2015 1st International Conference on Next Generation Computing Technologies (NGCT-2015)
Dehradun, India, 4-5 September 2015

T - Vigilant: To Unmask Radical Attacks and Halt the Innocents

Prasad J. Koyande, Kavita P. Shirsat
Computer Engineering, Mumbai University
Vidyalankar Institute of Technology, VIT
Mumbai, India
prasad.koyande@vpt.edu.in, kavita.shirsat@vit.edu.in

***Abstract*– Forbearance and wisdom should be the key now. The system to prevent terrorist attacks that will relay emergency alerts to all phones is set to begin. This system could warn people of terrorist strikes by text message. With the popularity of Social Networks, mostly news providers used to split their news in various social networking sites and web blogs. In India, many news groups stake their news on Twitter micro blogging service provider which provides real-nature to the system. The system is an early precursor that collects and analyzes real-time news of events such as terrorist attack, hijack, bomb blast etc. from Twitter and detects a target event. Objective behind is to ooze message to all the phones in a given area, providing them with up-to-date and accurate guidance on the specific threat and the best way to escape. 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 if category has a certain amount of the training samples which helps to guarantee the accuracy of classification. Large amount of feature will be collected for current research. The performance will speak the efficaciousness of the system.**

***Keywords*--KNN, Web mining, Text classification, Twitter, event detection, Terrorist attack.**

I. INTRODUCTION

Over past decades the terrorist attacks has been the major threats to international securities. While each of the attacks left a scar on the heart of every Indian the worst and the deadliest among all terrorist attack was the 26/11 Mumbai attack of 2008. Bringing Mumbai to a standstill. Though seven years has passed, the pain can still be felt. Targeting and killing around 183 and injuring more than 600 civilians. It was tactic that was very difficult to defend against. It precisely aimed to at getting international attention, intended to hurt the prosperity, creating fear in the minds of foreigners to venture, through various types of business. It had dubious objective of destroying it. Specially targeted to make a dent on country's economy. We salute all those who laid down their lives fighting terrorists and protecting our country. Nothing can defeat us. But that's not all we need to stand united against terrorism. We can't totally eradicate attacks but, we can take preventive measures to save the innocents. It could not be a solution to the problem but a step towards it. While the attack was on some innocent lives would have been saved who were heading to the pad or on the way like attack was doomed. With the development of web blogs and Social Networks we avail this facility to develop an application in order, it will emit an alert message to the users of a concerned location, if they are at nearby dwellings they can restrain themselves and furthermore inform others.

Now-a-days in India, many news groups allocate their news headlines as short messages in microblogging services such as Twitter. T-Vigilant stage these short messages in a timely fashion in order to find news related to terrorist attack, bomb blast, hijack or information about any other dismay. If news related to terrorist activity is dug up, an alert message is formed by T-Vigilant and is shoot in form of SMS (Short Message Service) to all the consumers. Sending message in a form SMS is concluded so, unvarying if the consumer is offline the alert message will reach through and prove beneficial to them and others.

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

1) Microblogging 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 ispossible to collect text posts of users from different social and significant groups.
4) Twitter's audience is represented by users from many countries. Although users from U. S. are prevailing, it is possible to collect data indifferent languages.

978-1-4673-6809-4/15/$31.00 ©2015 IEEE

We propose the terrorist attack detection and notification system that monitors tweets and delivers SMS. We usually follows three steps: first and foremost several active Twitter news groups such as 'abpnewstv', 'IBN7', 'ibnlive', 'ndtv', 'zeenews', 'timesofindia', 'timesnow', 'dna', and 'httweets' were chosen to extract the data; secondly, we use classifier to identify the target event and finally we developed an terrorist attack reporting system that sends a message in a form of SMS to registered users.

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 (the learning data will be provided by the developer) and unsupervised learning (the method will learn a clustering procedure by observing the distance among data) [3]. For the present study, supervised learning techniques will be used to detect terrorist attacked.

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 extract 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. Zipf's 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 terrorist activity with 90.13%. 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. Section 6 describes the T-Vigilant architecture in detail and the general discussion will be brief on section 7.

II. Data Gathering and Feature Selection

The classification will be applied into the short messages-news of Twitter microblog. Accordingly, twitter short messages are needed to be collected. For twitter, there was a character limit such that the length of one short message was limited to 140 characters [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's low [5] states that the n^{th} 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. The values X and Y were chosen as the frequency limit (lower cut off value and upper cut off value) which maximize the effectiveness.

Thus, irrelevant 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 there are basically two types of learning methods. Supervised learning 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 predefined and do not change regularly, supervised learning method is more applicable [3].

III. Methods

In this paper the usage of the k-nearest neighbor (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 classifier ensembles are briefly presented below.

A. The k-Nearest Neighbor 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 *k* 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 sethas 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 it tries 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:

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

4) Sorting the similarities calculated by the above formula and selecting the k-maximum values as k-nearest-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 terrorist group manually. One short message might be belong into several groups. Therefore, 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 Terrorist activities group short messages will be classified manually as "Terrorist Group". 90% data was used to train the system and 10% were used to test the system [4]. In order to train there will be training data set and testing will results 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 of relevant short messages that are retrieved [13]. As the system results the performance as in table 1, the

precision can be calculated using Equation 5 and recall could be calculated using Equation 6.

