ELECTION ANALYSIS USING MACHINE LEARNING

Rahul Kumar Singh, Md Kaif Mustafa

REVA UNIVERSITY BANGALORE

***Abstract:—*What other people think” has always been makes us curious while indulge in decision making process. Earlier, many of us asked our personal acquaintances to recommend a political candidate or political party to whom we can vote or consult a broker to decide in which share we should invest our money. With the invent of Internet and world wide web have now let it happen to find out large pool of online users to share recommendation, opinion, views about an event. Although these online users are never our personal acquaintance nor professional critics. An irony of situation is many people are sharing opinions available to completely unknown person via Internet. In the beginning of 21st century, social networking websites taken exponential growth in popularity among online users by providing them a platform for discussion and exchange information. Apart from posting photos and videos for event, now online users are posting their comments, views, opinion about event. With this platform, most researchers are analyzing posts or opinion, performing computational methods to come to conclusion about popularity of event. This analysis part is becoming more powerful day by day to gauge user’s opinion and would play a significant role for someone to make major decision in future. The social microblogging sites such as Pinterest, Tumblr, Apsense, Scoop and specially Twitter are leading pools of online users and contains massive information that can be useful by extracting knowledge from information and that knowledge become asset for business specialist, political specialist, marketing specialist etc.**

***Keywords: Sentiment Analysis, Machine Learning, Politics***

# I. Introduction

|  |
| --- |
| Copyright © IEEE–2021 ISBN: 978-1-6654-3970-1 |

Opinions are central to almost all human activities and influencers of our behaviors. Humans’ belief, perceptions of reality, choices that we make generally depends upon how others are evaluating things. For this reason, we often need to seek others’ opinions when we are in decision making prices. This is fundamental truth for individual and for any organization as well. Sentiment Analysis also consider as Opinion Mining is a field of study that analyze people’s emotions in most of the cases and a special field in textm ining. It is a study that analyze people’s emotions, views, opinion, attitudes towards entities such as product, individual, event, issue and their attribute. With the rapid development of Information and Communication Technologies (ICT), social networking sites now have potential to transform the whole political arena into computerized world by introducing the whole new concept of generate, regenerate and share information in a wide scale of political context[1]. In order to extract knowledge from information, Sentiment Analysis techniques can be used[2].

2021 10

th International Conference on System Modeling & Advancement in Research Trends (SMAR

: 10.1109/SMART52563.2021.9676303

# A. Sentiment Analysis Research

In the beginning of 21st century, Sentiment Analyses has become highly active area of Natural Language Processing (NLP) due to many factors. The major reason was rise of machine learning in NLP for information retrieval. Secondly, the availability of datasets on which many ML algorithms were trained. Last but not least, awareness of intellectual commercial challenges and applications[3]. Sentiment Analysis can be investigated at three levels:

* *Document-Level:*In this, Sentiment Analysis consider the whole single document as information unit and classify the document with express into positive or negative sentiment. For example, given a service review of automobile company, the system determines the overall expression of review in either positive or negative[4].
* *Sentence-Level:* In this level of classification, Sentiment Analysis determine a sentence and express an opinion into positive, negative or neutral classification. This analysis is closely related to subject classification to distinguished factual sentences from subjective sentences. However, sentence level SA are unable to provide the critical details about various aspects of sentence. For example, “Business laptops are heavy in weight but durable”. In this sentence, different opinion holders may have different opinions because aspects related to sentence does not analysed[4].
* *Aspect-Level:* This level of classification is more refined and more generous to express opinion itself. Instead of focus on language constructions like documents, clauses, paragraphs etc. aspect level classification focus on opinion itself. It is also called as feature-level which con- sider the opinion and specially target for whom opinion express. For example,” Although, it costs me an arm and leg to purchase Benz but this is my dream come true car.” clearly expressed a positive vibe even therei s some sort of negative sentiment. In fact, sentence is negative about cost of car but more emphasized on Dream come true. Hence, the main goal of aspect level is to identify opinions about entity in consideration of its related aspects[3]

# B. Sentiment Analysis Classification Techniques

In social websites, most of the information generated from online users and the major task of sentiment analysis is to extract knowledge from posted information. Later, compared the sentiment extract from these posts is again major challenge because generally social media posts contain unique characteristics of information. Further, text length limitation and informal text contents bring many difficulties and chal- lenges. In following figure 1, Sentiment Analysis classification techniques can roughly have divided into three major classes as Lexicon-based approach, Machine Learning approach and Hybrid approach.

