Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Text analysis framework for understanding cyber-crimes

Clinton Cardoza *, Rupali Wagh

Department of Computer Science, Christ University, Bengaluru-560029, India

ARTICLE INFO

Article history:
Received 17 February 2017
Received in revised form
11 August 2017
Accepted 19 August 2017

Keywords:
Natural language processing
POS tagging
Cybercrime
Text analysis
Term association analysis

ABSTRACT

The magnitude of impact of cyber-crimes is much greater as compared to other crimes and can be felt at personal, societal, national as well as global level. According to studies, developing countries are at a greater risk due to such crimes. Fight against cyber-crime requires a strategic and intelligent framework. This paper discusses text analysis framework using Natural Language Processing (NLP) and text mining techniques to extract crime related information which can be used for educating and spreading awareness and for further knowledge based analysis. News articles crawled from a leading newspaper website in India is used as the source of cyber-crime data. Parts of Speech (POS) tagging is used to extract important terms/concepts related to cybercrimes. Term association analysis on the other hand is used to understand the relationship of extracted terms of the data.

1. Introduction

Technological advents and World Wide Web (WWW) has revolutionized almost every sphere of human activities. Not surprisingly it is also the advent of cyber-crime activities worldwide. There is a steep surge in criminal activities that are carried out using internet and related technologies. Cybercrime is an umbrella term which includes varied set of criminal activities. Fraudulent operations that are carried online and compromise security; via bank transactions for financial fraud or misusing corporate information are the most widely reported cyber-crimes. Other forms of crime related to social media include harassment, fraud, theft, child trafficking, etc. Drug trafficking is another type of crime which falls into this category which involves illegal distribution and sale of drugs online. Cyber terrorism which includes anonymity, holding sensitive information, drug trafficking, uploading offensive content which includes sensitive content in terms of pictures and videos and harassment are emerging as threats to the society and are posing new challenges before law enforcement agencies.

In the past, cyber-crime attacks and threats were carried out by a group of individuals. Today we observe highly sophisticated cyber-criminal networks which bring together individual offenders from various parts of the world to engage in cyber-crime attacks in real time. Crime analysis is a law enforcement function that involves systematic analysis for identifying and analyzing patterns and trends in crime and disorder.

Literature defines it as the process in which a set of quantitative and qualitative techniques are used to analyze data valuable to police agencies and their communities (IACA, 2014). This paper describes various functions, processes and related agencies that work together to fight crime. Crime analysis today is supported by various intelligent computational techniques. This equips law enforcement agencies in fighting against crime in a better way. Crime can be analyzed based on the data recorded for committed crime. This data may be available as police record, as FIR or even as court case documents. This information can be structured as well as unstructured. Approaches for crime analysis are broadly divided into two categories – 1) Data mining based solutions that focus on mainly on available structured information and perform descriptive and predictive analysis. 2) Natural Language processing (NLP) based solutions that extract useful information from crime related documents to find patterns and trends in crime. Named Entity Recognition (NER) is the most common NLP technique used in crime analysis. Computational assistance provided in crime analysis ranges from being specific like predicting crime locations and identifying habitual criminals to very generic like information extraction from crime related documents. Predictive policing, one of the

* Corresponding Author.
Email Address: clinton.cardoza@mca.christuniversity.in (C. Cardoza)

https://doi.org/10.21833/ijaas.2017.010.010
2313-626X/© 2017 The Authors. Published by IASE.

key objective of crime analysis, involves identification of criminal activities by analyzing content extracted from various resources which help the authorities track and correct the growth of such activities. Organizations such as the Interpol, F.B.I and CAO use predictive policing to detect and fix crime related attacks and activities.

Cybercrimes are different from other crime types and cybercrime analysis plays an important role in taking proactive measures both for crime control and crime information dissemination. In India where digitization has gained momentum very recently, availability and access to detailed records about cybercrimes in public domain is relatively low. But considering the nature, impact and magnitude of cyber related crimes, social awareness regarding these types of crimes has become very important and is seen as a vital component in fighting against cybercrime. In this study, news articles related to cybercrime are used as data set for analysis. This work proposes text analytics framework for information and pattern extraction from crime related news articles. First step of preliminary uses natural language technique, POS tagging to extract cyber-crime related information. Basic text mining framework is then used to add more insights into the crime-data.

