Middle-East Journal of Scientific Research 23 (6): 1215-1221, 2015
ISSN 1990-9233

DOI: 10.5829/idosi.mejsr.2015.23.06.22271

Named Entity Recognition in Crime News Documents Using Classifiers Combination

Hafedh Ali Shabat and Nazlia Omar

Center for AI Technology, FTSM, University Kebangsaan Malaysia,
UKM, 43000 Bangi Selangor, Malaysia

Abstract: The increasing volume of generated crime information readily available on the web makes the process of retrieving and analyzing and use of the valuable information in such texts manually a very difficult task. This work is focus on designing models for extracting crime-specific information from the Web. Thus, this paper proposes an ensemble framework for crime named entity recognition task. The main aim is to efficiently integrating feature sets and classification algorithms to synthesize a more accurate classification procedure. First, three well-known text classification algorithms, namely Naïve Bayes, Support Vector Machine and K-Nearest Neighbor classifiers, are employed as base-classifiers for each of the feature sets. Second, weighted voting ensemble method is used to combine theses three classifiers. To evaluate these models, a manually annotated data set that is obtained from BERNAMA is used. Experimental results demonstrate that using ensemble model is an effective way to combine different feature sets and classification algorithms for better classification performance. The ensemble model achieves an overall F-measure of 89.48% for identifying crime type and 93.36% for extracting crime-related entities. The results of the ensemble model trained with suitable features outperform baseline models.

Key words: Crime domain · Classifiers Combination method · Named Entity Recognition · Machine Learning

INTRODUCTION

With the increasing volume of crime information available on the Web, a means to retrieve and exploit relevant information is needed to provide insight into criminal behavior and networks in order to fight crime more efficiently and effectively. Designing an electronic system for crime named entity recognition and analysis from online news documents is necessary to assist the authorities in reducing the crime rate. In the crime domain, police and crime analysts need immediate information on certain crime cases to solve the crime or prevent it from happening again. Crime news documents consist of details on crimes and these details make these documents beneficial. The named entity recognition models extract this beneficial information quicker and with high and reliable accuracy [1].

Crime type identification and Named entity recognition are important information extraction tasks that deals with the recognition and classification of documents or tokens or sequences of tokens related to a particular class or entity. In the crime domain, police and crime analysts need immediate information on certain crime cases to solve the crime or prevent it from happening again. Crime news documents consist of details on crimes and these details make these documents beneficial. Many studies have been conducted on the performance of these methods based on general Newswire articles. However, the studies are limited for crime-domain.

In this paper, we present named entity recognition system based on ensemble framework for both crime named entity recognition and Crime type identification tasks. This system can recognize crime types (e.g., theft, murder, sex crimes, kidnapping and drugs) and extract entities (e.g., crime weapons, locations and nationality) from crime documents. The main aim of using ensemble framework is to synthesize a more accurate classification procedure. First, three well-known text classification algorithms, namely Naïve Bayes, Support Vector Machine and K-Nearest Neighbor classifiers, are employed as base-classifiers for each of the feature sets. Second, weighted voting ensemble method is used to combine theses three classifiers.

Corresponding Author: Hafedh Ali Shabat, Center for AI Technology, FTSM, University Kebangsaan Malaysia, UKM, 43000 Bangi Selangor, Malaysia.

Related Works: Several popular NE models employ a variety of techniques for the extraction of NE in the crime domain. [2] developed a model that utilizes neural networks to acquire useful information from unstructured crime documents and reports. The information retrieved is introduced into a database as the subsequent stage for other data and text extraction models to identify crime-related patterns. The extracted information is in a structured form and is essential for data mining systems [3]. In order to identify crime patterns and accelerate the crime solving process, [4] utilized a clustering algorithm. Subsequent to upgrading measures, the k-means clustering procedure was utilized to reinforce the means for ascertaining crime patterns. Genuine law enforcement information from a sheriff's office was utilized for the application of this procedure.

