Analyzing News Summaries for Identification of Terrorism Incident Type[6]

Sarwat Nizamani[1], Nasrullah Memon[2]

[1]University of Sindh, PAKISTAN, [1]Maersk Mckinney Moller Institute, University of Southern Denmark, DENMARK,

[2]Mehran University of Engineering and Technology, PAKISTAN.

[1] sarwatnizamani@gmail.com

ABSTRACT

In this paper we present experiments for the detection of terrorism incident types from news summary. The news summaries from the global terrorism dataset have been analyzed using machine learning techniques. We have conducted experiments using different learning algorithms including Naive Bayes, decision tree and support vector machine. The results of the experiments show that decision tree learning algorithm can well learn incdent types from the news summary and achives high accuracy for detecting the type of incident from the news.

Keywords: GTD, Classification, Decision tree, Naïve Bayes, SVM

INTRODUCTION

After the tragic events of September 11, the academicians have been diverted towards the area of counterterrorism at large scale. This paper is presents our efforts for the noble cause. In the paper news summaries are analyzed for training the models which learn to detect the type of terrorism incident. The learning is then applied on unknown news summaries, in order to identify the type of terrorism incident. The experiments are performed on the part of Global terrorism database (GTD). The paper emphasizes on the use of text mining for extracting the critical information from the free text. In the paper, we present text mining experiments to detect terrorism incident type from news summary in the Global Terrorism Database (GTD). The purpose of the research is to emphasize that we can extract useful information according our query from free text using classification techniques. It is time consuming if one goes through the lengthy text to extract a specific kind of information. Classification techniques can be applied in different ways according to one's requirements to extract specific information from text. We have applied classification techniques for accomplishing the desired task. We experimentally show that we can extract this information from free text summary in the database. By using training data from the GTD, we train the classifiers to learn the patterns of the incident and classify the new incident from the news data as a specific type of terrorism incident. We have applied text mining on news summary, and trained the classifiers by providing training data. We performed experiments using three different classifiers i.e. decision tree (J48 WEKA implementation of C4.5), Naïve Bayes and Support Vector Machine (SVM). We present the experimental analysis of the classifiers. The evaluation method that we have used for experimental analysis is tenfold cross validation. In the experiments we show the empirical analysis of all the three classifiers on the GTD. For applying text mining techniques we have used *Waikato Environment* for Knowledge Analysis (WEKA) [14].

6 This paper was presented in its original form at 2011 International Conference on Computers and Advanced Technology in Education and published by Springer

ISSN: 2307-3721, e ISSN: 2307-3713

We show experimentally that a simple decision tree classifier can identify the incident type with adequate accuracy. SVM classifier also achieved reasonable accuracy but at the expense of long running time where Naïve Bayes classifier runs faster but with low accuracy. According to our findings we can reliably apply classification techniques on task like detecting terrorism incident type from news data summary using decision tree classifier. Below we present brief description of GTD.

Overview of Global Terrorism Database (GTD)

The Global Terrorism Database is an open source database that contains information regarding terrorism incidents that took place between 1970-2008 in all over the world. There are certain characteristics of the dataset defined on website [1] of the GTD.

Following is the brief description of the dataset:

Total number of incidents	Over 87000
Incident types include	38,000 bombings 13,000 assassinations 4,000 kidnappings
Minimum number of variables	45
Maximum number of variables	>120
Supervised by	12 Terrorism research experts
Sources of information	3,500,000 News articles, 2500 News sources

In the next section we present related work. Section 3 describes classification techniques; whereas in Section 4 we elaborate the terrorism incident type detection. We discuss preprocessing of data in Section 5 while we illustrate experimental results in Section 6 and conclusion and future work is presented in Section 7.

