Twitter Sentiment Analysis on Government Law Using Real Time Data

Sujata Patil[1], Bhavesh Wagh[2], Aditya Bhinge[3], Aakash Sahal[4], Prof. Madhav Ingale[5]

[1,2,3,4] Student, Department of Computer Engineering, Jayawantrao Sawant College of Engineering, Hadapsar, Savitribai Phule Pune University, Pune, Maharashtra, India

[5]Asst. Prof. of Department of Computer Engineering, Jayawantrao Sawant College of Engineering, Hadapsar, Pune, Maharashtra, India

ABSTRACT

Social media monitoring has been growing day by day so analyzing social data plays an important role in knowing people's behavior. So we are analyzing Social data such as Twitter Tweets using sentiment analysis which checks the opinion of people related to government schemes that are announced by the Central Government. This paper-based is on social media Twitter datasets of particular schemes and their polarity of sentiments. The popularity of the Internet has been rapidly increased. Sentiment analysis and opinion mining is the field of study that analyses people's opinions, sentiments, evaluations, attitudes, and emotions from written language. User-generated content is highly generated by users. The growing importance of sentiment analysis coincides with the growth of social media such as reviews, forum discussions, blogs, micro-blogs, Twitter, and social networks. It is difficult to analyze or summarize user-generated content. Most of the users write their opinions, thoughts on blogs, social media sites, E-commerce sites, etc. So these contents are very important for individuals, industry, government, and research work to make decisions. This Sentiment analysis and opinion mining research is a hot research area that comes under Natural Language processing. We plot and calculate numbers of positive, negative, and neutral tweets from each event.

Index Term : Sentiment Analysis, Opinion Mining, Natural Language Processing

I. INTRODUCTION

Social media have become an emerging phenomenon due to the huge and rapid advances in information technology. People are using social media on daily basis to communicate their opinions with each other about a wide variety of subjects, products, and services, which has made it a rich resource for text

mining and sentiment analysis. Social media communications include Facebook, Twitter, and many others. Twitter is one of the most widely used social media sites.

The number of Twitter messages sent per second worldwide. There is no standard method for mining and analyzing social media business data. Here, an open-source approach for text mining and sentiment analysis using a set of R packages for mining Twitter data and sentiment analysis is presented, which is applicable for other social media sites. A case study of Eight Indian Government Schemes is presented to show the importance of analyzing user-generated online opinions from Twitter. This is helpful for evaluating government performance monitoring from the People's perspective instead of making Peoples surveys which are expensive and time-consuming. Sentiment analysis has been first introduced by Liu, B. It is also known as opinion mining and subjectivity analysis is the process determine the attitude or polarity of opinions given by humans to a particular scheme. Sentiment analysis can be applied to any textual form of opinions such as blogs, reviews, and Microblogs. Microblogs are those small text messages such as tweets, a short message that cannot exceed 160 characters. These microblogs are easier than other forms of opinions for the sentiment. Sentiment analysis can be done on a document level or a sentence level. In the first case, the whole document is evaluated to determine the opinion polarity, where, the features describing the product/service should be extracted first. Whereas, the second one, the document is divided into sentences each one is evaluated separately to determine the opinion polarity.

II. LITERATURE SURVEY

Research is going on tremendously in the area of sentiment analysis. Some of the important reviewed papers are discussed below:

Gupta et al. [1] presented a novel method which deals with location-wise tweets and meanwhile compared the working efficiency based on accuracy. Author implemented Naïve Bayes and Support Vector Machine classifiers in order to group the twitter data into positive and negative tweets. At that point, the locations are put into categories and sentiment mapping is done which helped in analyzing the opinions of individual Indian states separately. Important features were considered for classification such as latitude, longitude, kilometers and number of tweets to distinguish the opinions based on region. Importance of dataset pre-processing is also demonstrated in the paper. However the author recognized state-wise people reactions to demonetization, because of population polarization, the general feeling of the citizens couldn't be caught.

Gautama et al. [2] focused on three supervised learning algorithms namely SVM, Naïve Bayes and maximum entropy for the twitter data classification. Tweets are classified based on the expressed sentiments. Obtained results are contrasted with respect to performance measures such as: accuracy, precision and recall. A semantic analysis is driven from WordNet sentiment dictionary subsequent to training and categorization. Consequently, a relative measurement is done on the classification through supervised machine learning techniques and semantic analysis.

Tsapatsoulis et al. [3] tried to label tweets manually for the training dataset preparation. Then these manually annotated tokens are tested to check whether this dataset can form an index of terms that can be utilized for creating viable tweet classification models. Then the manually labeled index of terms is compared with various automatically extracted feature sets in the categorization of tweets using machine learning framework. Author used three distinct algorithms for the justification purpose. Three approaches were identified namely lexicon approach, machine learning strategy and the social strategy. From the observation, it is found that the explicitly identified tokens gave better performance accuracy than the other feature extraction methods. However,

the feature set combination is not considered in the result observed by the author.

