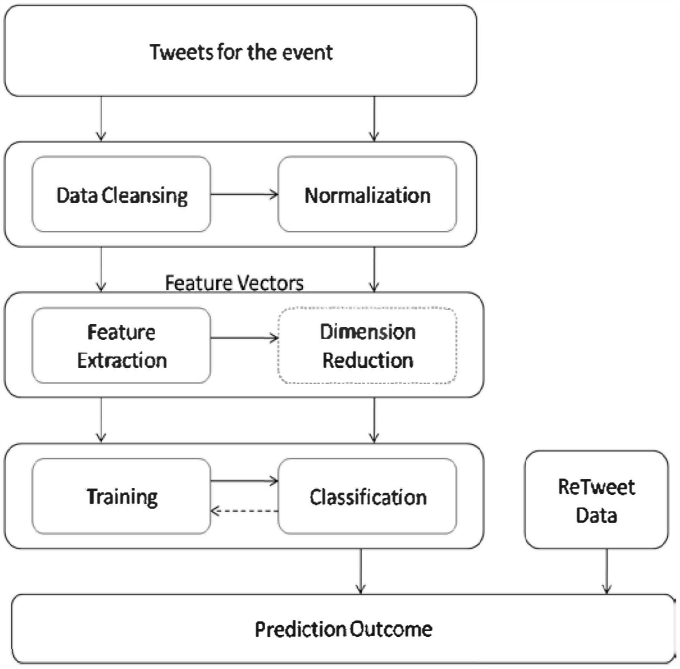

Figure 2: High Level System Flow

We fetch the tweets using Twitter API v 1.1. In order to remove stop words and extract features, we perform data cleaning and normalization as explained in section 3.2. We extract the target based extended features model [17] by modifying it and twitter user data from the normalized data. This feature vectors are used in part of chunks to train the classifier as a part of incremental training. After utilizing nearly $2/3^{rd}$ of the data we test it with $1/3^{rd}$ of the data. The sentiment analysis results are incorporated with influence factor to predict the results using prediction model established, as described in section 3.5. Algorthm-1 explains the basic flow of the data.

Algorithm - 1: Predicting Outcome of an Electoral Poll with Modified Extended Features

Input: Enter set of keywords {k} related to electoral event

Output: Outcome in terms of predictions

1. T = {{tweetId, tweetxText, userId, retweetCount, tweetDate}, Query}. Store tweet with attributes for given query.

2. T' = normalized(T). Perform data cleaning and normalization.

3. Extract the extended targets from T' with modified extended feature model. Store extended targets and features. Crate map M = <TweetId,Query_feature>.

4. FV = word_features(T') Create feature Vector of word sets.

5. FV'= PCA(FV). Perform dimension reduction of feature vector.

6. Provide an incremental training with subset of FV' as T_i with output class C ; C ϵ{positive, negative } and i is iteration.

7. Use trained dataset to classify using Supervised Learning methods. Store class C with tweet. Store sentiment values for each participant U_i as <U_i,Score_Pos,Score_neg>.

8. Evaluate user influential factor using model for participant U, for class P using associated tweet and retweet ount.

9. Feed achieved values from step-5 and step-6 to influence model.

Following subsections explains the each Figure-2 blocks in detail.

A. Data Collection Approach

We searched using twitter search API v 1.1 to collect data with various hash-tags like #USElections2012, #USElections, #Elections2012 for US Presidential Election 2012 for the time period of August,2012 to October,2012. We collected approximately 100000 tweets for the duration. In order to collect tweets with Candidate's name, we used terms "Barack Obama" ,"John Biden", "Democractic" and "Mitt Romney", "Paul Ryan", "Republican". We went through terms like #BJP, #Congress, #KJP , #KarnatakaElections for Karnataka state assembly election 2013 between April 2013 and May 2013 along with politicians' name related to all the parties. To ensure that tweets are fetched only from corresponding nations we considered UTC time-offset of the tweet. We do not check for the location of the tweet due to lack of information in majority of the dataset. However, the tweets with name of the states are collected but, it is natural that a tweet containing state name is not necessarily be originated from that state. As an example, a

tweet containing "California" might have been originated from state of Texas.