1. *Lexicon-Based Approach*: Lexicon-based approach which is often called rule-based approach depends upon sen- timent dictionary which contains some precompiled words or terms with their association strength. This approach fur ther divided into ”Dictionary-based” approach which extract the opinion from text and then identified its corresponding antonyms and synonyms from dictionary whereas in ”Corpus- based” approach extract opinion from text in context of spe- cific orientation and then find synonym opinion from corpus and then used computational statistics methods to find a sen- timent polarity [5],[6]. There are plenty of lexicons available for SA as WordNet, SentiWordNet, SentiTFIDF, SebticNet, Sentiful etc.
2. *Machine Learning Approach*: Machine Learning ap- proach plays significant role in Sentiment Analysis text clas- sification. Various Machine Learning algorithms are trained with large volume of training data before test models. For example, in text classification problem where set of training data-set S = X1, X2, X3......Xn where each X(record) is labelled to a class. A test classification model is related to the feature in the underlying record to one of the class labels. For given instance ’Y’ on unlabeled class, the model is used to predict class label for ’Y’. It is some time challenge to express opinion in a text when there is only single label assigned to an instance called as hard classification whereas in soft classification, probabilistic value of labels is assigned to NA instance[4]. Further, in text classification problem where topic-related words are key features which express opinions in positive or negative such as good, amazing, anger, anxiety, bad, worst etc. Machine learning further divided into” supervised learning” and” unsupervised learning”. In supervised learning where labelled training data consist pair of inputs and their corresponding correct outputs. During training, the algorithm search for pattern in dataset to correlate the inputs and their corresponding outputs. After training, trained algorithm tests with new and unseen inputs and determine their outputs in previous labelled training data[7]. Most common and frequent used machine learning classifiers are: • Linear Classifier
   * Naive Bayes Classifier
   * Support Vector Machine
   * Decision Tree
   * Random Forest

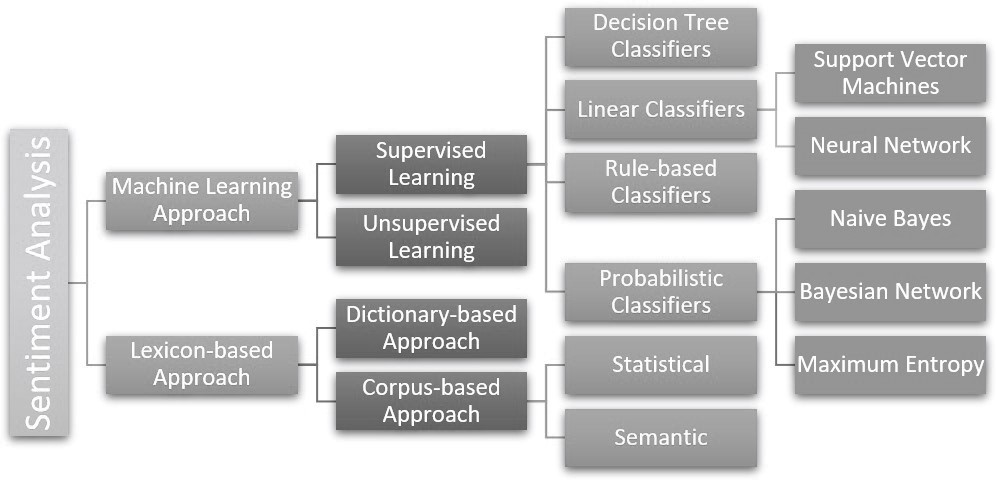


Fig. 1:Sentiment Analysis Classification Techniques

1. *Hybrid Approach:* In hybrid approach, any combination of above-mentioned approaches can use when require for accu- rate sentiment analysis. This approach generally implements in situation where algorithm can easily have fine-tuned by design rules for noise and conflict categories which have not been trained successfully[8].

## *C. Sentiment Analysis Applications*

Earlier, business corporates always wanted to know the consumer feedback about their products or services to analyze the popularity of their product or services among society. For this, organizations conducted offline surveys with limited people. For an individual need’s opinion from an existing users or friends before purchase a product. In another, people generally seek others recommendation for vote to political party or political candidate. But now a day, social media play significant role to make decision more accurate and precise with the involvement of big pool of online users and their massive opinions. Sentiment Analysis applications are not only bound with consumer products, healthcare, financial services, social services, political election but also plays a significant role in application orientated research. To illustrate, In Ger- man Federal Elections 2009, twitter was first microblogging website to predict the results[9]. In this study, authors claim high rate of accuracy in predict a winner on the basis of collection of tweets for each political party. Although, this study faced massive criticism globally by researchers because it was conducted post-election.

