International Journal of Database Theory and Application
Vol.8, No.4 (2015), pp.41-48
http://dx.doi.org/10.14257/ijdta.2015.8.4.05

# Twitter Crossfire: Terror Attack Detection via Probabilistic Classifiers

Herman Wandabwa[1], Liao Zhifang[2] and Korir Sammy[3]

*[1] School of Information and Communication Engineering, Central South University, Changsha, China*
*[2] School of Software, Central South University, Changsha, China*
*[3]School of Information and Communication Engineering, Central South University, Changsha, China*
*wandabwa2004@gmail.com,zfliao@csu.edu.cn,korirsammy@gmail.com*

### *Abstract*

*The advent of social computing brought with it different social networking platforms. The idea of surfers socializing with people of different backgrounds as well as geographical regions is quite fascinating. In our approach, we delved deeper in disaster discovery whereby we extracted panic related attributes and trained them with real data in three disaster scenarios in different parts of the world. Fine tuning of the final attributes led to accuracies above 91% proving the fact that with proper attribute selection and handling of sparse data balance, it's possible to detect related disasters as soon as related tweets appear. We believe that we are the first to use probabilistic classifiers approach as well as NLP in specifically human induced terror attacks detection as there is no known system currently that solely caters for these.*

***Keywords:*** *Disaster Detection, classification, Twitter analytics*

## 1. Introduction

Twitter has become one of the internet's most popular micro blogging platforms since its launch in 2006. Its premise lies in sending of 140 character messages to the digital space known as the Twitter sphere [1, 2]. Twitter service handled 1.6 billion search queries in 2012 alone. In 2013, it was one of the most visited websites and many have described it as the SMS of the internet. Currently it is ranked as the 8th most visited website by traffic volumes [3] and has more than 500 million users, out of which more than 271 million are active users.

The advent of GPS enabled smart devices has augmented the study of not only how users tweet but where they tweet from and the relation with the tweet content. Twitter has also emerged as a key resource of free and open volunteered geographic information [4]. While much of this data is noise, containing plain communication and chatter, some tweets contain attempts at citizen journalism [5], whereby users describe and provide information about the world around them.

This therefore forms the basis of our research as we attempt to dissect whether we are able to detect a disaster via the semantics of the tweets and to be specific terror attacks that have been rampant in recent days. Our contribution to this research is as below:-

1. Collection and integration of machine learning algorithms to clearly and with high accuracies identify a disaster in the streaming and noisy messages. This will specifically be in relation to human induced terror attacks.
2. Proof that with proper attribute selection and deployment of added information gain filters, accuracies in identification of terror related events data can be improved.

ISSN: 2005-4270 IJDTA

The rest of the paper is organized as follows:

Section 2 touches on related works where we'll dig deeper on what has been previously done in the field of disaster prediction be it man made or natural ones. Past works on the same will be elicited and their contributions to our research. Section 3 is dedicated to our approach i.e. how we came up with our findings including methods, algorithms as well as the performance evaluation. Section 4 and 5 include the conclusion and future work respectively.

## 2. Related Works

Tweets are being used in the detection and monitoring of natural as well as manmade events. An event is commonly considered as an occurrence at a specific time and place [1, 2, 4, 5, 6]. In the social media sphere therefore it's often described as an occurrence causing change in the volume of text data that discusses the associated topic at a specific time. This occurrence is characterized by topic and time, and often associated with entities such as people and location [7]. This will therefore be build around emerging keywords in a discussion. An emerging keyword can be viewed as a semantic unit which links to a very recent news event. Disasters whether natural or human induced will therefore be classified as events. Currently, the literature relevant to Twitter and disaster events uses a volumetric method of analysis.

