International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-1, October, 2019

Sentiment Analysis of Twitter for Election Prediction

Balika J. Chellia, Kartikaya Srivastava, Jishnu Panja, Ritwik Paul

Abstract: Elections are considered to be the most important feature of a democracy. In the past few years, election analysis and predictions have become very important for political parties and news organizations. The influx of various social media platforms such as twitter, Facebook and YouTube have drawn a large number of people that share their ideological and political thoughts and hence, it's become important to analyse them in a much more sophisticated manner. Various data mining algorithms have been used to extract tweets and perform sentiment analysis pertaining to a related topic. Sentiment analysis refers to the technique to identify positive, negative or neutral opinions from a text. Though the use of sentiment analysis we will analyse the sentiment score for the two main political parties of India. The paper will brief on various techniques that have been used for election predictions. Various results from different methods have been included in this paper along with precision, accuracy and validity of the final outcome. The main aim of this paper is to create a model for the better prediction that will help in the analysis of voting choices of users. To increase the validity of the final results, various refining techniques have been used so that only relevant tweets are analysed.

Index Terms - Data mining, Election prediction, sentiment analysis, Twitter

I. INTRODUCTION

India has around 900 million registered voters, making it the worldś largest democracy. With more than 18 assembly by-elections happening in just 2019, it becomes very interesting to analyse the mood of the voters and predict, which candidates have a higher chance of winning. There have been many traditional polls that have been held over the past many years [1]. But with the rise of social media, especially Twitter in India, the political discussion has not just remained limited to drawing rooms, newspapers and magazines. Twitter recorded 45.6 million tweets during a one-month period in the run-up to the polls and registered 1.2 million tweets on the first day of voting on Thursday. This was a substantial amount of information and could be used to know the mood of the voters if harnessed in an effective way. India has more than 7.75 million active Twitter users as of 2019.

Revised Manuscript Received on October 30, 2019.
* Correspondence Author
Balika J. Chellia, Department of Computer Science and Engineering ,SRM Institute of Science and Technology, Ramapuram, Chennai, India
Kartikaya Srivastava, Department of Computer Science and Engineering ,SRM Institute of Science and Technology, Ramapuram, Chennai, India
Jishnu Panja, Department of Computer Science and Engineering ,SRM Institute of Science and Technology, Ramapuram, Chennai, India
Ritwik Paul, Department of Computer Science and Engineering ,SRM Institute of Science and Technology, Ramapuram, Chennai, India

Even in areas, where these technologies are used less, the impact of the outreach that a politician can have through twitter and other social media websites is greatly remarkable [2]. As the interaction between people and the social media websites becomes greater, the trend of the internet from lengthy posts to short tweets give us a greater insight into a person's ideological, political and religious beliefs [3]. All these factors are important in context to Indian elections. Through twitter, we can effectively estimate and predict a personś voting preference based on the politician he or she follows, their religious and linguistic habits, retweets and topics that they have liked. The same techniques can be used for a consumer survey of certain products. Various people that buy items online post reviews on them on twitter, and through the use of methodologies discussed here, the same model can be applied to get consumer satisfaction score. For better consumer experience, various interactions on other social media platforms, including twitter can be analyzed [4]. One of the main features that make India unique is itś linguistic diversity. It also becomes one of the important factors that decide voting patterns. In [5], the authors of the paper have taken into account tweets in a certain language and have used that dataset to perform sentiment analysis. While it is difficult to analyze tweets in different languages, it greatly reduces the overall dataset and certain variations that may come because of linguistic diversity and the divide may not be reflected in the overall score. Also in the same paper, the authors have not used emoticons, that are also an important aspect while calculating the polarity of a specific tweet. In India, linguistic divisions also play an important role in determining the income levels of people.

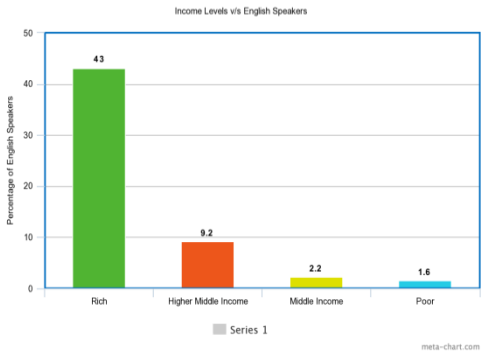

Fig 1. Percentage of English speakers compared to various income levels in India [21]

Thus, the percentage of English speakers varies greatly according to income levels. Hence making a dataset comprised of tweets just in English will not give us the full picture and the overall score may greatly vary as income level is also a crucial factor in making election predictions. The other factor that needs to be taken into account while making our dataset is to classify geotagged tweets.

