Microblogging Comments Classification

Swapnil Babaji Shinde
Department of Information Technology,
Pimpri Chinchwad College of Engineering Pune, India

Mohammad Muzammil Shaikh
Department of Information Technology,
Pimpri Chinchwad College of Engineering Pune, India

Sudeep Thepade, PhD
HOD, Department of Information Technology,
Pimpri Chinchwad College of Engineering Pune, India

ABSTRACT
Nowadays, microblogging sites like, Twitter, Pinterest is used by many people to share their sentiments. These comments can be classified and analyzed to find hidden patterns. The System needs to classify these comments into various classes which can be used to find the interest of users. These interests of users will be used for giving them personalized news and also for decision making in business. Twitter tweets having a limit of 140 characters. So, people share only important comments through tweets. Using text mining most important keywords can be found from tweets and classified accordingly in multiple classes.

General Terms
Information Retrieval, Classification

Keywords
Naïve Bayes Classification, Twitter Feeds Analysis, Text Mining, News Recommendation

1. INTRODUCTION
Microblogging services allow users to broadcast small messages over sites that can be accessed by other subscribers. Microblogging sites have a limit on message size. Most popular microblogging site Twitter limits message size up to 140 characters. This leads to the production of a large amount of text data that needs further analysis. Twitter generates big data from tweets posted by its users. These comments are in natural languages and having hashtags. So, we need to process this comments to remove unwanted data. There is a need of better text classification techniques to classify these microblogging comments to find the interest of users. There are various classification techniques but some of these having problems for text classification. Therefore in this paper, some analysis is done on various classifiers by using standard datasets.

2. LITERATURE SURVEY

2.1 Classification of microblogging comments
Nowadays, many people are using microblogging sites like Twitter, Pinterest, Tumblr. But Twitter is most popular than others according to statistics of many companies. According to Statista [5], Twitter has averaged monthly 319 million active users during the fourth quarter of 2016. These microblogging comments should be well classified in some definite classes. Using that we can find interested articles of users.

2.2 Multinomial Naïve Bayes
Multinomial Naïve Bayes is one of the Bayes family of classifiers. Bayes classifiers are based on Bayes probability theorem. Naive Bayes classifier is very popular for binary as well as multiclass classification. Naïve Bayes classifier used when the dataset is high dimensional. These classification algorithms use evidence which finds the probability of the given comment in the dataset.

$$P(c|d) = \frac{P(d|c)\,P(c)}{P(d)} \quad (1)$$

Here, P(c) is a prior which is the probability of specific class. P(d|c) is the probability of document for a specific class. P(d) is evidence i.e. probability of any document. And finally, P(c|d) is the probability of class for given features [12]. This algorithm is mostly used for multinomial distributed data. Multinomial naïve Bayes classifier uses frequency of words for text classification [11]. The data is represented as word vector counts.

2.3 Data Mining Classifiers
There are multiple classification algorithms used in data mining and every algorithm has specific features. Every family consists of several algorithms. Bayes family uses probabilistic approach for classification. It outputs a class which has a maximum probability for given microblogging comment. Naïve Bayes classifier is best for applications like spam filtering, diagnosis of diseases and news classification. Tree-based models create a tree in which intermediate nodes represents features and leaf nodes represents output classes. Lazy learning classifier contains famous algorithm KNN which completely stores training dataset in memory. It computes the distance between a test comment and all other comments and selects k-nearest comments to find output class. In text classification, each keyword represents the feature. Therefore, tree-based models have high dimensional attributes which build complex tree.

3. APPLICATIONS
Microblogging comments classification helps us to categorize people opinions into some finite classes. We can use tweets to find opinions of people and recent events or trends in society. In elections, we can track sentiments of people to predict exit polls. We can build a profile of users which can be used to recommend users with a set of products. Using comment classification we will get set of topics in which user is interested and we can provide only personalized news to the user. In this application, we need comment classification algorithms like naïve Bayes, tree models and Meta-algorithms. Using comments on twitter, the system can detect any criminal activity in the entire globe.