[5] created an information extraction IE system tailored for the crime domain. This system is capable of procuring crime information from police reports, witness reports and news-based documents. It retrieves information on people, weapons, vehicles, locations, time and clothes. An evaluation of this system was conducted with the utilization of two different formats of text documents, namely, police reports and witness reports. These reports were gathered from forums, blogs and news agency websites.

A prototype was developed for the identification of crime types in the Arabic context [6]. Two techniques were applied for the recognition processes: the first technique was completely dependent on direct identification with the utilization of gazetteers and the second technique is a rule-based model in which rules are constructed on the basis of a crime indicator list that includes various relevant keywords [7]. Developed a comparable procedure that recognizes other related information aside from the type of crime. This information includes the nationality of the victim and the location of the crime scene. This procedure includes an indicator to manage the Arabic language.

[8] Fashioned a technique that relies on natural language processing procedures and employs the Semantic Inferential Model. The system they developed is customized for using collaborative environments on the Internet. This method, called Wiki Crimes, is utilized to acquire two fundamental crime entities from online Web pages, namely, crime scene location and crime type. [9] Came up with an IE model that emphasizes solely on the extraction of information related to theft which includes the location of the crime scene, i.e. its address. The theft information is extracted from newspaper articles for three countries, namely, New Zealand, Australia and India. The model utilizes NER to determine if the sentence is inclusive of a crime scene location. The approach employed is the conditional random field which is a machine learning method used to verify the presence of information on crime scene location in a sentence.

The objective of this study is to develop a procedure for the mining of data from online crime news documents through the voting combination approach by merging the support vector machine (SVM), Naïve Bays (NB) and k-nearest neighbor (KNN) classifiers. This procedure can be employed for the identification of crime types (theft, murder, sex crime, kidnapping and drug) and the procurement of information from crime documents regarding the weapons used in the commitment of a crime, the location of the crime scene, the nationalities of those involved etc.

Research Design: This study presents an ensemble machine learning framework for both automatic crime text classification and crime NER system. The methodology consists of two main tasks: crime type identification and crime NER

The process starts with pre-processing procedures to eliminate flawed, noisy (meaningless) and sporadic data. In any case, the pre-processing of data is an operational requirement prior to the execution of other data mining procedures. To determine the superior perceptive terms for training and testing, several feature extraction techniques were put into operation. Lastly, a number of machine learning categorization procedures are used for the recognition of named entities and the identification of crime types. Figure 1 illustrates the architecture of the proposed framework in this work.

Language Resource Description: Using a supervised machine learning technique is dependent on the availability of annotated training data. Such data are usually created manually by humans or experts in the relevant field. To design a crime NER, the present study developed an annotated dataset from crime documents. Each word in the training corpus was labeled for crime weapons, crime location and nationality entities. The data used in this research were collected from the Malaysian National News Agency (BERNAMA). Type of crime, weapons, location of the crime and nationality involved were annotated and classified manually.