The paper is organized as follows – Section 2 provides an overview of work done in the field of crime analysis for both structured as well as unstructured data. It highlights major approaches, techniques and their significance in crime control. Section 3 explains the methodology of the proposed work with the description of all steps involved. Section 4 presents results of analysis and their interpretation. Section 5 mentions the future direction of the proposed framework.

2. Literature survey

In the past few years there is a paradigm shift from information processing to intelligent processing in the functioning of all domains. Crime analysis also has undergone various transitions. Law enforcing agencies and forensic experts are now banking on computational intelligence. Data mining framework proposed for crime data analysis (Chen et al., 2004) is considered as a pioneering step in application of intelligent techniques for crime data. The work emphasized on unsupervised techniques like clustering, association rule mining and outlier analysis for providing insights into crime information. While clustering is used for grouping similar crimes, crimes that are committed together frequently are detected using association rule mining. Outlier analysis is applied for deviation detection to identify fraudulent operations. Application of classification for crime pattern detection is also discussed. Based on this general framework, many intelligent solutions for particular crime category were proposed. Crime pattern detection based on prisoner's data using unsupervised techniques and k means algorithm is presented by Nath (2006). Interesting analysis on finding similar modus operandi in crime and detecting crime pattern is discussed by Wang et al. (2013). Crime pattern is defined with crime specific parameters and modus operandi parameters and a supervised learning approach to detect and crime pattern is suggested in the paper. Extensive analysis of crime data particularly related with murder cases in India is presented by Chakravorty et al. (2015). This work considered both structured and unstructured data from crime reporting agency, National Crime Record Bureau (NCRB). Structured data analysis was used for location based trends of crime and motives behind murder crimes. Unstructured data is processed using NLP techniques for extraction of crime related information.

Crime analysis using other sources of data like police reports, news articles and even victim interviews have been employed in various studies due to unavailability such extensive crime data. News articles are used most widely as crime data for analysis and thus NLP techniques along with data mining approaches in this field of crime data analysis are used prevalently. Crawling newspaper websites for crime news articles and further processing them for crime information extraction is proposed by Jayaweera et al. (2015). This paper suggests a multi-phase framework for crime analysis of news article data for extraction of documents, classification of documents as crime and non-crime documents, crime entity extraction and machine learning approaches for further crime analysis. Named entity extraction (NER) is obvious choice for crime information extraction from documents. Arulanandam et al. (2014) discusses two approaches NER and conditional random field and their performance in identification of crime location sentences. Tayal et al. (2015) presents simple statistics about cybercrimes in Taiwan and highlights the need to take measures to control cybercrime at various levels of society like schools, colleges social agencies and governments. This paper implicitly has emphasized on the impact of cyber or computer related crime. These crimes are different from other crime categories due to their spread and impact. Anybody from corporate giant to common man, government agencies to schoolchildren is vulnerable to these crimes. This requires generic analysis of cybercrimes which can aid in general awareness and education about these types of crimes. Cyber security capability model (CM^2) is an important initiative in today's technology connected world which allows organizations and countries to assess their preparedness against cybercrimes (Barclay, 2014). Society, technical, operational, business, legal and regulatory, and education capability building are stated as pillars of Cyber security capability model. This paper also emphasizes on the importance of education and capacity building for sustainable cyber security especially in developing countries. Parts of Speech tagging is an NLP technique that identifies syntactic category of words in a sentence. POS tagging is used in various domains for enhancing domain knowledge and ontology enrichment (Aubin and Hamon, 2006). It can be used for identifying important words in software bug reports (Tian and Lo, 2015). Researchers have suggested many variations in basic POS tagging approaches to enhance the performance of domain specific term extraction. Correlation analysis is used popularly in all domains to quantify associations among extracted features (Alsuhaibani et al., 2015; Omar et al., 2004). Park (2008) elaborated various analysis types and also mentions that correlation analysis is used as one of major quantitative technique in this domain.