Sentiment analysis of social media data i.e. Twitter helps businesses in decision making. The system can classify tweets of users into positive and negative classes about any event in business [10]. This will lead to effective strategic planning in business.

4. PROPOSED METHODS

The comments data need to be collected from microblogging site and then it should be classified using an appropriate methodology to find categories of test comments.

4.1 Data Collection

In this system, we have collected required Twitter data using Twitter4J which is the unofficial Java library to collect tweets from user's timeline. For this, we need to register to Twitter developer console and collect access keys. Twitter4J having built-in OAuth (Open Authentication) support and support Java version 5 or later [3]. The query functionality of Twitter4J provides access to public tweets of the string given to query.

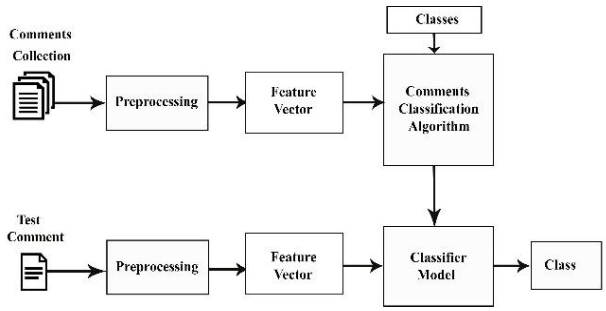

Figure 1: Block diagram of comments classification system

4.2 Preprocessing

Collected comments are in the raw form which consists of stop words like articles, conjunctions, and prepositions. Preprocessing includes removing stop words, stemming text and then weighting it [2]. In preprocessing standard stop lists can be used like SEO Stop List [7].

4.3 Comments Classification Algorithms

In filtering feature vector get generated which can be further given as input to the algorithm along with classes. For text classification, there are various algorithms from different classification families. For text classification, Bayes family works better than other families due to evidence. There is a need of training dataset with labels assigned to each comment.

4.4 Classifier Model

The classifier model is built by using classification algorithm. It can be trained by some standard datasets or user build datasets. In this system classifier model gets trained by using tweets dataset with some definite categories. After complete training test comment or tweet applied to the model to get output category for comment.

5. EXPERIMENT ENVIRONMENT

5.1 Experiment Platform

For experimentation of classification, we have used NetBeans 8.0 IDE and multiple Weka classification techniques. Java provides a wide range of services to process data using API's. Weka API provides an easier building of classifiers on various image and text datasets. It has the ability to provide feature extraction, transformation, vector quantization, image classification and much more [8] [9].

5.2 Test bed/ Dataset

In this system for evaluating many algorithms we have used two standard datasets as follows:

- 20 Newsgroup - Comprises of 18000 newsgroups posts on 20 topics split into two subsets for training and testing based on messages posted before and after a specific date. 20 Newsgroup has 18846 documents, with 11314 (60%) training and 7532 (40%) testing [4].

- Reuters 21578 (R8 and R52) – Documents in Reuters-21578 first appeared on Reuters newswire in 1987. The Reuters-21578 dataset is a standard and widely distributed collection of hand-labeled articles pulled from Reuters the magazine. It's a very well-known benchmark which has been a considerable aid in the development of algorithms for the task of text categorization and contains 21578 documents in 135 categories. There are 5946 training documents and 2347 testing documents [6].

6. RESULTS AND DISCUSSION

6.1 Generic comments

For finding results of comments classification we have used two standard datasets 20 Newsgroup and Reuters-21578. These datasets are divided into training and test instances. After building any text classifier model there is a need to test it over a test dataset in order to calculate its accuracy. These two standard datasets having some unique features which enable us to evaluate classifier and analyze results.

6.2 Proposed method with 20 Newsgroup

The 20 Newsgroup dataset consist of 20 classes like rec.sport.hockey, rec.sport.baseball, sci.med, sci.space which are relevant to each other [4]. The word vectors that are generated here contains some similarity. Therefore, finding optimum classifier is possible.

For calculating results using 20 Newsgroup dataset mainly two variations are used:

1. Five Classes with 300 train and 150 test instances

In this variation, five classes from 20 newsgroup dataset are selected with 300 instances of each. Using these 1500 training instances model was built and tested using 150 instances of each class.