Topics and discussions are found to greatly vary according to different geographical locations. Thus, it is essential that we take into account tweets from locations that we are focusing on. As this paper is made to analyze elections in the Indian context, we will refine tweets that come from India only. In this paper, we use the twitter dataset pertaining to the two main political parties, i.e the Congress and BJP. Based on the effectiveness, simplicity and accuracy we used Naive Bayes classifier to calculate to the polarity of the sentiment for these two political parties.

II. RELATED WORK

- Much research has been done in the field of sentiment analysis using tweets for election prediction. This part of the paper describes the various techniques that have been used for election prediction. Many papers have tried to identify the political and ideological leanings of the user to predict the voting choice of the given user. For e.g in [6], the authors used tweet containing parties' name in several political events to assign an ideological leaning of the user who posted the tweets. In [7], the authors use tweets from President Candidates of Indonesia, and tweets from relevant hashtags for sentiment analysis gathered from March to July 2018 to predict the Indonesian Presidential election result. The authors make an algorithm and method to count important data, top words and train the model and predict the polarity of the tweet. In [8], they also make use of the Naive Bayes algorithm to find the polarity of the sentiment. Though they have used other social media websites such as Facebook and Youtube that form their dataset. In the context of Indian elections, language becomes an important factor as explained above. Thus, we need to extract tweets in different languages too, [9] created a model to translate tweets in different languages for sentiment analysis. The method leveraged on lexical resources for sentiment analysis available in English (SentiWordNet). First, a document in a different language than English is translated into English using standard translation software. Then, the translated document is classified according to its sentiment into one of the classes "positive" and "negative". Other methods proposed by researchers are by:

(1) Utilizing interaction information between potential

voter and the candidates and

(2) Creating a trend line from the changes in the follower of the candidates.

Researchers, though still argue that a lot more still needs to go through a lot becoming a completely accurate and reliable model for election prediction. In [10], the authors argue that the power of citizens through the use of social media is still unknown. They analysed the case of the presidential candidate Enrique Peña Nieto PRI in Mexico who won the presidency with a large participation but without the support of Twitter users.

Methodology - This part of the paper focuses to explain the related mining of tweets and performing sentiment analysis on them.

III. DATA COLLECTION

There are basically two ways of collecting data from twitter. (1) By searching tweets matching to keywords and (2) Collecting all the tweets provided by twitter through streaming APIś. In the first method, the data is relatively small, while in the second method we can get many sets of keywords for an effective outcome [11]. APIś or Application programme interface is a way to acquire data from other services for our own functions. Likewise, Twitter has provided us with many of its own APIś. The first step would be to extract tweets in various languages which will at the end comprise of our primary dataset from where we will perform our sentiment analysis. For that, we need to install ¨Tweepy¨. Then we will need to authenticate our programme, for that twitter provides us with ´token keys' and ´consumer keys´. Twitter also has a predefined list of language codes that we can use to extract tweets from different languages. For querying tweets, we need an exhaustive list of keywords and catchphrases that are relevant to our project. Some examples of hashtags we can use are #IndianElections, #IndiaVotes, #BJP, #Congress, #Reelection etc. Apart from that, we can identify local election issues and also use them in our search queries as well, such as Economy, Healthcare, Jobs and Sanitation. This will help us in sentiment analysis later and tell us whether people are satisfied with the performance of the incumbent government on these issues or not. We will store our tweets obtained after querying in a JSON format which basically stores as a key-value pair. As twitter offers a lot of ways to provide information from their servers, we will store each and every attribute of the user in a JSON format. Using these key-value pairs we will be able to extract and filter relevant tweets later from our database. For that, we need to import jsonpickle. It is just another library that handles files in JSON format. The pseudocode for storing tweets in JSON format is given below-

PSEUDOCODE:

1. Import Json pickle
2. Create a JSON file that will hold all our tweets.
3. After using the twitter APIś and extracting the relevant tweets, store them by mentioning the file path.
4. Convert the tweets into JSON format by using JSON.encode function.

