International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 05 May| 2022 www.irjet.net p-ISSN: 2395-0072

A Review of machine learning approaches to mine Social Choice of voters.

Ankita Sengar

[1]*Assistant professor*

Abstract – *Mining the intention of user and predicating their future behavior is biggest tool for any commercial or noncommercial organization. Also Predicting election results is new area where intent mining is applied. In last decade, social platforms has been immensely used in elections. There are many statistical and AI based model which has predicted the result of a national election to a great accuracy. This paper presents a literature review of various machine learning based strategies used to analyze voter intent from their social media handle like twitter and predicting the results of elections.*

Key Words: **Social Choice theory, Intent Mining, Lexicon Model, Supervised Model, NLP, Ensemble, Naïve Bayes, SVM, Tfid vectorization, Election, Twitter**

1.INTRODUCTION

Today social media provide various platform where intellectuals of various field connect and interact their ideologies be it social, political or personal. These platforms with time have become like a data mine which holds people intent from several accepts, which if properly excavated and analyzed are worth the millions for several commercial and noncommercial organizations. Data Mining is the process of knowledge discovery from data (KDD) by sorting and analyzing large data sets. This data could extract knowledge like user opinion regarding there likeness or unlikeness towards a product, favor or against certain government policy, their sentiments toward something or their intent to vote in favor of certain party or person etc.

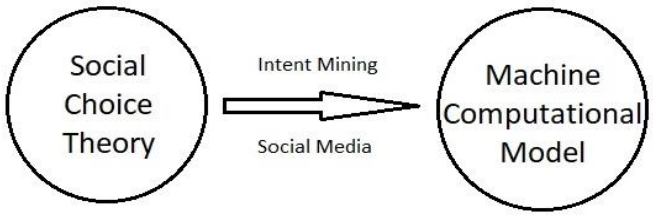


Fig 1: Block diagram representation of approach

Social choice theory is the field of microeconomics which study collective decision processes of Society. It concerns a cluster of actions and results concerning to that action (individual inputs like votes, preferences, judgments) into collective outputs (e.g., collective decisions, preferences of governance). Social choice is about a group of individuals or society choosing a winning outcome (e.g., policy, electoral candidate) from a given set of options. Many AI models are used these days to harness user opinion regarding certain policy or electoral choices.

Intent of a user can be positive, negative or neutral towards certain political party or representative. These intents of a voter are implicitly or explicitly shared by them on various social media platforms like twitter, Facebook etc. These implicit and explicit opinion is collected depending on well-articulated features. And several of these features of many people makes a dataset which is used to extract patterns and knowledge which could either directly predict voter intent or could be used to train a machine which will in future predict such results if provided with similar features.

There have been many early works for predicting election. Tumasjan[1] showed that by using only volume of tweets with mere mention of political parties can be used to statistically represent election polls favor. Wang[2] attempted to improve Tumasjan[1] proposed work by combining and separting tweets into positive sentiment and negative sentiment of voters. Franch F[3] proposed vote share prediction model using ARIMA (auto regressive integrated moving average) model, which exceeded the accuracy of traditional expensive polls. Barkha Bansal[4] used lexicon-based model to collect data against positive and negative sentiments and then statistically predicted election results for UP. On the other hand, Marozzo[5] used a supervised machine model (Random Forest algorithm) to find polarity of tweets and news texts related to political campaigns.

In next section literature review of machine learning and intent mining approaches which are generally used to predict social choice or election result are covered.

2. LITERATURE REVIEW

Any approach be it manual statistical prediction or building a model which will predict for us, both the approaches heavily depend on data sets used for this prediction. Data sets used to make conclusion can break or make a model. Therefore, before reviewing specification of various algorithm and model used to predict a social choice. It's necessary to consider several steps which go into making datasets useful. Data to make prediction or train a machine can be collected from social media platform like Twitter. Twitter provides

the opportunity to collect data from it for research purpose for which many tools and library are provided by python and google colob. But this data is raw and unstructured consisting large amount of information which could be removed for more optimize and quick model.

2.1 Natural Language Processing

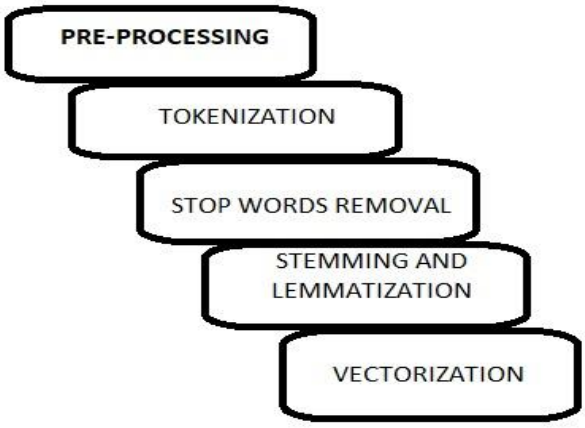


Fig 2: NLP Pipeline

NLP uses Language Processing Pipelines to read and decipher human languages to machine readable form. These pipelines consist of five sub processes. That breaks text into small chunks, reconstructs it to be analyzed, and processed to bring us the most relevant information when doing text analysis.