Most methodologies lean towards selection of words, phrases or hash tags that relate to an event. Tweet messages containing the words being searched are deemed relevant to the event. Studies on such have been done by Takeshi Sakaki [8] on detecting and precise location of where the earthquake strikes by analyzing the tweets to get the semantics as well as plotting the tweet location on the map by applying Karlman filtering algorithm. Messages such as "earthquake" or "earth is shaking" most likely will infer presence of an earthquake. A spike in such phrases therefore correlelates with an earthquake event. Twitter users in this case act as social sensors while tweets are the sensor information.

## 3. Our Approach

### 3.1 Methodology

Our approach consists of data sampling, data preprocessing, model construction, and model evaluation phases. In sampling a random set of such events data meeting the required threshold is selected. This is in relation to panic which manifests in terror related events. We realized that in most human induced events, panic in the messages is clearly evident. Accurate measure of the attribute therefore augments findings in relation to such events. Data cleaning removes the irrelevant information including but not limited to human errors, symbols etc. We followed this up with feature extraction steps whereby we extracted a set of features or variables to be evaluated during prediction. We then build a classification/prediction model meant to indicate whether presence of certain features in a dataset can lead to a conclusion that the messages are indeed terror related. The rest of the steps are as below.

### 3.2 Data Sampling and Preprocessing

We collected three different datasets and at different times via Twitter's streaming API [9]. They were all collected at times when such events were taking place hence the relevance. To be specific, the events are as below:-

| Name | Data Collection Duration | Tweets & Retweets |
|---|---|---|
| Mpeketoni | $18^{th}$ – $22^{nd}$ June 2014 | 87507 |
| Sydney | $16^{th}$ – $20^{th}$ Dec 2014 | 25992 |
| Peshawar | $15^{th}$ – $21^{st}$ Dec 2014 | 252648 |

Data was grouped into two homogeneous strata, NORMAL and DISASTER. Normal refers to any message that doesn't depict panic. On the other hand disaster as the name suggests represents disaster or panic in a message in relation to the situation. Random sampling is then used in each stratum independently to obtain data of the required sample size. We ended up with 4927 instances and 1602 attributes.

The below assumption was made prior to training of the datasets:-

1. The data collected in one way is disaster related i.e. was collected at a time one as about to happen or during a time interval when one was experienced. This was to provide leverage for a disaster related environment data.

### 3.3 Model Construction and Evaluation

Once appropriate attributes have been extracted, the next phase is to test their predictive importance. Three prediction techniques are used in this phase: Naïve Bayes Multinomial, Sequential Minimal Optimization (SMO) and the J48 decision tree as they were implemented in Waikato Environment for Knowledge Analysis (WEKA 3.6.4)[10].

**a) Multinomial Naïve Bayes**

Multinomial Naïve Bayes (MNB) computes class probabilities for a given document i.e. let the set of classes be denoted by $C$ and $N$ is the size of the vocabulary. The MNB assigns a test document to the class that has the highest probability $\Pr(C|t_i)$ which using the Bayes rule is given by:-

$$\Pr(C|t_i) = \frac{\Pr(c)\Pr(t_{i|C})}{\Pr(t_i)}, \quad c\in$$

The class prior Pr(c) can be estimated by dividing the number of documents belonging to the class *c* by the total number of documents, $\Pr(t_i|c)$ is the probability of obtaining a document like $t_i$ in class *c* and is calculated as:

$$\Pr(t_i|c) = (\textstyle\sum_n f_{ni})!\pi_n \frac{\Pr(w_n|C)f_{ni}}{f_{ni}!}$$