## II. Literature Review

In this, authors give close eye on summarization, en- hancements, challenges, future scope and categorized research papers published in this field since 2009. Research papers are categorized upon various parameters such as objective, algorithm, data scope, data set, their performance parameters and future scope. In Table:1 about 21 papers reviewed in this literature review organized in ten columns. The fourth column specifies whether the study detailed in this paper are domain oriented or not. The eighth column indicates each algorithm in paper tackle general analysis (GA) or binary classification (BC) with respect to positive or negative.

**In this paper** [10], Makazhanov *et. al*. developed an election outcome prediction model in 2012 Albertan and 2013 Pakistan General Elections on various behavioral features. Researchers claimed that their model outperformed in F-measures, Senti- ment and Text classification and further used to analyze voter’s behavior changes with respect to election campaign. In data collection process, researchers first differentiate political candidate and non-political candidate’s twitter account on the basis of frequency of tweet posted by each account’s category on daily basis. This way researchers collected 252 political candidate’s accounts along with four accounts each from four political parties. For non-political candidate’s account, researchers used selected 27,822 accounts on the basis on 27 keywords such as party hashtag, party names, candidate names and so on mentioning in their posts over a period of 10 days before actual elections. Total 181,972 tweets were collected from these accounts and final 3000 tweets were selected for analysis. Researchers used Support Vector Machine (SVM) and Logistic Regression (LR) to remove spammer accounts and performed 10-fold cross validation to validate dataset. Researchers claimed that LR classifier performed slightly better than SVM with F-measure of 85% and 94% for each classifier in classification of accounts. For prediction of political preferences, researchers used Decision Tree and Logistic Regression classifier as training classifier each for every party and present the result. Each result consists seven bars with respect to two classifiers (LR and DT), three human annotators and three baselines.

Researchers claimed that human annotators performed slightly better than classifiers for two out of four political parties whereas LR shows better predictions than DT.

**In this paper** [11], Grant *et. al*. analysed twitter’s utilization among Australian politician to find out how many politicians are actually tweets on twitter, how often they are tweeting, what benefits actually they are getting from tweets and finally the significance role of social media in Australian politicians. For this, authors collected politician’s name list from various resources such as twitter links on some politician’s websites, Google advance search and through crawling politician’s follow list in their tweets. However; with this author collected massive amount of twitter accounts because there was large dataset of fake accounts also. To remove those, authors considered only those accounts that reflects the identification of politicians such as standard politician image in their pho- tographs, their Twitter’s biographies and their web-links from twitter profile. Further, authors used ’random ()’ function to collect random sample of Australian Twitter and Australian politicians who were signed at same time period. This way authors claimed Australian politicians are more engaged than normal Australian users in posting tweets. Further, researchers also claimed that politician’s followers list is more than they actually follow. Most interesting some politicians got retweets by their followers which some time establish popularity among people. However, authors were unable to answer a question on uptake of social media in Australian politics.

**In this paper** [12], Casarin et. al. analysed total 69202 political tweets of 56 politicians using Bayesian method in Columbian political election held in year 2015 in which researchers found the probability of retweets on the basis of combination of sensitive words in original tweets which in turn had a huge impact on political outcomes. For data collection, researchers used ’twittR’ package has an interface to Twitter web API and created as single text document in Spanish language for each politician’s tweets. Further, tweets were tokenized in Spanish trigrams. During analysis, researchers analysed each document as an exchangeable set of tokens say unigrams or more general n-grams and further understood tokens as stemmed words such as ’taxation’, ’taxes’ and ’taxing’ become ’tax’ and then applied Bayesian approach to identified a maximum posterior for efficient calculation. In result, researchers computed text sentiment reduction score and plotted it against logarithm number of votes in order to represent logarithm regression that proved R(square) = 0.41a s high significant score with p less than 0.01.