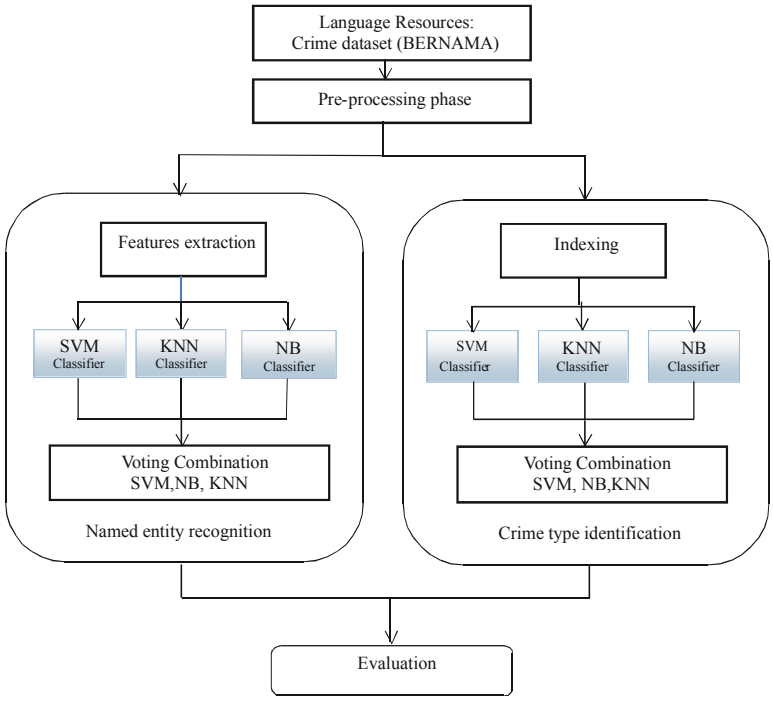

Fig. 1: Architecture of the proposed framework

Table 1: Sample of crime text annotated with POS tags

Word	Tag	Word	Tag	Word	Tag	Word	Tag
A	DT	by	IN	Bandar	NNP	tonight	RB
Security	NN	three	CD	Bare	NNP	Selangor	NNP
Guard	NN	armed	JJ	Ton	NNP	said	VBD
Was	VBD	men	NNS	Hussein	NNP	the	DT
Shot	VBN	shop	NN	On	NNP	……	……
Dead	JJ	in	IN	here	RB	……	…….

Pre-Processing: As the data were collected from Malaysian newspapers and social media sites, they generally included noisy data. Therefore, pre-processing the data is crucial using machine learning approaches. Before any NER and crime type can be identified, each crime document must pass through the pre-processing steps. In this system, during the pre-processing phase, crime type identification requires tokenization, stop word removal and stemming and NER requires tokenization and part of speech (POS) in the pre-processing phase. Table 1 shows a sample of crime text after the pre-processing phase.

Feature Extraction: In all classification procedures, feature extraction is crucial as it enhances the performance of classification tasks in relation to quickness and learning efficacy. The objective of feature extraction is the conversion of each word to a vector of feature values. An array of features was defined for the procurement of data from online sources regarding nationalities, weapons and crime scene locations. Subsequently, the grouping of these features under the three primary feature sets of (a) features established on POS tagging, (b) features established on word affixes and (c) features established on the context is carried out. These feature sets are also utilized for the representation of words in the corpus. Table 2 shows a summary of these feature sets used for the extraction of weapon, location and nationality, respectively.

Indexing: Type of crime is identified in this work through the classification of the crime documents. Therefore, the document is converted from a full text version to a document vector to make the document simpler and easier to deal with. Document representation a method used to decrease the complexity of documents and make them easier to handle. This process is accomplished by converting the complete text edition of the document into a document vector. The vector space model (VSM) is arguably the most frequently used document representation [10]. Text classification has a problem that automatically assigns unlabeled crime documents to predefined crime types. In the task of crime type classification, text representation transforms the content of the textural documents into a compact format, so that

Table 2: summary of features sets

Feature category	Feature name	Feature
Word affixes	F1	Prefix1
	F2	Prefix2
	F3	Prefix3
	F4	Suffix1
	F5	Suffix2
	F6	Suffix3
Context-based features	F7	Previous word(window size 2)
	F8	Next word (window size2)
	F9	Number of weapons indicator words before (size of window 7)
	F10	Number of weapons indicator words after (size of window 7)
	F11	Distance in words between the current word and indictor words before current word.
	F12	Distance in words between current word and indictor words after current word.
POS-based	F13	Is the part of speech of the word is Noun

the documents can be recognized and classified by a classifier [11]. In the VSM, a document is represented as a vector in the term spaces = (w_1,w_2,….,$w_{|V|}$), where |v| is the size of the vocabulary. The value of wi represents how much the term wi contributes to the semantics of the document d. Crime type classification as a text classification task borrows the traditional term weighting schemes from the information retrieval field, such as TF.IDF [12]. In the present study, identifying the type of crime is achieved by classifying the text.