This paper aims at extracting cyber-crime related information from news articles to provide understanding of crime to the general public in contrast to in depth crime analysis and predictive policing. POS tagging is used for identification and extraction of information which is further analyzed using statistical measures like term frequency and correlation analysis.

3. Proposed system

The objective of this work is to identify and extract cybercrime related information from news article corpus and perform preliminary statistical analysis to identify associations of terms in extracted data.

This section of the paper describes the methodology used in the study. Fig. 1 shows the steps of the proposed framework.

3.1. Data collection and corpus building

The dataset in terms of cybercrime incidents and reports is retrieved using a custom-built web crawler from a popular Indian news based website Times of India (http://timesofindia.indiatimes .com/). Since the cybercrime records are available in the form of elaborate articles, the contents of the same are extracted page wise and structured for better understanding. There are primarily two dedicated Python packages used in this work i.e. requests and BeautifulSoup. The "requests" package helps the user to analyze and obtain web content based on the website's URL while providing access to the response data. The "BeautifulSoup" package provides methods to navigate and search web content, structures the text content suitable for enrichment and further processing while removing all the web-based dependencies and presents legible information. The main objective of the web crawler is to extract all the articles pertaining to cybercrime attacks and related information from the current date to the last available date page wise and applying a set of formatting features to structure the contents of the articles. With the help of the web crawler, the data is extracted specific to the user's query which in this case is cybercrime records. The crawler then requires two pieces of information i.e. a container class and a navigation class. All the relevant cybercrime articles are bound to a specific area within the container class while the navigation class provides a set of hyperlinks to assist the crawler to navigate through all the pages of the search. Since the crawler is implemented in Python, it is possible to structure the information extracted using a method from Beautiful Soup named prettify which is suitable for further processing.

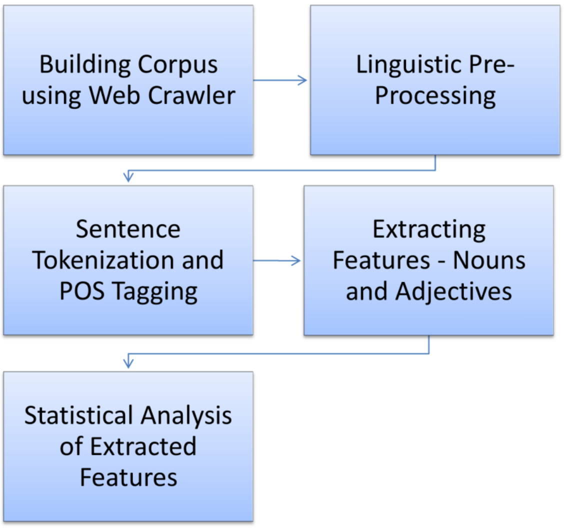
Building Corpus using Web Crawler
Linguistic Pre-Processing
Sentence Tokenization and POS Tagging
Extracting Features - Nouns and Adjectives
Statistical Analysis of Extracted Features

Fig. 1: Steps used in methodology

3.2. Linguistic pre-processing

Linguistic pre-processing to clean the document corpus and remove unwanted words is carried out. This primarily includes stop word removal and Document Stemming. Word categories such as symbols and connectors present in the articles are treated as irrelevant information and can be removed using a stop-word removal function. In this case a normal list of stop-words and another identical list converted to camel case are referenced in the function to remove all the stop-words from the articles. This is aid uniform processing of each article which serves as a pre-requisite in information extraction. In most cases, there might be words present in the articles which hold the same meaning but exist in different tenses. In order to simplify and normalize such kind of words, we apply stemming. For this work, we use Lancaster Stemmer which normalizes words with a suitably better accuracy which is much needed for analysis.

3.3. Sentence tokenization and POS tagging

Once all the stop words have been removed from the content of the articles, the information need to be split into individual sentences for tokenization. In this regard, we used the default training set of the Punkt Sentence Tokenizer package developed by Kiss and Strunk from the NLTK corpus. The advantage of using this package is that the entities with respect to names of persons, organizations and locations can be identified after the NER algorithm has been invoked on them.