Taiwo Kolajo et. al. [4] proposed an advanced method for pre-processing the social media tweets. He used Nigerian tweets as source text. With the help of Twitter API, tweets subjected to politics are collected. More concentration was given on local knowledge such as slang words are identified during pre-processing stage. Ambiguous words are resolved with the help of Lesk algorithm on context basis. Models are built using SVM, Multi-layer Perception (MLP) and Convolutional Neural Network (CNN). Obtained result proved that CNN performs well than SVM and MLP. Built model got an accuracy of 99.17% while classifying Naija tweets.

Parveen et al. [5] demonstrated on HDFS and MapReduce architecture. While implementing initially the dataset is pre- processed. After that, a supervised machine learning technique named Naïve Bayes is enforced on the dataset for the classification. A trained SentiWordNet dictionary is used while training the classifier. In this review, two methodologies have been used to implement Naïve Bayes technique. First method is based on map stage where SentiWordNet dictionary content is read from a file and transformed to Hashmap table for key-value based polarity extraction of words which makes the process faster. Second methodology consists of reduce stage, where the overall sentiment polarity of individual tweet is gathered. Then the identified polarity is grouped into five classes as strongly positive, positive, strongly negative, negative and neutral. However, the sentiment classification is divided among 5 different classes; the implementation part is carried out assistance of just a single classifier.

Abdelwahab et al. [6] studied the impact of training dataset size variation on SVM and Naive Bayes classifiers. Author also analyzed the impact of shifting the training dataset size on the learning curves to absorb information of both SVM and Naïve Bayes classifiers when utilized in twitter opinion examination. Along with this, the effectiveness of the training dataset on various ensemble types is tested. When two ensemble methods are compared with each other, it is found that AND combined result of Support Vector Machine and Naïve Bayes classifiers is good enough than the ensemble 2 method where SVM and Naïve Bayes algorithms are OR combined. However, consolidating the after effects of classifiers brought about equivocal outcomes for the correlation between Naive Bayes and SVM.

Neethu et al. [7] considered two unique methods for extricating the sentiments from dataset; symbolic approach and the machine learning strategy. Symbolic approach utilizes the lexical resources which are available in unsupervised opinion categorization. In machine learning strategy, various techniques such as Random Forest, Maximum Entropy, and Support Vector Machine are implemented to categories user opinions based on extracted features. The author identified that Machine Learning strategy is more efficient and simpler than Symbolic approach.

Sahni et al. [8] utilized the subjectivity of sentences to categorize dataset. Normally an absolutely objective sentence does not pass on any assumption. Therefore, just subjective sentences are used here in the training set. In the implementation stage, different natural language processing libraries are adapted to identify the subjective sentences. Thereafter, the various classifiers are applied on the pre- processed dataset. Distinctive feature removal methods namely n-grams, POS tagging etc are utilized to extract significant features.

III. SYSTEM OVERVIEW

A. Objectives:
- To implement an algorithm for automatic classification of text into positive, negative or neutral.
- Sentiment analysis to determine the public opinion is positive, negative or neutral towards the subject of interest.

- Graphical representation of the sentiment in form of pie-chart.

B. Problem Statement

- With the popularity of face recognition, criminals the problem in sentiment analysis is classifying the polarity of a given text at the document, sentence or feature level.
- To analyze whether the expressed opinion in a document, a sentence or an entity feature is positive, negative, or neutral.
- Sentiment Analysis is one of the approaches that are used for analyzing the opinion of people regarding any government law.
- In our work, we analyzed various Indian Government projects, schemes and laws by extracting the data from social networks.

C. Architecture Diagram
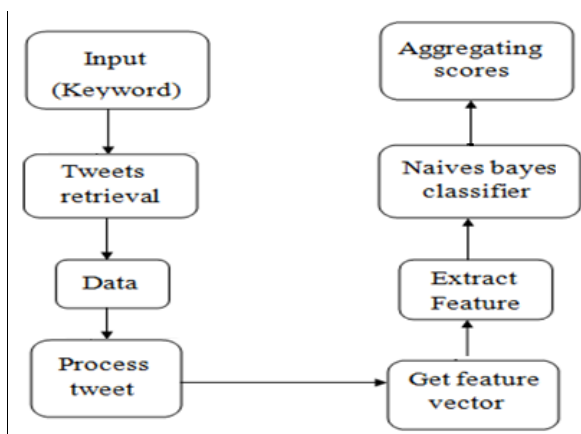
Fig.1: System Architecture

There are primarily two types of approaches for sentiment classification of

1) Using machine learning based text classifier such as naïve bayes.
2) Using NLP (Natural Language Processing)

We will be using machine learning and natural languages processing for sentiment analysis of tweet.

The machine learning based text classifiers are a kind of supervised machine learning paradigm where the classifier needs to be trained on some more labelled training data before it can be the original data hand labelled manually. After suitable training they can be used on the actual test data. The naive bayes (NB) is statistical classifier whereas support vector machine is ka kind of vector space classifier. The statistical text classifier scheme of naive Bayes can be adapted to be used for sentiment classification problem as it can be visualized as a 2-class text classification problem in positive and negative classes. Support vector machine is k kind of vector space model based classifier which requires that the text documents should be transformed to feature vectors before they can used for classification.

