TWITTER NEWS STRATIFICATION USING RANDOM FOREST

[1]PRASAD J. KOYANDE, [2]KAVITA P. SHIRSAT

[1,2]Computer Engineering, Vidyalankar Institute of Technology, Mumbai, India
Email: [1]prasad.koyande@vpt.edu.in, [2]kavita.shirsat@vpt.edu.in

Abstract: With the popularity of Social Networks, mostly news providers used to share their news in various social networking sites and web blogs. In India, many news groups share their news on Twitter micro blogging service provider. These data carries valuable information relevant to social research areas. Thus, the idea is to categorize the news into different groups so the news groups in India are identified. News groups are selected on their popularity to extract the short messages from Twitter Micro Blog. Short message extracted from Twitter was classified into 12 major groups. Machine learning techniques were used to train the data. In order to create the instances words from each short message were consider and bag-of-words approach was used to create feature vector. The data was trained using Random Forest machine learning techniques. Random forest is a best ensemble learning method, which is consist of multiple decision trees built on random inputs and separating nodes on a random subset of features. Because of its good classification and generalization ability, random forest is preferred in various domains. Large amount of feature will be collected for current research. The performance will speak the efficacious of the system.

Keywords: Random Forest, Web Mining, Text classification.

I. INTRODUCTION

With the development of web blogs and SocialNetworks, many organizations and News providersused to share their news headlines in various websites and web blogs. Now-a-days in India, many news groupsshare their newsheadlines as short messages in microbloggingservices such as Twitter. Once these short messages get processed, it carries out significant amount ofinformation which is relevant for many social research areas. Theidea is toclassify news short messages into different groups so that the user identifies the most popular newsgroup in India. This will help in availing information regarding development, war, education etc. thereby understanding the current state of the country.

There are several news portals currentlyavailable to retrieve short messages. Twitter microblog was opted due to 4 reasons [1].

1) Microblogging platforms are used by diversepersonalities to express their opinion about contrastingtopics, thus it is a valuable source of people'sspeculation.
2) Twitter contains an enormous number of textposts and it grows every day. The collectedcorpus can be arbitrarily large.
3) Twitter's audience varies from regular users tocelebrities, company representatives, politicians and even the country's president. Therefore, it ispossible to collect text posts of users fromdifferent social and significantgroups.
4) Twitter's audience is represented by users frommany countries.

The research was conductedand several active Twitter news groups such as'abpnewstv', 'IBN7', 'ibnlive', 'ZoomTV', 'ndtv', 'zeenews', 'timesofindia', 'timesnow',' economictimes',' starsportsindia', 'dna', and 'httweets' were chosen to extract the data.Twitter API provides the ability of extracting 180short messages with other necessary details in XMLformat. The short messages will be classified by thesystem into 12 groups: war-terrorist-crime, economy-business,health, sports, development-government,politics, accident, entertain, disaster-climate,education, society and international. These groupswere designated in order to cover the main areas of general news provider.

With the development of machine learningtechniques [2], now-a-days, many researchers tend touse machine learning techniques in text classification [1]. There are 2 types of machine learningtechniques as supervised learning [3] (the learningdata will be provided by the developer) andunsupervised learning [3] (the method will learn aclustering procedure by observing the distanceamong data). For the present study, supervisedlearning techniques will be used as the 12 groups donot change regularly.

In order to classify short messages usingmachine learning techniques, a proper set of featuresare required to extract from the short messages. Thebag-of-words approach [1], was used to extractfeatures from the short messages. The frequency of each word had been used as data. As there are largeamount of words in different short messages, usingall data will cause to increase the dimension. Thus,the common words are needed to identify and removefrom the dataset. Zipfs low [5] states that the utmost common word in a human language text occurswith a frequency inversely proportional to n. Rarely occurred words do not carrysufficient information but the noise. Thus, lowfrequent words and high frequent words will beremoved from the data in order to reduce thedimension up to certain level.

Once created the dataset, it is important to find asuitable classification method in order to classify

theshort messages. Random Forest [6] was used to classify the data as it is capable of dealingwith high dimensional dataset [1], [5]. The system provides an accuracy of identifying news whichbelongs to group accident with 100%,development-government with 100%, disaster-climatewith 100%, economy-business with 100%,education with 96.4%, entertainment with 100%, healthwith 95.2%, international with 80%, politics with 100%, society with 78.40%, sports with 100 % andwar-terrorist-crime with 100%.The next section will brief out the method of data gathering and feature selection. Section 3 will throw a light on the approach of data training and Section 4will briefout the evaluation criteria of the system. The generaldiscussion will be brief in section 5.