But before we convert our tweets into JSON format we need to extract the relevant tweets which will comprise our corpus. For that reason, the following pseudocode is given below-

PSEUDOCODE :

1. Import tweepy, jsonpickle
2. Create four variable -

   I. consumer_key

   II.consumer_secret

   III.access_token

   IV.access_token_secret
3. Using OAuthHandler, we pass these 4 variables as parameters.
4. Create a variable that allows us to use the API services for our project.
5. Create a query and set the upper limit of the number of tweets we need. (The query will be composed of the keywords mentioned above.)
6. Using the tweet.lang method we can use the language code that is provided by twitter and query the tweets we require in those languages.

A. Local Language - to - English Translation -

As we have tweets in various languages in our database we need to translate them into a specified format so that they can be easily classified as positive, negative or neutral. For that, we will use a Translator API. For example, if we have a tweet in Hindi

"जो सही है उसे ट्वीट किया जाएगा, जो गलत है उसका विरोध किया जाएगा"

The translation of this will be as follows:

¨What is right will be tweeted, what is wrong will be opposed¨

B. Filtering our Dataset:

One of the main challenges that come with data collection through streaming APIś is to identify the specific attributes that are required for our given problem. For eg. In [12], the authors predicted that Hilary Clinton would win the 2016 elections. Based on the data collected, they observed that positive tweets were more for Clinton than Trump and vice versa. Neutral tweets were also more in favour of Clinton, but as we know the final results of the elections were vastly different. One of the inferences that can be drawn from this wrong prediction is that the quality of the corpus that is going to be analysed needs to be properly refined so that it suits our model. In this paper, we have tried to filter our dataset as much as possible so that our targeted tweets satisfy our end results. For this, there are a couple of conditions that we have made for a tweet to enter our final dataset -

(i) Identifying users that will not participate in the voting process: In [13], the authors have used multiple search queries such as, "I am not voting;" "I'm not voting;" "I am not going to vote;" and "I will not vote". But we need to be careful not to include tweets that say, ¨I am not going to vote¨ - for a particular candidate. In our JSON file, we will remove the dictionaries of users that will not be part of the voting process.

(ii) Identifying the Location of the user: Another factor to consider is the location of tweet: Tweets coming for Indians living abroad should not be taken into account as they will not be eligible to vote, hence we have to make sure that the tweets that will be analysed are from a specified location which we will decide based on our requirements. Tweet data consists of two classes of geographical metadata (1) Tweet location: Provided when the user shares his location at the time of the tweet (2) Account Location: Location of the user on his/her profile. Both these parameters will not help us correctly identify the location of the user every time. Many times the user will not provide his location at the time of the tweet, which will limit our user coverage substantially. Also, many times multiple instances of the same toponym (places name) in different geographic regions can cause a disambiguation problem [14]. One way to solve these problems to use twitter geo search API, to get the place id and then perform a regular search using, place:place_id.

```
places = api.geo_search(query="INDIA", granularity="country")
```

(iii) Removing bots: Bots are automated twitter accounts that are used to generate interest in a specific topic. They can perform human-like interactions on twitter, such as retweets and DMś and spreads messages at a rate faster than humans. It was identified than around 75% of all of the tweeted links to popular websites are from automated accounts. These are widely used by political parties in India, especially during elections as a campaigning tool. Hence it is essential to identify these automated accounts and discard them from our dataset for better sentiment analysis. In [15], it was argued that by calculation ¨account reputation¨ score, twitter accounts can be classified as being either auto-generated or real. The account reputation score (AR) can be given by

$$Account\ Reputation = \frac{Follower_{number}}{Follower_{number} + Friends_{number}}$$

Bots usually have a low AR score, as they have fewer followers than friends. In [15], it was identified that bots have an AR score of less than 0.5 compared to actual human accounts that have a score close to 1. Using our JSON dataset, we can calculate the AR score of each account that we have extracted and classify them accordingly as being either a bot account or a human account.

