Detecting crime patterns from Swahili newspapers using text mining

George Matto* and Joseph Mwangoka

School of Computational
and Communication Science and Engineering,
The Nelson Mandela African Institution of Science and Technology,
P.O. Box 447, Arusha, Tanzania
Email: mattog@nm-aist.ac.tz
Email: josephwam@gmail.com
*Corresponding author

Abstract: The Tanzania Police Force, as many other law enforcement agencies in developing countries, relies mostly on manual, personal judgments, and other inadequate tools for analysis of data in its crime databases. This approach is inadequate and prone to errors. Moreover, research shows that more than half of all crimes committed in Tanzania are not reported to police and thus it is likely that they are not analysed by the police. In this study, we use text mining to extract crime patterns from sources of crime data outside police databases. In fact, we use four daily published Swahili newspapers. With the help of our developed patterns mining model we extracted several crimes reported in the newspapers, we mapped the distribution of the mined crimes country-wide, and with the use of FP-growth, we generated association rules between the mined crimes. Results from this study will contribute to crime detection and prevention strategies.

Keywords: crime; crime patterns; text mining; association rules; FP-growth.

Reference to this paper should be made as follows: Matto, G. and Mwangoka, J. (2017) 'Detecting crime patterns from Swahili newspapers using text mining', *Int. J. Knowledge Engineering and Data Mining*, Vol. 4, No. 2, pp.145–156.

Biographical notes: George Matto is a PhD scholar at the School of Computational and Communication Sciences and Engineering at the Nelson Mandela Institution of Science and Technology. He holds MSc in Computer Science and BSc (Hons) in Computer Science. His areas of research include database management systems, data mining, text mining, pattern recognition and big data.

Joseph Mwangoka is a Senior Lecturer at the Nelson Mandela African Institution of Science and Technology, Arusha, Tanzania. He received his PhD degree from the Tsinghua University, Beijing, China in 2009. Until 2012 he was a Senior Research Engineer at the Institute of Telecommunications, Aveiro, Portugal. His research interests include data science, cognitive radio technology, dynamic spectrum management, ICT4D/E, health informatics, and cloud computing. He has co-authored a number of peer-reviewed book chapters, journal articles and conference proceedings.

1 Introduction

In Tanzania, reports from the Tanzania Police Force and the National Bureau of Statistics show a slight decrease of crimes reported in police stations. For instance, as Figure 1(a) shows, the number of criminal offences reported in 2015 was 519,203, compared with 528,575 that was reported in 2014, a decrease which was equivalent to 1.8%. Similarly, there was a 6% and 1.1% decrease of crimes reported in 2014 and 2013 respectively (The United Republic of Tanzania, 2016; 2015; 2014; 2013). While such reports record a crime decrease, several researches indicate a continued rise of crime fear among Tanzanians. Crime fear is defined by Hale (1996) as the fear of being a victim of crime. In 2011, for example, 42% of Tanzanians were living with fear of becoming victims of crimes (Wambura, 2015a), 41% in 2012 (Gaddis et al., 2013), and over 45% in 2014 (Twaweza, 2014) as shown more in Figure 1(b).

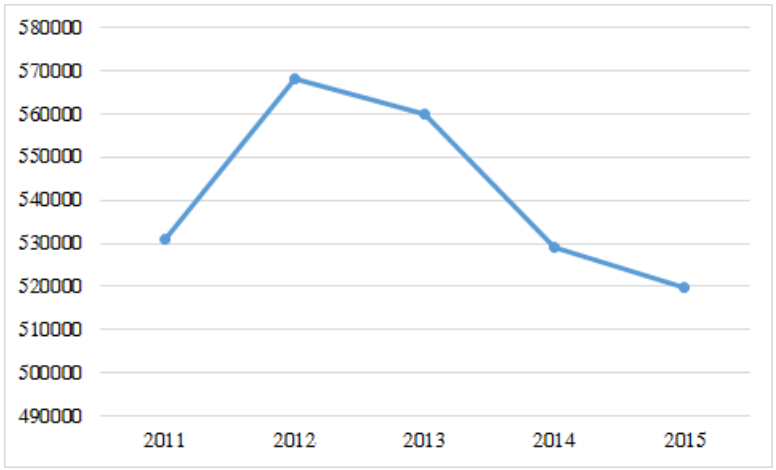
Reports about reduced number of reported crimes on one hand, and those of increased crime fear on the other hand presents a contradicting fact about the actual crime situation in the country. However, Jackson (2009) argues that there exists a link between fear of crime and likelihood of victimisation, and that, high crime fearing rate is a natural response to crime incidents as it is grounded on the reality of crime. This is why Twaweza (2014) pointed out that the increased crime fear in Tanzania is a result of the increased crime incidents. In connection, research shows a low tendency of crime reporting in Tanzania. In 2011 to 2013 for example, 54% of people who were victims of crime did not report the incidents to police (Wambura, 2015b). Therefore, although police records show a decreasing rate of reported crimes, the incidents are likely on the increase.

