International Journal of Scientific Research & Engineering Trends
Volume 6, Issue 2, Mar-Apr-2020, ISSN (Online): 2395-566X

Public Opinion Poll Generation using Twitter and Machine Learning Algorithms

Chelikani Satya Sankeerth,Sairam Sadu,Assistant Professor Choudari Lakshmi
Department of Computer Science Engineering
Gandhi Institute of Technology and ManagementVisakhapatnam,India
sankeerth.chelikani@gmail.com,sadusairam146@gmail.com,lakshmich533@gmail.com

***Abstract* – Today we live in an information age where a lot of data is being generated each minute. Because of quick increment in the measure of client created information via web-based networking media platforms like Twitter, several opportunities and new open doors have been prompted for associations that try hard to keep a track on data. Twitter is an immensely fast emanant micro-blogging social networking platform for clients to communicate their perspectives about various aspects like governmental issues,products,movies,sports etc. Because of the vast opinion rich web asset such as twitter a lot analysis is focusing on the area of sentiment analysis. Individuals are attempting to build a system that can identify and classify opinion or sentiment as given as an electronic text. Sentiment Analysis can be seen as a field of text mining and natural language processing.Sentiment analysis is a process of consequently recognizing whether a client created content communicates positive or negative opinion about an element.A right strategy to predict sentiments could be utilized to separate sentiments from the web and predict online client's preferences which may prove valuable for economic,market research and also for the government in different ways like keeping a track on customer reviews, government can get the public opinion on their policies easily and efficiently, it can also be used in election analysis and election prediction. Consequently, tweets can be utilized as a significant hotspot for mining public's sentiment.**

***Keywords* – Sentiment Analysis, Opinion Mining, Text Classification, Twitter, Preprocessing, Tf-idf vectorizer, Support Vector Machine, multinomial NB, Grid Search, Randomized Search.**

I. INTRODUCTION

In today's world with the rising technology the data generated by the technology is also getting increased day by day. This data is being collected by the organizations and being used in many ways. There are a lot of micro-blogging platforms available in the market today while twitter being first of its kind. Even today twitter is among the top of them maintaining about 330 million active users. Many people who write their activity, thoughts or opinions every day on Twitter. If we collect and process the opinions from a huge amount of data such as tweet data on Twitter, we can get insight information which is useful.

Opinion polls and surveys are the bridge between customers and product manufacturers. Assessment of public sentiment is a sort of review or request intended to quantify the public perspectives with respect to a specific point or an item.[5] Various strategies, including face to face and personal meetings, telephone meetings, and overviews sent via mail or email or accessible online have been utilized to accumulate and find out about individuals' contemplation about fundamental issues.There are two principal approaches for mining open political conclusions: subjective and quantitative. Quantitative techniques, for example, surveys, give a greater number of information than the subjective strategies, for example, interviews.Among different technologies available to get feedback\review social media plays a very crucial role. Individuals share their emotions and sentiments on Twitter on such an enormous scope, that it very well may be utilized as an important, openly accessible asset for both scholarly community and industry.Unlike traditional surveys, collecting and analyzing Twitter data is a cost-effective way to survey a large number of participants in a short period of time. It is particularly an interesting platform because of its concept of hashtags[1].

Along with the short messages, users can use the hashtag symbol '#' before a relevant keyword or phrase in their Tweet to categorize those Tweets and help them show more easily in Twitter Search. The use of hashtags makes the problem of text classification relatively easier since the hash tag itself can convey an emotion or opinion. While these tweets can be retrieved with the API key from twitter. By giving a keyword what we are currently surveying on we can get the current tweets from twitter. These tweets are now used as information or data to get the public poll on the current topic[2].For classifying the retrieved data we need to have a trained algorithm. The efficiency of the poll depends on this training. In this paper we are going to compare two algorithms namely SVM and multinomial NB and also we are going to

compare two hyperparameter tuning algorithms namely Grid search and Randomized search.[3] .The training will happen with the help of a movie review data set and after training we will be classifying tweets and will display public opinion with the help of a bar graph and pie chart. Out of the available algorithms we noted high performance with SVM and Gridsearch combined. While grid search will tune hyperparameters in SVM and give best parameters at which svm will work efficiently[8].