**In this paper** [13], Tsakalidas *et. al*. analysed 1077526 Twitter posts posted by politicians only during EU Elections 2014 in Germany, Greece and the Netherlands on time-series features. As elections were held on different time-series in these countries so authors created three different prediction models to predict the results. For this, researchers created 11 twitter-based and 01 poll-based feature for each party and later inputted all these features to three different prediction models to predict final result. This pre-poll twitter-based data aggregation contains party’s name, its abbreviation, its Twitter account name and some misspelled names. Later authors extracted Twitter-based features based upon past works for each party and then correlate them with opinion polls. Instead of train a classifier with labelled corpus of tweets, researchers used lexicon-based approach to generate generic algorithm applied of three different cases. In light of less sentiment lexicon for three different languages, researchers used Google Translator for English translator such as SentiWordNet, Opin- ion Lexicon and Subjectivity Lexicon and assigned values according to each lexicon in terms of polarities. For Sentiment Analysis, researchers used linear regression, Guassian process sequential minimal optimization in association of WEKA. Authors claimed that their Twitter-Poll-Based (TPB) model with respect to Guassian Process (Lowest MAE-1.31), Sequential Minimal Optimization (MAE-1.35), Linear Regression (MAE- 1.42) outperformed in terms of lowest Mean Absolute Error (MAE) and Mean Square Error (MSE).

**In this paper** [14], Kruikemeier *et. al*. studied interpersonal relationship between online Twitter campaigning and political involvement effects on citizens of country. In this study, researchers experimented on two key style characteristics of social media such as two dimensional communication (low vs high interactive communication the form of Direct Communication between political candidate and citizen) and three dimensional communication (depersonalized (Party focus), Individual (Politician focus) and Private communication (Private life focus). Experiment conducted on twitter accounts of 66 Dutch democrats with their political leader and young citizens of country. Researcher’s analysis revealed that highly interactive communication on different aspects of political as- pects specifically private information perceived social presence among citizens.

**In this paper** [15]**,** Jungherr *et. al*. analysed interpersonal relationship of Twitter posts with election outcomes during German Federal Election held in year 2009.In this study, re- searcher focused on four key metrics for analysis such as total number of hashtags mentioning political party, total number of hashtags mentioning leading candidate, relationship between explicitly positive and explicitly negative tweets of a political party and number of users who used hashtag mentioning any political party’s candidate. For data collection, researcher target only those candidates using hash tag whose tweet contents mentioned name of party/candidate or campaign related hashtags. With this way, researcher collected 10,085,982 tweets from 32,731 users over a period of four months before election day and after filtering with Wahlgetwitter website (tracked name of parties with combination of suffixes” +/-” (#party+ means supportive party and #party- means opposition party) in hashtags), authors analysed 224,551 messages from 11,212 users. After data collection, author predicted vote share on the basis of number of positive hashtags corresponding to political party/ candidate which was unlikely to happen in actual results. Authors identified a key point during analysis such that share of hashtags mentioning party/candidate was fluctuate on daily basis.

**In this paper** [16]**,** Ibrahim *et. al*. experimented 10 mil- lion twitter data over the period of 60 days and predictedt he results of Indonesian Presidential Election held in year 2014. In this, researchers performed automatic buzzer detection on twitter to remove any tweets posted by paid users, computer bot and fanatic user. Further, researchers divided each tweet into various sub tweets and associate each with subsequently candidate along with polarity because in single tweets multiple candidates might be discuss with different polarity. For automatic buzzer detection of computer bot or paid user, researchers employed machine learning approach in their model with define set to characteristics to detect noise such as average time period, daily frequency, age of account, number of URLs mentioned in tweets, number of retweets and its frequency, followers list and following list. Researchers employed lexicon and keyword-based approach to identify the sentiments in political tweets and achieved 86% accuracy to remove noise from dataset. For sentiment analysis,r esearchers divided tweets with help of delimiters and assign candidate name with polarity. Researchers leveraged positive sub-tweets towards candidate and claimed experiment shows Mean Absolute Error (MAE) of twitter-based prediction is 0.61%.