B. Normalization & Feature Reduction

Structured Tweets are generally in sentence format, with URLs specified for images or blog articles. To get data that is in usable format we remove the stop words that contains general terms like a, the, etc and emoticons. We performed following operations on tweets in cleansing and normalizing phase.

- Username: We remove the usernames from tweets for parsing purpose; we embed USER_NAME in place of the username. Username are stored with number of re-tweets for later use in prediction model.
- Internet Acronyms and Emoticons: We convert internet acronyms like <3 to "love" or "gud" to "good" in order to make the meaning out of the symbols posted as part of tweet. Famous emoticons are also replaced by mood words.
- Duplicate Tweets: We check for duplicate tweets from a sliding window of last 100 tweets. The duplicate tweets occur due to sharing a tweet using RT or quotation marks or plagiarism. All the duplicate tweets are scrapped.
- Candidate Accounts: We also omit the tweets from official accounts of Presidential election candidates and their party as they ought to be biased and doesn't count as public opinion.
- Word Expansion: We expand the famous acronyms as well. The expansion is considered for standard English words and specific to the electoral terms as well. As an example, AFAIK – As Far As I Know. BO – Barack Obama.
- URLs: We removed image or article urls as well as embedded tweet URLs from the original tweet. However, embedded tweets were stored if it contains the original search query.
- Repeated words: If a word is being repeated in a tweet for more than two times consecutively, occurrence of the word had been limited to two. e.g very very very very good has been replaced by very very good.
- Repeated characters: If a character is being repeated for more than 2 times consecutively, character occurrence has been limited to two occurrences . e.g. "lonnngggg" has been replaced by "long" after smoothing.

C. Feature Extraction and Extended terms

We use the unigram, bigram and a Unigram + Bigram (hybrid) feature extraction method for study purpose. Hybrid features are taken for absolute positive words like "wonderful", "awesome", "always" etc and negative words such as "never", "not", "hardly" etc.

To extract the extended features, we use the modified model explained by Long Jiang et al [17]. Any term that occurs minimum of 500 times has been taken as an extended target. We use K=20 in extracting top K nouns from the terms that appeared for more than the threshold value. This method helps fetching related terms that can be mapped with original target queries.

As an example in following tweet user explains positive sentiment for target "Obama" by explaining "healthcare policies".

"Healthcare policy was a boon, Thanks Mr. Obama."

We also ignore the future transitive verbs when followed by the query terms. e.g "Obama will be swearing in January" does not make any value addition to the prediction.

We also make sure that generic terms like "USA", "voters" are ignored during extended features extraction to ensure exclusion of non-subjective data. We use Stanford POS[4] tagger to extract the significance of the word.

D. Machine Learning Methods

We initially applied polarity based classification methods using the set of positive and negative words provided by NLTK API for python.

$$Polarity = \frac{P(\text{Positive\_Words}) / P(\text{Total\_Words})}{P(\text{Negative\_Words}) / P(\text{Total\_Words})} \quad (1)$$

However, this method works only for independent features based on Standard English dictionary and failing to capture query specific sentiments. Table-1 provides example of positive, negative and neutral tweets.

TABLE I. EXAMPLE OF TWEET CATEGORIZATION

Sentiment	Tweet
Positive	@sandrasays Obama gives a hope of bright future with his leadership skills.
Negative	Obama war policies affected the economy like hell
Neutral	@BarackObama to address Philadelphia tomorrow.

In domain of text classification and sentiment analysis, machine learning algorithms are widely used. We use the most proven text classification supervised algorithms like Multinomial Naïve Bayes, SVM, MaxEnt and feed-forward ANN.

i. Naïve Bayes: Naïve Bayes is a simple probabilistic model based on the Bayes rule with independent feature selection, which worked well on text categorization [18]. Naive Bayes does not restrict the number of classes or attributes to deal with. Asymptotically Naive Bayes is the fastest learning algorithm for the training phase. In this paper, we make use of multinomial Naive Bayes model [19]. Class c* is assigned to tweet d, where

$$C^* = \arg Max_c P_{NB}(\text{C} \mid \text{D}) \quad (2)$$

$$P_{NB}(\text{C}/D) = \frac{(\text{P(c)}\sum_{i=1}^{m} P(\text{f} \mid \text{c})^{n_i(\text{d})})}{P(\text{d})} \quad (3)$$

4. http://nlp.stanford.edu/software/tagger.shtml

In this formula, f represents a feature and $n_i(d)$ represents the count of feature f_i found in tweet d, m represents the number of total features taken into consideration. Parameters P(c) and P(f|c) are obtained through maximum likelihood estimates [20].