II. DATA GATHERING AND FEATURE SELECTION

The classification will be applied into the shortmessages-news of Twitter microblog. Thus, twittershort messages are needed to be collected. Fortwitter, there was a character limit such that thelength of one short message was limited to 140characters [7]. Thus the user bound to provide thenews by using few amount of words. This caused tolimit the words of the short messages into key words.Twitter API provides the ability of retrieving suchshort messages for a given user in XML file format.

Each XML file could carry out 180 short messages at once.Once gathered the data, the features are need toextract from the short messages. These features arerequired to learn the patterns amount the groups. Thewords are used as features. Thus the bag-of-wordsapproach [1], [5] was used to extract the features. This will pool the words from all short messages andwill create a document vector, containing words.Some researchers had used n-gram instead of words [8]. However, n-gram method cause to increase thedimension of the dataset, as it uses unigram, bigram,trigram which make complex for the system torecognize the pattern [8]. Thus, the words are chosenas the features.

In order to create the dataset, the frequencies ofwords were used. All words of the documents do notcarry out useful information. To avoidusing very low frequent words which do not carry outany valuable information regarding the group. Zipfslow [5] states that the nth most common word in ahuman language text occurs with a frequencyinversely proportional to n. Thus, the common wordswere removed from the dataset by removing highfrequency data. Therefore, a lower cut off value and upper cut off value were required to choose in orderto obtain best set of features. The frequencies ofselected data range were shown in figure 1. Thevalues X and Y were chosen as the frequency limit(lower cut off value and upper cut off value) whichmaximize the effectiveness.

Thus, unimportant features were removed fromthe dataset. This caused to reduce the dimensionwhich makes the training method more effective.

FIGURE I: RANGE OF FREQUENCIES OF CHOSEN DATA

In machine learning [2], [3], there are basicallytwo types of learning methods. Supervised learning [2] and unsupervised learning [2]. In supervisedlearning, the developer provides learning data to thesystem in order to train the system. In unsupervisedlearning, the system itself learns patterns from thedata. For the current situation, as the group are predefined and do not change regularly, supervisedlearning method is more applicable [3].

III. DATA TRAINING

In machine learning [2], [3], there are basically two types of learning methods. Supervised learning [2] and unsupervised learning [2]. In supervised learning, the developer provides learning data to the system in order to train the system. In unsupervised learning, the system itself learns patterns from the data. For the current situation, as the group are pre-defined and do not change regularly, supervised learning method is more applicable [3].

Ensemble learning [6] refers to the algorithms that producecollections or ensembles of classifiers which learn to classifyby training individual learners and fusing their predictions.Growing an ensemble of trees and getting them vote for themost popular class has provided a good enhancement in theaccuracy of classification. Often, random vectors are built thatcontrol the growth of each tree in the ensemble.The ensemble learning methods can be divided into twomain groups: bagging and boosting. In bagging, models arefit in parallel where successive trees do not depend on previoustrees. Each tree is independently built using bootstrapsample of the dataset. A majority vote determines prediction.In boosting, models are fit sequentially where successive trees assign additional weight to those observations poorly predictedby previous model. A weighted vote specifies prediction.

A random forest [6] adds an additional degree of randomness to bagging. Although each tree is constructed using a differentbootstrap sample of the dataset, the method by which the classification trees ate built is improved. A random forest predictoris an ensemble of individual classification tree predictors. Foreach observation, each individual tree votes for one class andthe forest predicts the class that has the plurality of votes. The user has to specify the number of randomly selected variables(m_{try}) to be searched through for the best split at each node.

Whilst a node is split using the best split among all variablesin standard trees, in a random forest the node is split using thebest among a subset of predictors randomly chosen at that node. The largest tree possible is grown and is not pruned. The rootnode of each tree in the forest contains a bootstrap samplefrom the original data as the training set. The observationsthat are not in the training set, are referred to as "out-of-bag"observations.

