BAUSTJ
BAUST Journal

Crime Information Extraction and Classification from English Newspaper to Explore a Horizon for Bangla Language: A Systematic Literature Review

Salma Tabashum[1*], Ariful Islam[2], Fahmida Naznin Fami[3], Mun Yea Mahafi Taz Zahara[4] and Md. Mamun Hossain[5]

[*1, 2, 3, 4, 5]Department of CSE, Bangladesh Army University of Science and Technology, Saidpur, Bangladesh

emails: *[*1]baridhi.tabassum@gmail.com; [2]anjum.arif@gmail.com; [3]nazninfami@gmail.com; [4]rgcdola@gmail.com; and [5]mamunsust12@gmail.com ('*' - corresponding author)*

ARTICLE INFO

Article History:

Received: 11th September 2019
Revised: 25th July 2020
Accepted: 27th July 2020
Published online: 25th August 2020

Keywords:

Deep Learning
Word2Vector
Steamer
CNN
Tokenize
Bangla Crime Classifier

ABSTRACT

In this research, firstly we have reviewed that how crime related data can be analyzed and extracted the crime information from online English news articles with the intention to identify crime information for Bangla language. Secondly, we have tried to categorize the tools, techniques, technologies and algorithms required to analyze crime data for English language. Then we analyzed whether these technologies can be applied to extract crime information from Bangla newspaper. Finally, we propose a model to analyze crime data for Bangla language.

1. INTRODUCTION

Crime data analysis is a systematic analysis for detecting and analyzing various types of crimes and classify its patterns (Dasgupta et al., 2017) (Alshuwaier et al., 2013). These patterns play an important role for solving different crime types, problems and in making different strategies to solve the crime problems. Different types of news articles of online newspapers publish thousands of crime news which contain the details of victims, crime type, criminals, locations etc. By extracting relevant information from the news sources, the enforcement agencies analyze these phenomena for all relevant factors, such as evaluation of the crime dictionaries, prediction of the occurrence of future crime, alert to their agencies for future crimes (Jayaweera et al., 2015)(Sathyadevan et al., 2014). Many studies have discovered various techniques to investigate the crime data.

Information extraction can be done using Name Entity Recognition (NER) (Chen et al., 2004) (Zhang et al., 2013). NER classify named entities into pre-defined categories such as the person names, organizations, locations, time expressions, quantities, percentages, etc. NER is a part of natural language processing (NLP). An article may contain detailed information. But to choose the informative data and information we may need classifier. A classifier can be built using classic machine learning algorithms like Naïve Bayes, Logistic Regressions, Support Vector Machine (Michailidis et al., 2006) (Wang et al., 2006) or deep learning algorithms like Convolutional Neural Network (CNN) or Recurrent Neural Network (RNN).

2. REVIEW METHODOLOGY

We categorize the review methodology into three main sections. First section describes the research work related to the crime data analysis. Second section involves the comparisons among the methods that are used in the reviewed papers. Third section involves the presentation of summary of researches in crime data analysis.

A. Searching Procedure

A wide range of view is necessary for obtaining the maximum coverage of the literature. Before starting the search, a reliable and suitable source must be chosen in order to find the most relevant articles. In this case Google Scholar is used as a search engine and also observed these following digital libraries:

- IEEE Explore (http://ieeexplore.ieee.org)
- ACM Digital Library (www.acm.org/dl)
- Springer (www.springerlink.com)

This search engine and libraries searches various journals and research papers based on given key words. We used the following keywords:

- Crime data analysis.
- Crime data extraction.
- Crime analysis using machine learning.
- Crime analytics from newspaper.

B. Selection Criteria

At first, we read the titles of the journals that were found based on the keyword search result. From the list we selected those journals which seemed appropriate for our research. Then we narrow downed the list by reading the abstract of the selected papers. The papers which were interesting to us, we collected the whole paper by downloading these. Then the papers were read entirely to select the final list of papers based on their contents.

3. CLASSIFICATION AND DATA EXTRACTION METHODS

The research papers are classified according to the use of different technologies, methods, models, updated tools etc.