**In this paper** [17]**,** Singh *et. al*. employed machine learning algorithm to analysed sentiment inside Twitter data during Punjab Assembly Election held in 2017. For tweet collection, researchers used ASP.Net integrated with Twitter streaming API tweetinvi on the basis of hash (#) tag corresponding to political party or candidate name and collected 9157 tweets (English and native language) a month before election day.T o make reliable pre-processing authors used R Language to remove unwanted tweets, translated tweets in native language (Punjabi) to English language, remove stop words/web links, remove extra blank spaces and then converts all tweets into lower case. Further, authors performed hash (#) tag analysis, mention () analysis and network analysis in order to make detailed summary of tweet statistical. For polarity analysis of tweets, authors used WEKA 3.8 and built classification model using Support Vector Machine (SVM), kNN (k-Nearest Neighbor) and Decision Tree (DT). Authors claimed SVM outperformed against other classification models with 10- fold cross validation. For winning seat predictions, authors proposed forecasting method with an algorithm which calculate Actual Sentiment Score (ASS) for each party and also identified historic data on the basis of vote sharing and then applied Linear regression for actual seat prediction. With this, authors claimed that their model achieved 43.2% accuracy in ASS against actual vote share (%) of INC (38.5%).

**In this** **paper** [18]**,** Joe *et. al*. introduced the methodology to predict outcome of Indian General Election held in year 2019. Researcher used decision tree classifier to train and test the data in a form of tweets collected from Twitter. Jupyter “tweepy” library was used to extract tweets from Twitter in association of Twitter API connectivity. Further, tweets were stored in MongoDB using “pymongo” library. In preliminary experimentation, regular expressions and emoji were filtered as these symbols not interpreted to any sentiment. This preliminary experiment was conduct using Artificial Neural Network, Naive Bayes (NB) and Support Vector Machine (SVM). However, these classifiers were not good enough to experiment by creating a tree of grammar from Parts of Speech (POS tag). Then, authors used decision tree classifier to evaluate the polarity, gauge the public mood and calculate popularity score with respect to classified tweets in positive, negative and neutral. This score was calculated for fifty days in routine before election day. Authors showed the result that ruling party (BJP) obtained highest popularity score 72% as compared to opposition parties as 49%. Researchers claimed that proposed methodology achieved 97% accuracy prediction result using Decision Tree Classifier.

**In this paper** [19], Agarwal *et. al*. proposed a semi automated system hashtag-based approach to separate the tweets to develop a sentiment classification model to generate training data. Further, authors combined deep learning classification models to improve prediction accuracy. Subsequently, the trained model is then used to categorized tweets into target political party’s sentiment classes. For this, researchers firstly collected 41 million tweets in months of April and May, 2019 from Twitter streaming API by using hashtags (corresponding to BJP and INC) and bounding box. In their study, researchers used Long-short term memory network (LSTM) for classification and then classification models augmented with neural network word embedding such as “Word2vec”, “Glove” and “fasttext” in order to achieve better prediction accuracy. Authors claimed that the proposed model achieved 87.30% accuracy in classification of tweets corresponding to political party’s sentiment classes.

**In this paper** [20]**,** Singh *et. al*. studied displeasure among common people from different states of India during demon- etization of high Indian currency by Indian government on November 2016. In this study, researchers used total 30,220 Twitter posts collected from “Tweetinvi” API associate with VS2012 with parameters such as tweet IDs, tweets, dates, senders and location. Researchers analysed sentiments behind tweets and classified into five categories (Highly Positive, Positive, Neutral, Negative, Highly Negative) using SA APIi n association of MS Excel and conducted analysis on two phases as banking and nonbanking days for unbiased results. Researchers claimed that people supported this decision on announcement day and following day that was no-banking day by tweeting 38.31% high number of positive tweets against 26.03% negative tweets but on banking day situation turned problematic because of non-availability of new banknotes and received 32.42% positive tweets and 33.32% negative tweets. Due to diverse geographical size of Indian states, researchers used scoring method to classified states into further six states (Very Happy, Happy, Neutral, Sad, Very Sad, No Data) to analysed supportiveness of policy.

**In this paper** [21]**,** Pathak *et*. *al*. analysed total 8,877,275 social media buzz posted during 100 days for twelve Indian political parties during Indian General Election held in year 2014.In this study, researchers collected massive social media buzz from simplify 360*◦* https://simplify360.com/ January 1, 2014 to April 09,2014 for analysis with inspiration of [9] where merely counting of tweets against predicting seats would be an only indicator. Further, researchers used standard error estimation, root mean squared error, and R2 in theirm odel where R2 with model achieved 72.9% variance in the data and indicated that political party can achieve higher number of seats if party has well developed plan for social media campaign. Although researchers mentioned social media buzz may not always give accurate results because of diverse state- wise political strategies, different literacy rate among states and low social media penetration in India but separate IT cells, paid agents who can tweet and retweet in association of partly goal can increase seat winning chances.

