Election Tweets Prediction Using Enhanced Cart and Random Forest

Ambati Jahnavi, B. Dushyanth Reddy, Madhuri Kommineni, Anandakumar Haldorai, and Bhavani Vasantha

Abstract In this digital era, the framework and working process of election and other such political works are becoming increasingly complex due to various factors such as number of parties, policies, and most notably the mixed public opinion. The advent of social media has deployed the ability to converse and discuss with a wide range of audience across the globe, whereas gaining a sheer amount of attention from a tweet or post is unimaginable. Recent advances in the area of profound learning have contributed to the use of many different verticals. Techniques such as long-term memory (LSTM) perform a sentiment analysis of the posts. This can be used to determine the overall mixed reviews of the population towards a political party or person. Several experiments have shown how to forecast public sentiment loosely by examining consumer behaviour in blogging sites and online social networks in national elections. This paper has proposed a model of machine learning to predict the chances of winning the upcoming election based on the common people or supporter views on the web of social media. The supporter or user shares their opinion or suggestions about the group or opposite group of their choice in social media. It has been required to collect the text posts about election and political campaigns, and then the machine learning models are developed to predict the outcome.

Keywords Sentiment analysis · Decision tree · Random forest and logistic regression

A. Jahnavi · B. Dushyanth Reddy · M. Kommineni · B. Vasantha
Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India
e-mail: madhuri.cbit@gmail.com

A. Haldorai (✉)
Department of Computer Science and Engineering, Sri Eshwar College of Engineering, Coimbatore, Tamil Nadu, India
e-mail: anandakumar.psgtech@gmail.com

S. Smys et al. (eds.), *Inventive Computation and Information Technologies*, Lecture Notes in Networks and Systems 173,
https://doi.org/10.1007/978-981-33-4305-4_61

1 Introduction

The online platform has become an enormous course for individuals to communicate their preferences. Using various assessment techniques, the ultimate intent of people can be found, for example, by eviscerating the content of the tendency, positive, negative, or truthful. For instance, assessment appraisal is always noteworthy in a relationship to hear their client's insights on their things by imagining eventual outcomes of races and getting ends from film ponders. The data snatched from the opinion evaluation is helpful for predicting the future choices. Rather than associating individual terms, the relation between the set of words is also considered. While selecting the general assumption, each word's ending is settled and united using a cap. Pack of words will also ignore word demands, which prompt phrases with invalidation should be erroneously described. In the past decades, there has been a massive improvement in the use of small-scale blogging stages, for instance, Twitter. Nudged by that advancement, associations and media affiliations are continuously searching for ways to analyze the information about what people ponder about their things and organizations in the social platforms like Twitter [1]. Associations, for instance, Twitratr, tweetfeel, and social mention are just an uncommon sorts of individuals who advance tweet presumption examination as one of their organizations [2].

Although a significant proportion of work has been performed on how emotions are expressed in different forms such as academic studies and news reports, where significantly less study has been done [3]. Features, for instance, customized linguistic component marks and resources, for instance, idea vocabularies have exhibited the accommodation for supposition examination in various spaces, and anyway will they also show significance for evaluation assessment in Twitter? This paper begins to analyse this request [4].

2 Literature Survey

Notwithstanding the character goals on tweets, working out the concept of Twitter messages is basically close to the sentence-level assumption evaluation, the welcoming and express language used in tweets, as well as the general idea of the local micro-blogging allows Twitter's thinking evaluation to extend beyond the expectation [5]. It is an open solicitation on how well the highlights and procedures are utilized on continuously well-shaped information that will move to the micro-blogging space [6].

Ref. [7] It involves measures such as data collection, pre-processing of documents, sensitivity identification, and classification of emotions, training and model testing. This research subject has grown over the last decade with the output of models hitting approximately 85–90% [8].

Ref. [9] Firstly, in this paper, they have presented the method of sentiment analysis to identify the highly unstructured data on Twitter. Second, they discussed various

techniques in detail for carrying out an examination of the sentiments on Twitter information [10].

Ref. [11] They suggested a novel approach in this paper: hybrid topic-based sentiment analysis (HTBSA) for the task of predicting election by using tweets.

Ref. [12] Using two separate versions of SentiWordNet and evaluating regression and classification models across tasks and datasets, it offers a new state-of-the-art method for sentiment analysis while computing the prior polarity of terms. The research investigation is concluded by finding the interesting differences in the measured prior polarity scores when considering the word part of speech and annotator gender [13].

Ref. [14] This paper proposed a novel hybrid classification algorithm in this paper that explains the conventional method of predictive sentiment analysis. They also integrated the qualitative analysis along with data mining techniques to make sentiment analysis method more descriptive [15].

