KNN BASED TWITTER NEWSCASTER

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

[1,2]Computer Engineering, Mumbai University,
Vidyalankar Institute of Technology,
VIT Mumbai, India
E-mail: [1]prasad.koyande@vpt.edu.in, [2]kavita.shirsat@vit.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 KNN (K – Nearest Neighbor) machine learning techniques. The KNN is a typical learning algorithm based on analogy, so each category has a certain amount of the training samples which helps representatives guarantee the accuracy of classification. Large amount of feature will be collected for current research.The performance will speak the efficacious of the system.

Keywords- KNN, Web Mining, Text Classification

I. INTRODUCTION

With the development of web blogs and Social Networks, many organizations and News providers used to share their news headlines in various websites and web blogs. Now-a-days in India, many news groups share their news headlines as short messages in micro blogging services such as Twitter. Once these short messages get processed, it carries out significant amount of information which is relevant for many social research areas. The idea is to classify 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 currently available to retrieve short messages. Twitter micro blog was opted due to 4 reasons [1].

1) Micro blogging platforms are used by diverse personalities to express their opinion about contrasting topics, thus it is a valuable source of people's speculation.
2) Twitter contains an enormous number of text posts and it grows every day. The collected corpus can be arbitrarily large.
3) Twitter's audience varies from regular users to celebrities, company representatives, politicians and even the country's president. Therefore, it is possible to collect text posts of users from different social and significant groups.
4) Twitter's audience is represented by users from many countries.

The research was conducted and several active Twitter news groups such as 'ABP News Tv', 'IBN7', 'IBN live', 'ZOOMTV', 'NDTV', 'Zee news', 'Times of India', 'time snow',' economic times',' star sports India', 'dna', and 'httweets' were chosen to extract the data. Twitter API provides the ability of extracting 180short messages with other necessary details in XML format.

The short messages will be classified by the system into 12 groups: war-terrorist-crime, economy-business, health, sports, development-government, politics, accident, entertain, disaster-climate, education, society and international. These groups were designated in order to cover the main areas of general news provider. With the development of machine learning techniques [2], now-a-days, many researchers tend to use machine learning techniques in text classification [1]. There are 2 types of machine learning techniques as supervised learning [3] (the learning data will be provided by the developer) and unsupervised learning [3] (the method will learn a clustering procedure by observing the distance among data). For the present study, supervised learning techniques will be used as the 12 groups do not change regularly.

In order to classify short messages using machine learning techniques, a proper set of features are required to extract from the short messages. The bag-of-words approach [1], was used to extra ct features from the short messages. The frequency of each word had been used as data. As there are large amount of words in different short messages, using all data will cause to increase the dimension. Thus, the common words are needed to identify and remove from the dataset. Zipfs low [5] states that the utmost common word in a human language text occurs with a frequency inversely proportional to n. Rarely occurred words do not carry sufficient information but the noise. Thus, low frequent words and high frequent words will be removed from the data in order to reduce the dimension up to certain level. Once created the dataset, it is important to find a suitable classification method in order to classify the

short messages. KNN [6] was used to classify the data as it is capable of dealing with high dimensional dataset [1], [5]. The system provides an accuracy of identifying news which belongs to group accident with 100%,development-government with 100%, disaster-climate with 100%, economy-business with 100%,education with 96.4%, entertainment with 100%, health with 95.2%, international with 80%, politics with 100%, society with 78.40%, sports with 100 % and war-terrorist-crime with 100%. The next section will brief out the method of data gathering and feature selection. Section 3 will brief about the methods used for classification. Section 4 will throw a light on the approach of data training and Section 5 will brief out the evaluation criteria of the system. The general discussion will be brief on section 6.

II. DATA GATHERING AND FEATURE SELECTION

The classification will be applied into the short messages-news of Twitter micro blog. Thus, twitter short messages are needed to be collected. Fortwitter, there was a character limit such that the length of one short message was limited to 140characters [7]. Thus the user bound to provide the news by using few amount of words. This caused to limit the words of the short messages into key words. Twitter API provides the ability of retrieving such short 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 to extract from the short messages. These features are required to learn the patterns amount the groups. The words are used as features. Thus the bag-of-words approach [1], [5] was used to extract the features. This will pool the words from all short messages and will create a document vector, containing words. Some researchers had used n-gram instead of words [8]. However, n-gram method cause to increase the dimension of the dataset, as it uses unigram, bigram, trigram which make complex for the system to recognize the pattern [8]. Thus, the words are chosen as the features.

In order to create the dataset, the frequencies of words were used. All words of the documents do not carry out useful information. To avoid using very low frequent words which do not carry out any valuable information regarding the group. Zipf slow [5] states that the nth most common word in a human language text occurs with a frequency inversely proportional to n. Thus, the common words were removed from the dataset by removing high frequency data. Therefore, a lower cut off value and upper cut off value were required to choose in order to obtain best set of features. The frequencies of selected 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) which maximize the effectiveness. Thus, unimportant features were removed from the dataset. This caused to reduce the dimension which makes the training method more effective.

Figure I: Range Of Frequencies Of Chosen Data

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 un supervised learning, the system itself learns patterns from the data. For the current situation, as the group are predefined and do not change regularly, supervised learning method is more applicable [3].

III. METHODS

In this paper the usage of the k-nearest neighbour (KNN) algorithm as a base classifier for the classifier ensembles is discussed. This approach was tested in combination with a group of the most popular classifier ensembles, i.e. Bagging. The concepts of the KNN classifier and the four classifier ensembles are briefly presented below.

