ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

Technique for Sentiment Analysis from Twitter Data

Chandan Arora [1], Dr. Rachna [2]

[1]Research Scholar, Department of Computer Science, Global Institutes of Management &Emerging Technologies, Amritsar, India

[2]Associate Professor, Department of Computer Science, Global Institutes of Management & Emerging Technologies, Amritsar, India

ABSTRACT: Sentiment analysis (SA) enables online users to check details regarding a particular product if it is good enough or not. Various organizations such as e-commerce giants like Flipkart, Amazon, etc. publish reviews of products online. Users can analyse these reviews using Sentiment Analysis before buying anything. The four steps are followed for the sentiment analysis in the first step, the first step is applied in which data pre-processed. In the second step feature of the data will be extracted which is given as input to the third step in which data is classified for the sentiment analysis. In this paper, pattern based technique is applied for the feature extraction in which patterns are generated from the existing patterns which increase the accuracy of data classification. The proposed algorithm is implemented in python using the nltk tool box and it is analysed that execution time is reduced and accuracy is increased at steady rate.

KEYWORDS: Sentiment Analysis, Twitter, classifiers.

I. INTRODUCTION

Social media generates a large amount of sentiment rich data in the form of tweets, notices, blog posts, remarks, reviews, etc. Also, social media provides opportunity to businesses by giving a platform to associate with their customers for promotions. People mostly rely on user produced content over online, all things considered, for decision making. Micro-blogging websites have advanced to wind up a source of changed sort of information [1], because people post and share real-time opinions, feelings and express their sentiments about numerous subjects of day to day life. Most organizations have started surveying these blogs to understand general sentiments posted by the users [2]. Twitter is a social networking and micro-blogging administration that allows its users to post ongoing messages, called tweets. Tweets have numerous unique characteristics, which implicate new challenges and shape up the method for conveying sentiment analysis on it as compared to various domains. Sentiment analysis is a process that automates mining of opinions, views and emotions from text, speech, tweets and databases through Natural Language Processing (NLP). Sentiment analysis includes classifying opinions into positive, neutral or negative sentiments [3]. There are various types of views that can help in determining the already existing studies related to the sentiment analysis. They are the technique user, view of the text, level of detail of text analysis, rating level and so on. There are basically three categories that are broadly classified for all the existing techniques. Machine learning technique utilizes some learning algorithms regarding the assumption via training on a known dataset. The machine learning approach is utilized for predicting the extremity of assessments in view of trained and also test data sets. It applies the ML algorithms and utilizations linguistic features [4]. The lexicon-based approach includes calculating opinion polarity for a review utilizing the semantic orientation of words or sentences in the review. The "semantic orientation" is a measure of subjectivity and feeling in content. It takes a predefined list of words under consideration, where each word points to a particular opinion [5]. In the hybrid approach, the combination of both the machine learning and the lexicon based approaches can possibly enhance the conclusion classification execution. There are a few advantages and limitations in utilizing these distinctive approaches relying upon the motivation behind the examination [6].
There are different sorts of classifiers that are generally utilized for text classification which can be likewise utilized for twitter sentiment classification.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

1. Naive Bayes: Naïve Bayesian classifier learns the pattern of a set of documents that have been ordered. It compares the contents with a predefined set of words to classify the documents to a specific category. Give d a chance to be the tweet and c* be a class that is assigned to d, where

$$C^* = argmac_c P_{NB}(c|d)$$

$$P_{NB}(c|d) = \frac{(P(c))\sum_{i=1}^{m} p(f|c)^{n_{i(d)}}}{P(d)}$$

Here, "f" is a 'feature', count of feature (fi) is represented with ni(d) and is available in d which represents a tweet. Here, m denotes no. of features. Parameters P(c) and P(f|c) are computed through maximum likelihood estimates, and smoothing is utilized for unseen features.

2. Support Vector Machine: This classifier analyses the data; characterizes the decision boundaries and utilizes the kernels for computations that are performed. The input data are two sets of vectors of size m each [7]. Data given as input is classified into a class. It separates the tweets utilizing a hyper plane. SVM utilizes the discriminative function defined as

$$g(X) = wT\varphi(X) + b$$

"X" is the feature vector, "w" is the weights vector and "b" is the bias vector. Φ() is the non-linear mapping from information space to high dimensional feature space. "w" and "b" are found out automatically on the training set.