- The authors used data sets and data sources which were maximum from online newspapers and their languages are English and Arabic for classification and extraction of data (Sun et al., 2008) (Alruily et al., 2009)
- After collecting data from different sources, the authors used different methods and techniques to classify appropriate news and crime data. Some used CRF based classifier, some SVM etc.
- Then they used NER models for training and getting the name and entities like persons involved in crimes (victims, criminals), crime types, locations, date, time etc.
- In some researches, they also visualized the crime locations in map which are informative in various charts like graph charts, pie charts, bar graph etc.

4. COMPARISION OF TECHNOLOGIES AND TOOLS USED IN CRIME ANALAYSIS AMONG THE PAPERS

The procedure and all the methods used in the papers for data extraction and classification are categorized in four stages.

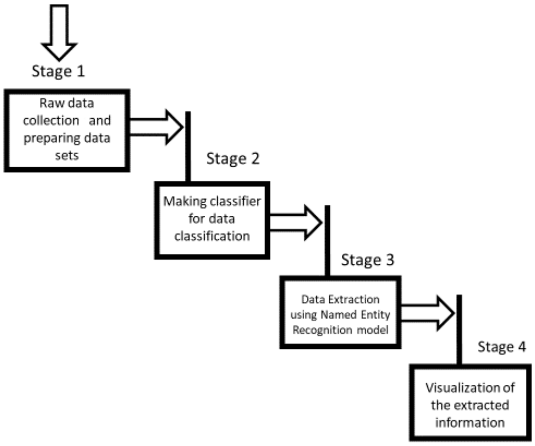
Stage 1
Raw data collection and preparing data sets
Stage 2
Making classifier for data classification
Stage 3
Data Extraction using Named Entity Recognition model
Stage 4
Visualization of the extracted information

Figure 1: Working procedure for information extraction.

Raw Data Sources:

- **Open Sources and Online Newspapers:** Open source intelligence findings are extracted (Dasgupta et al., 2017) from the net online newspapers. This data is in unstructured format.
- **Intelligence Agency Reports:** Intelligence agencies maintain information concerning the criminals. Some Intelligence agencies (Chandra et al., 2017) are Research and Analysis (RAW), Intelligence Bureau (IB), Narcotics Control Bureau (NCB).
- **Previous Investigation:** Files and Police Reports after distinguishing the suspects the police evoke previous investigation files of the intended suspect (Sathyadevan et al., 2014) which are available in text, photo, CCTV bank account phone call, email send-receive records, witness and victim statements etc. Police reports contain info concerning the crime suspect etc. These are one of most the reliable sources for assembling crime data.
- **WebCrawler:** Lots of researchers (Das & Das, 2017) (Rohini & Isakki, 2016) used web crawler to crawl data from online sources.

Classification Techniques:

- Semi-supervised learning technique to learn Different categories of crime events from the News documents (Dasgupta et al., 2017).
- SVM based classification models to learn crime related concept (Hassan & Rahman, 2017)

(Rohini & Isakki, 2016) (Alruily et al., 2009) (Al-Shoukry & Omar, 2015) (Cortes & Vapnik, 1995) (Joachims, 1998) (Yang & Liu, 1999) (Joachims, 2002) (Michailidis et al., 2006).

- Document clustering is an unsupervised process of grouping documents into different groups called cluster. The uncategorized documents are grouped into the meaningful clusters without any prior information (Hassan & Rahman, 2017).
- The CNN applies a linear transformation to all K -windows in the given sequence of vectors (Das & Das, 2017).
- K-Nearest Neighbor (KNN) This procedure can be employed for the identification of crime types (Al-Shoukry & Omar, 2015).

Information Extraction Technique:

- The Stanford Named Entity Recognizer (Dasgupta et al., 2017) to extract potential named entities like Organization, Location, Date, Time etc. The Stanford Deterministic Co-reference Resolution System, which implements a multi-pass sieve algorithm, is used via the Stanford Core NLP tool suite (Dasgupta et al., 2017) (Maynard et al., 2001).
- The named entity recognition method by collecting location names (city, district, division) from downloaded newspaper dataset (Hassan & Rahman, 2017).
- A Python program has been implemented where Noun Phrase Chunking (NPChunking) (Pedregosa et al., 2011) is taken into account and it searches for chunks related to each noun phrase (Das & Das, 2017). The result is also compared with the data collected from National Crime Records Bureau.

