International Journal of Engineering &Technology, 7 (3.12) (2018) 722-725

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



Categorisation of Tweets Using Ensemble Classification Methods

S. Mohanavalli[1], S. Karthika[2], Srividya[3], K.R.Uthayan[4], N. Sandya[5]

[1,2,3,4,5]Department of Information Technology, SSN College of Engineering, Kalavakkam
**Corresponding Author Email: [1]mohanas@ssn.edu.in, [2]skarthika@ssn.edu.in*
[3]srividhyav@ssn.edu.in, [4]uthayankr@ssn.edu.in, [5]sandyan14097@it.ssn.edu.in

Abstract

Twitter is a micro-blogging site that facilitates users to exchange short messages. Twitter is predominantly used in fields like business, healthcare, education and nation security. Twitter is being used by a large number of users for updating real time information and sentiment expression. The objective of this paper is to automate the classification of tweets into particular category using various machine learning algorithms like naïve bayes, SVM, and linear regression model. The proposed ensemble model aims to improve performance metrics of these algorithms. A comparative study of the algorithms used for tweet classification is done and results are discussed in the paper.

Keywords: *Categorization, Twitter Analysis, Liblinear, Naïve bayes, SVM.*

1. Introduction

Social networking service has become a powerful tool for expressing people's views with other users around the world and also to connect with those who share similar or related personal interests, career interests or real-life connections. Some Social networks like Twitter, Facebook, etc. involve a large amount of information, opinions [13] on different aspects of their everyday life which includes reviews, emotions and feelings.

Twitter in recent times has emerged as popular micro blogging site and a medium for real-time information [12].Messages posted in twitter is known as tweets. These messages limit is up to 280 characters. As Twitter is very much popular site in all around the globe, huge amount of population in the world uses Twitter and generate millions of tweets each day. In order to gain opinion of speaker behind certain topics sentiment analysis is been done. This analysis mainly classifies "tweets" into positive, negative and neutral sentiment[8][15].

Twitter streams are un-categorized and noisy data. If a user wants to search tweets about a particular topic then the categorization of tweets must be done. Insufficient data in training set, Twitter abbreviations, Poor spelling and poor grammar, incomplete sentences are some of the problems. In order to address the above problems and to get useful and meaningful information from the vast amounts of tweets, tweet data needs to be preprocessed

To classify trending topics into particular category like: arts, books, health, music, politics, religion, science, sports common techniques used are Bag-of-Words technique, social network information, and network-based classification. Various machine learning algorithms [9] like Naive Bayes, SVM, random forest and Maximum Entropy are also used for categorization of tweets.

This paper mainly focuses on classification of tweets into particular category by improving the performance metrics of various machine learning algorithms and a comparison between machine learning algorithms and liblinear classifier is also performed. For the classification of tweets, some of the features proposed in past literature are used.

The rest of the paper is grouped as follows. In section 2 literature works relevant to this paper is discussed. In section 3 details about the model is given .In section 4 experiments are presented and results are discussed. Conclusion and future works of research is given in section 5.

2. Literature Survey

Many works like twitter sentiment analysis, twitter trending topic detection, twitter sentiment classification have been done. Some of the works have been discussed Wilson [2], used hash tags to collect tweets and has proved to train classifiers better when compared to J.Read [1] who used emoticons to collect tweets .Pak and Paroubek [3] suggested a method to collect tweets from twitter API and proved that classifier using bigram features has given highest classification accuracy.

In the research of Cambria [4], they used lexical-based approach for sentiment classification. Major drawback of this technique is poor recognition of tweets when contrary tweets were present. Tara [5] used naïve bayes for categorization of tweets and applied Map-Reduce paradigm to existing Naïve Bayes algorithm to perform a better categorization. Stavrianou [6] proposed automated tweet classification of political tweets using decomposition of hash tags to overcome the drawbacks of Stanford NLP techniques. Tapan [7] purposed technique of EFWS (Effective Word Score) of tweets to improve sentiment analysis along with machine learning algorithm. Authors [9]-[11][16] compares classification performances of machine learning algorithms namely SVM, Decision tree and naïve bayes and concluded that SVM approach works best.

Yun [17] used ensemble sentiment classification based on majority vote principle of ensemble classification and results have shown that ensemble approach improves the accuracy in twitter sentiment classification. Cui [18] has analyzed the twitter data by combining

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

ensemble method (LLDA, Naive Bayes) classifier outputs and including the individual probability distributions with a cloud taxonomy service as an additional learning layer. Xia [19] used two types of feature sets: Part-of-speech information and word relation and three base classifier: Naïve Bayes, Maximum Entropy and Support vector machine for Sentiment classification.