Robust proactive measures are needed to support the prevention of further criminal incidents. In fact, this is the primary objective of an efficient police (Zaman, 2013). The advancements in science and technology plays a major role on this. Technologies can help in analysing crime datasets to find emerging patterns, series, and trends. This will help police force to understand the current crime trends and predict or forecast future occurrences (Usher and Rameshkumar, 2014).

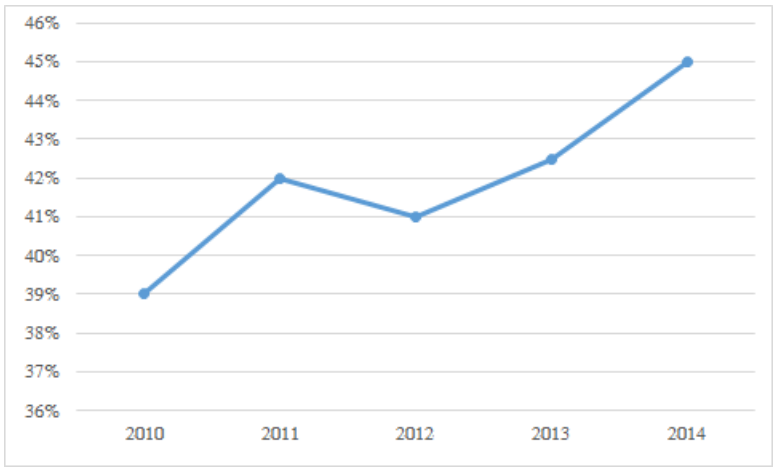
Unfortunately, Tanzania Police Force, as many other law enforcement agencies in developing countries, rely mostly on manual or personal judgments and other inadequate tools for inspection, exploration and analysis of crime data. Moreover, as said by Isafiade and Bagula (2013), the volume of data that can be processed simultaneously within a reasonable time frame is limited thus results into omission of complex and crucial relationships between different crimes attributes. Apart from the challenge of maximally exploring crime datasets, the Tanzania Police Force is faced with another challenge of capturing and analysing unreported crimes.

Newspapers, social media, and other similar platforms can be a useful source of crime data that are not necessarily reported in police stations. Text mining and other data mining techniques are capable of providing police with useful insights from such platforms. It is on this same line that this research was carried out to employ text mining to analyse and extract crime patterns from Tanzania's Swahili newspapers. Swahili is a Tanzanian official language. Specifically, our objectives were three fold. To mine frequently reported crimes, to investigate on the distribution of crimes per regions, and to generate association rules between the mined crimes. Results from this study will be helpful to police and other law enforcement agencies in the process of crime detection and prevention.

Figure 1 (a) Trend of crimes reported to police stations, from 2011 to 2015 (b) Trend of crime fearing rate in Tanzania, from 2010 to 2014 (see online version for colours)



(a)



(b)

The remainder of this paper is organised as follows: Section 2 provides related studies; Section 3 presents the methodology used; Section 4 presents results and discussions; Section 5 gives the conclusion and puts forward the future work.

2 Related studies

Data mining is a process of extracting knowledge from huge amount of data stored in databases, data warehouses and data repositories (Jani, 2014). It can be achieved by association, classification, clustering, predictions, sequential patterns, and similar time sequences (Hipp et al., 2002). In Association, the relationship of a particular item in a data transaction on other items in the same transaction is used to predict patterns. The

idea of mining association rules originated from the analysis of market-basket data where rules like 'if a customer buy bread he is 85% likely to purchase butter also' are generated. Today the generation of association rules is one of the most popular data mining methods. Moreover, association rules are not restricted to dependency analysis in the context of retail applications but are successfully applicable to a wide range of business problems.