C. Tweets Pre-processing

After we have filtered our dataset, we can perform sentiment analysis, but before that, we need to pre-process our tweets. Tweets generally contain a lot of noise, because of excessive usage of the acronym, irregular grammar, ill-formed words and non-dictionary terms. Unstructured twitter data will greatly affect the performance of sentiment classification [16]. Thus, it is essential to remove this noise from our data before we can use our algorithms on them:

1. Removal of all non-ASCII characters and non-English characters.
2. Removal of URL Links. AS URL links do not help in sentiment analysis, they will be removed [17].
3. Remove numbers. To refine the tweet, they will be deleted.
4. Transforming words like "won't" and "can't" into "will not", "cannot", and "not" separately [17]. Negation words play an important role in polarity calculation, tweets containing these words should be transformed.
5. Slang translation and expansion. As many tweets contain various slang, we will use the Internet and text slang dictionary to find the meanings of different slang. For example, afaicr4:¨as far as I can remember for¨ and cayc:¨call at your convenience¨
6. Removing Stopwords: A stop word is a commonly used word (such as “the”, “a”, “an”, “). They can be removed by obtaining stopwords from precompiled lists.
7. Elongated words containing character repetition and shorthand notations like¨sorry¨ v/s¨sorryyyy¨ and¨seriously¨v/s ¨srslỳ can be removed using Lexical normalization [18].
8. To handle twitter content along with hashtags, emoticons etc. we can use Standard Tweet Normalization tools: - We can use tools like GATE Twitter NLP, which can normalize tweets.
9. P-O-S tagging: It is known as Parts of Speech tagging and is used in sarcasm detection. It basically identifies a word as a noun, verb, adjective, etc, based on its definition and relation with other words [19].

D. Sentiment Analysis for Twitter

Sentiment Analysis is the process through which we can make identifying the emotion in a text-based on many semantic clues. It refers to the use of Natural Language Processing to determine the attitudes and opinions of a speaker, writer, or other subjects within an online mention. We can classify the text as positive or negative using itś polarity score. For the classification of tweets, we use the Naive Bayes algorithm, due to its high precision, recall and F-score values. In [20], they compared the Naive Bayes with Maximum Entropy and Support Vector Machine. The following results from that experiment were found, which clearly showed that the accuracy for the Naive Bayes algorithm is much better than the other two.

Table 1. Accuracy comparison between different text classification methods

Method	Accuracy
Naive Bayes	88.2
Maximum Entropy	83.8
SVM	85.5

Naive Bayes algorithm is based on Bayes’ probability theorem. Itś primary use is for text classification that involves high-dimensional knowledge sets [3].

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

Where,

I. P(A|B): Probability (conditional probability) of occurrence of the event given the event B is true.
II. P(A) and P(B): Probabilities of the occurrence of the event A and B respectively.
III. P(B|A): Probability of the occurrence of event B given the event A is true.

To build the classifier we use the following steps:

1. Import nltk
2. Build the vocabulary: It will basically be a list of all the words in our pre-processed training data.
3. Create a world_list which consists of the frequency of each word in the list.
4. Compare each word in the word_list with the tweets at hand and associate a label against the tweet if those words are in the vocabulary are resident in that tweet or not.
5. After that, we train our classifier using the built-in library function nltk.NaiveBayesClassifier.train()
6. Testing the model to get the majority vote of the labels given by the classifier.

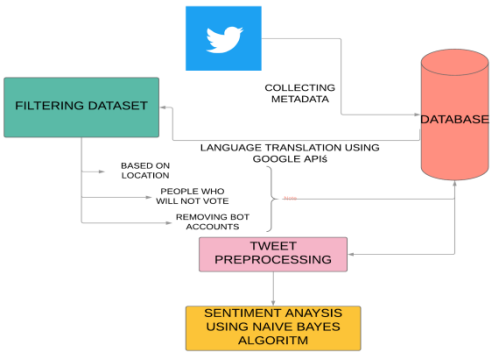

Fig 2. System Architecture

IV. RESULT AND DISCUSSION

After using tweets collected from twitter APIś we got the sentiment score for the two main political parties of India, i.e. BJP and the Congress. We refined our tweets by measuring both the parties on similar issues such as jobs, infrastructure, economy and healthcare. The corpus of our tweets was taken in context with the Maharashtra and Haryana Election. Hence, we also included tweets written in Hindi and Marathi languages. Special emphasis was given to tweets belonging to these two states using our api.geo_search function as discussed above. In fig (3), we visualised the sentiment score for the BJP into 3 parts - positive, negative and neutral. We analysed tweets from the last 30 days with a total count of 1032 tweets.