Classification Algorithms: The majority of machine learning methodologies involve two stages. During the initial stage, training is conducted for the generation of a trained machine, while the subsequent stage entails classification. An appraisal of selected machine learning methodologies was conducted during the course of this study. However, for the acquirement of information on crime which includes the nationalities involved, the weapons used and crime scene locations through crime documents available online, this study settled on the following machine learning classifiers:

Support Vector Machine (SVM): This innovative machine learning procedure was recommended by [13]. In line with the conviction to lower structural threats related to computational learning conceptions, a decision surface utilized by the SVM bifurcates the training data points to come up with decisions that are ascertained by the support vectors. These support vectors are recognized as active components in the training set. While the generation of quite a few variations of the SVM have been recorded [14], this investigation has opted to focus solely on linear SVM. The choice of this technique is attributed to its reputation for good quality text classification [15]. SVM optimization (dual form) is expressed as such:

$$\overline{\alpha} = \arg\min\left\{-\sum_{i=1}^{n}\alpha_i + \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \langle \vec{x_i}, \vec{x_j}\rangle\right\} \quad (1)$$

$$Subject\ ot: \sum_{i=1}^{n}\alpha_i y_i = 0; \quad 0 \le \alpha_i \le C \quad (2)$$

Naive Bayes (NB): The NB procedure is frequently utilized for classifying reviews. Given a feature vector table, the algorithm determines the rearward possibility in which the review is linked to a range of classes and the algorithm allocates it to the class with the greatest rear potential. The NB procedure, which corresponds to a stochastic model of document fabrication, utilizes the Bayes rule. For the purpose of classifying class c* with the highest potential for a new document d, the calculation is as follows:

$$C^* = \text{argmaxc } P(C \mid d). \quad (3)$$

Exhibited below is the NB classifier computation for the posterior probability.

$$p(c_j \mid d_i) = \frac{p(c_j)p(c_j \mid d_i)}{p(d_i)} \quad (4)$$

K-Nearest Neighbour (KNN): The principal function of this supervised learning algorithm is the categorization of data according to its likeness to the predefined data. With the utilization of a variety of distance measurements the classifier gauges the likeness between an unclassified data object and the predefined data. Subsequently, the algorithm calculates the gap between the unclassified data object and the k nearest objects located in the predefined training dataset. The class majority of the k nearest neighbours is deemed the determined class for the unclassified data objects [16]. The Euclidean distance is frequently employed for distance measurement:

$$D_{Euclidean}(x,y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2} \quad (5)$$

where x = (x_1, x_2,..., x_m) and (y_1,y_2 ,..., y_m) signifies the m attributes of two samples.

Voting Combination: Voting algorithms take the outputs of some classifiers as the input and select a class that has been selected by most of the classifiers as the output. The voting rule counts the predictions of component classifiers and then assigns test sample x to class i with the most component predictions.

$$O_J = \sum_{k=1}^{D} I(argmax(O_{kj}) = j), \quad (6)$$

where I(•) is the indicator function. The sum rule combines the component outputs using the following equation:

$$O_J = \sum_{k=1}^{D} O_{kj}, \quad (7)$$

which is equivalent to the averaging of outputs over classifiers (average rule).

RESULTS AND DISCUSSION

To evaluate the proposed models, the data of the corpus were collected from the Malaysian National News Agency (BERNAMA). The types of crime, weapons, location of crime and nationality contained in these documents were annotated and classified manually. In this study using the cross-validation process, the corpus was randomly partitioned into 10 equal subsamples. A single subsample was retained as the validation data for testing the model and the remaining 9 subsamples were used as the training data. The cross-validation process was then repeated 10 times (10 folds). The training set represents the input values for the classification model of NB, SVM and KNN. The corpus represents the data entries in this model.