RELATED WORK

Global Terrorism Database is a large collection of terrorism incident data in all over the world. It is a good source for counterterrorism and criminology research. A number of researchers have analyzed the dataset and presented their useful findings in the literature. In this paper we discuss some of them. Dugan et al. [2] have used GTD for analyzing hijacking incidents before 1986. The authors used continuous time survival analysis to estimate impact of counter-hijacking interventions on the hazard of differently motivated hijacking attempts and logistic regression analysis to model the predictors of successful hijackings. The authors found that the policy interventions examined significantly decreased the likelihood of non-terrorist but not that of terrorist hijackings.

Greenbaum *et al.* [3] have used the GTD to analyze the impact of terrorism on Italian employment and business during 1985 to 1987. The authors concluded that terrorist attacks reduced the employment following the year of attack. The authors [4] used terrorist attacks data from 1970 to 2004. In the article, the authors have tried to show the characteristics of global terrorism. The authors also included an analysis showing the link between the terrorism and political affairs in the country.

The article [5] discusses the impact of governmental counter-terrorism policies on the violence in the country. They show that it has positive as well as negative impact. The authors [6] have studied the GTD for domestic terrorism in the United States. They used group-based trajectory analysis to examine the different developmental trajectories of U.S.

ISSN: 2307-3721, e ISSN: 2307-3713

target and non-U.S. target terrorist strikes. The authors concluded that four trajectories best capture attack patterns for both. The authors [7] have used spatial (country name, place name) and temporal (date, month, year) information from the GTD and found a number of useful patterns from the database. The authors have presented the patterns using visualization.

Paper deals with the analysis of news summary, therefore, we also here discuss some related work in connection to the news analysis. A study by Nizamani and Memon [16] presented a semantic based news analysis method using a technique known as semantic role labeling [25]. The study dissects the new reports in order to highlight important information from the reports.

In this paper, we apply text mining approach to the major variable of the dataset that is the summary of terrorism incident. We try to extract information about type of terrorism incident from the summary. We experimentally show that classification techniques can learn from news summary to detect the incident type. The next section presents various classification algorithms used in the experiments

CLASSIFICATION ALGORITHMS

Classification [15] is a kind of supervised machine learning algorithm. It takes training examples as input along with their class labels. It can be defined by following equations:

$$D = \{t_1, t_2, \ldots t_n\} \quad (1)$$

$$t_i = \{a_1, a_1, \ldots, a_m\} \quad (2)$$

$$C = \{c_1, c_2, \ldots, c_k\} \quad (3)$$

Where D is a dataset consisting of n training examples, t_i is a training example, each a_i is an attribute, m is the total number of attributes and c_i is a class and k is the total number of classes. With respect to our terrorism incident type detection D is collection of *22235 terrorism incidents*, each terrorism incident t_i comprises of *5345 attributes* a_i and C is a set of terrorism incident type and total number of incident types k is 9.

Decision Tree

Decision tree is a kind of divide and conquer algorithm. A decision tree consists of finite number of nodes—internal and external nodes. Each internal node corresponds to an attribute selected by some measure of algorithm like information gain or gain ratio that divides the training examples into the parts according to the values of that attribute. For example if the attribute has three possible values then there will be three branches going out from that node. The choice of attribute at particular level of hierarchy usually depends on the class distinction ability of that attribute. External nodes in the decision tree contain decisions or the class value. ID3 (Iterative Dichotomiser 3) is a kind of decision tree algorithm by Quinlan [9]. The algorithm suffers from over fitting and also the algorithm can only work on nominal values and discrete values and also ID3 does not deal with missing value. To overcome these issues of the ID3, Quinlan [10] proposed C4.5 algorithm. It uses pruning to overcome over fitting problem, uses discretization at a certain threshold to deal continuous data and ignores missing value attributes while making decisions.

Naïve Bayes (NB)

Naïve Bayes [11] is a simple and efficient technique used by data mining community for classification task. It uses Bayes theorem to estimate probabilities for each class to decide the class of an instance. NB assigns the maximum probability class label to a test instance.