**In this paper** [22], Khan *et. al*. proposed a framework to analyze tweet data for classification in order to gauze trends. For this, researchers collected tweets from Streaming API for real time data capturing with geo-location parameter and then stored in text file for further analysis. For tweet pre-processing, researchers used Apache Spark to create Resilient Distributed Datasets (RDD), split tweets as BOW, removed stop words and performed stemming of words. For tweets classification into political and non-political tweets researchers performed calculative formulas and tagged 20,000 Indian tweets as polit- ical or non-political using starting set of 387 words in political word list. In their study, researchers took 1000 pre-processed tweets for training and test dataset each and compared their algorithm’s result with Naive Bayes algorithm and claimed that their algorithm achieved 85% accuracy whereas NB achieved 55%.

**In this paper** [23], Jose *et. al*. used set of Machine Learning classifiers with lexicon-based classifier in order to deter- mine the accuracy of classifier on unseen data. Researchers used combination of Naive Bayes (NB), “SentiWordNet” and Hidden Markov Model (HMM) classifiers to classify the sentiments in political tweets and sentiments inside tweets for upcoming movies. For data acquisition, researchers used Twitter streaming API to collect 12000 tweets about Arvind Kejriwal and Kiran Bedi for 3 weeks before Delhi Election Day then classified data by removal of URL tags, removal of personal identification by @ user tag and at last removal of Hash (#) symbol and special characters. Further, to han- dle negation researchers used algorithm with state variables (stores negation state) and bootstrapping. With this, algorithm transforms word follow by “nt” or “not” into” not” + word. For sentiment classification for political tweets, researchers used various classifiers such as “SentiWordNet”, “NB” and “HMM”. SentiWordNet and WordNet and assign sentiment numeric scores, positive, negative and objectivity whereas NB compute posterior probability of a class. For sentiment classification for movie’s tweets, researchers used ensembles approach. In their study, researchers claimed that classification using ensembles approach achieved high accuracy (71.48%) than each individual classifier (SentiWordNet:21.05%, NB: 69.92%, HMM: 64.06%).

I**n this paper** [24], Ahmed *et*. *al.* investigated use of social micro blogging site ’Twitter’ during election campaign in Indian General Election 2014.Authors presented multi-level manual and computerized analysis of 98,363 tweets from eleven political parties, two months prior to election day out of which 4% tweets were in non-English language. Re- searchers analysed political party’s election campaign strate- gies on Twitter by identified four major concern such as frequency of usage by parties, dominant issues tweeted by parties, platform of integration data collection, researchers used ”Symosos”(http://www. symoses.comn website and from Election Commission of India (ECI) website of which analysis was conducted on 30% sample of every Twitter’s account except BJP for which authors sampled 4000 out of 80,981 tweets. To analysed dominant issued discussion, researchers used probabilistic techniques and found that AAP (Diverse in issues) cited slogan of Indian Independence, AITC cited impression of strong candidate leader, BJP (most consistent) cited thankful messages to its users for donating to party funds. Most significant, INC cited” Rahul Gandhi” as party supreme in most of its tweets and reiterate allegations on opposition parties, especially BJP. For interaction among parties, re- searchers claimed that BJP primarily used Twitter for re-tweets (61%), AITC primarily used Twitter for normal posts (89%), AAP for re-tweets (60%) and INC seems more balanced with new post (36%), mentioned party leaders (30%) and re-tweets (27%) and least focus on interaction (6%)

**In this paper** [25]**,** Sharma *et. al*. analysed twitter posts in different language to predict the outcome of state assembly election in year 2016.For this researcher collected 42,235 tweets political tweets for five national parties over a period ofm onth. Researchers built a classifier model using SVM, Naive Bayes, Decision-Tree and Dictionary based to classify the tweets into positive, negative and neutral. For data collection, researchers used google Spreadsheet to establish connection with Twitter using Google script by identified key-details from twitter account and importing all the search result into Spread- sheet. Researchers used Hash tag (#) to collect tweets for and against the different political parties. After pre-processingo f dataset, researchers analysed the polarity of 23,998 using SentiWordNet and performed different classifiers to predict result outcomes. Authors claimed that SVM achieved 78.4% accuracy outperformed other classifiers.