- **Sentence cleaning**: Clear all white space rows, html tags, smiley and segment para into sentence.
- **Tokenization**: Break Sentence into works (Token)
- **Stop Words removal:** Remove prepositions and postpositions from tokens.
- **Stemming and Lemmatization:** Stem token into base form and check grammar or if word is meaningful.
- **Vectorization:** Convert textual data into numeric vector form.

2.2 Approaches to predict Social Choice

Once properly structured and clean data is achieved voter intent is positive, negative or neutral can be predicted using either of lexicon based model or a supervised model or an unsupervised model. In this paper we will limit ourself to prior two approaches.

2.2.1 lexicon based Model

The lexicon model based approach, make use of pre-prepared sentiment lexicon to score a document. The pre-prepared sentiment lexicon should contain a word and corresponding sentiment score to it or a words or multiword tagged as positive, negative or neutral. The lexicon may be developed manually, for example, Taboada[6] or semi-automatically deriving sentiment values from resources such as WordNet, for example, Esuli and Sebastiani[7].

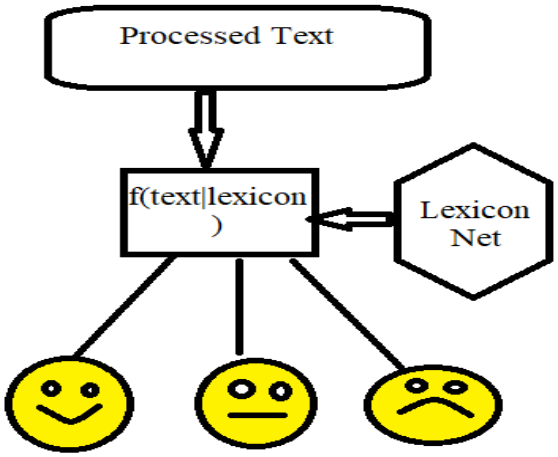


Fig 3: Lexicon Model

To predict the overall sentiment of a document, a formula, a function or an algorithm is needed to calculate the polarity of intent of voter.

After finding sentiment score (sentiment polarity and magnitude) of each tweet of the datasets, Barkha Bansal[4] used below formula to calculate total positive volume and total positive magnitude in each dataset. Finally, vote share for each party was statistically calculated using three methods presented in following equations.

$$VS_1 = \frac{TTV_i}{\sum_{i=1}^{l} TTV_i}$$

$$VS_2 = \frac{TPV_i}{\sum_{i=1}^{l} TPV_i}$$

$$VS_3 = \frac{TPM_i}{\sum_{i=1}^{l} TPM_i}$$

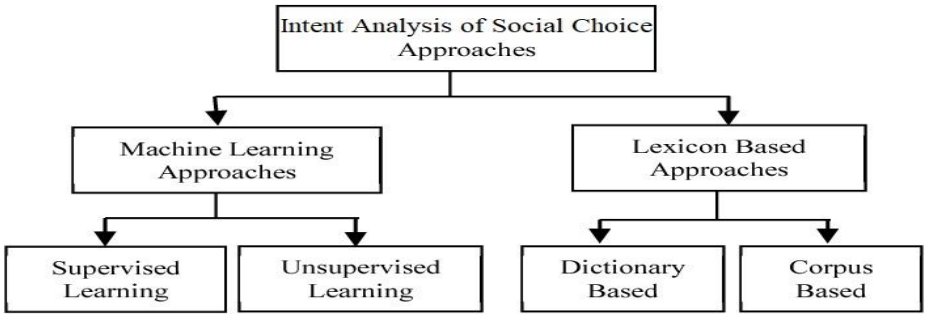


Fig 4: Intent Analysis Approach

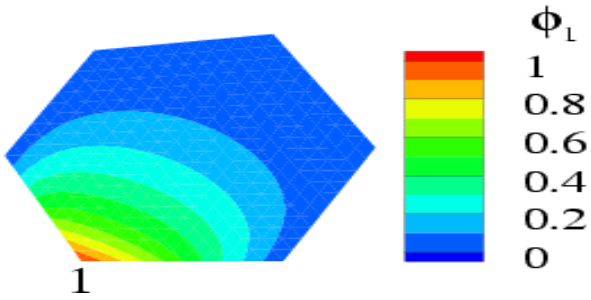
2.3 Machine Learning Approaches

Machine-learning methods to model a machine which can predict user intent of social choice can be supervised learning based or unsupervised learning based. Most research till now has been constricted to deploying a supervised model for intent prediction which after first four preprocessing steps either use a 'bag-of-words' approach (usually cleaned stemmed and lemmatized) or feature vectorization of corpus as independent feature vector (Counter Vectorization) or Multiword terms feature vector (Tfidf vectorization or hashing vectorization) to convert text into numerical format for further calculation and training. Supervised learning approaches which we intend to cover in this literature review is used to develop a classifier which can learn from mined features from social media and classify intent of user into positive or negative or neutral depending on model is bi-classifier or tri- classifier.