Ref. [16] This research work chose to use two automated classification learning methods in this paper: support vector machines (SVM) and random forest for incorporating a novel hybrid approach to classify the Amazon's product reviews.

Ref. [17] Here, the proposed research work aims to build a hybrid sentiment classification model that explores the basic features of the tweet and uses the domain-independent and domain-related lexicons to provide a more domain-oriented approach for analysing and extracting consumer sentiment towards popular smartphone brands in recent years.

3 Methodology

The following figure shows the steps followed in the proposed model (Fig. 1).

Decision Tree

As the implementation of machine learning algorithms is mainly intended to solve problems at the industry level, the need for more complex and iterative algorithms is becoming as an increasing requirement. The decision tree algorithm is one such algorithm used to solve problems in both regression and classification.

Decision tree is considered as one of the most useful algorithms in machine learning because it can be used to solve many challenges. Here are a few reasons why decision tree should be used:

(1) It is considered the most comprehensible machine learning algorithm and can easily be interpreted.
(2) This can be used for problems with classification and regression.
(3) It deals better with nonlinear data as opposed to most machine learning algorithms.
(4) Building a decision tree is a very quick process since it uses only one function per node to divide the data.

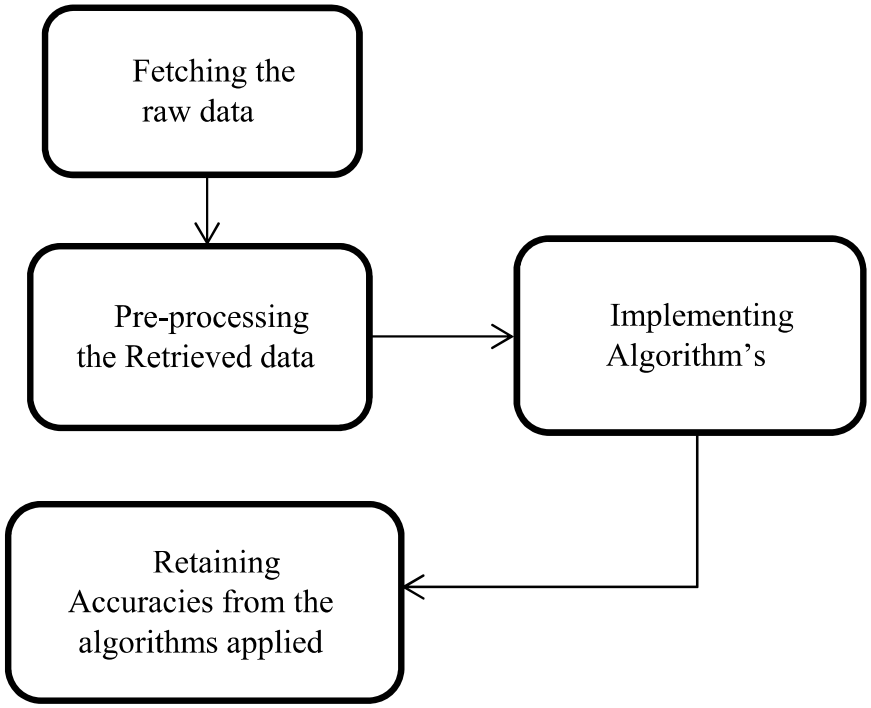

Fig. 1 Flow chart of the proposed work

Recursive partitioning is an important instrument in data mining. It lets us explore the structure of a collection of data while making decision rules simple to imagine for predicting a categorical (classification tree) or continuous (regression tree) outcome. This section explains the modelling of the CART, conditional inference trees (Fig. 2).