II. ALGORITHMS

We will be using only Supervised Learning algorithms in this paper, when we are training a supervised learning algorithm the training data will consist of inputs with their respective outputs.In training phase algorithm will search for patterns in the data that match with the desired outputs. In the next phase algorithm will take a new unseen input and determine which label it belongs to based on the prior training data[8]. The object of Supervised Learning algorithm is to predict the correct label for the unseen input data. Supervised learning algorithm can be simply represented as

$$Y=f(x)$$

Where x is the unseen input and Y is the predicted output

1. Algorithms for Classification

Classification is a supervised learning technique which is used to predict the category of the observation based on the training data.In classification an algorithm is trained based on the dataset, the trained algorithm will classify new observations. Such as Yes or No, 0 or 1, Spam or Not Spam , Positive or Negative etc. We use 2 algorithms for classification in this paper[12].

- **Support Vector Machine(SVM)**

A Support Vector Machine(SVM) is a supervised learning algorithm that can be used for both classification and regression problems as well. While it is used commonly for classification problems. SVMs work on the idea of finding a hyperplane that divides a data set into two classes effectively.

Ex:class-1:positive class-2:negative. Where support vectors are the data points which are present nearest to the hyperplane, the points in the dataset are plotted in training phase and a hyperplane is assumed. For a classification task with only 2 features a single hyperplane can separate the data. The distance between the hyperplane and the nearest data point to the hyperplane is considered as margin. The goal is to choose a hyperplane with the greatest possible margin between the hyperplane and any point in the training set which gives a greater possibility of classifying new data correctly[11]. A small change in the data set does not greatly affect the hyperplane so svm is considered to be stable.The performance of svm can be tuned using its hyperparameters like C and gamma.SVM can even handle non-linear data using Kernel trick.

- **Multinomial NB**

This is a classification technique based on Bayes Theorem with an assumption of independence among predictors.It is a probabilistic classifier that learns the probability of an object with certain features[9]. It can be used in spam filtration,sentiment analysis etc.Naive Bayes rule is termed “naive” because it makes the idea that occurence of a precise feature is freelance of the prevalence of the other options.It calculates the probability of each tag and outputs with the highest one.It finds the probability using Bayes Theorem.While Bayes theorem finds the probability based on the prior knowledge.It is a simple algorithm but it can outperform some sophisticated classification methods[11]. There are 3 types of Naive Bayes models.They are Gaussian,Multinomial,Bernoulli. Multinomial Naive Bayes is suitable with discrete features .It is used for multinomial distribution that generally requires integer feature counts.

2. Algorithms for Hyperparameter Tuning

Machine learning model is termed as a mathematical model with a number of parameters that need to be learned from the data. There are another kind of parameters that cannot be learned directly which are termed as Hyper-parameters.Hyper parameter tuning is the problem of selecting a set of optimal hyperparameters for training an algorithm. Hyperparameter is used to control the learning process.

- **Grid Search**

SVM also has some hyperparameters in which most used hyperparameters are “c” and “gamma” . These values are tuned to find the optimal hyper-parameter with which algorithm performance will be maximum.The main goal behind grid search is to create a grid of hyper parameters and all the combinations in the grid[5] . Hence this method is called Gridsearch.GridSearchcv takes a dictionary that describes the parameters that will be tried on a model in order to train it.In that dictionary keys are the parameters and the values are set to be tested.

```
param_grid = {‘C’: {0.1,1,10,100,1000],
’Gamma’:[1,0.1,0.01,0.001,0.0001],
’Kernel’:[‘rbf’]}
```

- **Random Search**

Random search is another kind of technique where random combinations of the hyperparameters are tried to find the best solution for the given model.It is similar to that of grid search but it takes less time compared to grid search and is reported to give better results than grid search.The selection of the parameters is completely random and no intelligence is used luck plays a key role. As random values are selected at each instance it is considered that the whole action space has reached because of randomness. In grid search it takes a very long time to cover every combination in search ,while in this search random combinations of parameters are considered in every iteration.It works best under the assumption that

not all the hyperparameters are equally important.The chances of finding an optimal parameter is higher in random search because of random search pattern when compared to that of grid search method[4].

III. DATASET

Data set plays a very important role in training any algorithm while the trained algorithm will predict the output. The performance of our project will mainly concentrate on the dataset and the algorithm we choose. In this paper a data set is needed for the purpose of training an algorithm, with this trained algorithm we can classify our retrieved tweets. 'Movie review' dataset is chosen for the purpose of training with 6000 rows of information and 2 fields which are labelled as 'label' and 'review'. The review field consists of english sentence which is a review while the label field consists of the classification of the respective review either positive - 'pos' or negative - 'neg'.