Text mining refers to using data mining techniques for discovering useful patterns from texts. Data mining and text mining have become a powerful technique with great potential to help criminal investigators focus on the most important information on crime datasets, as such they help police investigating officers to identify hidden patterns from crime data (Varghese et al., 2010). A great deal of scientific research have consequently been performed on crime patterns mining (Gangavane and Nikose, 2015).

Several of such research, however, have been concentrated on identifying crime patterns from crime databases and other structured data. For example, Zubi and Mahmmud (2014) proposed model for crime and criminal data analysis using data mining techniques. They used Libyan national criminal record data for their experiment which was based on association rule mining and clustering. Another research by Isafiade and Bagula (2013) which focused on creating a flexible and effective solution to crime situation recognition used crime incident reports that were stored on various crime databases. Wang et al. (2012) proposed a pattern detection algorithm called series finder that grows a pattern of discovered crimes from within a database, starting from a 'seed' of a few crimes. Series finder used data collected by the Crime Analysis Unit of the Cambridge Police Department to detect patterns of crime committed by the same individual(s).

Same trend of generating crime patterns from structured crime data is observed in Jani (2014) and Varghese et al. (2010). Elyezjy and Elhaless (2015) attempted to investigate crime patterns using text mining and network analysis by mining offenders' names from unstructured text data in the Arabic language. However, they again did the mining from investigations documents that were obtained from police department. These, and other similar literature, show that there is a little research in methods and techniques that extract crime patterns from unstructured texts in other sources outside police crime datasets. There is similarly little research that consider the same extractions from datasets in local languages.

3 Methodology

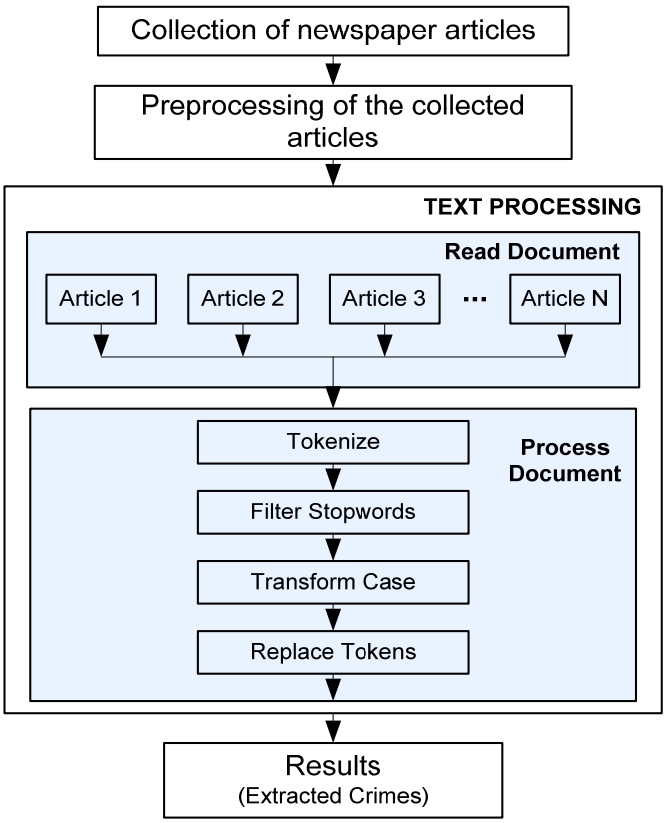
3.1 Source of data

In this study we used datasets obtained from four reputable Swahili newspapers. Selected newspapers were Majira, Mtanzania, Mwananchi and Nipashe. At least eight articles from each of the newspaper were collected and analysed. The articles were those with crime related news reports published in May 2016. Swahili newspapers were used because of two reasons. First, most of the newspapers in Tanzania are published in Swahili and second, the selected newspapers had news reporters from all over the country and hence countrywide coverage.