Since an individual tree is unpruned, the terminal nodes cancontain only a small number of observations. The training dataare run down each tree. If observations iand j both end up in the same terminal node, the similarity between i and j isincreased by one. At the end of the forest construction, thesimilarities are symmetrized and divided by the number oftrees. The similarity between an observation and itself is setto one. The similarities between objects form a matrix whichis symmetric, and each entry lies in the unit interval [0,1].Breiman defines the random forest as [6]:

> A random forest is a classifier consisting of a collectionof tree-structured classifiers {h (x, e k), k = 1, ...} where {e k} are independent identically distributedrandom vectors and each tree casts a unitvote for the most popular class at input x.

A summary of the random forest algorithm for classification is given below [9]:

- Draw n_{tree} bootstrap samples from the original data.
- For each of the bootstrap samples, grow an unprunedclassification tree, with the following modification: ateach node, rather than choosing the best split among allpredictors, randomly sample m_{try} of the predictors andchoose the best split from among those variables. Baggingcan be thought of as the special case of the random forestobtained when m_{try} = p, the number of predictors.
- Predict new data by aggregating the predictions of then$_{tree}$ trees, i.e., majority votes for classification, averagefor regression.

The generalization error of a forest of tree classifiers dependson the strength of the individual trees in the forest and thecorrelation between them. Using a random selection of features to split each node yields error rates that compare to AdaBoost[9]. An estimate of the error rate can be obtained, based on thetraining data, by the following [9]:

- At each bootstrap iteration, predict the data that is not inthe bootstrap sample, called "out-of-bag" data, using thetree which is grown with the bootstrap sample.
- Aggregate the out-of-bag predictions. On the average, eachdata point would be out-of-bag around 36% of the times,so aggregate these predictions. Calculate the error rate, andcall it the "out-of-bag" estimate of error rate.

The random forest performs well compared to several otherpopular classifiers. In addition, it is userfriendlyas it has only two parameters: (i) the number ofvariables in the random subset at each node, and (ii) the numberof trees in the forest. The random forest is not usually verysensitive to the values of these parameters.

To create the training data and testing data, each short message was classified to a group manually. One short message might be belong into several groups. Therefore, each category was considered as a separate binary classification problem [4]. The training process was developed in order to recognize whether the selected short message belong to the group A short messages will be classified manually as "Group A" or "other". 90 % data was used to train the system and 10 % were used to test the system [4]. In order to train the 12 groups, there will be 12 training data sets and testing will results 12 tables, each table describing their performance as in table 1.

TABLE 1

	Expected Group	
Observed classes	True positive (tp)	False positive (fp)
	False negative (fn)	True negative (tn)

IV. EVALUATION

The evaluation was carried out in order to measure the effectiveness. Effectiveness is purely a measure of the ability of the system to satisfy the user in terms of the relevance of short messages retrieved [12]. It is assumed that the more effective the system, the more it will satisfy the user [12]. The effectiveness of the retrieval system was measuredusing precision and recall values [12].Precision is the fraction of retrieved shortmessages that are relevant. Recall is the fraction ofrelevant short messages that are retrieved [13]. As thesystem results the performance as in Table 1, theprecision can be calculated using Equation 1 andrecall could be calculated using Equation 2.

$$Precision = \frac{t_p}{t_p + f_p} \quad (1)$$

$$Recall = \frac{t_p}{t_p + f_n} \quad (2)$$

In order to measure the performance, theperformance will be depending on the biasness of thetraining data. In order to avoid the biasness, crossvalidation [14] was applied to the testing process.Koveri [14] suggest that the best number of fold is 10. Thus, 90 % data were randomly chosen to trainthe data. The system was tested 10 times per groupand the average precision and

recall values werecalculated [12]. The results were briefed in table 2.

These measures were used to figure out bestfrequency limit for the feature selection. Thus, it isimportant to calculate a single measurement insteadof 2 values [14]. Many alternative methods wereproposed over the years and Harmonic mean [13] hadidentified as the best single value summaries [13].The harmonic mean (F-measure) was given inEquation 3 and weighted harmonic mean was givenin Equation 4. Values of P less than 1 emphasize theprecision whereas values of P greater than 1emphasizes recall [13].

$$F = \frac{2 \times P \times R}{P+R} \qquad (3)$$

$$F_\beta = \frac{(\beta^2+1) \times P \times R}{\beta^2 \times P+R} \qquad (4)$$

For the current situation, the main focus will beon the fraction of retrieved short messages that arerelevant. Thus $F_{0.5}$ will be used to select the bestfrequency range and to measure the effectiveness ofthe system. Table 3 provides the values obtain for F measure and $F_{0.5}$.