In order to handle occurrence of unknown words, which had not been encountered during training phase, we use Laplacian smoothing with k=1 to allocate equal probability.

ii. Maximum Entropy: Maximum entropy classification (MaxEnt, or ME, for short) is another method which has proven successful in a number of natural language processing applications. MaxEnt sometimes outperforms the Naive Bayes at standard text classification, however this phenomena does not hold true in all the cases [18]. MaxEnt prefers a uniform classification model that satisfies a given constraint [24]. The model is represented by the following:

$$P(C \mid D) = \frac{1}{Z(d)} \exp(\Sigma_i \lambda_{i,c} F_{i,c}(d,c)) \quad (4)$$

Where, c is a class, d is the tweet, λ is the weight vector, $Z(d)$ is a normalization function. $f_{i,c}$ is a feature/class function for feature f_i and class c, defined as follows:

$$f_{i,c}(d,c') = \begin{cases} 1, n_i(d) > 0 and c' = c \\ 0, Otherwise \end{cases} \quad (5)$$

We used Gaussian-prior with ten iterations to gain the sufficient weight accuracy as suggested by Pang and Lee [2].

iii. Support Vector Machines: Support vector machines (SVMs) are well proven to have high efficiency in the domain of text categorization, and generally provides better accuracy compared to Naive Bayes [20]. SVMs are large-margin classifiers differing to probabilistic classifiers like Naive Bayes and MaxEnt [21]. SVM follows the philosophy of achieveing maximum margin hyper-plane for given data.

We use the NLTK SVM^{light} [5] API with a linear kernel, all parameters set to default values. We use the input data as sets of vectors of size m. Every record in the vector represents the presence of a feature. As an example, in case of a unigram feature extractor, a word in a tweet is considered as a feature. If the feature is present, the value is 1, else the value is 0. In order to avoid scaling of data and increase the overall speed we use feature presence, rather than use of a count [23][20].

iv. Artificial Neural Network (ANN): Artificial Neural network relies on the idea of derive features from linear combinations of the data provided as an input, and model the output as a nonlinear function of these features [25]. This has resulted in one of the most popular and effective forms of machine learning system [26]. We used feed-forward neural network to garner the benefits of its advanced learning capability.

Among various implantations of feed-forward networks, BPN is known to be the best and it remains one of the most useful ones. Published papers and rule of thumb suggests the hidden layers to be one or two in general [12].

5. http://svmlight.joachims.org/
6. http://www.numpy.org/

The feed-forward network training by back-propagation pseudo-code algorithm can be summarized as follows by Algorithm-2 [12].

Algorithm - 2: Feed Forward Network Training by Back Propagation Learning

While error ε is larger than convergence value
for each training input Ti:
1.1. Apply the inputs to the network.
1.2. Calculate the output for every neuron from the input layer, through the hidden layer hi, to the output layer Oi.
1.3. Calculate the error at the output layer Oi.
1.4. Use the output error to compute error signals for pre-output layers.
1.5. Use the error signals to compute weight adjustments.
1.6. Apply the weight adjustments.
end for
2. Calculate the network performance periodically.

We achieve all the optimal settings, number of hidden neurons, system learning rate etc for the neural networks are achieved by trial-and-error. We realize the output in terms of positive and negative values from the output layer.

v. SVM with PCA: With large set of dataset the algorithm becomes intractable under certain situations. In order to deal with curse of the dimension, we performed dimension reduction using PCA. Dimension reduction is essentially selecting M features out of N for given value set X [27].

$$[x_1\ x_2\ x_3\ \ldots\ x_n] \longrightarrow [x_1\ x_2\ x_j\ \ldots\ x_m] \quad (6)$$

We used default parameters provided with numpy API of python. We Applied the PCA on feature vector of hybrid feature extractor and provided the reduced dataset as training set to SVM.