3. Maximum Entropy(ME): In ME Classifier, no assumptions are taken regarding the relationship in the middle of the features extracted from dataset [8]. This classifier always tries to maximize the entropy of the system by assessing the conditional distribution of the class label. The conditional distribution defined as MaxEnt makes no independence assumptions for its features, not at all like Naive Bayes. Maximum Entropy classifier is expressed in the equation below:

$$P_{ME}(c|d,\lambda) = \frac{\exp\left[\sum_i \lambda_i f_i(c,d)\right]}{\sum_c \exp\left[\sum_i \lambda_i f_i(c,d)\right]}$$

Where c is the class, d is the tweet and λi is the weight vector. The weight vectors decide the importance of a feature in classification.

II. LITERATURE REVIEW

Rincy Jose, et.al, (2015) proposed in this paper [9] Natural Language (NLP) based approach to enhance the sentiment classification by adding semantics in feature vectors and thereby using ensemble methods for classification. Adding semantically similar words and context-sense identities to the feature vectors will increase the accuracy of prediction. The conducted experiments show that the semantics based feature vector with ensemble classifier performs better than the traditional bag-of-words approach with single machine learning classifier. Also, the ensemble method outperforms the traditional classification methods by about 3- 5%. Among the ensemble methods Extremely Randomized Trees classification performs better than others.

Nehal Mamgain, et.al, (2016) proposed in this paper [10], an effort to enter into the novel domain of analyzing sentiments of individuals' opinions regarding different colleges of India. Not only a probabilistic model based on Bayes' theorem was utilized for spelling revision, which is disregarded in other research contemplates, additional preprocessing measures like the expansion of net lingo and removal of duplicate tweets were taken. Moreover, a contrast was illustrated among four distinct kernels of SVM: RBF, linear, polynomial and sigmoid. Multilayer Perceptron Neural Network surpasses the results yielded by the machine learning algorithms owing to its exceptionally accurate approximation of the cost function, ideal number of hidden layers and learning the relationship among input and output variables at every progression.

Aldo Hernández, et.al, (2016) proposed a sentiment analysis method on Twitter content to predict future attacks on the web [11]. The method is based on the daily gathering of tweets from two sets of users; the individuals who utilize the platform as a method for expression for views on relevant issues, and the individuals who utilize it to present contents

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

identified with security attacks in the web. The goal is to predict the response of specific groups involved in hacking activism when the sentiment is sufficiently negative among various Twitter users. For two contextual analyses, it is demonstrated that having coefficients of determination greater than 44.34% and 99.2% can figure out whether a significant increase in the percentage of negative opinions is identified with attacks.

Anurag P. Jain, et.al, (2015) proposed in this paper [12], an approach for examining the sentiments of users utilizing data mining classifiers. It additionally compares the performance of single classifiers for sentiments analysis over ensemble of classifier. Experimental results acquired demonstrate that k-nearest neighbor classifier gives high predictive accuracy. Results likewise demonstrate that single classifiers out-performs ensemble of classifier approach. It can be seen from the test results that data mining classifiers is a decent decision for sentiments prediction utilizing tweeter data.

Manju Venugopalan, et.al, (2015) proposed in this paper [13], a half and half model for sentiment classification to explore the tweet specific features to offer a domain oriented approach for investigating and extracting the sentiments of shoppers towards popular smart phone brands during previous few years. The analyses illustrated that results enhance by around 2 points on an average over the unigram baseline. The SVM accuracy has improved in the range 1.5 to 3.5 and J48 could provide an accuracy improvement ranging from 1.5 to 4 points across domains. The improved lexicon which have adapted polarities learning the domain and the tweet specific features extracted have added to the improvement in classification accuracies.

III. RESEARCH METHODOLOGY

The base paper is based on sentiment analysis from the social network sites. We extract the features and are subsequently classified using the classification techniques. The technique of N-gram is applied which extracts color features from the web sites. To extract the textural features technique of wavelet transformation is applied. The steps involved in sentiment analysis are: -

1. Input Data: In the first step, tweets are given as input, which can be either in the excel sheet or the real-time data, and then it is extracted using the twitty application.
2. Pre-processing: In the pre-processing phase, input data is pre-processed, i.e., data is tokenized and stop words are removed.
3. Feature Extraction: The pre-processed data is then given as input to the feature extraction algorithm in which n-gram algorithm is applied and priority to each word is assigned that needs to be classified.
4. Classification: In the last step, the classification technique is applied on the feature extraction data for the sentiment analysis. In this work, SVM classifier has been applied for data analysis.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