The standard evaluations of precision, recall and F-measure were used to evaluate the efficiency of named entity recognition capabilities available in the proposed model.Precision (P), Recall, F-Measure and Macro-average (F1) are the main criteria of assessing the effectiveness of the crime NERsystems [17].

Experimental Setting: In this study, a number of experiments were conducted to evaluate the performance of the proposed models. These experiments were conducted to identify the type of crime and to recognize the related named entities from the crime documents. In crime type identification, four experiments were conducted. The first experiment was performed using the NB classifier; the second experiment used the SVM classifier; the third experiment applied the KNN classifier; and the fourth experiment used the voting combination method. All these experiments were applied to identify five types of crime (theft, murder, sex crimes, kidnapping and drugs) by classifying the documents in the corpus, depending on the contents of the documents. In the NER, four experiments were also conducted. The first experiment was performed using the NB classifier; the second experiment used the SVM classifier; the third experiment applied the KNN classifier; and the fourth experiment used the voting combination method. All these experiments were applied to identify the entities (weapon, nationality and location of crime) from the crime documents. These experiments applied a set of features that include three types: word affixes, context-based features and POS-based.

Experiments for Crime Type Identification: In the crime type identification module, a training corpus consisting of 1402 crime documents was used. The data in the corpus were divided into five crime categories: theft, murder, sex crimes, kidnapping and drugs. The theft class contained 383 documents, accounting for 27.31% of the corpus; the murder class contained 298 documents, accounting for 21.25% of the corpus; the sex crimes class contained 299 documents, accounting for 21.32% of the corpus; the kidnaping class contained 200 documents, accounting for 14.26% of the corpus; and the drugs class contained 222 documents, accounting for 15.83% of the corpus. Four experiments are applied: the first experiment evaluates the NB classifier, the second experiment evaluates the SVM classifier, the third experiment evaluates the KNN classifier and the fourth experiment evaluates the classifier combination method. Table 3 shows a summary of the experimental results using the NB, SVM and KNN classifiers, as well as the voting algorithm, in identifying the type of crime.

Experiments for Classifying Named Entity: In this experiment, the overall performance of each individual classifier and combination classifiers in crime entity extraction was examined. The three classifiers, NB, SVM and KNN are applied to the entire feature space. After that, the classifiers combination method is evaluated.

Table 3: performance for each type of crime

	NB (%)	SVM (%)	KNN (%)	Voting Algorithm (%)
Theft	69.5	83.2	82.0	87.4
Murder	61.6	88.9	72.9	90.1
Sex Crimes	68.7	87.1	83.3	88.6
Kidnapping	64.4	81.2	84.7	87.4
Drugs	69.1	89.8	87.8	93.9
Macro-F-Measure	66.7	86.04	82.15	89.48

Table 4: performance for each entity type

	NB (%)	SVM (%)	KNN (%)	Voting Algorithm (%)
Weapons	86.73	91.08	82.35	93.3
Nationality	94.02	96.25	84.23	97.5
Location	87.66	89.28	81.62	89.28
Macro-F-Measure	89.47	92.20	82.73	93.36

Table 4 shows a summary of the experimental results using the NB, SVM and KNN classifiers, as well as the voting algorithm, in extracting nationality, weapon and crime location.

DISCUSSION

According to the experiments of identifying the type of crime, the highest result yielded by individual classifiers was by the SVM classifier with 86.04% accuracy and the lowest result was yielded by the NB classifier with 66.6% accuracy. The voting combination method gave a result of 89.48% accuracy, which was higher than that of all individual classifiers. Furthermore, according to the experiments of the crime named entity covered (weapon, nationality and crime locations), the highest result yielded by individual classifiers was by the SVM classifier with 92.2% accuracy and the lowest result was yielded by the KNN classifier with 82.73% accuracy. The voting combination method resulted in 93.36% accuracy, which was higher than that of all individual classifiers.