Where $f_{ni}$ is the count of word *n* in our test document $t_i$ and $\Pr(w_n|C)$ the probability of word *n* given class *C*. The latter probability is estimate from the training documents as:

$$\widehat{Pr}(w_n|c) = \frac{1 + F_{nc}}{N + \sum_{x=1}^{N} F_{xc}}$$

where $Fxc$ is the count of word *x* in all the training documents belonging to class *c* and the laplace estimator used to prime each word count with one to avoid the zero frequency problem. The normalization factor $\Pr(t_i)$ can then be computed as below:-

$$\Pr(t_i) = \sum_{k=1}^{|C|} \Pr(k)\Pr(t_i|k)$$

**b) Sequential Minimal Optimization (SMO)**

Sequential Minimal Optimization (SMO) is an algorithm that quickly solves the SVM Quadratic Programming (QP) problem without any extra matrix storage and without using numerical QP optimization steps at all. SMO decomposes the overall QP problem into QP sub-problems, using Osuna's theorem [11]. The problem is broken into a series of the smallest possible sub-problems which are then analytically solved. Because of the linear equality constraint involving the Lagrange multipliers $\alpha_i$, the smallest possible problem involves two such multipliers. The constraints of any two multipliers $\alpha_1$ and $\alpha_2$ are reduced to

$$0 \leq \alpha_1, \alpha_2 \leq C,$$
$$y_1\alpha_1 + y_2\alpha_2 = k,$$

This is then reduced analytically. This implementation globally replaces all missing values and transforms nominal attributes into binary ones. It also normalizes all attributes by default.

**c) J48 decision tree**

J48 classifier creates a binary tree. Once the tree is built, it is applied to each tuple in the database and results in classification for that tuple [12, 13]. The training data is a set say $T = T_1, T_2 \ldots.$ of already classified samples. Each sample $S_i$ consists of a p-dimensional vector $(x_1, i, x_{2,i,} \ldots. x_{p,i})$ where $x_j$ represent attributes or features of the sample, as well as the class in which $T_i$ falls.

### 3.4 Evaluation Criteria and Findings

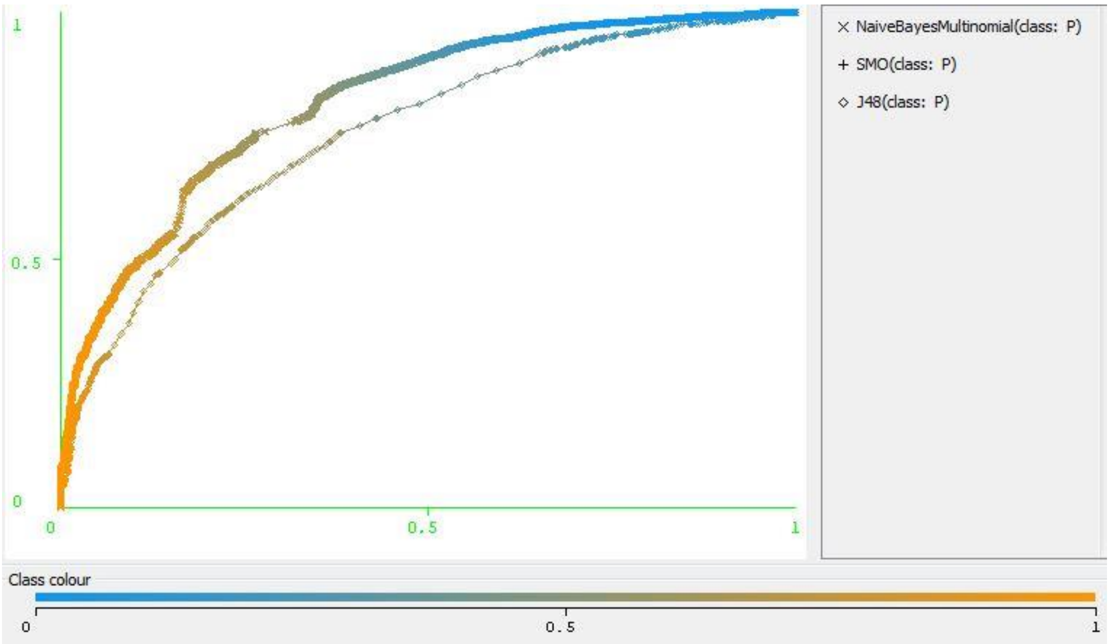
With balanced datasets, overall accuracy is enough to measure the performance of classifiers in predictive models especially in cases with balanced datasets. For imbalanced datasets, overall accuracy is not appropriate because it gives wrong results for models that do not generalize well. The situation is the same even in datasets with no direct negation i.e. in cases of the negative classes. Therefore, other evaluation measures need to be deployed.