Table 1

Comparison of technologies and tools used in crime analysis)

References Paper	Technologies	Tools	Accuracy
T. Dasgupta, A. Naskar, R. Saha and L. Dey, "Crime Profiler: Crime Information Extraction and Visualization from News Media".(Dasgupta et al., 2017)	syntactic parser module morphological analysis module named entity recognizer (NER) module pronoun resolution module Entity resolution module Parts-of-Speech (POS) tagging relevant information extraction BMI measure	N/A	60% -71%
M. Hassan and M. Rahman, "Crime News Analysis: Location and Story Detection", in *2017 20th International Conference of Computer and Information Technology (ICCIT)*. (Hassan & Rahman, 2017)	Data Collection module Preprocessing module crime pattern recognition (Named Entity Recognition) module Recognizing the Basic Named Entities modus operandi features	Selenium and newspaper NLTK (Bird et al., 2009) Support Vector Machine (SVM) Hierarchical clustering cosine similarity The Term Frequency-Inverse Document Frequency	NA
P. Das and A. Das, "A Two-stage Approach of Named-Entity Recognition for Crime Analysis", in 8th ICCCNT 2017, India. (Das & Das, 2017)	Data Collection module Preprocessing module crime pattern recognition (Named Entity Recognition) module Recognizing the Basic Named Entities modus operandi features	Web Crawling Google Geocoding API Noun Phrase Chunking	NA

T. Dasgupta, A. Naskar, R. Saha and L. Dey, "Automatic Curation and Visualization of Crime Related Information from Incrementally Crawled Multi-source News Reports", in 27th International Conference on Computational Linguistics: System Demonstrations,2018.(Dasgupta et al., 2018)	Convolution Recurrent Neural Network (CRNN) Crime data extraction module	GloVe word vector representation Bi-LSTM BMI tool	NA
D. ROHINI and D. DEVI, "Crime Analysis and Mapping through Online Newspapers: A Survey", 2016. (Rohini & Isakki, 2016)	Document Classification Module Information Extraction Duplicate Detection Crime Analysis and Mapping Crime prediction Module	SVM (Hsu et al., 2003) (support vector machine) GATE (General Architecture for Text Engineering) murmur # implementation Hamming distance Hot spot detection, Crime comparison, Crime pattern visualization	NA
M. Alruily, A. Ayesh and H. Zedan, "Crime Type Document Classification from Arabic Corpus", 2009. (Alruily et al., 2009)	Data Collection Module Named Entity Recognition (NER) Crime Type Recognition System (CTRS) pattern identification Parts of Speech (POS) tagger Multilingual Crime Type Recognition System (MCTRS) Crime domain	Corpus Gazetteers Rule-based approach Poibeau, Nath designed Byaraktar & Temizel developed system Support Vector Machine (SVM) Knowledge Base (KB)	60% - 77%
H. Shabat and N. Omar, "Named Entity Recognition in Crime News Documents Using Classifiers Combination", in Middle-East Journal of Scientific Research, 2015. (Al-Shoukry & Omar, 2015)	Data Collection Module Named Entity Recognition (NER) Information extraction (IE) system Crime type identification Module Pre-Processing Module	Mozenda, Corpus PunktTokenizer from NLTK entity chunker, Stanford NER, Gazetteer LBJ NER Tagger, Nadeau & Sekine-2007	78.5% - 82%

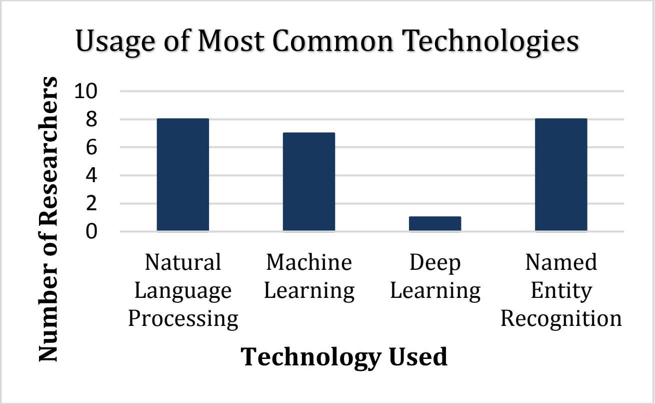

Figure 2: Use of most common technologies in crime analysis.