IV. PREPROCESSING

Preprocessing our data will improve our performance. There is a great scope for preprocessing our retrieve tweets as they come directly from the web. They contain a lot of metadata which will be not helpful for us[6]. The best way to eliminate that unnecessary data.Initially when we are retrieving tweets a lot of data like user id, user information, different kinds of text information along with font ,color,etc will be retrieved. All that information is useless and only the text tweet is only what we need[7]. All the data will be stored in a dictionary of dictionaries format while text will be presented in the dictionary of dictionaries. We need access to it and use the data in that. Within that tweets also sometimes other languages will be retrieved. We need to eliminate those tweets with other languages. For that purpose we use the 'langdetect' package in python. Along with the english tweets some tweets will include hyperlinks which will affect classification. So with the help of the 're' package in python and the regular expression 'http\S+' we can detect hyperlinks and we can eliminate them.

- **Vectorizer**

Tf-idf stands for Term Frequency-Inverse Document Frequency. It's a statistical measure which evaluates the relevance of a word to a document in a collection of documents. It generally consists of two metrics and the relevance depends on these 2 metrics: number of times a word appears in a document and the inverse document frequency of the word across a set of documents. It is widely used in automated text analysis, and it is very useful for scoring words in machine learning algorithms for Natural Language Processing[2]. It works by incrementing the proportionality to the number of times a word appears in a document, but is offset by the number of documents that contain the word.Due to this the common words in a document like this,that,if,what etc are ranked low even though they appear many times in the document because they don't mean much to the document in particular.

After multiplying these two terms we will get the tf-idf score of a word in a document.When we are passing our text data in the form of numbers the performance will be increased a lot when compared to using an algorithm directly without our vectorizer. This obtained score can be fed to machine learning algorithms[1].

V. TWYTHON

Twython is one of the libraries in python it is mainly used to access twitter. With twython we can access our twitter account programmatically. Along with using twitter we can also stream live tweets on a specific topic. For doing this we initially need a twython library installed in our python along with a specific API Key and Access token which is linked with our twitter account. With the help of these two along with necessary inbuilt methods we can access live tweets which we will be using further in our paper for gathering public opinion on a specific topic[1].

VI. IMPLEMENTATION AND RESULTS

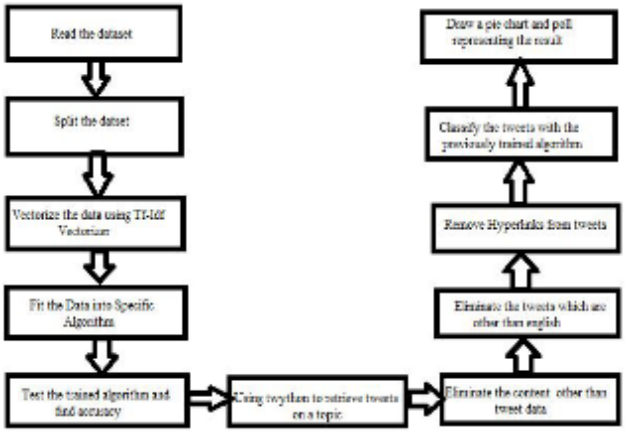

Fig 1. Process flow of complete project.

The algorithms used and the preprocessing steps to be used are listed and explained before. Initially the dataset needs to be read and then we need to train and test our data with the algorithms[6]. The input to the algorithms is in the format of digits or vectorized text which is vectorized with the help of tfidf vectorizer. The tested data also need to be in that format only. Initially the svm algorithm is used and then multinomial NB algorithm is used and their performance is compared and analyzed. SVM best fits the data set, so Hyperparameter tuners grid search and random search are used with svm and the performance of both the tuners are compared and analyzed. Out of both the tuners grid search gave us a better and efficient resultWith the help of twython along with twitter API key and API secret key we need to

retrieve tweets on a topic from twitter and perform preprocessing steps.To the retrieved tweets we need to use our trained algorithm for classifying them[10]. The classified tweets are then displayed in the format of pie chart and bar graph.