TABLE 2: PRECISION AND RECALL VALUES

Group Name	Avg. Precision	Avg. Recall
Accident	0.824	1
Development-Government	1	1
Climate-Disaster	0.733	1
Economy-Business	0.955	1
Education	0.964	0.964
Entertainment	0.846	1
Health	0.714	0.952
International	1	0.8
Politics	0.8	1
Society	0.965	0.784
Sports	0.857	1
War-Terrorist-Crime	1	1

TABLE 3: RESULTS OF F-MEASURE AND F_β-MEASURE

Group Name	F-MEASURE	F_β-MEASURE
Accident	0.903	1
Development-Government	1	1
Climate-Disaster	0.846	1
Economy-Business	0.977	1
Education	0.964	1
Entertainment	0.917	1
Health	0.816	0.994
International	0.889	0.999
Politics	0.889	1
Society	0.865	0.961
Sports	0.923	1
War-Terrorist-Crime	1	1

With the results obtain from Table 3 it is clearthat Random Forest provides good results for most of groups.

V. DISCUSSION

A system which is able to classify newsheadlines will be useful in various social researches.With the development of web technologies, peopleget involved in many social networks and web blogs.Twitter is a micro blog which allows many famousnews suppliers to publish their news headlines.Twitter API supports user to retrieve available shortmessages. These retrieved files will be in XML fileformat and each file could retrieve maximum numberof 200 short messages per once.

In order to applymachine learning, a properfeature set was required. The feature set was createdby pooling the words and creating a document vector.Thisapproach was named as bag-of-words approach. The frequency of each word was chosen as data. When considering all words together, it create hugedimension of instances. In order to reduce the dimension, a lower cut off frequency and an upper cut off frequency value were chosen. The value was chosen as the frequency range which maximizes the accuracy.

There are 12 groups defined and each group was treated as separate binary classification problem as same short message could be belong into several groups. System was trained using Random Forest.The effectiveness of the training system can be measure using recall and precision values. Precision is the probability of retrieving relevant short messages. Recall is the probability of the relevancy of retrieved short messages. The harmonic measure (F-measure) was used to obtain a single value for recall and precision. The weighted F -measure ($F_\beta$measure) was used as precision was needed to be emphasizing in current situation. The system provides best results for Accident, Development-Government, Climate-Disaster, Entertainment, Health, Education, Sports, War-Terrorist-Crime, Politics and Economy-business groups. It provides reasonable effectiveness more than 96%.

REFERENCES

[1] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion," in Analysis, 2010.

[2] N. J. Nilsson, INTRODUCTION TO MACHINE LEARNING. 1998.

[3] J. K. M. Han, Data Mining : Concepts and Techniques, 2nded. 2006.

[4] InoshikaDilrukshi, Kasun De Zoysa, Amitha Caldera. “Twitter News Classification Using SVM”, Computer Science &Education (ICCSE 2013 IEEE).

[5] K. G. Zipf, Human Behaviour and the Principle of Least Effort. Oxford, England: Addison-Wesley, 1949.

[6] L. Breiman, “Random Forests,” Machine Learning, Vol. 40.No. 1.2001.

[7] (2012, Apr.) Counting Characters.[Online]. https://dev.twitter.com/docs/counting-characters

[8] W. B. Cavnar and T. .I.M., "N-Gram-Based TextCategorization," in Proceedings of SDAIR-94, 3rd AnnualSymposium on Document Analysis and Information Retrieval, pp. 161-175, 1994.

[9] AndyLiawand Matthew Wiener, “Classification and Regression by Random Forest,” R News,Vol. 2/3, December 2002.

[10] Yoav Freund and Robert E. Schapire, "A Short Introduction to Boosting", Journal of Japanese Society for Artificial Intelligence, 14(5):771-780, September, 1999.

[11] A.Z. Kouzani, S. Nahavandi, K. Khoshmanesh,"Face Classification by a Random Forest",TENCON 2007 - 2007 IEEE Region 10 Conference,Oct. 30 2007-Nov. 2 2007.

[12] C. 1 Rijsbergen, Information Retrieval, 2nd ed. London: Butterworths, 1979.

[13] C. D. Manning, P. Raghaban, and H. Schutze, Introduction to Information Retrieval.Cambridge University Press, 2008.

[14] Y. Baeza and B. R. Neto, Modern Information Retrieval.Boston, 1999.

★★★