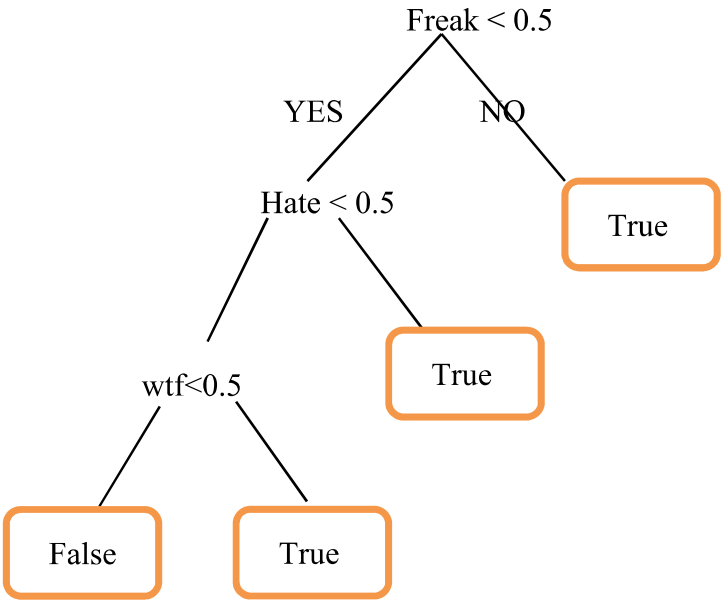

Fig. 2 Sample tree that appears after the implementation of cart algorithm

Random Forest

The random forest algorithm works by aggregating the predictions from different depths of multiple decision trees. Decision tree in the forest will be trained on a dataset subset called the bootstrapped dataset.

The portion of samples left out when constructing each decision tree in the forest is referred to as the Out-Of-Bag (OOB) dataset. As observed later, the model can automatically determine its own output by running each of the samples through the forest in the OOB dataset.

Remember how the impurity measurement is generated with each feature by using the Gini index or entropy when deciding on the criteria with which to split a decision tree. Nonetheless, a predefined number of features are randomly chosen as candidates in random forest. The above would result in a greater difference between the trees which would otherwise have the same characteristics.

If the random forest is used for classification, and a new sample is provided, the final prediction is made by taking most of the predictions produced in the forest by each individual decision tree. In the event, it is used for regression and a new sample is provided; the final prediction is made by taking the average of the predictions produced in the forest by each individual decision tree.

Logistic Regression

Key backslide is a genuine model which uses a determined ability to display a parallel subordinate variable in its main structure, but there are many dynamically complex developments. Determined backslide is surveying the parameters of a critical model in backslide analysis.

Numerically, a double-determined model has a dependent variable with two possible characteristics, for example, pass / bomb, which is represented by a marker variable called "0" and "1″. In the vital model, the log-risks (the odds logarithm) for the value "1″ tested is an immediate mix of one independent variable ("markers") in any event; the free factors can be either a double factor (two classes, coded by a pointer variable) or a constant variable (any authentic value). The relative probability of the value called "1″ will shift from 0 (irrefutably the value "0″) to 1; from now on the limit that changes to probability over log opportunities is the defined limit, hence the name.

The unit of estimation for the scale of the log-chances is known as a logit, from a given unit, hence the elective names. Essentially proportionate to models with a different sigmoid limit as opposed to the defined limit, the probit model, for example, can use it similarly; the usual explanation for the main model is that it increases one of the self-sufficient factors that multiply the odds of the result at a predictable rate, with each free factor having its own parameter; this summarizes the odds magnitude for a double bad variable.

The twofold determined backslide model has extensions to different degrees of the dependent variable: straight out yields with various characteristics are shown by multinomial vital backslide, and if the various classes are mentioned, by ordinal vital backslide, for example, the comparing chances of ordinal key model.

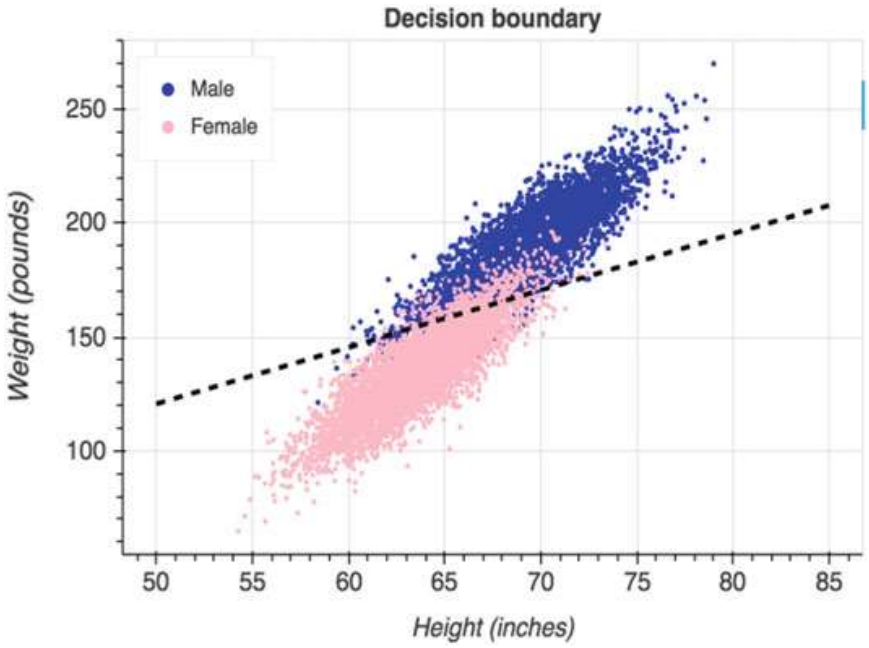

Fig. 3 Classification of the retrieved data using logistic regression, represented in a form of plot