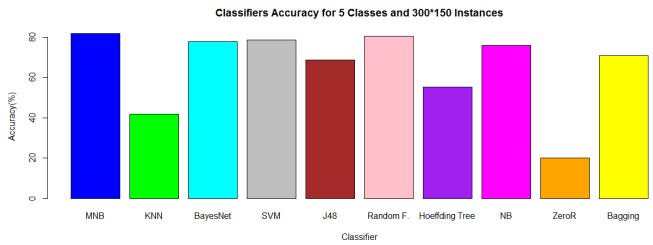

Figure 2: 20 Newsgroup Results for 5 Classes with 300 training and 150 testing in instances.

As a result of 5 classes Multinomial naïve Bayes working good along with Random Forest, Bayes Network, and SVM. Multinomial Naïve Bayes having accuracy around 81 %.

2. Five Classes with 200 train and 50 test instances

In this variation, training and testing instances are minimized by keeping five classes as it is. By decrementing a number of instances, accuracy of a classifier got reduced.

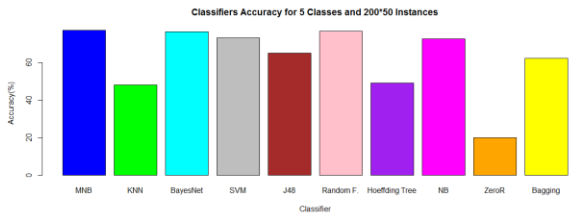

Figure 3: 20 Newsgroup Results for 5 Classes with 200 training and 50 testing in instances.

As shown in figure 4, Bayes family classifiers work better for comment classification. Multinomial NB classifier having accuracy around 77 % and Random Forest, SVM having accuracy around 74 % to 76 %.

3. Sixteen Classes with 300 train and 150 test instances

In this classification number of classes increased up to 16 and instances were kept same to 300 by 150. By increasing number of classes, many classifiers failed to provide accuracy like five class classification.

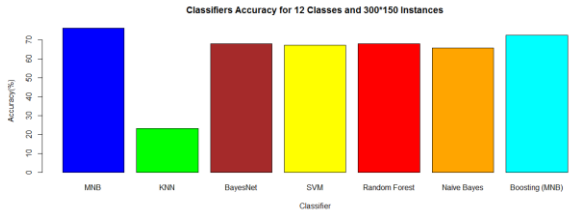

Figure 4: 20 Newsgroup Results for 16 Classes with 300 training and 150 testing in instances.

According to these results, Multinomial NB provides the stable model with about 76 % accuracy. But, SVM, Random Forest does not perform well. They have accuracy around 67%.

4. Sixteen Classes with 200 train and 50 test instances

In this classification variation, a number of classes kept 16 and number of instances reduced. Due to less number of instances accuracy get degraded. But still, some classifiers in Bayes family works better. For comments classification, if the number of instances reduced then output feature vectors having more chances of similar words which degrade the performance of the model.

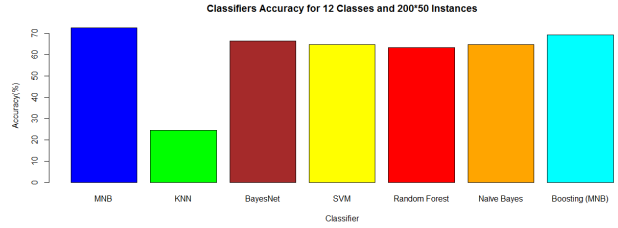

Figure 5: 20 Newsgroup Results for 16 Classes with 200 training and 50 testing in instances.

Multinomial NB having accuracy around 72 % and SVM, Bayes Network, Naïve Bayes provides around 74 % accuracy.

6.3 Proposed method with Reuters Dataset

Reuters dataset provides two kinds of datasets R8 and R52 having 8 and 52 classes simultaneously. The main feature of the Reuter dataset is that it has a variable number of instances in classes [6]. It contains following two datasets:

1. R8 Classification

The R8 dataset contains eight classes e.g. crude, earn, grain etc. Using R8 documents classifiers can be tested well for variable text instances.