E. Incorporating Sentiment Analysis

Sentiment distribution for the tweets relevant to a single party is indicative of the sentiment towards that party but, the single party distribution is not sufficient to predict an outcome. For an event like an election, relative sentiment between the parties conceivably has as much of an influence.

On twitter, the users with more followers get to be more influential over other users due to their widespread by Evan Rosenman et. al. [28]. They also established that retweets and mentions have very high correlation value using Spearman's rank correlation equation.

$$\rho = 1 - \frac{6\sum (x_i - y_i)^2}{N^3 - N} \quad (7)$$

Following the results from Meeyong cha et. al [4] that follower or mention count does not necessarily show the influence, we decide to consider retweets count as an influence factor.

Retweets are done in two ways over twitter. First way allows the user to share the original tweet after embedding a comment along with quotation mark or "RT" preceding the

original tweet. Second is the default "retweet" button given by twitter. Count of retweets done by second method is available as part of meta-data of original tweet. We consider retweets count as, the tweets that start with RT without embedding any comment in a window of 100 to the original tweet and the retweet count which is available as a part of meta-data.

We consider the ratio of summation of total positive tweets and corresponding retweets to the total positive tweets as a positive influence factor - C for the party or the target query. We count the negative influence factor - D on similar way.

We use the Sentiment based Volume method as described in [28], we evaluate the influence factor with the linear regression method as,

$$y = \beta t_i(\mathrm{A}-\mathrm{B}) + \beta t_j(\mathrm{C}-D) \qquad (8)$$

Where, βt_j represents the weight coefficient for tweet sentiment volume for party j, βi_j represents influence coefficient for influence. A, B represents positive sentiment scores where C,D represents influence scores to positive and negative factors.

F. *Analysing Gender based Votes*

We analyze gender based result trends for main presidential candidates based on term frequency. NLTK Names data list made up of almost 10000 male-female names is used to achieve the task. This list provides approximately 80% accuracy in guessing the gender of the twitter user correctly. Term frequency for presidential election candidate names and the related terms have been taken into consideration.

IV. Results and Discussion

We use the dataset collected using twitter API containing tweets in English language, for the original terms and extended features.

Data cleaning and normalization process leaves us with 47.8% of original data. Table-II shows the results after performing normalization step.

TABLE II. Effects of Data Cleaning and Normalization

Reduction Method	# of Features	% of Original
None	978857	100
UserNames	319890	32.68
URLs	180011	18.39
Repeated Words	35238	3.6
Repeated Characters	15172	1.55
Duplicate Tweets	121084	12.37
Official Account Tweets	160238	16.37

We used bag-of-words, bigrams, and Unigram + Bigram in order to achieve the sentiment values for each model. Hybrid model provides the best results with all the classifiers when mixed with modified extended feature vector model.

Prediction model explained in equation (8) is independent of the classifier employed or the feature extraction methods used for sentiment analysis, however it depends on the outcome of the sentiment analysis.

We follow 10-fold accuracy matrix with datasets taken at random. Each time the results are compared with prediction results of uselectorals.org[7] for popular votes. Like any other method, our method also faces the constraints of real-time voting parameters compared to social media voting scenario. We accept the result as valid if the results match with difference of 2.5%. We achieve best sentiment accuracy with SVM, which also leads us to the best accuracy results. Among feature extraction model, Hybrid model of Unigram + Bigram provides best accuracy results. Figure-3 provides the results obtained with our methods.

SVM outperforms MaxEnt and Naive Bayes for the process of feature presence verification for sentiment analysis [2]. Notwithstanding the structural similarities in the output function of SVM and ANN, divergence found in the way the solutions is obtained. An important advantage of SVM over ANN lies in the use of optimization approach; SVM obtains the support vectors in a convex optimization problem by always finding the global minimum with an unique solution, whereas ANN is trained with gradient descent methods, which may not converge to an optimal/global solution [25].