It is a pre-requisite to identify all the entities based on their grammar classification. POS tagging enables the article to be tagged by parts of speech. Consider the sentence-"Experts predicted digital awareness 2018". If this were to be subjected to P.O.S tagging, it would be converted to "PERSON Experts/NNS predicted/VBD digital/JJ awareness/JJ 2018/CD" where NNS is a plural noun, JJ is an adjective and CD is a cardinal number. It is observed that every word in the sentence bears a unique tag which helps to narrow down search results. By committing to POS tagging, entities bearing credentials of persons, locations and organizations with respect to the cyber-crime history can be identified and used for further analysis.

3.4. Noun and adjective extraction

Proper nouns, common nouns and adjectives are extracted into a text file as step 1 from the refined content of the articles processed. This step is carried out to ensure that all the relevant nouns are identified before further processing and to establish meaning and connectors with the help of adjectives. Table 1 depicts the details of the dataset considered for analysis.

Table 1: Dataset details

Document Attribute	Value
Documents	30
Total number of sentences	107
Average number of words per document	65
Average number of nouns (Proper / Common) identified per document	27
Average number of adjectives (Comparative / Superlative) identified per document	6

3.5. Statistical analysis

Correlation analysis is a simple yet very powerful statistical measure which is popularly used for studying the strength of relationship among various terms in your data. In this study correlation coefficient is used to measure association between various terms. Adjectives and Nouns extracted in step 2 forms the final feature set for further analysis. The following steps are ensured to get the end results.

- Construction of document term matrix
- Finding frequent terms
- Finding strongly associated words that appear together frequently.

4. Results and discussion

4.1. Visualization of prominent terms from the data

Terms represent concepts of a document. Simple techniques and visualization namely most frequent words and word cloud are used to get first level understanding of prominent concepts/terms.

Fig. 2 indicates youth, cheat, sexual and Facebook are a few important concepts that can be highlighted from the data.

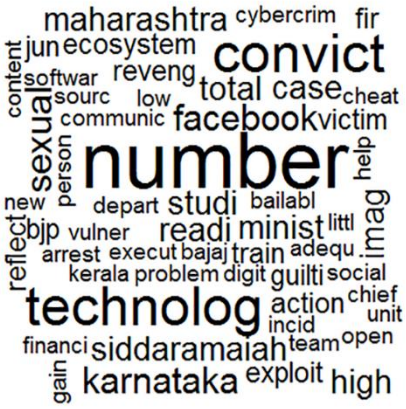

Fig. 2: Representation of terms as Wordcloud

Most frequent words plotted along with their frequency emphasize on concepts such as Facebook, online, sexual crimes and rise of their numbers as important terms.

Fig. 3 also shows that few state names (Uttar Pradesh) appear very frequently in the data which can be due to high cybercrime rates or high case reporting related to cybercrimes in that state.

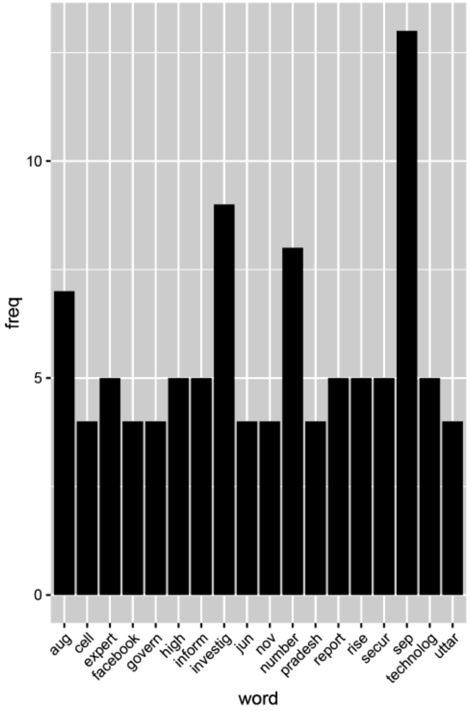

Fig. 3: Frequent words / terms

These preliminary representations of data suggest certain types of cybercrimes are more prevalent in the case reported.

To identify which two concepts are strongly related, term association analysis (which uses correlation coefficient) was applied to major concepts identified in previous step. Following graphs in Fig. 4 show terms highly correlated with “youth”, ”Facebook”, ”cheat”, and “sexual”.