3. Proposed Framework

The proposed system collects data from the twitterAPI and preprocesses data to remove stop words, tokens, etc. The training data is developed followed by categorization of tweets into particular category. Then performance of these classifiers is **analyzed. The proposed system has the process flow as shown in Figure 1.**

The system is composed of the following modules

A) Tweets extraction
B) Tweets Preprocessing
C) Training data and testing data
D) Performance analysis of classifiers

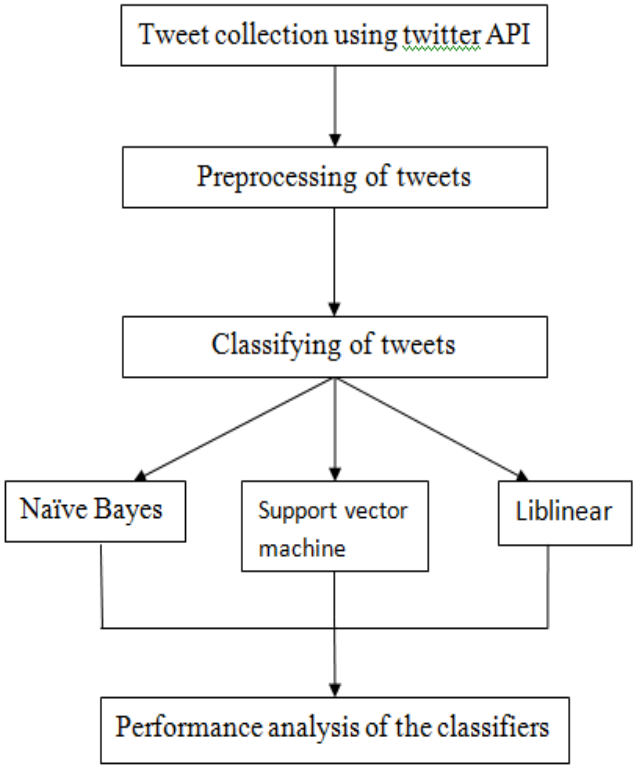

Fig. 1: Proposed Framework

A. Tweet Extraction

Tweet extraction is to fetch the tweets using Twitter APIv 1.1. [3].Based on the design and access method APIs are generally of two types:

• REST API is based on the pull strategy for data retrieval.
• Streaming APIs provides a continuous flow of public information from Twitter.

Tweets were collected using various hash-tags like #movies, #books, #science, #sports, #politics, #law #environment etc.

B. Tweet Preprocessing

The process of cleaning and removal of noisy data from the tweets is known as preprocessing. Reducing the noise in the text helps to refine the performance of the classification process [22].

Tokenization: Tokenization is a process of breaking into pieces called tokens and removing certain characters such as punctuation marks, URLs, hashed words, special characters etc. Alphanumeric words and words starting with numbers are removed.

Stop words removal: A stop-list are those words which do not carry any meaning. It is language specific. Some of the commonly used stop words are “a”, “of”, “the”, “I”, “it”, “you”, “and”, etc.

Stemming: It is a process for reducing derived words to their root form. e.g. “repeated”, “repeating”, “repeats” is reduced to the stem “repeat”.

Removal of Repeated tweets: If a tweet collected is being repeated more than once, occurrences of the tweet had been limited to single occurrences.

Removal of Internet Acronyms and Emoticons: Emoticons are facial expressions pictorially represented using punctuation and letters. In this process it removes internet acronyms like <3,☺ ☹ which are posted as part of the tweets.

C. Training and Testing

The key objective of the system is to classify the tweets into particular category .Training data must contain a class label. Around 1000 tweets were used as training data samples and varied range of data sample where used to perform testing using various Machine learning algorithms. Some of the machine learning algorithms that were used is naïve bayes and SVM.