A. The k-Nearest Neighbour Classifier

The most commonly used version of the classifier can be described as follows [9]. For a given query instance z the output of the KNN classifier is the most probable class:

$$kNN(z) = \max_{c_i \in C} p(c_i, z) \qquad (1)$$

In the case of the standard KNN classifier the class probabilities are calculated as follows:

$$p(c_i, z) = \frac{\sum_{x \in Kz} l(x_c = c_i).K(d(x,z))}{\sum_{x \in Kz} K(d(x,z))} \qquad (2)$$

where *l*() is a function that returns 1 when the condition is true, zero otherwise; *d(x,z)* is the distance between two points, in this case Euclidean distance and *Kz*is the nearest neighbours of *z*. *K()* is the kernel function defined below:

$$K\big(d(x,z)\big) = \frac{1}{d(x,z)} (3)$$

A lot of effort was put in trying to determine the optimal value of the *k* parameter. Some studies (e.g. [9]) suggest, based on experiments performed on a number of datasets, that for the most part the optimal value of *k* is close to 5.That is why this value was used in the study.

B. Classifier Ensembles

Bagging: One of the most popular ensemble methods is Bagging (or bootstrap aggregating) developed by

Breiman [10]. This algorithm works in the following way. A dataset consisting of X instances is given, each belonging to one of M classes. The method generates T new versions of a learning set by taking repeated bootstrap samples from the original dataset (sampling with replacement). Each new set has the same size as the original (although the size can be adjusted) and therefore, some of the instances can appear more than once. The algorithm trains each classifier D(t)based on one of the samples (where t = 1, 2, …, T). The outcome given by the final classifier D(*) is an aggregation of the results provided by T classifiers.

IV. 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].

The basic idea of KNN algorithm is transforming the text into weighted feature vector in the feature space according to the Vector Space Model. At first, it doesn't need to do anything on training samples. But when the test sample comes, this algorithm will compare the test sample with all the training samples and calculate their similarities. Then ittries to find the k-nearest samples from all the training samples. The set of these samples is called k-nearest neighbor of the test sample. The category of test sample is decided by the category that often appears in the k-nearest neighbor. The process is as follows [11]:

1) Setting the value of k, that is, the number of nearest neighbor.
2) Transforming the test sample q and the training sample dj into corresponding feature vector according to the Vector Space Model.
3) Using vector cosine similarity to calculate the similarity between the test sample q and the training sample dj, the formula is as follows:

$$\cos(d_j, q) = \frac{\sum_{i=1}^{|V|} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{|V|} w^2{}_{ij}} \times \sqrt{\sum_{i=1}^{|V|} w^2{}_{iq}}} \quad (4)$$

4) Sorting the similarities calculated by the above formula and selecting the k-maximum values as knearest-neighbor of test sample q.
5) Getting the categories of k-nearest-neighbor and dividing the test sample into the category that often appears.

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)

V. 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 measured using precision and recall values [12].Precision is the fraction of retrieved short messages that are relevant. Recall is the fraction ofrelevant short messages that are retrieved [13]. As the system results the performance as in table 1, the precision can be calculated using Equation 1 and recall could be calculated using Equation 2.

$$Precision = \frac{t_p}{t_p + f_p} \quad (5)$$

$$Recall = \frac{t_p}{t_p + f_n} \quad (6)$$

In order to measure the performance, the performance will be depending on the biasness of the training data. In order to avoid the biasness, cross validation [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 train the data. The system was tested 10 times per group and the average precision and recall values were calculated [12]. The results were briefed in table 2.

These measures were used to figure out best frequency limit for the feature selection. Thus, it is important to calculate a single measurement instead of 2 values [14]. Many alternative methods were proposed over the years and Harmonic mean [13] had identified as the best single value summaries [13].The harmonic mean (F-measure) was given in Equation 3 and weighted harmonic mean was given in Equation 4. Values of P less than 1 emphasize the precision whereas values of P greater than 1emphasizes recall [13].

$$F = \frac{2 \times P \times R}{P + R} \quad (7)$$

$$F_\beta = \frac{(\beta^2 + 1) \times P \times R}{\beta^2 \times P + R} \quad (8)$$

For the current situation, the main focus will be on the fraction of retrieved short messages that are relevant. Thus $F_{0.5}$ will be used to select the best frequency range and to measure the effectiveness of the 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

With the results obtain from Table 3 it is clear that KNN provides good results for most of groups.

VI. DISCUSSION

A system which is able to classify news headlines will be useful in various social researches. With the development of web technologies, people get involved in many social networks and web blogs. Twitter is a micro blog which allows many famous news suppliers to publish their news headlines. Twitter API supports user to retrieve available short messages. These retrieved files will be in XML file format and each file could retrieve maximum number of 200 short messages per once. In order to apply machine learning, a proper feature set was required. The feature set was created by pooling the words and creating a document vector. This approach was named as bag-of-words approach. The frequency of each word was chosen as data. When considering all words together, it create huge dimension with 3569 instances. In order to reduce the dimension, a lover cut off frequency and an upper cut off frequency value was chosen. According to the current system, these values are respectively 10 and 38. The value was chosen as the frequency range which maximizes the accuracy. This caused to reduce the dimension up to 126.

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 KNN. 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_β 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, 2nd ed. 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] Mateusz Budnik, IwonaPozniak-Koszalka, LeszekKoszalka, "The Usage of the k-Nearest Neighbor Classifier with Classifier Ensemble", 12th International Conference on Computational Science and Its Applications, 2012 IEEE

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

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

[9] T. K. Ho, "The Random Subspace Method for Constructing Decision Forests," Transactions on Pattern Analysis and Machine Intelligence, 832-844, 1998.

[10] L. Breiman, "Bagging predictors," Machine learning 24, 123-140, 1996.

[11] Jiang Tao; Chen Xiao-li; Zhang Yu-fang; XiongZhong-yang. Improved kNN using clustering algorithm. Computer Engineering and Application, 2009, 45(7): 153-158.

[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.

★★★