The results obtained are indicative of the fact that many interrelationships exist in the terms extracted as important terms related to cybercrime. Since the corpus size is small and the above relationships are just representative, detailed association analysis can be used for understanding their relations.

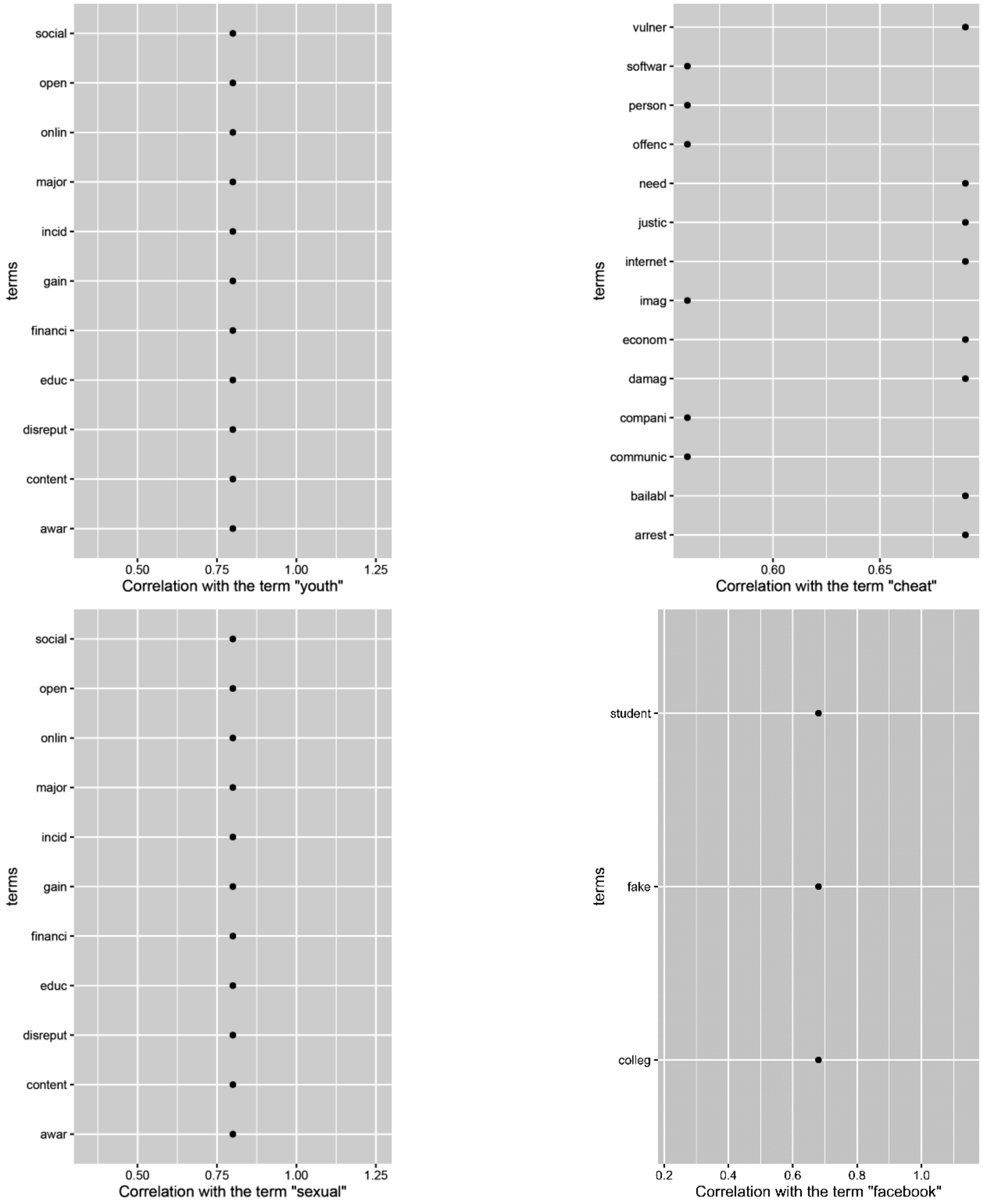

Fig. 5: Term associations from the data

5. Conclusion and future work

The work presented in this paper is a first step towards intelligent analysis of news articles for pattern detection from crime data. This work mainly focuses on extraction of crime information using simple techniques – POS tagging and statistical analysis. Results show that this proposed framework can be used for getting knowledge from raw data presented in the news articles.

