Vote Recommendation System using Aspect based Machine Learning Approach

Swati Sharma, Mamta Bansal

Abstract: Over time, the information on WWW has escalated exponentially, paramounting to embryonic research in the field of Data Analysis using Natural Language Processing (NLP) and Machine Learning (ML). As data is increasing day by day there is huge demand for data analysis to get subjective information and analyzing government data is very useful and demanding task. So, in this paper, an application is being developed which will recommend the user to which party to vote will be benignant for themselves and for country, depending on the area of interest of different users. The data is collected from various governmental websites of multiple areas like women empowerment, education, employment, child labor etc. which will enhance the authenticity of the output. The main ground of this research is to lubricate common people and politicians as well. For common people; is for deciding their precious vote, to which party to give will be good for themselves and nation too. For politicians; they will have an idea about themselves and other politicians that which party is preferable and which is not preferable in respective areas, so that the politicians can work accordingly.

Keywords: Artificial Intelligence, Machine Learning, Naïve Bayes, Natural Language Processing, N Gram, Support Vector Machine.

I. INTRODUCTION

IN this era, Data Analysis is very demanding and useful topic where World Wide Web (WWW) is exceedingly freighted with cosmic amount of data [1]. This huge amount of data possesses abundant precious information when analyzed according to once requirement. In this work, a medium has been submitted for endorsing the exquisite votes of user to which party to vote will be good by probing the tweets publish by various users using ML-Machine Learning technique. Data is accumulated of numerous of areas like death rate, education, literacy rate, women empowerment, divorce rate, military force, health department, mortality rate, birth rate, employment, vulnerable employment, marriage rate, population growth, refugee population, forced labor, sex ratio, science and technology, research and development, trademark applications, financial and Non financial assets, taxes, revenues, grant, insurance, statistical coverage, warehouse, subsidies, transportation, logistics, travel and tourism, water and irrigation, sanitation, agriculture, merchandise import or export, ICT goods, smoking, alcohol consumption, family planning, catastrophic diseases, usage of condoms, immunization, medical treatment, health workers, health expenditure and so on. Data is accumulated from numerous government web sites of past years.

Data collected is since year 1964 till date. A huge size of warehouse is managed on various government web sites. It is advantageous for practitioners to use the data for analyzing in numerous areas [2].

A model is being deployed asking for users details like the name, the age and category in which user has keen interest. The area plays a marvelous role here, as everyone has their own respective interest with which they prefer to go for party to whom to vote. Once filling details, the model will perform analysis on data by accounting the work done by numerous parties using ML-Machine Learning, as it validates the model to get train by self learning with not any external support. When it is exposed to more data, the system applications will sanction to acknowledge, deploy, change and expertise by their own. Thus, the deployed model will suggest the user to which party to vote will be good [3].

II. PROPOSED SYSTEM

The system architecture presented in Fig. 1 outlines the ML perspective for the Sentiment Analysis (SA), in this part system is designed asking for users details like the name, the age and area where user has interest. The area plays a supreme role here, as everyone has their own interests and in their engrossment they prefer party for their vote as all parties had worked in numerous field. Once filled the asked details, the model will perform analysis on the data about the work done by them using Machine Learning approach, it entitles the model to train by them with no external support. When it is exposed to more data, the model programs will sanction to acknowledge develop, moderate and expertise by own; when the sample data is trained by own and expertise by using ML-Machine Learning technique. Thus, the generated model will suggest the user to which party to vote will be good. Machine Learning-ML technique is used for executing the idea of suggesting vote to user to who to vote will result as good. This model is enveloping highest number of categories on which government had paid off. However, it is wrapping many election parties [4].
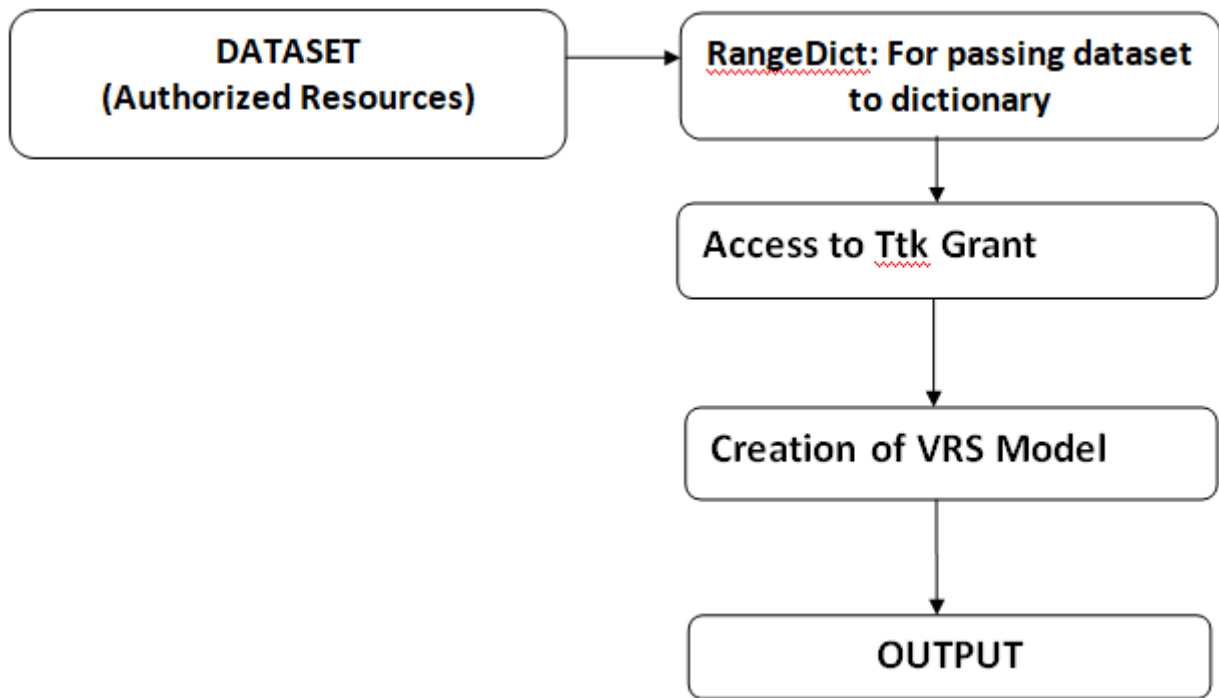
Fig. 1.Architecture of Proposed Methodology