Machine Learning Algorithms: Machine learning algorithms automatically trains and improve from experience without being explicitly programmed. Some of the algorithms are

Naïve Bayes: Naïve Bayes classifier uses the Bayes Theorem. It predicts probabilities for each class such as the probability that given record or data point belongs to a particular class [20]. Naïve bayes can be calculated using equation (1)

$$P\left(\frac{T}{C}\right)=\frac{P(T)\times P\left(\frac{C}{T}\right)}{P(C)} \quad (1)$$

T: Tweet
C: Category
P (T): Prior probability of hypothesis T
P(C): Prior probability of training data C
P (T/C): Probability of T given C
P(C/T): Probability of C given T

Support Vector Machine:

Support vector machines are supervised learning models that are used for classification of data. The SVM is to train a model and then assign new tweets into a particular category. This can be achieved by creating a linear partition of the feature space. A hyperplane is a line that splits the input variable space. Categorization can be done by placing tweets either above or below the separation plane [19].

If an object is placed above the plane, it is represented by: $\vec{b}\cdot\vec{x}+b_0>0$

If an object is placed below the plane, it is represented by: $\vec{b}\cdot\vec{x}+b_0<0$

LIBLINEAR:

Liblinear is a linear classifier with several features [14]. Liblinear supports

Linear regression and linear SVM .Main features of liblinear include Multi-class classification, Cross validation, Automatic parameter selection, Probability estimates, Weights for unbalanced data.

D. Performance Analysis

The overall performance of the system is computed in terms of accuracy, error rate, precision, recall and f-measure.

Accuracy: It is the ratio of correctly classified tweet samples (true positives plus true negatives) [23] to the total tweet samples. It is evaluated by the formula given in equation (2)

$$accuracy = \frac{accurately\,classified\,tweets}{totaltweet} \quad (2)$$

Precision: It is ratio of the total relevant tweets to the total retrieved tweets [21].Precision can be calculated using equation (3)

$$PRECISION = \frac{TP}{(TP + FP)} \quad (3)$$

TP: true positive, FP: false positive

Error Rate: It is percent of misclassified data out of the total validation data. That can also be computed using the following formula [17]. Error rate can be calculated using equation (4)

$$ERRORRATE = \frac{totalmisclassified}{totalinput} \quad (4)$$

Recall: Recall is the ratio of relevant tweets that have been retrieved to the total amount of relevant tweets [21]. Recall can be calculated using equation (5)

$$RECALL = \frac{TP}{(TP + FN)} \quad (5)$$

TP: true positive, FN: false negative

F-Measure: It is the harmonic mean of precision and recall [17]. F-measure can be calculated using equation (6)

$$F - MEASURE = \frac{2 \times PRECISION \times RECALL}{PRECISION + RECALL} \quad (6)$$

4. Experiment and Result

For collection of tweets twitter API was used. Around 4, 00,000 tweets were crawled using various hashtags like art, book, science, sports, cinema, politics etc. The resulting tweets were preprocessed and the feature vectors were obtained. Varied range (200, 400, 700, and 1000) was used as testing data sets. To compare the performance of Liblinear classifier against other machine learning algorithms like SVM and Naive Bayes various experiments were conducted and observations were noted. The system is tested for performance with these parameters: Precision, Recall, F –measure, accuracy and error rate.

Naive Bayes classifier considers that all the features are unrelated to each other. It predicts a probability that given tweet belongs to a particular category. The category with the highest probability is considered as the most likely category. **Figure 2** shows classification of tweets of varied range of data sets using naïve bayes classifier. In this classifier accuracy decreases as data set range increases. When 200 testing data is used accuracy is maximum and when 1000 testing data is used accuracy is least

SVM works by mapping tweets to a high-dimensional feature space so that tweets can be categorized, even when they are not linearly separable. The tweets are transformed in such a way that the separator could be drawn as a hyperplane. Thus data on sides of hyperplane is considered as class. **Figure 3** shows classification of tweets of varied range of data sets using SVM classifier. When 700 testing data is used accuracy is maximum.

Liblinear is targeted to primarily work with linear kernel .It uses one vs rest strategy. One-vs.-rest strategy involves training the samples of particular class as positive samples and all other samples as negatives. Thus base classifiers is required to produce a confidence value for its decision, rather than just a class label. It predicts class label using highest confidence score. **Figure 4** shows classification of tweets of varied range of data sets using liblinear classifier. When 700 testing data is used accuracy is maximum.

Figure 5 shows comparison of performance metrics of different classifiers. When overall performance of these classifiers is evaluated by taking the average of performance metrics of different series of data sets of these classifiers, liblinear has an average accuracy of 50% while naïve bayes and SVM had 35% and 33% respectively. Liblinear uses L2 regularization.L2 regularization will help you reduce the amplitude of model coefficients and leads to a more stable solution. Due to above reason Liblinear outperformed other classifiers.