*True positive (TP):* Number of positive instances correctly predicted.
*False negative (FN):* Number of positive instances wrongly predicted as negative.
*False positive (FP):* Number of negative instances wrongly predicted as positive
*True negative (TN):* Number of negative instances correctly predicted.
We evaluated the above prediction techniques with our dataset and below is the performance summary.

**Table 1. Performance Evaluation Metrics**

| | Predicted Class | | | |
|---|---|---|---|---|
| | | Positive | Negative | |
| Actual Class | Positive | TP | FP | Prediction Value = TP/(TP+FP) |
| | Negative | FN | TN | Prediction Value = TN/(FN+TN) |
| | | Sensitivity =TP/(TP+FN) | Specificity =TN/(FP+TN) | |



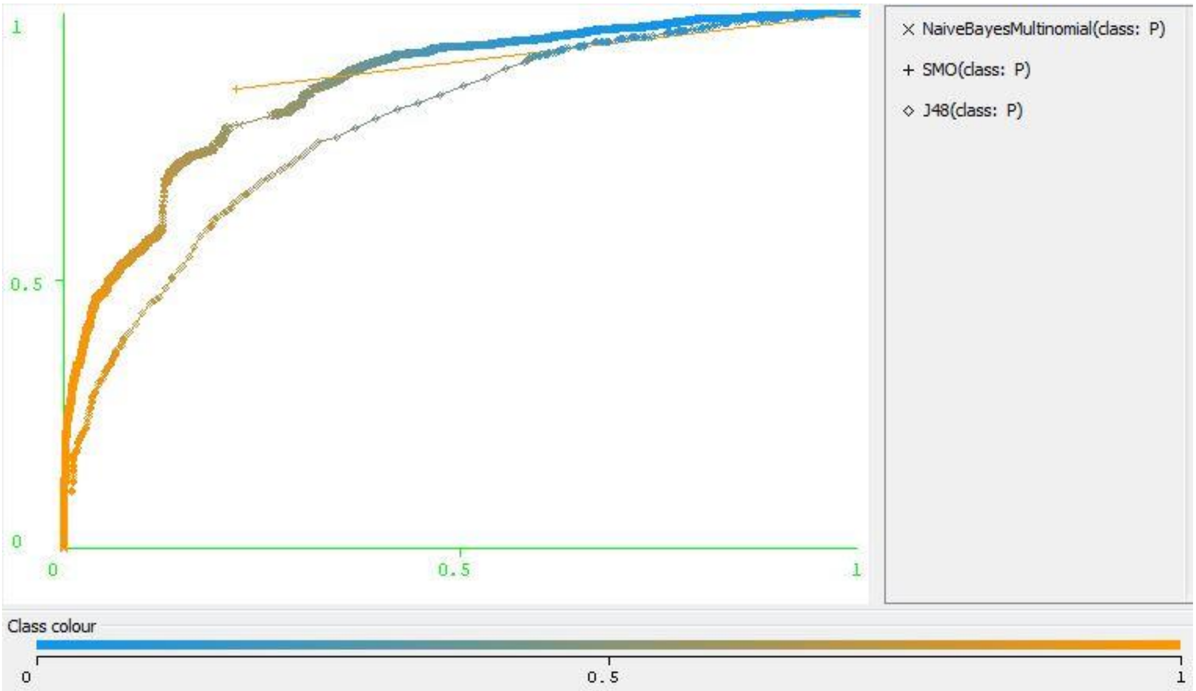

**Graph 1. ROC Showing Algorithms Performance Evaluation with an Unbalanced Dataset**

With the unbalanced data as earlier described, MNB algorithm performs slightly better than the other two. This is because of the independent probability of each class which is derived from the test class sizes. The higher the number of attributes in one class automatically means the probability of an almost same accuracy value will be derived at the end. This explains why the J48 algorithm performed a bit dismally compared to the others as individual attributes dependency probabilities had to be computed regardless of the assigned class.