III.   METHODOLOGY

A.   Data Collection

The data is accumulated from assorted government websites for forming the authencity of the designed application. The data accumulated is of numerous categories where the government has worked and is keep on working like employment, women empowerment, death rate, child labor etc. The data can differ based on requirement of designed application. Fig. 2 shows the steps to be followed in aspect based machine learning approach.

B.   Feature Collection

Bag-of-words technique is used for illustration of text. The thrust of this plea is the ignorance of absolute location; it possesses unordered set of words without referring to absolute location. This plea is very much effectual in grabbing the theme of discussion; where the earmark work has forecasted. The stop words necessitates to be refined or separated but no doubt stop words play active part in partitioning the sentences i.e. they helps in figuring out the sentence [5]. Below mentioned bag-of-words are taken care here:

N-gram

An outlook of assembling features. It plays on the postulation of n-word expressions. For example, 'I Love Python', 'Python is best programming language'. The prior mentioned sentences have different lengths of the expressions. It bestow more components to document and so appropriate for the word sense uncertainty. The sentences with greater than three words are not as important as the dataset gets enormous [6].

Lemmatization

It is tactic of merging numerous forms of words comprising of identical meaning and can be scanned as a individual item [7]. This tactic is selected over the stemming as the lemmatization pact with morphological perusal of words and also with part-of-speech tags.

C.   Feature Subset Selection

After accumulating required features, forward step is feature subset selection. It assists in identifying subset of required features from the uninterrupted dataset. It disregards the features which has no prime importance. Relevant selection of attributes can notably heighten the intellect of classifier models [8]. Following stairs to focus while choosing features: First are the wrappers; in accordance to predictive ability, they prefer learning machines for marking subset of features. If a algorithm has number of features to choose and examine, it will intricate the job but the output will maintain more accuracy and also by heighten size of data, it will obtain best results. Forward Selection elimination and Backward Selection elimination are two direction of performing greedy search [9]. In Forward Selection successive accumulation of attributes one-by-one and become bigger subsets. In Backward Elimination, the outset step is set incorporating of entire features and those with minimum importance are discarded from the dataset [10].

Second are the filters; they pick feature sub sets as supplant step independently of chosen predictors. Analogously, filters are contemplated speedier than wrappers; oblige with very common selection of features which are neither trained for respective learning machine. It is powerful for reducing the space dimensions.

Third are embedded techniques, they pick features while training the process and are very inclusive for particular learning machine [11].

136

D. Training

Hand tagged sub set of dataset is being assembled, by incorporating crowd sourcing technique [12]. It is plan of expanding the data set with the assistance of large count of People. By assembling data using crowd sourcing technique, Specialist can calculate disperse, valuable and plenty quantity of data in very low cost [14].