5. EXPLORING A HORIZON FOR BANGLA LANGUAGE

All the researches in crime data analysis and extracting information are performed in English, Arabic language etc. But no researches are done in Bangla language. If we analyze our Bangla news sources like online newspapers, we can see that from just one newspaper site, with a seven-day range, we find almost 75 crime news. This shows us the excessive crime related problems occurring in our country. As most of the criminals and crimes have Bangla as a common communication media, we have decided to take the path of solving Bangla crimes, as Bangla newspaper is an easy and valid data source of Bangla crimes. For analyzing the crime data from Bangla Newspaper, the whole procedure can be divided into some categories:

- Dataset: As no work is done in Bangla Language, so we will make our own data set.
- Classifier: We will make a crime detecting classifier using Deep learning algorithms like

Convolutional neural Network (CNN), Recurrent Neural Network (RNN) etc.

- NER Model: To extract the crime information from newspaper, we need NER (Name Entity Recognizer) model. As we don't have any NRE for Bangla language, we will build our own trained NER model to extract relevant information like location, time, crime types, Crime related Persons etc.

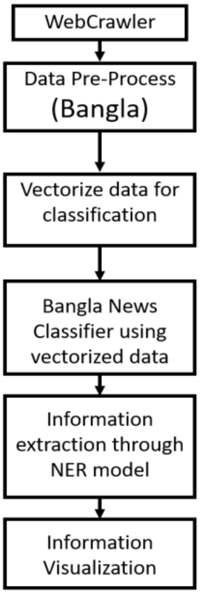
WebCrawler
Data Pre-Process (Bangla)
Vectorize data for classification
Bangla News Classifier using vectorized data
Information extraction through NER model
Information Visualization

Figure 3: Use of most common technologies in crime analysis.

6. CONCLUSION

With increasing number of populations, increasing number of crimes is happening everywhere around us. So, crime has become very common in our country as well as in the world. If we open a newspaper, we find crime is occurring every now and then throughout the globe. So, crime information extraction has become a rudimentary need; especially in our country where sophisticated technology to detect unlawful activities is very rare. Due to the emergence in the field of data mining, especially in the field of crime data analysis, the research has become widespread. In the newspapers, we can find the similar crime is occurring again and again in the same place and sometimes the suspected is the same person. So, we can analyze, investigate and predict crime type, crime zone and also the criminal for future protection. With this view, this paper reviews so many online newspapers, articles, books, journals, and magazine and tried to identify the tools, techniques, technologies, algorithms required to achieve the goal for Bangla as most of the research are carried on for English newspaper. An efficient method for crime prediction can be developed using these analyses. This information is also needed to propose a model to extract crime information from online Bangla newspaper.

REFERENCES

Alruily, M., Ayesh, A., & Zedan, H. (2009). Crime type document classification from Arabic corpus. *2009 Second International Conference on Developments in eSystems Engineering*. https://doi.org/10.1109/dese.2009.50

Al-Shoukry, S. A., & Omar, N. (2015). Arabic named entity recognition for crime documents using classifiers combination. *International Review on Computers and Software (IRECOS)*, *10*(6), 628. https://doi.org/10.15866/irecos.v10i6.6767

Alshuwaier, F. A., Almutairi, W. A., & Areshey, A. M. (2013). Smart search tools using named entity recognition. *2013 International Conference on Information Technology and Applications*. https://doi.org/10.1109/ita.2013.78

Arulanandam, R., Purvis, M. A., & Savarimuthu, B. T. (2014). Extracting crime information from online newspaper articles. In *Proceedings of the second Australasian web conference (AWC 2014): Auckland, New Zealand, 20 - 23 January 2014* (pp. 31-38).

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media.

Chandra, B., Gupta, M., & Gupta, M. P. (2017). Adaptive query interface for mining crime data. *Databases in Networked Information Systems*, 285-296. https://doi.org/10.1007/978-3-540-75512-8_20

Chen, H., Chung, W., Xu, J., Wang, G., Qin, Y., & Chau, M. (2004). Crime data mining: A general framework and some examples. *Computer*, *37*(4), 50-56. https://doi.org/10.1109/mc.2004.1297301

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

Das, P., & Das, A. K. (2017). A two-stage approach of named-entity recognition for crime analysis. In *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-5). IEEE.