ISSN: 2307-3721, e ISSN: 2307-3713

Support Vector Machine (SVM)

SVM is considered to be the state of art classification algorithm. SVM is a supervised machine learning technique used for classification. SVM is based on Vapnik's statistical learning theory [13]. SVM has some unique features due to which it is considered as state-of-the-art in classification. It is considered well suitable for the task of text classification and hand written digit recognition. Its unique features for text categorization are [12]: (i) It works well with high dimensional data; (ii) It can make a decision boundary by using only a subset of training examples called support vectors; (iii) It can also work well on non-linearly separable data by transforming the original feature space into a new feature space that is linearly separable by using the kernel trick. Joachims [12] has defined some properties of text classification for which SVM is the ideal choice of solution. SVM has a main limitation that it suffers from long running time when runs on large datasets

PREPROCESSING DATA

In the paper, we used terrorism incidents from GTD which took in the period of 2001 to 2008. The required part of the dataset is transformed into the ARFF file, in which each instance represents an incident from GTD. ARFF is Attribute Relation File Format used by WEKA [14]. From the GTD we only used two fields of each incident namely; summary (a text field) which presents the description of the incident; and type of incident which takes a value from one of the type of terrorism incidents. The summary field is further preprocessed using text mining processes because it involves the free text. This further preprocessing is applied using WEKA utility (String To Word Vector). This utility performs text mining steps such as, tokenization, stop word removal and feature weighting, etc.

DETECTION OF TERRORISM INCIDENT TYPE

Terrorism incident type detection is considered as a text classification problem. We carried out the task using classification algorithm. The process involves the training data, from which the patterns of the incident type are learned by the learning algorithms.

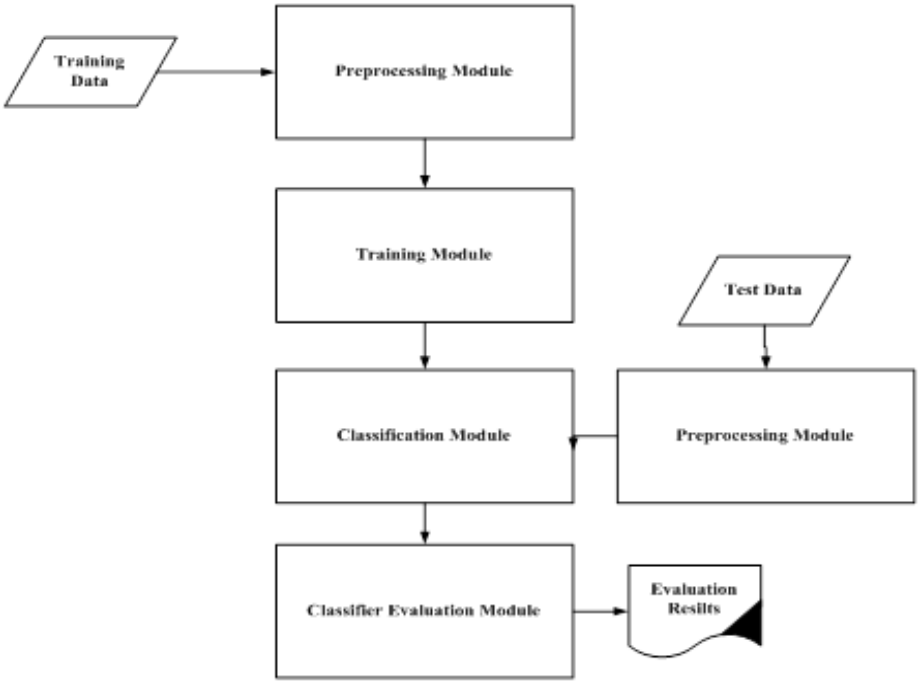

Figure 1. Work flow of terrorism incident type detection

ISSN: 2307-3721, e ISSN: 2307-3713

We provided the training incidents as training data which is comprised of the summary of incidents as well as type of the incident. After the training data is provided, preprocessing is performed, which makes the data in appropriate form for different classification algorithms. After the classifier is trained on training data, the learning is applied for predicting the type of terrorism incident using only the summary of incident. The complete process is demonstrated in Figure 1.