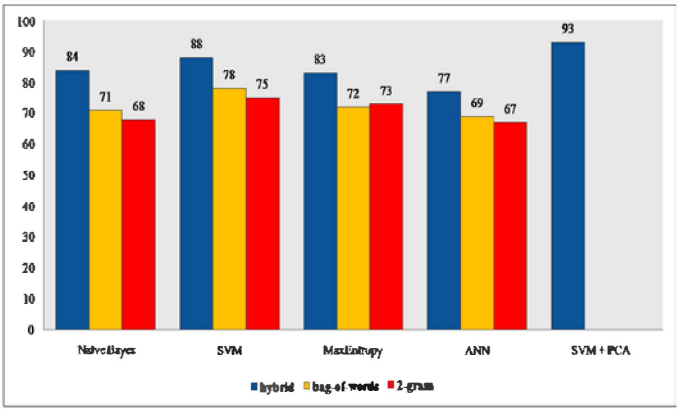

Figure 3: Prediction of US Presidential Elections 2012

We selected a subset of our dataset to test SVM ensemble with PCA, which resulted in providing rise in accuracy of approximately 5% with our model in the best case scenario. Overall, PCA collaborated with SVM provides better accuracy of 2% approximately on an average. However, reducing dimensions affects the accuracy in an inconsistent manner.

As per twitter data collected for terms specific to US election, Democratic candidate Mr Barack Obama was ahead by margin of 10.8 % on popular votes by getting 55.4% votes on twitter to Republican candidate Mr Mitt Romney's vote share of 44.6%.

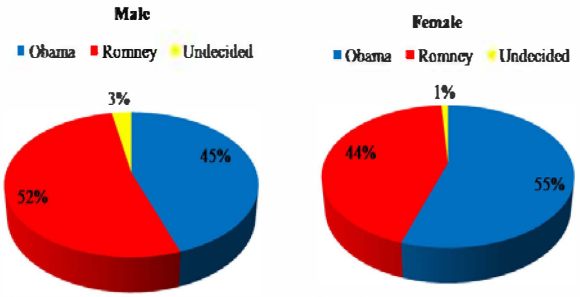

Figure 4: Gender Based Vote for Presidential Candidates

7. http://uselectorals.org/

Figure-4 shows the gender based results for USA Presidential Candidates of 2012 and it is observed that 55% of females inclined towards the Democratic candidate Mr Barack Obama; whereas, 52% male voters inclined towards Mr Mitt Romney, and 3% remains undecided.

Twitter based sentiment results when incorporated with twitter features for influence factor, provides a reasonable amount of accuracy for election predictions for US presidential elections 2012. However, Karnataka Assembly Elections 2013 results show entirely different picture of the twitter based prediction method with mere accuracy of 58% over 50 iterations and the vote-share results swinging in favour of Bhartiya Janta Party (BJP).

We followed similar steps for Karnataka elections of 2013 prediction attempt as applied for US Elections 2012 results with gathering over 25000 tweets. We employed Hinglish (Hindi words written in English fonts) word lists in order to parse Indian tweets written in regional language. Table-III shows comparison of the vote share in percentage on twitter to the actual results.

TABLE III. VOTE SHARE BASED ON TWITTER SENTIMENT ANALYSIS

Party Name	Experiment Results (%)	Original Results (%)
BJP	37	20
INC	23	37
JDS	12	20
KJP	28	11
Other	0	12

We performed the classification using SVM with extended features model and hybrid extractors, trusting the results achieved by US Presidential election outcome. We performed the SVM with PCA as well which yields improved results in the case scenario, which can be seen in Figure-5.