5. Conclusion and Future Work

This paper has focused on classification of real-world tweets into particular category by using existing classifiers like naïve bayes, svm. In this paper tweets were collected and preprocessed. Processed tweets are classified using existing classifiers and liblinear classifier and performance metrics are compared. Results have proven that liblinear is more robust when compared with other existing classifiers as liblinear has used L2 regularization to give better results. The unique characteristic of paper is the usage of large data sets and significant improvements over the performance metrics of the existing machine learning algorithms and linear classifiers.

Our work can be extended in several dimensions. Feature extraction can be done in more enhanced way to improve the relevance and categorization of tweets. Fuzzy logic , deep learning and additional variants of the ensemble classifier such as Artificial Neural Networks, Decision Trees can be explored for more detailed classification.

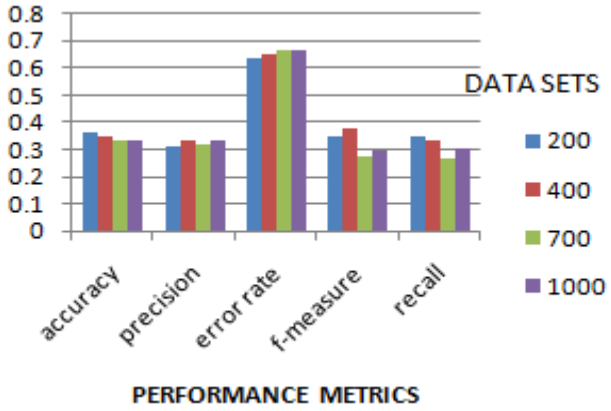

Fig. 2: (performance analysis of naïve bayes)

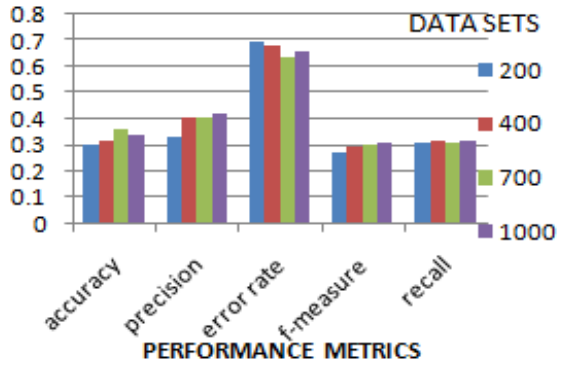

Fig. 3: (performance analysis of SVM)

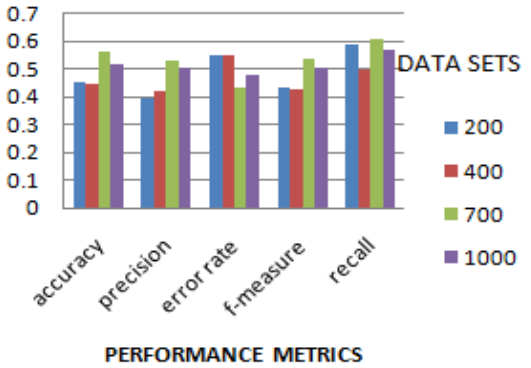

Fig. 4: (performance analysis of liblinear)

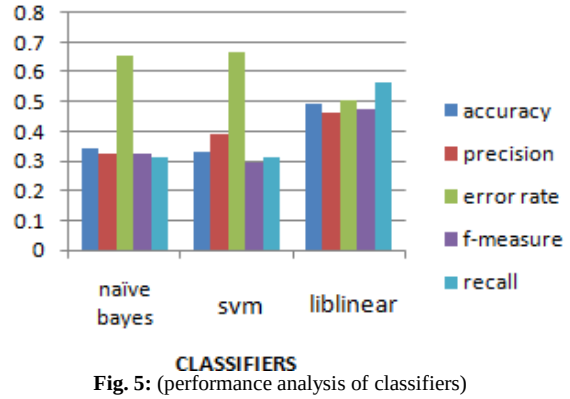

Fig. 5: (performance analysis of classifiers)

References

[1] Read, Jonathon "Using emoticons to reduce dependency in machine learning techniques for sentiment classification". Proceedings of the ACL Student Research Workshop. Association for Computational linguistics, pp 43-48 (2005).

[2] Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann. "Recognizing contextual polarity in phrase-level sentiment analysis." Proceedings of the conference on human language technology and empirical methods in natural language processing, Vol. 7,pp 347-354 (2005).