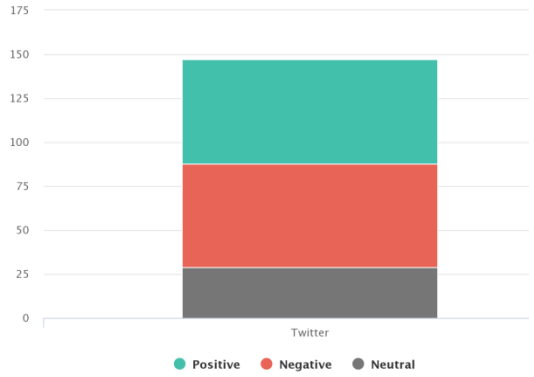

Fig 3. Sentiment analysis chart for BJP

Similarly, we got a sentiment score pie-chart for the congress over the last 30 days with 1018 tweets. Fig (2) represents the sentiment score for the Congress party on twitter.

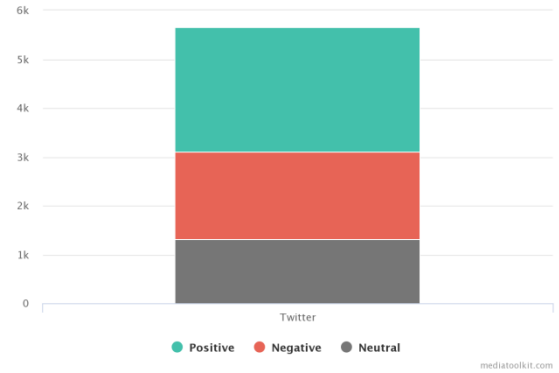

Fig 4. Sentiment analysis chart for Congress

V. CONCLUSION

A large amount of information contained in microblogging websites makes them an attractive source of data for opinion mining and sentiment analysis. In this paper, we prepared a model to analyse and predict elections in India. Our main goal was to accurately predict the political choice of twitter users correctly and based on that predict the election outcome. For that, we used streaming APIś from twitter to gather data. We created a multilingual database with tweets from different languages for better local results. From that, we refined our corpus by removing the unwanted tweets that did not suit our purpose. Using Naive Bayes classifier, we were able to analyse the tweets and classify them as being positive or negative. From the results, we can clearly see that the percentage of positive tweets for Congress are slightly higher than the BJP. This is a big deviation from the exit polls that were conducted before the election outcome that clearly showed BJP increasing its seat tally in Haryana [22], as well as in Maharashtra [23]. The final election outcome was vastly different as in both states the BJP lost seats and the Congress increased its seat share compared to the previous elections. These results reflect in our final sentiment score for both the parties.

FUTURE WORK

- In the future, we would like to identify the age of the user, so that during data filtering process we can eliminate the tweets that come from twitter handles that have an age of less than 18. Dividing our corpus-based on gender, caste and community based on user identification algorithms can help us to analyse voting patterns in India better.

REFRENCES

1. Studying Elections in India: Scientific and Political Debates by Stéphanie Tawa Lama-Rewal.
2. Political Tweets and Mainstream News Impact in India: A Mixed Methods Investigation into Political Outreach Sunandan Chakraborty Indiana University Indianapolis, Indiana sunchak@iu.edu Joyojeet Pal University of Michigan Ann Arbor Ann Arbor, Michigan joyojeet@umich.edu Priyank Chandra University of Michigan Ann Arbor Ann Arbor, Michigan prc@umich.edu Daniel M. Romero University of Michigan Ann Arbor Ann Arbor, Michigan drom@umich.edu.
3. Twitter Based Election Prediction and Analysis Pritee Salunkhe, Sachin Deshmukh 1,2 Department of Computer Science and Information Technology, Dr Babasaheb Ambedkar Marathwada University, Maharashtra, India.
4. U. R. Hodeghatta, ¨Sentiment analysis of Hollywood movies on Twitter," in Proc. IEEE/ACM ASONAM.
5. Prediction of Indian Election Using Sentiment Analysis on Hindi Twitter Parul Sharma Department of Computer Science San Jose State University San Jose, CA, USA parul.sharma@sjsu.edu Teng-Sheng Moh Department of Computer Science San Jose State University San Jose, CA, USA teng.moh@sjsu.edu.
6. Quantifying Political Leaning from Tweets and Retweets Felix Ming Fai Wong, Chee Wei Tan†, Soumya Sen, and Mung Chiang Princeton University, †City University of Hong Kong {mwthree, Soumya,chiangm}@princeton.edu,cheewtan@cityu.edu.hk.
7. Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis Widodo Budiharto and Meiliana Meiliana.
8. POLL PREDICTION BASING ON SENTIMENT USING NAÏVE BAYES AND DICTIONARY BASED CLASSIFIERS Tarun Matta1, Kishore Sabbavarapu, Ch. Swapna Priya and Dr Narasimham Challa Student, Assistant Professor and Professor, Department of CSE, Vignan's Institute of Information Technology, Visakhapatnam.
9. Using SentiWordNet for Multilingual Sentiment Analysis Kerstin Denecke Research CenterL3S Appelstrasse 9a, D-30167 Hannover, Germany denecke@l3s.de.
10. Using Twitter in Political Campaigns: The Case of the PRI Candidate in Mexico Rodrigo Sandoval-Almazan, the Autonomous University of the State of Mexico (UAEM), Toluca, México.
11. Boutet, A. K. (2012). What's in your Tweets? I know who you supported in the UK 2010 general election. Proceedings of the International AAAI Conference on Weblogs and Social Media.
12. Quantifying Political Leaning from Tweets and Retweets Felix Ming Fai Wong, Chee Wei Tan†, Soumya Sen, and Mung Chiang Princeton University, †City University of Hong Kong {mwthree, Soumya,chiangm}@princeton.edu,cheewtan@cityu.edu.hk
13. Using Twitter for Demographic and Social Science Research: Tools for Data Collection Authors: Tyler McCormick1,2,4,5, Hedwig Lee1,5, Nina Cesare1 Ali Shojaie3,2,4 1 Department of Sociology, University of Washington 2 Department of Statistics, University of Washington 3 Department of Biostatistics, School of Public Health, University of Washington.