The model itself fundamentally models probability of yield similar to information and does not perform verifiable gathering (it is definitely not a classifier), anyway it might be used to make a classifier, for instance by picking a cutoff regard and orchestrating commitments with probability more noticeable than the cutoff as one class, underneath the cutoff as the other; this is a commonplace technique to make a twofold classifier. The coefficients are overall not handled by a shut structure verbalization, not in the least like straight least squares (Fig. 3).

4 Performance and Result Analysis

Considering that election outcomes are very difficult to predict using other methods, including public opinion polls, and with social media such as Facebook and Twitter increasingly prevalent, the authors chose to use Twitter's sentiment analysis to forecast Indian general election results (Table 1).

Table 1 Accuracies of algorithm's

Algorithm	Accuracy of tweets	
	Neg	Pos
Cart	0.8789	0.9437
Random forest	0.9155	0.9493
Logistic regression	0.8845	0.9268

From the above table, the overall accuracies of the tweets obtained using cart algorithm are 91.13%, random forest algorithm is 93.24%, and logistic regression is 90.56%. So, from the results, it is observed that the random forest algorithm works better on election tweets data.

5 Conclusion

Hence, in this research work, in order to expand the data set size, there may be several other prospective fields to perform this analysis, where it includes the data from other major social networking sites, such as Twitter. It is also found that there is a dedicated research space to work with the training dataset by considering the model dataset that already specifies a certain number of algorithmic features. The major downside of this research work is that it fails to recognize the significant parameter called emotion, when defining the polarity of a tweet. Since the data was labelled manually, the volume was not high enough to have more precise information, so more tweets can be obtained and marked. As a continuation of this research work, the network size will be increased.

Reference

1. Malika M, Habiba S, Agarwal P (2018) A novel approach to web-based review analysis using opinion mining. In: International Conference on Computational Intelligence and Data Science (ICCIDS 2018) , Department of Computer Science and Engineering, Jamia Hamdard, New Delhi-110062, India
2. Agarwal A, Xie B, Vovsha I, Rambow O, Passonneau RJ (2011) Sentiment analysis of Twitter data
3. Popularity analysis for Saudi telecom companies based on Twitter Data. Res J Appl Sci Eng Technol (2013)
4. Liu B (2012) Sentiment analysis and opinion mining, Morgan & Claypool Publishers
5. Joshi S, Deshpande D (2018) Twitter sentiment analysis system. Department of Information Technology, Vishwakarma Institute of Technology Pune, Maharashtra, India
6. Agarwal A, Xie B, Vovsha I, Rambow O, Passonneau RJ (2011) Sentiment analysis of Twitter data. Department of Computer Science Columbia University New York, NY 10027 USA
7. Gupta B, Negi M, Vishwakarma K, Rawat G, Badhani P (2017) Study of Twitter sentiment analysis using machine learning algorithms
8. Umadevi V (2014) Sentiment analysis using weka. IJETT Int J Eng Trends Technol 8(4):181–183
9. Techniques for sentiment analysis of Twitter data: a comprehensive survey. In: 2016 International Conference on Computing, Communication and Automation (ICCCA)
10. Caetano JA, Lima HS, Santos MF, Marques-Neto HT (2018) Using sentiment analysis to define twitter political users' classes and their homophily during the 2016 American Presidential election
11. Bansala B, Srivastavaa S (2019) On predicting elections with hybrid topic based sentiment analysis of tweets. Department of Applied Sciences, The NorthCap University, Gurugram, India

12. Guerini M, Gatti L, Turchi M (2013) Sentiment analysis: how to derive prior polarities from SentiWordNet
13. Barahate SR, Shelake VM (2012) A aurvey and future vision of data mining in educational field. In: Proceedings 2nd International Conference on Advanced Computing and Communication Technology, pp 96–100
14. Chen X, Vorvoreanu M, Madhavan K (2014) Mining social media data to understand students' learning experiences. IEEE Trans 7(3):246–259
15. Twitter data sentiment analysis and visualization.Int J Comput Appl 180(20)
16. Al Amrani Y, Lazaar M, Kadiri KE (2018) Random forest and support vector machine based hybrid approach to sentiment analysis. Author links open overlay panel
17. Venugopalan M, Gupta D (2016) Exploring sentiment analysis on Twitter data, IEEE 2015