It allows gathering real time data and can even acquire lot more countess and substantial observations. It allows practitioners to simply approach people and numerous places and making cognizance to researchers for numerous tasks. It gives assurance to large count of contributors; for perfect crowd sourcing.
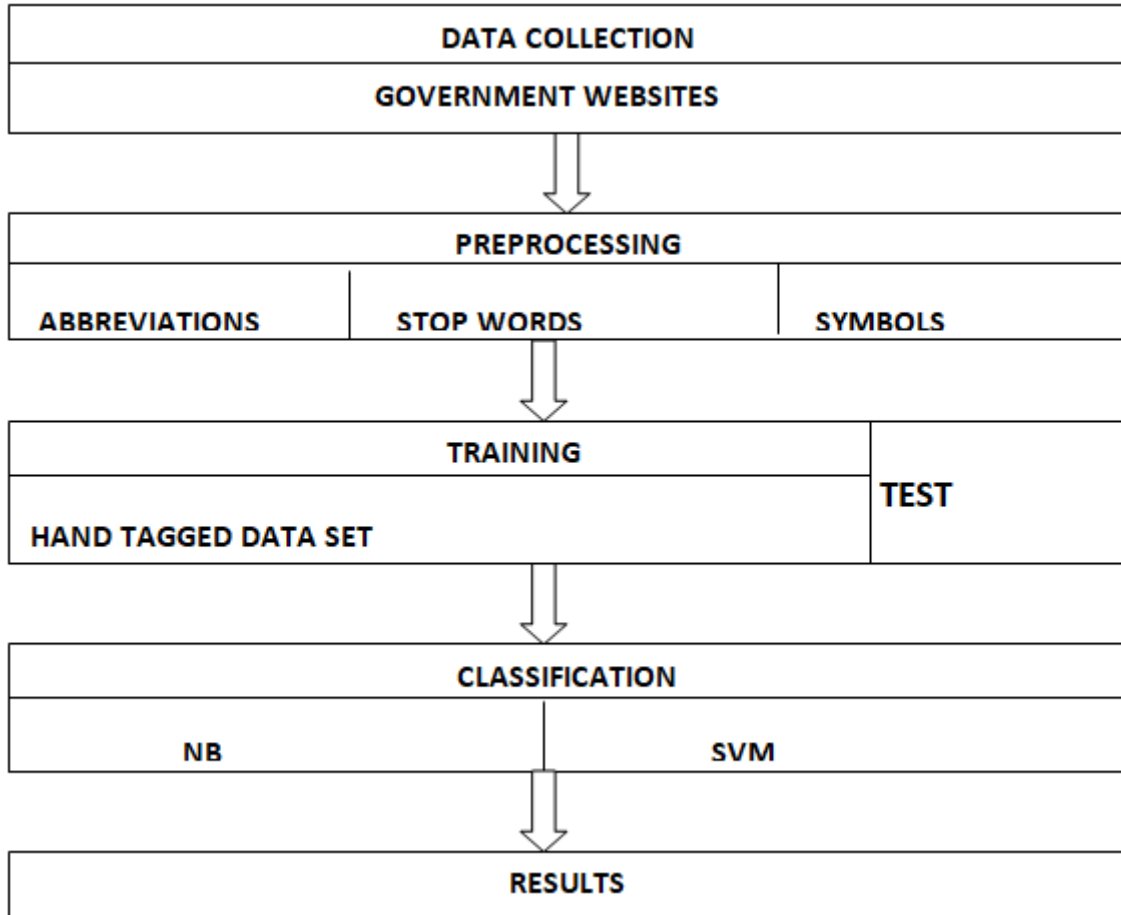

Fig. 2.Steps in Aspect-based Machine Learning Approach

E. Classification

On the fortuity of the necessity of our application, classification is essential; Multi Noval Bayes and Support Vector Machine are inherited here. Behind training the data set, the classifier is regulated to utilize the government provided data [13].

Multi Noval Bayes Theorem lies in the area of probabilistic classifier. It fringes number of algorithms which accompanies the principle i.e. entire pair of the attributes is quarantined released among each other. To flow with this algorithm, a data set needs to be inherited. The data set is segregated in two different parts i.e. Feature Matrix and Response Vector. In Feature Matrix, entire rows of data set are present possessing value of dependent attributes whereas in Response Vector, value of prediction i.e. Class Variables for complete vectors of attribute matrix are available .

The postulation of algorithm is individuality and same contribution of entire features for producing output. It has to be accepted, all time it is not mandatory to be correct with real time dataset. Support Vector Machine is type of the Linear Classifier Supervised Learning (SL) technique.