D. Proposed Algorithms

1) NLP (Natural languages processing).
2) Naive Bayes classifier.

1) NLP (Natural languages processing):

NLP is a field in machine learning with the ability of a computer to understand, analyze, manipulate, and potentially generate human language.NLP algorithms are used to provide automatic summarization of the main points in a given text or document. NLP algorithms are also used to classify text according to predefined categories or classes, and is used to organize information, and in email routing and spam filtering. NLP algorithms are typically based on machine learning algorithms. Instead of hand-coding large sets of rules, NLP can rely on machine learning to automatically learn these rules by analyzing a set of examples (i.e. a large corpus, like a book, down to a collection of sentences), and making a statistical inference.

2) Naive Bayes Classifier:

Naive Bayes is a classification algorithm that is suitable for binary and multiclass classification. It is a supervised classification technique used to classify future objects by assigning class labels to instances/records using conditional probability. Naive Bayes is a simple and powerful algorithm for predictive modeling. Naive Bayes is called naive because it assumes that each input variable

is independent. This is a strong assumption and unrealistic for real data; however, the technique is very effective on a large range of complex problems.

E. System Requirements

- Software Requirements
  - ➢ Python IDE (PyCharm)
  - ➢ Operating system (windows 7 or further)
- Hardware Requirements
  - ➢ Intel i3 64bit or further
  - ➢ Working computer.

IV. RESULT AND DISCUSSION

The details about the obtained result are discussed here. In the proposed work, we have created a dataset of 15000 tweets, which is distributed in 80:20 ratios between training and testing. Table 1, Table 2 show the performance measures of Random Forest, KNN, Logistic Regression, Naïve Byes and Support Vector Machine based classifiers respectively. Model efficiencies are compared with respect to precision, recall and F1-score. Similarly, Table 6 shows the performance of the classifier models based on accuracy.

Table 1: Naive Bayes Classification Measurements

Performance Measures (%)	
Positive Recall	56
Negative Recall	93
Positive Precision	81
Negative Precision	81

Table 2: Random Forest Classification Measurements

Performance Measures (%)	
Positive Recall	76
Negative Recall	93
Positive Precision	84
Negative Precision	89

V. CONCLUSION AND FUTURE SCOPE

Thus, we can conclude that this twitter sentimental analysis would be a great added feature and would greatly benefit the government in order to track the negative, foul and violent comments. Also after introduction of this sentiment analysis, users would use this in daily life in social media platform.

VI. REFERENCES

[1] Rajkumar S. Jagdale, Vishal S. Shirsat, Sachin N. Deshmukh," Sentiment Analysis of Events from Twitter Using Open Source Tool International Journal of Computer Science and Mobile Computing ISSN 2320–088X Impact Factor: 5.258 IJCSMC, Vol. 5, Issue. 4, April 2016

[2] Xing Fang and Justin Zhan."Sentiment analysis using product review data". Journal of Big Data 2015. DOI: 10.1186/s40537-015-0015-2

[3] Seyed-Ali Bahrainian, Andreas Dengel, "Sentiment Analysis and Summarization of Twitter Data", CSE, 2013, 2013 IEEE 16th International Conference on Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on Computational Science and Engineering (CSE) 2013, pp. 227-234, doi:10.1109/CSE.2013.44

[4] Prashant Raina, "Sentiment Analysis in News Articles Using Sentic Computing", ICDMW, 2013, 2013 IEEE 13th International Conference on Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on Data Mining Workshops (ICDMW) 2013, pp. 959-962, doi:10.1109/ICDMW.2013.27.

[5] Geetika Gautam and Divakar yadav, "Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis", 978-1-4799-5172-7/, pp.437-442 2014 IEEE.[6] Federico Neri , Carlo Aliprandi , Federico Capeci, Montserrat Cuadros, Tomas By, Sentiment Analysis on Social Media, Proceedings of the 2012 International Conference on Advances in Social Networks

Analysis and Mining (ASONAM 2012), p.919-926, August 26-29, 2012

[6]   [6] Godbole, N.; Srinivasaiah, M. & Skiena, S. (2007), Large-Scale Sentiment Analysis for News and Blogs, in 'Proceedings of the International Conference on Weblogs and Social media (ICWSM)'.

[7]   H. Tang, S. Tan, and X. Cheng. A Survey on Sentiment Detection of Reviews. Expert Systems with Applications. Elsevier. 2009, doi:10.1016/j.eswa.2009.02.06

[8]   Yulan He, "Learning sentiment classification model from labeled features. "Proceedings of the 19th {ACM} Conference on Information and Knowledge Management, {CIKM} 2010, Toronto, Ontario, Canada, October 26-30, 2010.

[9]   Long-Sheng Chen, Cheng-Hsiang Liu, Hui-Ju Chiu, "A neural network based approach for sentiment classification in the blogosphere", Journal of Informetrics 5 (2011) 313–322.

Cite this article as :