CONCLUSION

By evaluating the set of results obtained each time by applying a classifier, the highest accuracy was 89.48% for identifying crime types and 93.36% for identifying entities achieved using the combination method. The result illustrated that the proposed model was significant for identifying the type of crime and extracting the related named entities from crime documents. The research results were compared with those of the individual classifiers and confirmed the higher accuracy.

REFERENCES

1. Kumar, N. and P. Bhattacharyya, 2006. Named entity recognition in hindi using memm. the proceedings of Technical Report, IIT Bombay, India.
2. Chau, M., J.J. Xu and H. Chen, 2002. Extracting meaningful entities from police narrative reports. In Proceedings of the 2002 annual national conference on Digital government research, Digital Government Society of North America., pp: 1-5.
3. Hauck, R.V., H. Atabakhsb, P. Ongvasith, H. Gupta and H. Chen, 2002. Using Coplink to analyze criminal-justice data. Computer., 35: 30-37.
4. Nath, S.V., 2006. Crime pattern detection using data mining. InWeb Intelligence and Intelligent Agent Technology Workshops, 2006, WI-IAT 2006 Workshops, 2006 IEEE/WIC/ACM International Conference on, pp: 41-44.
5. Ku, C.H., A. Iriberri and G. Leroy, 2008. Crime information extraction from police and witness narrative reports. In Technologies for Homeland Security, 2008 IEEE Conference on., pp: 193-198.
6. Alruily, M., A. Ayesh and H. Zedan, 2009. Crime type document classification from arabic corpus. In Developments in eSystems Engineering (DESE), Second International Conference on., IEEE, pp: 153-159.
7. Alruily, M., A. Ayesh and H. Zedan, 2009. Crime type document classification from arabic corpus. In Developments in eSystems Engineering (DESE), Second International Conference on., IEEE, pp: 153-159.

8. Pinheiro, V., V. Furtado, T. Pequeno and D. Nogueira, 2010. Natural language processing based on semantic inferentialism for extracting crime information from text. In Intelligence and Security Informatics (ISI), 2010 IEEE International Conference on., pp: 19-24.
9. Arulanandam, R., B.T.R. Savarimuthu and M.A. Purvis, 2014. Extracting Crime Information from Online Newspaper Articles. In Proceedings of the Second Australasian Web Conference., 155: 31-38. Australian Computer Society, Inc.
10. Aas, K. and L. Eikvil, 1999. Text Categorization: A Survey. ISBN 82-539-0425-8.
11. Ko, Y., 2012. A Study of Term Weighting Schemes Using Class Information for Text Classification. In Proceedings of the 35th international ACM SIGIR conference on Research and Development in Information Retrieval, pp: 1029-1030.
12. Salton, G. and C. Buckley, 1988. Term-weighting approaches in automatic text retrieval. Information Processing and Management:an International Journal Process, Manage., pp: 513-523.
13. Cortes, C. and V. Vapnik, 1995. Support vector networks, Machine Learning, 20: 273-297.
14. Joachims, T., 1998. Text categorization with support vector machines: Learning with many relevant features, Springer Berlin Heidelberg, pp: 137-142.
15. Yang, Y. and X. Liu, 1999. A re-examination of text categorization methods. In Proceedings of the 22nd ACM annual International ACM SIGIR Conference on Research and Development in Information Retrieval., pp: 42-49.
16. Inyaem, U., P. Meesad and C. Haruechaiyasak, 2009. Named-entity techniques for terrorism event extraction and classification. In Natural Language Processing, 2009, SNLP'09, Eighth International Symposium on,. IEEE, pp: 175-179.
17. Manning, D.C., P. Raghavan and H. Schütze, 2008. Introduction to Information Retrieval. Cambridge University Press. ISBN 978-0-521-86571-5.