14. Geocoding location expressions in Twitter messages: A preference learning method Wei Zhang and Judith Gelernter School of Computer Science, Carnegie Mellon University, Pittsburgh, USA.
15. Detecting Automation of Twitter Accounts: Are You a Human, Bot, or Cyborg? Zi Chu, Steven Gianvecchio, Haining Wang, Senior Member, IEEE, and Sushil Jajodia, Senior Member, IEEE.
16. Jianqiang Z, Xiaolin G. Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis. IEEE Access, 5,pp.2870-2879,2017.
17. China Deep Convolution Neural Networks for Twitter Sentiment Analysis Zhao Jianqiang1) 2), Gui Xiaolin1),2)* 1. School of Electronics and Information Engineering, Xi'an Jiaotong University 2. Key lab of Computer Network of Shaanxi Province, P. R. China.
18. Lexical Normalisation of Twitter Data Bilal Ahmed Department of Computing and Information Systems The University of Melbourne Victoria, Australian bahmad@student.unimelb.edu.au.
19. An Efficient Approach for Sarcasm Recognition on Twitter Pattern-Based Method Khan ShehlaKulsum, Prof.S.G. Vaidya.
20. Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis Geetika Gautam Department of Computer Science & Engg. Jaypee Institute of Information technology Noida, Indigeetikagautam16@gmail.com Divakar Yadav Department of Computer Science & Engg. Jaypee Institute of Information technology Noida, India divakar.yadav@jiit.ac.in.
21. https://www.livemint.com/news/india/in-india-who-speaks-in-english-and-where-15578141 01428.html
22. https://www.ndtv.com/india-news/haryana-poll-of-polls-election-in-haryana-2019-sweep-for-manohar-lal-khattar-ml-khattar-exit-poll-2120433
23. https://www.indiatoday.in/elections/maharashtra-assembly-election/story/maharashtra-exit-poll-results-2019-poll-of-polls-predicts-bjp-shiv-sena-victory-1611626-2019-10-21

AUTHORS PROFILE

Balika J Chelliah, is an Associate Professor in Department of Computer Science & Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai. He received his Master and Ph.D degrees in Computer Science & Engineering from SRM Institute of Science and Technology, Chennai. His areas of interest are Service oriented Architecture, Web services, Cloud services and Software Engineering..

Kartikaya Srivastava is a third year B.Tech student. He is pursuing Bachelor degree in Computer Science & Engineering from SRM Institute of Science and Technology, Ramapuram, Chennai. He has a keen interest in the domain of Data Analytics.

Jishnu Panja, is a third year B.Tech student. He is pursuing Bachelor degree in Computer Science & Engineering from SRM Institute of Science and Technology, Ramapuram, Chennai. He has keen interest in machine learning, database management.

Ritwik Paul, is a third year B.Tech student. He is pursuing Bachelor degree in Computer Science & Engineering from SRM Institute of Science and Technology, Ramapuram, Chennai. He has keen interest in web development and front end.