As mentioned in the earlier sections this information can be vital in educating/spreading awareness in the society, though low cybercrime reporting rate in India remains a challenge till date.

These insights obtained pave way for more intelligent analysis in this domain. This framework can further be extended using NER to study/analyze related concepts. Unsupervised learning approaches can also be used on the data for understanding characteristics/concepts related to specific crime types.

In India, the pace at which digitization is making its impact and awareness about one of its side effect in terms of cybercrime needs to be accentuated. The work proposed in this paper entitled “Text Analysis

Framework for Understanding Cyber Crimes from New Articles" can be considered as a small step in this direction.

References

Alsuhaibani RS, Christian D, Newman, Collard ML, and Maletic JI (2015). Heuristic-based part-of-speech tagging of source code identifiers and comments. In the 5th IEEE Workshop on Mining Unstructured Data, IEEE, Bremen, Germany: 1-6. https://doi.org/10.1109/MUD.2015.7327960

Arulanandam R, Savarimuthu BTR, and Purvis MA (2014). Extracting crime information from online newspaper articles. In the 2nd Australasian Web Conference, Australian Computer Society, Auckland, New Zealand, 155: 31-38.

Aubin S and Hamon T (2006). Improving term extraction with terminological resources. In: Salakoski T, Ginter F, Pyysalo S, and Pahikkala T (Eds.), Advances in Natural Language Processing, 4139: 380-387. Lecture Notes in Computer Science, Springer Berlin Heidelberg, Heidelberg, Germany.

Barclay C (2014). Sustainable security advantage in a changing environment: The Cybersecurity Capability Maturity Model (CM 2). In the ITU Kaleidoscope Academic Conference on Living in a Converged World-Impossible without Standards?, IEEE, Saint Petersburg, Russia: 275-282. https://doi.org/10.1109/Kaleidoscope.2014.6858466

Chakravorty S, Daripa S, Saha U, Bose S, Goswami S, and Mitra S (2015). Data mining techniques for analyzing murder related structured and unstructured data. American Journal of Advanced Computing, 2(2): 47-54.

Chen H, Chung W, Xu JJ, Wang G, Qin Y, and Chau M (2004). Crime data mining: A general framework and some examples. Computer, 37(4): 50-56.

IACA (2014). Definition and types of crime analysis (White Paper 2014-02). International Association of Crime Analysts, University of Maryland, College Park, USA.

Jayaweera I, Sajeewa C, Liyanage S, Wijewardane T, Perera I, and Wijayasiri A (2015). Crime analytics: Analysis of crimes through newspaper articles. In the Moratuwa Engineering Research Conference, IEEE, Moratuwa, Sri Lanka: 277-282.

Nath SV (2006). Crime pattern detection using data mining. In the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops, IEEE, Hong Kong, China: 41-44. https://doi.org/10.1109/WI-IATW.2006.55

Omar N, Hanna JRP, and McKevitt P (2004). Heuristic-based entity-relationship modelling through natural language processing. In: the 15th Conference of Artificial Intelligence and Cognitive Science (AICS'04), Galway-Mayo Institute of Technology (GMIT), Castlebar, Ireland: 302-313.

Park KSO (2008). Exploring crime analysis. International Association of Crime Analysts, BookSurge, South Carolina, USA.

Tayal DK, Jain A, Arora S, Agarwal S, Gupta T, and Tyagi N (2015). Crime detection and criminal identification in India using data mining techniques. AI and Society, 30(1): 117-127.

Tian Y and Lo D (2015). A comparative study on the effectiveness of part-of-speech tagging techniques on bug reports. In the IEEE 22nd International Conference on Software Analysis, Evolution and Reengineering, IEEE, Montreal, Canada: 570-574. https://doi.org/10.1109/SANER.2015.7081879

Wang T, Rudin C, Wagner D, and Sevieri R (2013). Learning to detect patterns of crime. In the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer Berlin Heidelberg, Heidelberg, Germany: 515-530.