The most used Supervised machine-learning methods in intent analysis are the support vector machine (SVM) [8,9] ,Naive Bayes method [10] , Max Ent and logistic regression.

2.3.1 Maximum Entropy Model

Model based on maximum entropy adhere to Probabilistic classification. In this model instead of making inferences based on incomplete information, we draw them from that probability distribution that has the maximum entropy permitted by the information.



Maximum entropy distributions function could be represented as.

$$\underset{p(X)}{\text{maximize}} \; -\sum_X p(X)\log p(X) \;\; \text{subject to} \;\; \sum_X p(X) f_i(X) = c_i \;\; \text{for all constraints } f_i$$

$$\text{with solution: } p(X) = \exp\left(-1 + \lambda_0 + \sum_i \lambda_i f_i(X)\right)$$

2.3.2 Naïve Bayes Model

Naïve Bayes is a supervised machine classification algorithm, built on Bayes Theorem. Due to its Naïve assumption that predictors hold Conditional Independence among themselves it is called naïve bayes. It is based on assumption that all the features in a class are unrelated.

Bayes Theorem:

Let's take two events A and B. Then formula to calculate posterior probability P(B/A).

$$P\left(\frac{B}{A}\right) = \frac{\mathrm{P}(B) * P\left(\frac{A}{B}\right)}{\mathrm{P(A)}}$$

Posterior probability of single feature is easy to calculate but when we have two or more features, we can get a zero-probability problem. Therefore, we start by making a Naïve assumption of conditional independence among features to calculate posterior probability.

Suppose we have, Predictors: [X1, X2] and Target: Y
The formula to calculate the posterior probability is-

$$P\left(\frac{Y}{X1,X2}\right) = \frac{\mathrm{P}(Y) * P\left(\frac{X1,X2}{Y}\right)}{\mathrm{P(X1,X2)}}$$

Taking Conditional Independence for P(X1, X2/Y=1),

$$P\left(\frac{Y}{X1,X2}\right) = \frac{\mathrm{P}(Y) * P\left(\frac{X1}{Y}\right) * P\left(\frac{X2}{Y}\right)}{\mathrm{P(X1,X2)}}$$

This is the formula to calculate the posterior probability using Naïve Bayes Classifier.

2.3.3 Support Vector Machine

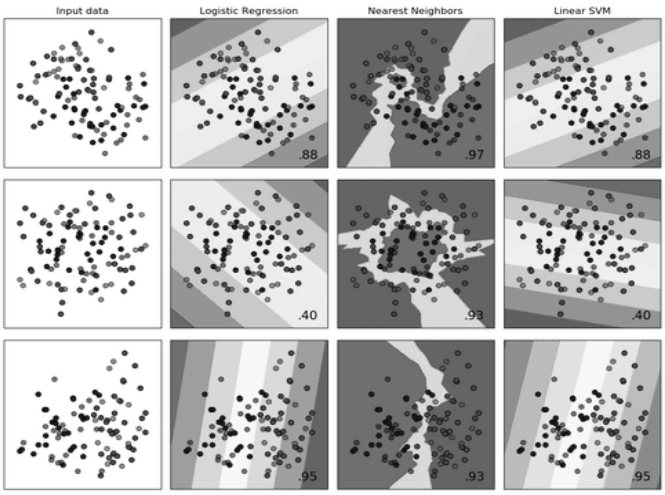


Fig:5 Data Visualization of various Algorithms

SVM is a regression and classification supervised machine learning algorithm. In SVM, we aim to find the best decision boundary called hyperplane among the data points that are plotted in n-dimensional space where n is the number of features. For a n-dimensional space i.e. n-feature visualization of data set, the hyper-plane has (n-1) dimensions. SVM can be used for Linearly separable data called Linear SVM model and also non linearly Separable data using a special SVM called Kernel SVM. Below Expression is representation of one such kernel function which could be used to separate non linear data.

$$K(x_i, x_{i'}) = exp(-\gamma + \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2)$$