EXPERIMENTAL ANALYSIS

The experiments are conducted using terrorism incident records from GTD between the period of 2001 and 2008. Each terrorism record is comprised of a news summary and a number of other features describing terrorism incident including the type of incident. For experiments we have taken total 22235 records. After preprocessing we have total *5345* distinguished features. A short description of the dataset is provided in Table 1. A detail of all the incident types, including the number of incidents of each type is demonstrated in Table 2. The experiments are conducted using three well-known classification algorithms, namely; Decision tree J48 (WEKA's implementation of C4.5), Naïve Bayes (NB) and Support Vector Machine (SVM). These are widely used classification algorithms very famous among research community [8]. The evaluation method and evaluation measures used in the experimentation are described in the following sub-section.

Table 1. General information about dataset used in experiments

Total number of incidents	22235
Total number of feature	5345
Total number of classes/ Types of incident	9
Incident period	2001-2008

Table 2. Incident type distribution in training data

Type of Incident	*No of Incidents*
Amed Assault	6797
Assassination	1167
Bombing Explosion	10731
Facility Infrastructure Attack	1820
Hijacking	59
Hostage Taking Barricade Incident	134
Hostage Taking Kidnapping	1111
Unarmed Assault	275
Unknown	141
Total	22235

Evaluation Method

We have used is 10 fold cross validation method for the purpose of evaluation of results. Tenfold cross validation splits the dataset in 10 subsets. It runs for 10 rounds, in each round 9 subsets are used for training and one of them is used for testing. In each round a new subset is chosen for testing. After 10 rounds the average accuracy of all the rounds is measured.

Evaluation Measures

The evaluation measures that we have used are accuracy, precision and recall. These measures are calculated as follows:

Accuracy = (Tp+Tn) / (Tp+Tn+Fp+Fn.) ……. (4)

Precision = Tp/ (Tp + Fp) ……. (5)

Recall = Tp/(Tp + Fn) ……. (6)

Where *Tp* is the number of incidents correctly classified as particular class, *Fp* is the number of incidents that were incorrectly classified as a particular class. *Tn* is the number of incidents that were correctly classified as other class and *Fn* is the number of incidents that were incorrectly classified as another class.

The experimental results (see Figure 2) clearly illustrate that from the news summary data we can successfully detect terrorism incident type. The classification algorithms can extract this information successfully. It is clearly depicted in the figure that decision tree correctly detects 83% of incidents with a balance of precision and recall.