[3] Pak, Alexander, and Patrick Paroubek. "Twitter as a corpus for sentiment analysis and opinion mining." In LREc, Vol. 10. (2010).

[4] Cambria, Erik, Bjorn Schuller, Bing Liu, Haixun Wang, and Catherine Havasi. "Statistical approaches to concept-level sentiment analysis." IEEE Intelligent Systems , Vol. 28, pp 6-9 (2013).

[5] Tare, Mohit, Indrajit Gohokar, Jayant Sable, Devendra Paratwar, and Rakhi Wajgi. "Multi-class tweet categorization using map reduce paradigm." International Journal of Computer Trends and Technology (IJCTT), Vol. 9, p 78 (2014).

[6] Stavrianou, Anna, Caroline Brun, Tomi Silander, and Claude Roux. "NLP-based feature extraction for automated tweet classification." Interactions between Data Mining and Natural Language Processing , Vol. 145 (2014).

[7] Sahni, Tapan, Chinmay Chandak, Naveen Reddy Chedeti, and Manish Singh. "Efficient Twitter sentiment classification using subjective distant supervision." In Communication Systems and Networks (COMSNETS), 2017 9th International Conference, Vol. 1 pp 548-553 (2017).

[8] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. "Sentiment analysis of twitter data". In Proceedings of the workshop on languages in social media (pp. 30-38) (2011, June).

[9] Kurniawan, Dwi Aji, Sunu Wibirama, and Noor Akhmad Setiawan "Real-time traffic classification with Twitter data mining." Information Technology and Electrical Engineering (ICITEE), 2016 8th International Conference on, pp. 1-5, (2016). IEEE.

[10] Yi, Jeonghee, and Wayne Niblack (2005). "Sentiment mining in Web Fountain." Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on, pp 1073-1083, (2005). IEEE.

[11] Tan, Songbo and Jin Zhang An empirical study of sentiment analysis for Chinese documents". Expert System with application, Vol. 4, pp 2622-2629 (2008).".

[12] Sakaki , Takeshi and Makoto Okazaki and Yutaka Matsuo." Earthquake shakes Twitter users: real-time event detection by social sensors". Proceedings of the 19th international conference on World wide web, pp 851-860. (2010).

[13] Luo, Zhunchen and Miles Osborne and Ting Wang " An effective approach to tweets opinion retrieval". World Wide Web, Vol. 18, pp 545-566 (2015)..

[14] Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R. and Lin, C.J.," LIBLINEAR: A library for large linear classification". Journal of machine learning research, 9(Aug), vol.9, pp.1871-1874 (2008).

[15] Chung, Jessica Elan, and Eni Mustafaraj. " Can collective sentiment expressed on twitter predict political elections?". AAAI ,Vol. 11, pp 1770-1771 (2011).

[16] Deshwal, Ajay, and Sudhir Kumar Sharma"Twitter sentiment analysis using various classification algorithms." Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO), 2016 5th International Conference on, pp 251-257, (2016)..

[17] Wan, Yun, and Qigang Gao"An ensemble sentiment classification system of twitter data for airline services analysis." Data Mining Workshop (ICDMW), 2015 IEEE International Conference on, pp 1318-1325 (2015). .

[18] Cui, Renhao, Gagan Agrawal, Rajiv Ramnath, and Vinh Khuc"Ensemble of Heterogeneous Classifiers for Improving Automated Tweet Classification." Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on, pp 1045-1052, (2016)..

[19] Xia, Rui, Chengqing Zong, and Shoushan Li."Ensemble of feature sets and classification algorithms for sentiment classification." Information Sciences ,Vol.181, pp 1138-1152 (2011)..

[20] Melville, Prem, Wojciech Gryc, and Richard D. Lawrence "Sentiment analysis of blogs by combining lexical knowledge with text classification." In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 1275-1284, (2009)..

[21] Kanakaraj, M. and Guddeti, R.M.R.." NLP based sentiment analysis on Twitter data using ensemble classifiers". In Signal processing, communication and networking (ICSCN), 2015 3rd international conference on (pp. 1-5). (2015).

[22] Gull, Ratab, Umar Shoaib, Saba Rasheed, Washma Abid, and Beenish Zahoor"Pre Processing of Twitter's Data for Opinion Mining in Political Context." Procedia Computer Science ,Vol.96, pp 1560-1570 (2016). .

[23] Shrivastava, S. and Nair, P.S., "Mood prediction on tweets using classification algorithm". Int J Sci Res (IJSR), Vol4 (11), pp.295-299 (2015).