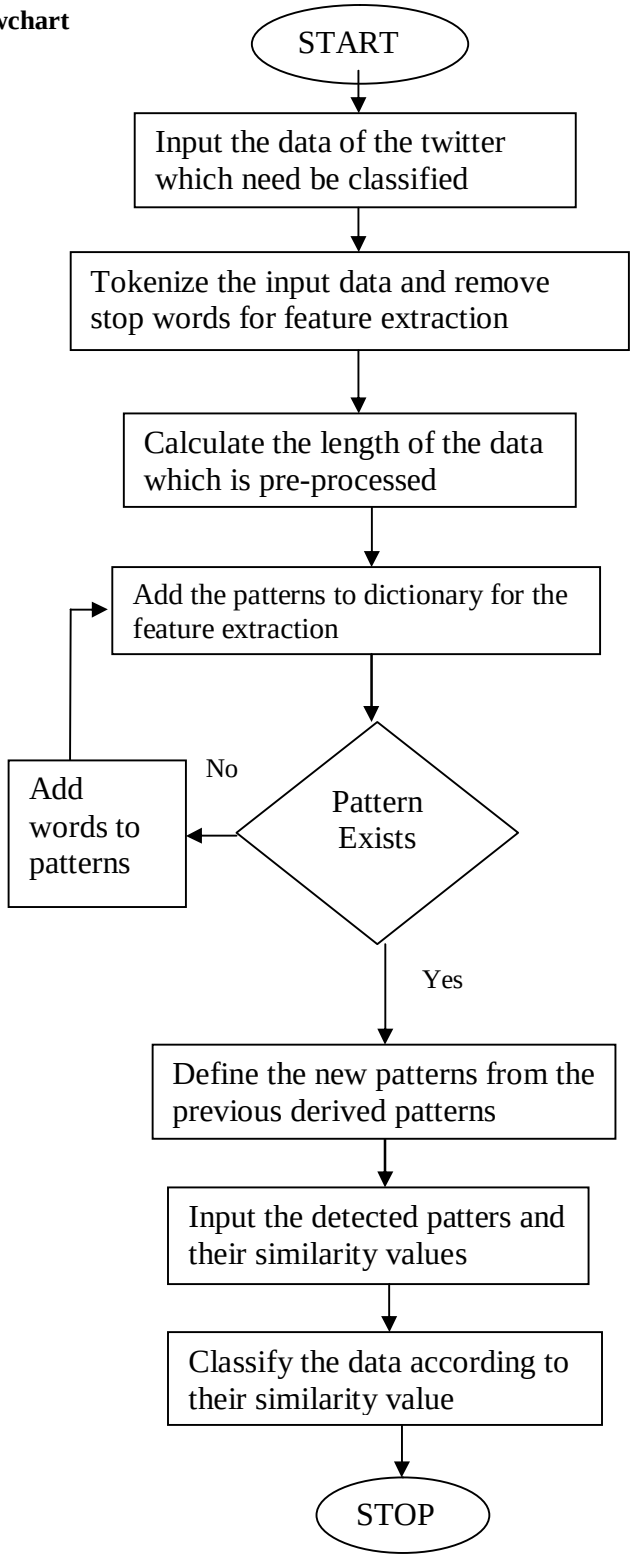
International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

Fig 1: Proposed Flowchart



ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

Experimental Results

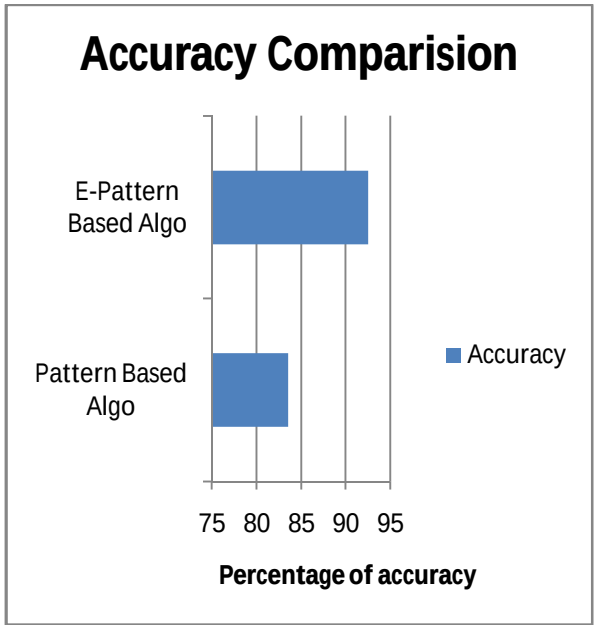


Fig 2: Accuracy Comparison

As shown in figure 2, the accuracy of pattern based algorithm and E-patterns based algorithm is compared in terms of accuracy and it is analysed that accuracy of enhanced algorithm is more due to better analysis of the data.