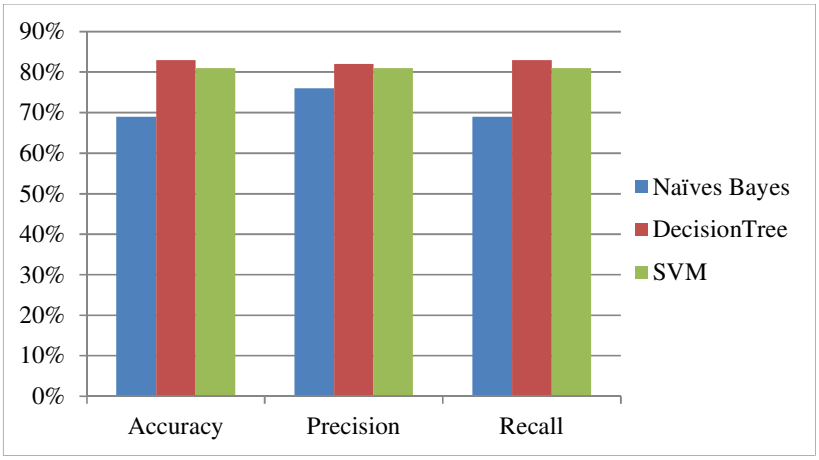

Figure 2. Experimental results

CONCLUSIONS AND FUTURE WORK

In this paper we applied text mining techniques and presented the experimental results for detecting terrorism incident types from the news summaries of terrorism incidents. We performed experiments using classification algorithms, such as decision tree, Naïve Bayes and state of the art SVM. Experimental results illustrate that the task can be successfully carried out using classification algorithms with satisfactory results. The results show that a high accuracy is achieved using J48 (decision tree) algorithm with a balance of precision and recall. SVM also achieved high accuracy, but it takes long running time when there is large dataset. The accuracy achieved using Naïve Bayes is lower comparatively but it runs faster. The current work employs the use of words as features without using any semantic knowledge. In future, we intend to incorporate the semantic knowledge which will have

ISSN: 2307-3721, e ISSN: 2307-3713

positive effect on the accuracy of the task. It is also included in our future plans to make use of spatio-temporal features from the dataset, in order to find the correlations among incident type, time and geo space.

ACKNOWLEDGMENTS

This work was carried out during PhD study of first author at University of Southern Denmark.

REFERENCES

[1]. http://www.start.umd.edu/gtd/about/

[2]. Laura et al. (2005). "Testing a Rational Choice Model of Airline Hijackings." *Criminology, 43*, 1031-1065.

[3]. Robert et al. (2007). "The Impact of Terrorism on Italian Employment and Business Activity." *Urban Studies, 44*, 1093-1108.

[4]. Gary, L., & Laura, D. (2009).*"Tracking Global Terrorism, 1970-2004.' In To Protect and to Serve: Police and Policing in an Age of Terrorism*", David Weisburd, Thomas Feucht, Idit Hakimi, Lois Mock and Simon Perry (eds.). New York: Springer.

[5]. Gary et al., (2009). "The Impact of British Counter Terrorist Strategies on Political Violence in Northern Ireland: Comparing Deterrence and Backlash Models." *Criminology, 47*, 501-530.

[6]. Gary, L. (1970 to 2004). Sue-Ming Yang and Martha Crenshaw. (2009). "Trajectories of Terrorism: Attack Patterns of Foreign Groups that have targeted the United States, 1970 to 2004." *Criminology and Public Policy, 8*, 445-473.

[7]. Guo, D., Liao, K., Morgan, M., (2007). "*Visualizing patterns in a global terrorism incident database. Environment and Planning B: Planning and Design, 34*, 767 – 784.

[8]. Wu, X. , Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A, Liu, B., Yu, P.S., Zhou, Z.H., Steinbach, M., Hand, D.J., Steinberg, D., "Top 10 algorithms in data mining". *Survey paper, Springer (2007).*

[9]. Quinlan, J. R. (1986). Induction of decision trees. *Journal of Machine Learning, 1*, 81-106.

[10]. Quinlan, J. R. (1993). C4.5: Programs for machine learning. *Machine Learning, 16,* 235-240. Springer.

[11]. McCallum, D. J., & Nigam. K. (1998). "A Comparison of event models for Naive Bayes text classification". Technical Report. *Workshop on learning for text categorization*. pp. 41–48.

[12]. Joachims, T. (2001). A statistical learning model of text classification for Support Vector Machines". *International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[13]. Vapnik, V. (1995). "*The nature of statistical theory*". Springer

[14]. Hall et al. (2009). The WEKA Data mining software: An Update. *SIGKDD Explorations, 11*(1).

[15]. Sebastiani, F. (2002). "Machine learning in automated text categorization. *ACM Computing surveys", 34*(1) pp. 1-47.

[16]. Nizamani, S., & Memon, N. (2012, January). *Semantic analysis of FBI news reports. In Neural Information Processing* (pp. 322-329). Springer Berlin Heidelberg.

[17]. Gildea, D., & Jurafsky, D. (2002). Automatic labeling of semantic roles. *Computational linguistics, 28*(3), 245-288.

ISSN: 2307-3721, e ISSN: 2307-3713