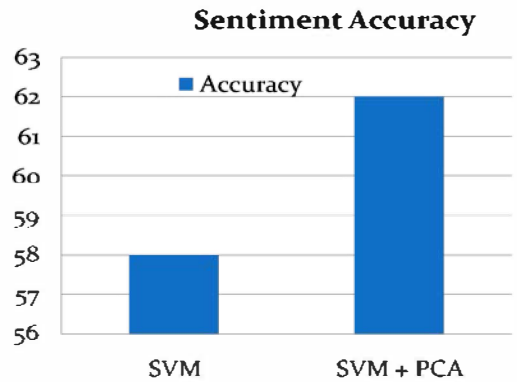

Figure 5: Sentiment Accuracy Comparison for SVM with PCA

The large variation in results is caused by the moderate level of accuracy for sentiment analysis of the twitter data. Large chunk of tweets skewed towards two parties BJP and Indian National Congress (INC) on twitter damages the accuracy for twitter votes; along with lack of presence of other parties like Janta Dal Secular (JDS), Karnataka Janata Paksha (KJP) and others yields reduced accuracy. Overall more number of negative tweets and corresponding re-tweets add up to negative value and lesser prediction accuracy towards the results, as it can be seen from Figure-6.

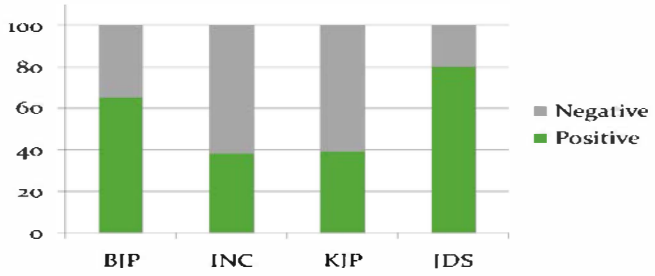

Figure 6: Sentiment Variation for Parties of Karnataka

Twitter based poll prediction results also depend on what percentage of the actual population that is going to vote is using twitter to express their opinions. As it is evident from figure-7, United States has very high population count active in twitter whereas, countries like India doesn't have sufficient amount of users on twitter. Deficiency of accurate locations at state level for the tweets also leads to consideration of twitter votes across the nation rather than opinions from a single state wide geographic area. As an example, a person sitting in other state of India like Delhi, may also tweet related to elections of Karnataka assembly; but, the user will not be allowed to cast his vote in actual elections.

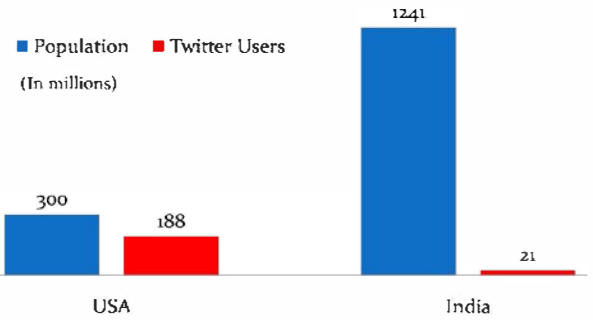

Figure 7: Population and Twitter Users of USA and India

V. CONCLUSION

Overall, we conclude that social network based behavioral analysis parameters can increase the prediction accuracy along with sentiment analysis. However, presence of all the entities in unbiased and equal manner is necessary to provide accurate results. To understand the influential parameters that effect the results, semantic features are also very useful from pointview of the entity itself. Twitter based social networks provides a great platform in measuring the public opinion with the reasonable accuracy of ~88% in case of US Presidential Elections 2012 and Karnataka Assembly Elections 2013 with SVM based supervised machine learning algorithms for sentiment analysis. We observed that PCA incorporated with SVM is helpful in reducing dimensions and achieving better accuracy but does not necessarily provide consistent output. Further, the inclusion of more influential factors based on the personal details such as age, educational background, employment, economic criterion, rural & urban and social development index will further enhance the poll prediction process to even higher prediction levels. Since the twitter data is increasing exponentially, the future requirement of opinion mining should be based on efficient, fast and hybrid classifiers using parallel computing.

REFERENCES

[1] Alexander Pak and Patrick Paroubek. "Twitter as a corpus for sentiment analysis and opinion mining", Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), may 2010.

[2] Bo Pang. Lilliam Lee, "Seeing Stars: Exploiting class relationships fpr sentiment categorization with respect to rating scales", 2002

[3] Brendon O'Connor and Balasubramanyan et al,From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series , Proceedings of the International AAAI Conference on Weblogs and Social Media, Washington, DC, May 2010

[4] Meeyoung, C. etl al., Measuring User In uence in Twitter: The Million Follower Fallacy. In Fourth International AAAI Conference on Weblogs and Social Media, May 2010.