2.3.4 Logistic Regression

A logistic regression is like a generalized linear model but with a canonical link function. i.e. output of generalized linear function is squashed in range of [0,1] using the sigmoid function (logistic function). Sigmoid-Function is an S-shaped curve that is discrete unlike continuous output result in linear classification. If output value is greater than threshold value or lower than a minimum threshold. we assign it as label 1, else we assign it a label 0 respectively.
In terms of easy computation, it is the best model among other generalized linear models for binary classification or regression. Function used to calculate the Logistic regression.

$$h_\theta(\mathbf{x}) = g(\theta^\top \mathbf{x})$$

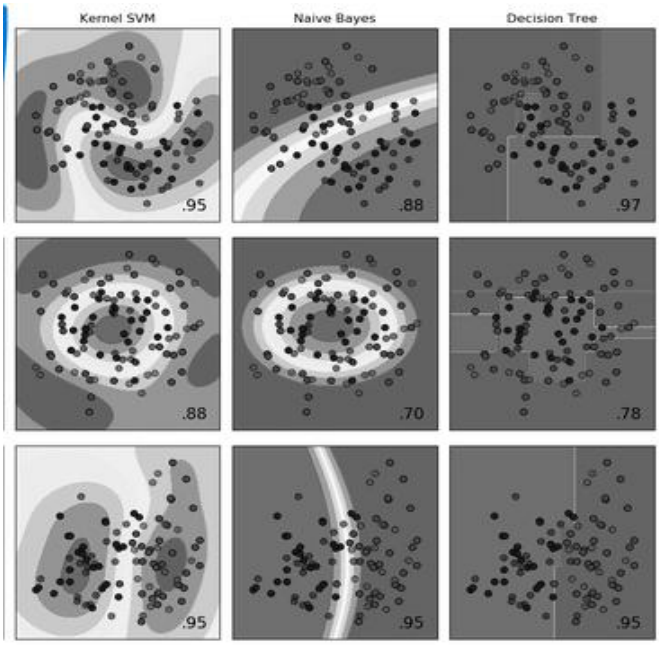
$$g(z) = \frac{1}{1 + e^{-z}}$$

2.3.5 Deep Learning

Deep learning is an advance representation of unsupervised learning which offers a set of algorithms inspired by how the human brain works, they are called neural nets. Deep learning architectures are being used heavenly for text classification because they are proving to perform at super high accuracy and precision.

The two deep learning architectures which are mostly used for text classification are Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN).

Deep learning works similar to how the human brain works to make decisions, using different techniques simultaneously to process huge amount of data collecting from social media.



3. CONCLUSIONS

There are lots of model using Artificial intelligence to predict election up to acceptable accuracy but still each day with new ensemble or better data processing this accuracy can be further increase. Deep learning model such a RNN which can train from unlabeled data and identify the hidden pattern and relational among that data, holds the future of machine model which can predict social choice of voters by using their mined intent.

REFERENCES

[1] Tumasjan A, Sprenger TO, Sandner PG, Welpe IM. "Predicting elections with twitter: What 140 characters reveal about political sentiment." Icwsm. 2010;10(1):178–185.

[2] Jungherr A, Jurgens P, Schoen H. "Why the pirate party won the german election of 2009 or the trouble with predictions: A response to ¨ tumasjan, a., sprenger, to, sander, pg, & welpe, im "predicting elections with twitter: What 140 characters reveal about political sentiment?" Social science computer review. 2012;30(2):229–234.

[3] Franch F. (Wisdom of the Crowds) 2: 2010 UK election prediction with social media. Journal of Information Technology & Politics. 2013;10(1):57–71.

[4] Barkha Bansala, Sangeet Srivastava. "On predicting elections with hybrid topic based sentiment analysis of tweets". Procedia Computer Science,Volume 135, 2018, https://miro.medium.com/max/1400/1*o4Dy1w4n2kDOLA8UEwGC9g.pngPages 346-353

[5] Marozzo F, Bessi "A. Analyzing polarization of social media users and news sites during political campaigns. Social Network Analysis and Mining". 2018;8(1)

[6] Taboada M, Brooke J, Tofiloski M et al. "Lexicon based methods for sentiment analysis". Comput Linguist 2011; 37: 267–307.

[7] Esuli A and Sebastiani F. SentiWordNet: "a publicly available lexical resource for opinion mining". In: Proceedings of the 5th

[8] Cortes C and Vapnik V. "Support vector networks". Mach Learn 1995; 20(3): 273–297.

[9] Vapnik VN and Vapnik V. "Statistical learning theory". New York: John Wiley & Sons, 1998.

[10] Zhang H. "The optimality of naive Bayes". In: Proceedings of the seventeenth Florida artificial intelligence research society conference, Miami Beach, FL, 12–14 May 2004, pp. 562–567. Palo Alto, CA: American Association for Artificial Intelligence

BIOGRAPHIES

Ankita Sengar
Assistant Professor
+8 year Experenice
Area of Interest:
Machine learning, IOT