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

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

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 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 7 and weighted harmonic mean was given in Equation 8. Values of P less than 1 emphasize the precision whereas values of P greater than 1 emphasizes 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_β.

TABLE 2: PRECISION AND RECALL VALUES

Group Name	Avg. Precision	Avg. Recall
Terrorist-Group	0.769	0.909

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

Group Name	F-MEASURE	F_β-MEASURE
Terrorist-Group	0.833	0.99

With the results obtain from table 3 it is clear that KNN provides good results for Terrorist Related Activities.

VI. T-VIGILANT ARCHITECTURE

The architecture exhibits the T-Vigilant (Figure 2). The main purpose is to extracts news from active news groups, analyze and spot the news linked to terrorist group and denounce it to the consumers.

A. Modules of Proposed System

1. Tweets Extraction: The system muster news from assigned and dynamic news group in Twitter. If the assemblage fade its leads to fallacious messages to user which can create havoc among the citizens. The news Groups preferred are collaborative and widely accepted (abp news, IBN7, ndtv, zee news…. etc.). Twitter API provides the ability of retrieving such short messages for a given user in XML file format. Each XML file could carry out 200 short messages at ever.

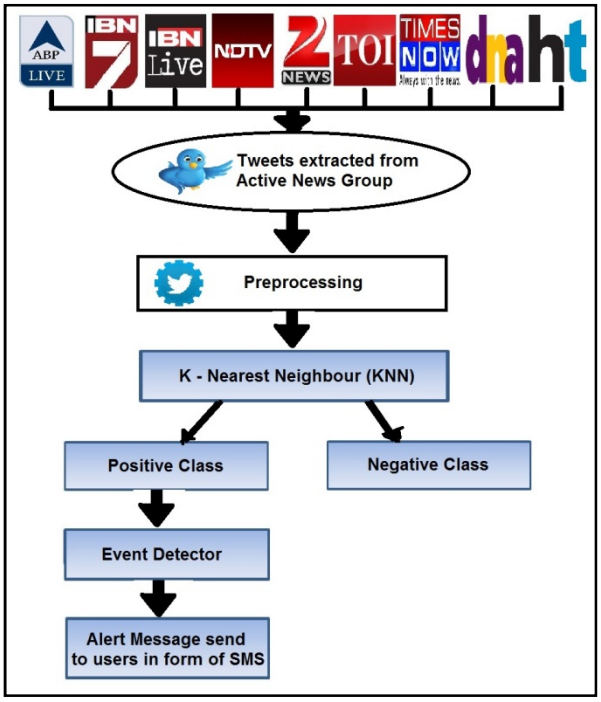

FIGURE 2: T-VIGILANT ARCHITECTURE

2. Pre-processing: Once the data is accumulated, the features need to be extracted from the short messages. The words are used as features. Thus the bag-of-words approach was used to extract the features. This will pool the words from all short messages and will create a document vector, containing words.

3. KNN Classifier: Document Vector containing the tweets are passed as an input to KNN classifier The news will be catalogue in order to detect it's affinity to Terrorist Activity Group. To create the training data and testing data for KNN classifier, each short message is classified to a group manually. Terrorist Group category is considered as a separate binary classification issue. The training process was developed in order to acknowledge, the selected short message connect to the Terrorist Group, 90% data was used to train the system and 10% was used to test the system.

4. Event Detector: After classification if the tweet connects to True Positive class in course tweets convey terrorist related activity news. News is processed in order to find the city where the attacks are on.

5. Sending Alert Messages: After finding city in Event Detector section actual Alert Message is formed and system circulates SMS to all the users of the system linked to the identified city by using SMS Text Messaging API.

VII. DISCUSSION

A system logically able to locate terrorist news from news headlines provided by the entrusted and dynamic news groups in Twitter. The news related to terrorist activity is dug up, an alert message is formed by T-Vigilant and is shoot in form of SMS (Short Message Service) to all the consumers. Sending message in a form SMS is concluded parallely if the consumer is offline the alert message will reach through and prove beneficial to them and others. Twitter API supports user to retrieve available short messages. These retrieved files will be in XML fileformat and each file could retrieve maximum number of 180 short messages ever. In order to apply machine learning, a proper feature set is 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.

Terrorist group was treated as separate binary classification issue. System was trained using KNN. The potency 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. It provides 99% potency. The system will work real magic if we maintain consistency. We will succeed in our endeavors even if one civilian feels he's/her life is valued. And our drudge towards it. .

VIII. 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] Inoshika Dilrukshi, 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, Iwona Pozniak-Koszalka, Leszek Koszalka, "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 TextCategorization," in Proceedings of SDAIR-94, 3rd AnnualSymposium 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 andMachine Intelligence, 832-844, 1998.
[10] L. Breiman, "Bagging predictors," Machine learning 24, 123-140, 1996.
[11] Jiang Tao; Chen Xiao-li; Zhang Yu-fang; Xiong Zhong-yang. Improved kNN using clustering algorithm. Computer Engineeringand 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.