[5] Antonie Boutet et al., What's in your tweet: I know Who You Supported in the UK 2010 general elections, Association for the Advancement of Artificial Intelligence, 2012

[6] Johan Bollen, Alberto Pepe, and Huina Mao. 2011. "Modeling public mood and emotion: Twitter sentiment and socioeconomic phenomena" . In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011), Barcelona, Spain

[7] Marko Skoric, Nathaniel Poor, Palakorn Achananuparp, Ee-Peng Lim, Jing Jiang et al. "Tweets and Votes: A Study of the 2011 Singapore General Election" In proceedings at 2012 45th Hawaii International Conference on System Sciences.

[8] Romero, D. M., Meeder, B., and Kleinberg, J., "Differences in the Mechanics of Information Diffusion Across Topics: Idioms, Political Hashtags, and Complex Contagion on Twitter", Proceedings of the 20th International Conference on World Wide Web (WWW), 2011, doi 10.1145/1963405.1963503

[9] Cozma, R., and Chen, K., "Congressional Candidates" Use of Twitter During the 2010 Midterm Elections: A Wasted Opportunity?" 61st Annual Conference of the International Communication Association, 2011

[10] Pew Research Center, "Parsing Election Day Media: How the Midterms Message Varied by Platform", Pew, 2010

[11] Rodrigo Moraes, Joao Francisco Valiati, Document-level sentiment classification: An empirical comparison between SVM and ANN, Elseveir Transactions for Expert Systems with Applications 40 (2013), pages 621-633

[12] Long-Sheng Chen, Cheng-Hsiang Liu, Hui-Ju Chiu, A neural network based approach for sentiment classification in the Blogosphere, Elseveir,Journal of Informatics 5 (2011), pages 313-322

[13] Song M, Kim Chul M, RT^2M : Real-time Twitter Trend Mining System, 2013 International Conference on Social Intelligence and Technology, DOI 10.1109/SOCIETY.2013.19

[14] Parikh and Movassate , Sentiment Analysis of User-Generated Twitter Updates using Various Classification Techniques , Stanford University, 2009

[15] Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on Twitter from biased and noisy data.In Proc. of COLING

[16] Peter D. Turney. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In proceedings of ACL 2002.

[17] Long Jiang et al, Target-dependent Twitter Sentiment Classification, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics

[18] N. Cristianini and J. Shawe-Taylor. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, March 2000.

[19] M Ashraf et. al. "Multinomial Naive Bayes for Text Categorization Revisited" , University of Waikato

[20] Alec Go et al , Twitter sentiment classification using distant supervision, Stanford University

[21] D. O. Computer, C. wei Hsu, C. chung Chang, and C. jen Lin. A practical guide to support vector classification chih-wei hsu, chih-chung chang, and chih-jen lin. Technical report, 2003.

[22] Pang and Lee, 2002,Sentiment Classification using Machine Learning Techniques , Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, July 2002

[23] Adam bermingham and Alan F. Smeaton , On using Twitter to monitor polticial sentiment and predict election results, Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), IJCNLP 2011, pages 2–10, Chiang Mai, Thailand, November 13, 2011

[24] K. Nigam, J. Laverty, and A. Mccallum. Using maximum entropy for text classification. In IJCAI-99 Workshop on Machine Learning for Information Filtering, pages 61 to 67

[25] Hastie, T., Tibshirani, R., & Friedman, J. (2001). The Elements of Statistical Learning. New York, NY, USA: Springer New York Inc.

[26] Russell, S. J., Norvig, P., & Davis, E. (2010). Artificial intelligence: A modern approach. Prentice Hall.

[27] Gutierrez R., Dimensionality reduction (PCA). http://courses.cs.tamu.edu/rgutier/cs790_w02/l5.pdf

[28] Evan T.R. Rosenman, Retweets,but Not Just Retweets, 2011

[29] Asur and Huberman ,Predicting the future with Social Media, WI-IAT '10 Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01, Pages 492-499