It is inherited for Classification and Regression circumstances. It provides the best output for selecting highest margin hyper-plane classes. Over-fitting issue is mainly handled by SVM model. Classes which comprise of single attribute are classified by linear dimension facet. Classes which comprise of binary attributes are classified by sketching line between classes and Class which comprise of multiple features; they are differentiated by sketching hyper planes.

The non-linear data is differentiated by transforming attribute space in high-dimensional space which is differentiated by drawing hyper-planes. Figure shows the steps involved in aspect-based Machine Learning (ML) approach

IV. RESULT AND DISCUSSION

An application is developed asking for user details like the name, the age and the area of the interest. There are numerous areas in which various parties had done massive work.

Resting on the choice of user, they will select the area like individual's have their own interest and on the dependency of interest; wishes to select election party who had done remarkable job in asked areas. After entering the details, it will recommend the vote to user. Fig. 3 shows for Rahul of age 20 years; according to his selected category, "Indian National Congress" is the best to vote. Fig. shows for Jolly of age 40 years; according to his selected category, "Janta Dal" is the best to vote.

Fig. 3.Screenshot 1 of developed model

138

Fig. 4.Screenshot 2 of developed model

Fig.4 shows for Neha aged 40 years; according to his selected category, "Indian National Congress" is the best to vote.

V. CONCLUSION AND FUTURE SCOPE

In this paper, we have developed a mode; which is recommending votes to users based on their area of interest. The area plays a supreme role here, as everyone has their own interests and in their engrossment they prefer party for their vote as all parties had worked in numerous field. Once filled the asked details, the model will perform analysis on the data about the work done by them using Machine Learning approach, it entitles the model to train by them with no external support. The work can be extended considering all the election parties.

REFERENCES

1. Moravcsik, M. J., & Murugesan, P. (1988). Some Results on the Function and Quality of Citations: Social Studies of Science 3(4), 538.
2. Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In LREc (Vol. 10, No. 2010, pp. 1320-1326).
3. Liu B., "Sentiment Analysis and Opinion Mining", in Synthesis Lectures on Human Language Technologies", pp. 1-167, 2012.
4. Turney P.D., "Thumbs UP or Thumbs Down: Semantic orientation applied to unsupervised classification of Reviews", In Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistic, pp.417-424, 2002.
5. Pang B., and Lee L.," A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts", In Proceedings of the 42nd annual meeting on Association for Computational Linguistics, Associa tion for Computational Linguistics, p. 271, 2004.
6. Ye Q., Zhang Z., & Law R.," Sentiment classification of online reviews to travel destinations by supervised machine learning approaches", Expert Systems with Applications, pp. 6527-6535, 2009.
7. Liang H., Xu Y., & Li Y.," Mining Users' Opinions Based on Item Folksonomy and Taxonomy for Personalized Recommender Systems", In Data Mining Workshops (ICDMW), IEEE International Conference on pp. 1128-1135. December, 2010.
8. Y. Gil, F. Wang and S. Kambhampati, " Listening to the crowd: automated analysis of events via aggregated twitter sentiment", Proc. of the Twenty-Third international joint conference on Artificial Intelligence, Beijing, China, pp. 2640–2646, 2013.
9. Whittle R., Medd W., Deeming H., Kashefi E., Mort M., Twigger Ross C., Walker G., Watson," After the Rain – learning the lessons from flood recovery in Hull Flood, Vulnerability and Urban Resilience: a real-time study of local recovery following the floods",Lancaster University, Lancaster UK, 2007.
10. Paokanta, Wang H., Can D., Kazemzadeh, "A system for real-time twitter sentiment analysis of 2012 US presidential election cycle", in Proceedings of the ACL 2012 System Demonstrations, Associations for Computational Linguistics, pp. 115-120, July, 2012.
11. Bucur Cristian, "Aspects regarding detection of sentiment in web content", International Journal of Sustainable Economies Management (IJSEM), Volume 3, issue 4, p.24-32, ISSN: 2160-9659, 2014.
12. T. Simon, A. Goldberg, L.A. Daniel, D. Leykin and B. Adini, "Twitter in the cross fire-the use of social media in the westgate mall terror attack in Kenya", PloS one, vol. 9, 2014.
13. Asad K. I., Ahmed T., & Rahman M. S.," Movie popularity classification based on inherent movie attributes using C4. 5, PART and correlation coefficient", In Informatics, Electronics & Vision (ICIEV), 2012 International Conference, pp. 747-752, May, 2012.

AUTHORS PROFILE

Swati Sharma, B.tech(Honors.),M.Tech (Honors.), Ph.D. pursuing from Shobhit university.Assistant professor in MIET, Meerut since 2010.

Dr. Mamta Bansal, Ph.d from Shobhit University, Meeut, Professor in Shobhit University, Meerut.