### 3.5 Class Imbalance, Information Gain and Ranking

Our dataset was unique in that it comprised of three subsets of almost related human related events. The common factor among the three was is the fact that they are terror related in terms of the disaster type. Since the data was collected prior to and during the event, a large chunk of the messages were panic related but since the events happened in three different places, choice of words were different though depicting the same thing. We therefore needed to make sure the attributes were balanced to get the best result which actually is the axis of our research i.e. maximizing the chances of detecting a disaster given a set of incoming messages.

We therefore had to add some filters to the dataset for better performance of the classifiers as the above output wasn't good enough. The dataset was first resampled using the Resample filter where the samplesizepercent remained the entire set of attributes from both classes and random seed set to 1. We also employed the Ranker search method and InfoGain as the attribute evaluator to class the influential attributes in the dataset. This is critical as most of the attributes in the previous output worked down on the probability of other attributes influencing the results in the test set. The best 100 in attributes in descending order were generated. The output therefore had the best attributes in terms of higher probabilities in both classes. Below is the ROC curve showing increased accuracy.




**Graph 2 Algorithms Performance with Information Gain filter and Rank Search Method**

## 4. Conclusion

With the nature of our data, there is a high chance that attributes will be enormous. However, not all these are paramount in predictive models. Only few of them are important in disaster prediction. Careful selection and labeling of attributes is key to achieving best results.

Performance evaluation overall had pretty good results in all the algorithms. SMO stood out as the best performing classifier when we resampled the two classes. This was because of the class imbalance in the original dataset. With resampling a random subsample of the two classes is generated without replacement. It is an unsupervised approach and number of instances is specified. The class distribution of the subsample is maintained towards a uniform distribution of the otherwise unbalanced class. SMO is an improved training algorithm for SVMs. Like other SVM training algorithms, SMO breaks down a large QP problem into a series of smaller QP problems. Unlike other algorithms, SMO utilizes the smallest possible QP problems, which are solved quickly and analytically, generally improving its scaling and computation time significantly. Its performance on sparse inputs is great, even for non-linear SVMs, because the kernel computation time can be reduced, directly speeding up SMO. Usually chunking spends a majority of its time in the QP code, it cannot exploit either the linearity of the SVM or the sparseness of the input data.

Once the attributes have been identified, rigorous evaluation should be performed in order to ascertain the predictive importance of the selected attributes. Many techniques can be used to achieve this, including filter technique like information gain. Information gain calculates the information content of an attribute. A ranker can then be used to rank the attributes and those with the highest information gain can be selected and used in predictive modeling.

Disasters especially man made ones happen daily across the world. Recently we have witnessed the Peshawar, Sydney as well as Tunis terror attacks. Disaster prediction is paramount in disaster management. Lives can be saved as well as saving so much on related costs if they are detected early. Twitter just like any other social media platform is critical in such information dissemination.

In this research we have studied disaster prediction in the context of improving disaster recognition patterns. Social media datasets have immense attributes in addition to noisy features. Careful selection of feature sets should be done to improve on the prediction rates as not all the features and attributes are of importance especially in disaster prediction. We therefore carefully selected features that exerted influence in a disaster related perspective in the observed datasets as we combined three of them that were related in nature.