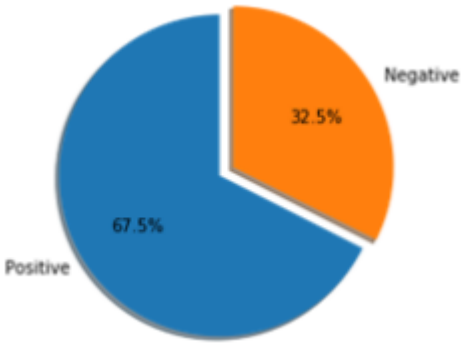

Fig 2. Result poll on a topic in pie chart.

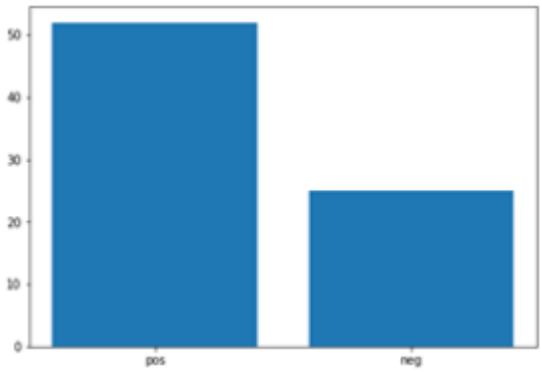

Fig 3. Result poll on a topic in bar graph.

VII.CONCLUSION

In this paper we proposed a model which gives the public opinion poll on any topic or subject by gathering tweets from twitter and using it as a source of information.There are a lot of ways for gathering public opinion out of which gathering it from social media is an ideal form because it have minimum cost and also most accurate information will be available. Out of all the social media platforms available we choose twitter.because it is used widely for expressing users' information. We choose different classification algorithms for classifying the retrieved tweets. Initially we used SVM and then multinomial NB out of which svm gave better results. Then we tuned SVM using hyperparameters grid search and random search out of which grid search gave us better results with highest accuracy of 92.4% at c =10 And gamma = 0.1.

The proposed model can be used to get a public opinion poll easily and efficiently.This model can be used for identifying public response by the government when they impose new laws or when they introduce new schemes. This model can also be used for election analysis which gives the public opinion poll on the particular political party. This can also be used by different industries where they can get the public response on their product.

REFERENCES

[1]. Karami, Amir et al. “Mining Public Opinion about Economic Issues: Twitter and the U.S. Presidential Election.” ArXiv abs/1802.01786 (2018): n. Pag.

[2]. J. Ramteke, S. Shah, D. Godhia and A. Shaikh, "Election result prediction using Twitter sentiment analysis," 2016 International Conference on Inventive Computation Technologies (ICICT), Coimbatore, 2016, pp. 1-5.

[3]. Jumadi, D. S. Maylawati, B. Subaeki and T. Ridwan, "Opinion mining on Twitter microblogging using Support Vector Machine: Public opinion about State Islamic University of Bandung," 2016 4th International Conference on Cyber and IT Service Management, Bandung, 2016, pp. 1-6.

[4]. Jadav, Bhumika M. and Vimalkumar B. Vaghela. “Sentiment Analysis using Support Vector Machine based on Feature Selection and Semantic Analysis.” (2016).

[5]. Liu, Bing. “Sentiment Analysis and Opinion Mining”. Morgan & Claypool Publisher. May, 2012.

[6]. Anstead, N. & O’Loughlin, B. (2015), ‘Social media analysis and public opinion: The 2010 uk general election’, Journal of Computer-Mediated Communication 20(2), 204– 220

[7]. Alexander Pak and Patrick Paroubek. "Twitter as a corpus for sentiment analysis and opinion mining". In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10), may 2010.

[8]. Y. Yang and F. Zhou, "Microblog Sentiment Analysis Algorithm Research and Implementation Based on Classification", 2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES) ,2015

[9]. Xiaohui Yu,Yang Liu, Aijun An” An Adaptive Model for Probabilistic Sentiment Analysis”, IEEE Computer Society ,Volume, Issue No. : 4191-4/10, pp-661-667, November 2010

[10]. Mouthami, K. et al. “Sentiment analysis and classification based on textual reviews.” 2013 International Conference on Information Communication and Embedded Systems (ICICES) (2013): 271-276.

[11]. Saraswati, Ni Wayan et al. “Naïve Bayes Classifier dan Support Vector Machine untuk Sentiment Analysis”. National Seminar of Information System. Bali, Indonesia. 2013.

[12]. Jahanbakhsh, K. & Moon, Y. (2014), ‘The predictive power of social media: On the predictability of us presidential elections using twitter’, arXiv preprint arXiv:1407.0622.