**In this paper** [26], Khatua *et. al*. collected 0.4 million tweets by using ’twitteR’ and Twitter steaming API during period from March to April, 2014 in Indian general election held in year 2014. Researchers identified set of keywords while collection of data such as political parties name, their prominent candidates, important constituencies. However, researchers later realized that this static data categorization lost its impact in long period of data collection and couldn’t capture the trending topics in a country. For this, researchers identified 15 political sensitive cities across countries and developed a program to identified latest ten trending topics in routine to make research more dynamic. Researchers then classified tweets into positive, negative or neutral polarities and explore the relationship between sentiment score and vote swing rather predict the result outcome on the basis of mention candidate/party name in tweets volume.

**In this paper** [27], Srivastava *et. al*. analysed tweets dataset to predict outcome of Delhi Assembly Election held in year 2014.Researchers used minimum root mean square error (RSME) to map the sentiment in tweets to the seat counts between AAP, INC and BJP. Researchers also trained event specific and time specific classification model with manually annotated to capture sentiment of users. Researchers collected and cleansed 3,52,730 tweets a month before election day for three major parties AAP, INC and BJP. For classify tweets into positive and negative polarities, researchers constructed training data set by use of IMDB training dataset, manually annotated from collected data and Automatic annotated from collected data. For prediction result, researchers used SENTIWORDNET for subjective test about any party irrespective of polarity and used SVM on WEKA to determine polarity of each tweet corresponding to a party. Researchers then obtained positive sentiment share of each party from total positive sentiment shares in whole dataset. In order to map vote, share to seat share, researchers analysed positive sentiment share (PSS) and RMSE for each party and claimed SVM achieved 93.7% accuracy with 10-fold cross validation for polarity classification.

**In this paper** [28], Almatrafi *et. al*. analysed twitter trends based upon their geographical locations during Indian General Election held in 2014.For analysis, authors collected 650,000 tweets for two mainly political parties AAP and BJP over a period of five days during campaign and built binary task classifier (Positive or Negative) using Naive Bayes algorithm. For data collection, authors used Tweepy API in Python and Collect tweets, re-tweets and favorite tweets based upon search term in hash tag. Researchers AWS EC2 for monitoring tweets, store it in csv file and later used MySQL database to load tweets from CSV file. For data pre-processing, researchers removed white spaces, transform’@’ symbols into ’AT USER’, convert all links into word ’URL’, replaced emoticons and abbreviations like ’OMG’, ’LOL’ into proper meanings etc. Researchers used Natural Language Toolkit 2.0 to build a classifier in association of NB. In order to train classifier, researchers used two balanced datasets representing positive and negative data sets for extract features. Authors claimed that their model achieved 70% accuracy to classify tweets, 80% recall and achieved 66% precision.

**In this paper** [29], Mehndiratta *et. al*. analysed 0.25 mil- lion political tweets posted before Indian Election in April 2014.Researchers used Twitter Streaming API called as twit- ter4j for collection of tweets and clean them by removal of hyperlinks, Hash (#) symbols, URLs, and removal of other lan- guage’s tweets except English. Researchers also used regular expression (regex) to remove other special characters in tweets. For tweet classification, researchers used logistic regression and conditional probability from LingPipe library to classify tweets into positive, negative polarities and generate training dataset. In their study, researchers showed that AAP party was most discussed party (gathered more than 0.12 million tweets) than other parties such as BJP gathered less 0.1 million tweets than AAP and INC stood third with 50,000 tweets. Researchers showed their analysis on positive and negative tweets by said that BJP got 82% positive and 18% negative tweets out of total 97,872 tweets, AAP got 40% in positive and 60% in negative tweets out of total 124387 tweets and INC got only 18% positive and 82% negative tweets out of total 54914 tweets.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: Prominent Research Publications in Recent Years   |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | **S. No.** | **Reference** | **Country** | **Year** | **Algorithm Used** | **Dataset** | **Tools** | **Polarity** | **Parameters** | | 1 | [17] | India | 2020 | K-NN  SVM  DT  LR | Political  Tweets | WEKA | Positive/  Negative | Precision | | 2 | [18] | India | 2019 | DT  NB  SVM | Political  Tweets | Tweepy | Positive/  Negative | Precision | | 3 | [19] | India | 2019 | LSTM | Political  Tweets | Twitter  Streaming  API | GA | Precision | | 4 | [21] | India | 2017 | LR  VM  NB | Political  Tweets | WEKA | Positive/  Negative | F-Measure  Recall  Precision | | 5 | [27] | India | 2015 | SVM | Political  Tweets | WEKS | Positive/  Negative | Precision |   III. Methodology Adopted by Prominent  Researchers in Recent Years *Tweet Preprocessing:*  **In paper** [17], researchers performed following set of i. Removed unrelated tweets from collected corpus. steps to reach at predict the election outcomes of Punjab ii. Convert tweets written in regional language into Assembly Election held in year 2017: English language and then convert all tweets into  *Tweet Collection* lower case.  i. Collect tweets of each political party from Twitter iii. Removed punctuations, numerical data, stop using Hashtag (#) symbol. words, web-links. |