Dasgupta, T., Dey, L., Saha, R., & Naskar, A. (2018). Automatic Curation and Visualization of Crime Related Information from Incrementally Crawled Multi-source News Reports. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations* (pp. 103-107).

Dasgupta, T., Naskar, A., Saha, R., & Dey, L. (2017). CrimeProfiler: crime information extraction and visualization from news media. *Proceedings of the International Conference on Web Intelligence*. https://doi.org/10.1145/3106426.3106476

Feldman, R., & Sanger, J. (2007). *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge University Press.

Gu, Y., Shen, S., Wang, J., & Kim, J. (2015). Application of *NoSQL* database MongoDB. *2015 IEEE International Conference on Consumer Electronics - Taiwan*, 158-159. https://doi.org/10.1109/icce-tw.2015.7216831

Hassan, M., & Rahman, M. Z. (2017). Crime news analysis: Location and story detection. *2017 20th International*

Conference of Computer and Information Technology (ICCIT). https://doi.org/10.1109/iccitechn.2017.8281798

Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). A practical guide to support vector classification.

Jayaweera, I., Sajeewa, C., Liyanage, S., Wijewardane, T., Perera, I., & Wijayasiri, A. (2015). Crime analytics: Analysis of crimes through newspaper articles. *2015 Moratuwa Engineering Research Conference (MERCon)*. https://doi.org/10.1109/mercon.2015.7112359

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *Machine Learning: ECML-98*, 137-142. https://doi.org/10.1007/bfb0026683

Joachims, T. (2002). A statistical learning model of text classification for SVMs. *Learning to Classify Text Using Support Vector Machines*, 45-74. https://doi.org/10.1007/978-1-4615-0907-3_4

Maynard, D., Tablan, V., Ursu, C., *Cunningham*, H., & Wilks, Y. (2001). Named Entity Recognition from Diverse Text Types. In *Proceedings of the Recent Advances in Natural Language Processing 2001 Conference* (pp. 257-274).

Michailidis, I., Diamantaras, K. I., Vasileiadis, S., & Frère, Y. (2006). Greek Named Entity Recognition using Support Vector Machines, Maximum Entropy and Onetime. In *LREC* (pp. 47-52).

Ou-Yang, L. (2013). Newspaper: Article scraping & curation. *Python Library. Retrieved*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

Rohini, D. V., & Isakki, P. (2016). Crime analysis and mapping through online newspapers: A survey. *2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16)*. https://doi.org/10.1109/icctide.2016.7725331

Sathyadevan, S., S, D. M., & S., S. G. (2014). Crime analysis and prediction using data mining. *2014 First International Conference on Networks & Soft Computing (ICNSC2014)*, 406-412. https://doi.org/10.1109/cnsc.2014.6906719

SeleniumHQ Browser Automation. (Extracted on Sept. 14, 2017). Source: https://www.seleniumhq.org

Shaalan, K., & Raza, H. (2008). Arabic named entity recognition from diverse text types. *Advances in Natural Language Processing*, 440-451. https://doi.org/10.1007/978-3-540-85287-2_42

Shah, F. P., & Patel, V. (2016). A review on feature selection and feature extraction for text classification. *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*. https://doi.org/10.1109/wispnet.2016.7566545

Sun, H., Liu, Z., & Kong, L. (2008). A document clustering method based on hierarchical algorithm with model clustering. *22nd International Conference on Advanced Information Networking and Applications - Workshops (aina workshops 2008)*. https://doi.org/10.1109/waina.2008.45

Wang, G., Chen, H., Xu, J., & Atabakhsh, H. (2006). Automatically detecting criminal identity deception: An adaptive detection algorithm. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *36*(5), 988-999. https://doi.org/10.1109/tsmca.2006.871799

Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '99*. https://doi.org/10.1145/312624.312647

Zhang, H. P., Liu, Q., Yu, H. K., Cheng, X., & Bai, S. (2013). Chinese named entity recognition using role model. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 8, Number 2, August 2003* (pp. 29-60).