The features were evaluated using three data mining algorithms i.e. Multinomial Naive Bayes, SMO and J48 tree. In order to further assess the impact of the proposed features, we ranked the feature sets of both the original and the modified datasets using information gain attribute selection filter. Values of TP and FP as well as overall accuracy recorded. Experimental results show improved prediction rates in all the algorithms. Multinomial Naive Bayes achieved better results without information gain features but overall accuracy was best with SMO especially after ranking features as per their influence. Several techniques addressing the class imbalance problem were also studied. Resampling was performed on the class sets to balance the features in the two classes. Experimental results greatly improved especially with the ranking of features in place. SMO performed much better thus having better accuracy overall.

## 5. Future Work

In general, it is impractical to achieve 100% accuracies in disaster prediction due to the shifty nature of the data. However, it is possible to improve on the current rates achieved by prediction models. In order to do that, other methods for feature extraction and feature combination will need to be studied in future. Better methods of handling class imbalance need to be unraveled. Of particular interest are the ensemble methods like boosting and bagging and whether they can be combined with other methods like resampling to achieve better performance even with increased attribute numbers.

## References

[1] C.-P. Wei and Y-H. Chang, Discovering Event Evolution Patterns From Document Sequences, Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on , vol.37, no.2, **(2007)**, pp.273-283.

[2] C. C. Yang, X. Shi and C.-P. Wei, "Discovering event evolution graphs from news corpora", Trans. Sys.Man Cyber, vol.A, no.39, **(2009)**, pp.850-863.

[3] Alexa, "The Web Information Company", **(2011)**.

[4] G. Kumaran and J. Allan, "Text classification and named entities for new event detection", In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, **(2004)**; ACM, New York, USA.

[5] T. Brants, F. Chen and A. Farahat, "A System for new event detection", In Proceedings of the 26th annual international ACM SIGIR conference on Research and developmentin informaion retrieval, **(2003)**; ACM, New York, NY, USA.

[6] Y. Yang, J. Zhang, J. Carbonell and C. Jin, "Topic-conditioned novelty detection", In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, **(2002)**; ACM, New York, NY, USA.

[7] W. Dou, X. Wang, D. Skau, W. Ribarsky and M. X. Zhou. LeadLine: Interactive Visual Analysis of Text Data through Event Identification and Exploration. IEEE Conference on Visual Analytics Science and Technology, **(2012)**.

[8] T. Sakaki, M. Okazaki and Y. Matsuo, "Earthquake shakes Twitter users: real-time event detection by social sensors", Proceedings of the 19th international conference on World wide web, **(2010)**.

[9] The Streaming APIs :Twitter Developers, https://dev.twitter.com/streaming/overview as at **(2015)**.

[10] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I.H Witten, "The WEKA Data Mining Software: An Update, SIGKDD Explorations", vol.11, **(2009)**.

[11] E. Osuna, R. Freund and F. Girosi, "Improved Training Algorithm for Support Vector Machines," Proc. IEEE NNSP '97, **(1997)**.

[12] H. Margaret, Danham and S. Sridhar, Data mining, Introductory and Advanced Topics, Person education , vol.1, **(2006)**.
[13] A. K. Sharma and S. Sahni, "A Comparative Study of Classification Algorithms for Spam Email Data Analysis", IJCSE, vol.3, no.5, **(2011)**, pp.1890-1895.

## Authors

**Herman Wandabwa**, he received Bsc degree in Information Technology from Jomo Kenyatta University of Scie nce and Technology (JKUAT), Kenya in 2007. He is currently working towards Msc degree in Computer Application Technology at Central South University, Changsha, China. His main research interests are in social and web mining.

**Zhifang Liao**, she is an associate professor in School of Software, Central South University, Changsha, Hunan, P.R.China. Her main research areas include data mining, social network and recommendation system.

**Korir Sammy**, he received Bsc Information Technology degree from Jomo Kenyatta University of Science and Technology (JKUAT), Kenya in 2007. He also received his Msc Computer Application Technology degree at Central South University, Changsha, China in 2012. He currently works for Cognizant as a programmer in Shanghai, China. His main research interests are social media mining; sentiment analysis and opinion mining for customer profiling.