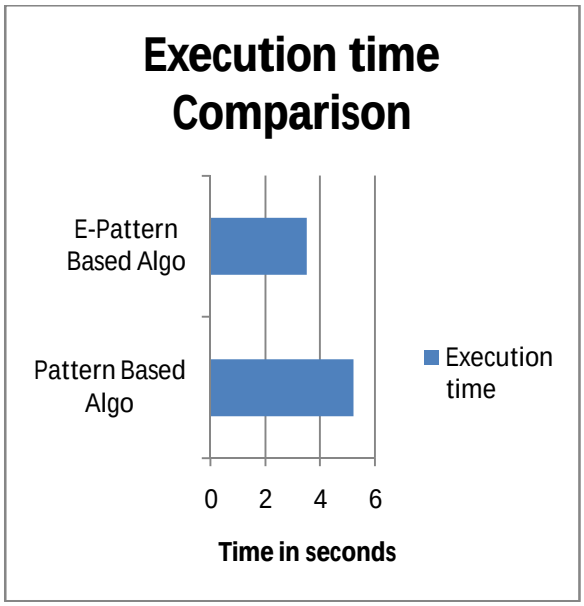


Fig.3 Execution time comparison

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

IV. CONCLUSION

In this work, it has been concluded that sentiment analysis is an efficient technique to analyse the user behaviour. The sentiment analysis contains four steps and in this work improvement in the feature extraction phase is done using the pattern based technique. The proposed improvement is implemented in python and it is analysed that execution time is reduced by 10 percent and accuracy is increased by 20 percent.

The proposed technique can be applied for opinion mining to analyse the behaviour of the users from the different fields in the future work. The proposed technique can be further extended for the sarcasm detection on the twitter data.

REFERENCES

[1] Eunjeong Ko, Chanhee Yoon, Eun Yi Kim," Discovering Visual Features for Recognizing User's Sentiments in Social Images", 2016, IEEE, 978-1-4673-8796, vol. 42, pp. 773-779

[2] Sara Tedmori, Rashed AI-Lahaseh," Towards a Selfie Social Network with Automatically Generated Sentiment-Bearing Hashtags", 2016, IEEE, 978-1-4673-8914, vol. 19, pp. 54-59

[3] F. Ciullo, C. Zucco, B. Calabrese, G. Agapito, P. H. Guzzi, M. Cannataro," Computational Challenges for Sentiment Analysis in Life Sciences", 2016, IEEE, 978-1-5090-2088, vol. 10, pp. 478-482

[4] Marie Katsurai, Shin'ichi Satoh," IMAGE SENTIMENT ANALYSIS USING LATENT CORRELATIONS AMONG VISUAL, TEXTUAL, AND SENTIMENT VIEWS", 2016, IEEE, 927y592-3453909-768-67, vol. 12, pp. 678-682

[5] Eunjeong Ko and Eun Yi Kim," Recognizing the Sentiments of Web Images using Hand-designed Features", 2015, IEEE, 918-1-4613-1290-9, vol. 14, pp. 784-788

[6] Stuti Jindal and Sanjay Singh," Image Sentiment Analysis using Deep Convolutional Neural Networks with Domain Specific Fine Tuning", 2015 International Conference on Information Processing (ICIP) , vol.46, pp. 654-663

[7] Dua'a Al-Hajjar, Afraz Z. Syed," Applying Sentiment and Emotion Analysis on Brand Tweets for Digital Marketing", 2015, IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), vol. 7, pp. 153-159

[8] Parinya Sanguansat," Paragraph2Vec-Based Sentiment Analysis on Social Media for Business in Thailand", 2016, IEEE, 348765342-43565-65765-555, vol. 9, pp. 231-236

[9] Rincy Jose, Varghese S Chooralil," Prediction of Election Result by Enhanced Sentiment Analysis on Twitter Data using Word Sense Disambiguation", 2015, IEEE, 978-1-4673-7349-4, vol. 4, pp. 123-128

[10] Nehal Mamgain, Ekta Mehta, Ankush Mittal, Gaurav Bhatt," Sentiment Analysis of Top Colleges in India Using Twitter Data", 2016, IEEE, 978-1-5090-0082-1, vol.14, pp. 67-72

[11] Aldo Hernández, Victor Sanchez, Gabriel Sánchez, Héctor Pérez, Jesús Olivares, Karina Toscano, Mariko Nakano and Victor Martinez," Security Attack Prediction Based on User Sentiment Analysis of Twitter Data", 2016, IEEE, vol. 13, pp. 673-678

[12] Anurag P. Jain, Mr. Vijay D. Katkar," Sentiments Analysis Of Twitter Data Using Data Mining", 2015 International Conference on Information Processing (ICIP), 978-1-4673-7758-4, vol. 19, pp. 432-438

[13] Manju Venugopalan, Deepa Gupta," Exploring Sentiment Analysis on Twitter Data", 2015, IEEE, 978-1-4673-7948-9 vol.6, pp. 34-39