**In this paper** [30], Caldarelli *et. al*. analysed quantify mea- sures of tweets posted in time evolution during Italian National Political Election in 2013. Researchers also focused on topics associated with the co-occurrence of two politicians in same tweet. Authors collected 3,378,442 tweets corresponded to one month prior to election and applied” Relative Support” parameter to examined relative strength of one political party against another political party by measuring the ratio of their temporal change with respect to election campaign or event. This parameter used the slopes of the time evolution in which tweets collected mentioning the name of political parties or their candidates to measure their support on Twitter. Authors claimed that” Twitter” is one way to get indication to election result outcome however they admitted that it over-predicted the political party’s vote number.

# Performed Social Media Analytics

1. Identified number of tweets, tweets type and number of users
2. Performed hashtag (#) analysis (Unique hashtag in collected tweets)
3. Performed mention (@) analysis (Unique candidate’s name mentioned in tweets)
4. Performed network analysis (Identified proportion of tweets counts for each political party)
5. Performed sentiment analysis vi. Performed E-Motion vii. Performed polarity analysis

# Seat Share Based on Polarity Analysis

**In paper** [18], researchers performed following set of steps to reach at predict the election outcomes of Indian General Elections held in year 2019:

# Tweet Collection

1. Collected tweets from Twitter database using Twitter API(Tweepy) and then stored in MongoDB using pymongo library.
2. At set time in routine, tweets collected from then Prime Minister’s twitter handles, ruling and opposition party’s tweeter handles, ruling party and opposition leaders handles with the conditions most retweets, most recent and in English language only.

# Tweet Preprocessing

i. Removed regular expression, emoji’s, non-English words, stop words and punctuation. ii. Tweets were sorted on based upon retweets

# Sentiment Analysis

1. Performed polarity analysis with respect to tweet classification using Decision Tree classifier in Textblob library.
2. Calculated popularity score of each party using formula:

Popularity = ((0 x Negative Tweet) + Neutral

Tweet / 2) + Positive tweets)/Total tweets

# Calculate Seat Winning Count on Average of Popularity Score for Each Poll Phase

*Compare the Obtained Result with Pre Poll Survey and Actual Result.*

**In paper** [19], researchers performed following set of steps to reach at predict the election outcomes of Indian General Elections held in year 2019:

# Tweet Collection

1. Collect tweets of each political party from Twitter using Hashtag (#) symbol and bounding box. Crawled the data set using Hashtag (#) corresponding to each political party, their prominent leaders.
2. Stored only useful information about tweets such that userID, tweetID, tweet, tweet location, date/ time, RT count, Fav count, hashtags and mentions.

# Tweet Preprocessing

1. Remove mention @ in tweets, hashtag # in tweets, URLs, Stop words except UserID, TweetID, Date/ Time, repeated words.
2. Replacement of abbreviation and slang words iii. Convert tweets into English language and into

lower case

# Hashtag Based Tweets Segregation

i. Identified and extract trending hashtag related to current topic of interest. ii. Performed targeted hashtag classification to sentiment analysis.

1. Build and train polarity classification model
2. Utilize trained model on real time tweets classification

# Sentiment Classification

1. Applied step “C” for each political party and then used each political party data set to train independent classification model.
2. Performed LSTM for tweets classification and then augmented with different vector representation.

**In this paper** [21], researchers performed following set of steps to reach at predict the election outcomes of Indian General Elections held in year 2014:

# Twitter Buzz

i. Collected twitter buzz from simplify 360 degrees (A market research company).

# Measure Accuracy in Predict Vote Share

i. Performed statistical analytical techniques such as R2, standard error of estimate, root mean square to measure accuracy.