Table 1. R8 Classification Results

Sr. No.	Classifier	Accuracy
1	Multinomial NB	96.2
2	SVM	95.2
3	Random Forest	93.05
4	Naïve Bayes	92.87
5	Bayes Network	91.82
6	J48	91.45
7	KNN	87.8

For R8 text classification, many classifiers work well because R8 has dissimilar classes with very small relevance to each other.

2. R52 Classification

The R52 contains 52 text classes with a variable number of instances [6]. Using Weka following results are calculated on R52 dataset:

Table 2. R52 Classification Results

Sr. No.	Classifier	Accuracy
1	SVM	90.14
2	Bayes Net	87.53
3	Multinomial NB	86.64
4	KNN	79.32

The results of R52 classification are as given in above table shows that when there are the distinct large number of classes with less similarity, then SVM performs well than Bayes family classifiers. But still, Bayes Network and Multinomial Naïve Bayes works well.

6.4 Overall Observations

All the results calculated for two datasets are important for selection of best classifier for text classification. To improve the performance of classifier not only algorithm selection but also appropriate processing of dataset is important. In 20 Newsgroup dataset classes are very similar to each other and therefore feature vectors are very similar. Because of this, the accuracy of classifier degrades in many cases as discussed above. For 20 Newsgroup dataset, Multinomial NB classifier performs very well.

In Reuters R8 and R52 dataset, classes are different from each other and number of training and testing instances are variable. Therefore, the accuracy of classifiers improves due to the difference in word vectors generated by classifiers. Due to dissimilarity in classes Support Vector Machine, Bayes Network, KNN are also performing well classification as compared to 20 Newsgroup results. In both cases, Multinomial Naïve Bayes has given better accuracy than others.

7. CONCLUSION

Microblogging comments were used in business analytics and social relationships finding. From results of testbeds, we came to know that Multinomial Naïve Bayes performs better for text classification. We can improve its performance by well preprocessing of comments and also ensembling of multiple better working classifiers. Now tweets analysis can be used to improve security in defense of the country. In a business, the system can detect the polarity of any launch by analyzing tweets of people. The frequency of tweets is increasing day by day, therefore, real-time analysis of tweets is required to deal with recent trends in the entire globe.

8. REFERENCES

[1] Nirmal Jonnalagedda and Susan Gauch, Personalized News Recommendation using Twitter, IEEE, WIC and ACM conference, 2013.

[2] Shokoufeh salem minab and mehrdad jalali, Online Analyzing of Texts in Social Network of Twitter. ICTCK, 2014.

[3] Konstantinos Semertzidis, "Crawling Twitter Data"

[4] The 20 Newsgroup Dataset Classes and Instances information in detail: http://qwone.com/~jason/20Newsgroups/

[5] Statista Twitter statistics: https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

[6] R52 and R8 datasets : http://csmining.org/index.php/r52-and-r8-of-reuters-21578.html

[7] SEO Stop List : https://www.link-assistant.com/seo-stop-words.html

[8] Sudeep Thepade, Dimple Parekh, Unnati Thapar, Vandana Tiwari, LBG Algorithm for Fingerprint Classification, IJAET, Nov. 2012. ISSN: 2231-1963.

[9] Dr. Sudeep Thepade, Rik Das, Saurav Ghosh, Content Based Image Classification with Thepade's Static and Dynamic Ternary Block Truncation Coding, IJER, Volume No. 4, Issue No. 1, pp: 13-17, Jan. 2017. ISSN: 2319-6890.

[10] Sagar Bhuta, Avit Doshi, Uchit Doshi, Meera Narvekar, A Review of techniques for Sentiment analysis of Twitter Data, ICICT, 2014.

[11] Andrew McCallum, Kamal Nigam, A Comparison of Event Models for Naive Bayes Text Classification.

[12] Naïve Bayes Classification with formulae – About Learning with some examples URL: https://amitranga.wordpress.com/machine-learning/naive-bayes-classification/