3.2 Workflow of the crimes mining process

The first step in the process of crimes mining was the collection of articles. The collected articles were then pre-processed before loaded to the crimes mining model that was built on RapidMiner Studio. Pre-processing involved moulding the articles obtained from various newspapers platforms into a suitable format.

Figure 2 Workflow of the mining process (see online version for colours)

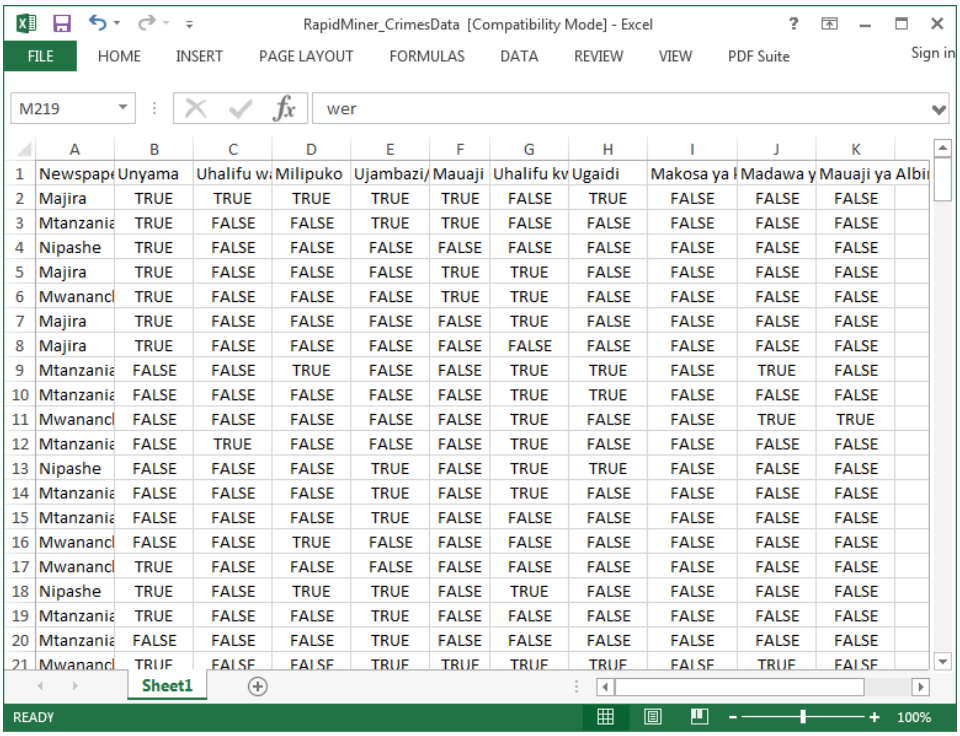


The next step was 'TEXT PROCESSING'. This is actually what we trained our text mining model to do. To accomplish this, it reads and process document. In the 'read document', the model loads pre-processed newspaper articles in .txt format. Several articles (of the same newspaper) were then combined together as one document and taken to the 'process document' step.

We trained the model to do four things in the 'process document' step. First was 'tokenize' in which words in the articles were grouped together and counted. Second was 'filter stopwords'. Stopwords are the most common words in a language. For example, stopwords in the English language are such as the, is, at and which. RapidMiner has built-in dictionaries in several languages to find and filter stopwords out. Unfortunately, the articles that were used in this work were in Swahili, the language which dictionary is not available in RapidMiner. Third step was 'transform cases'. Since RapidMiner is case sensitive where letters that are uppercase do not match with the same letters in lowercase,

we opted to use lower cases. So, in this step, all letters were transformed into lower case. The fourth and last step was 'replace tokens'. Tokens are words, phrases, symbols or other meaningful elements in the articles. Several of such tokens can be presented differently but meaning the same thing. In this step, similar tokens were replaced by more common ones, and their total occurrences were recorded. In this way, we were able to obtain various reported crimes, their frequency of occurrences, and the regions in which they occurred. This mining process workflow is summarised in Figure 2.

Figure 3 Sample data prepared for rules generation (see online version for colours)



	A	B	C	D	E	F	G	H	I	J	K
1	Newspap	Unyama	Uhalifu w	Milipuko	Ujambazi/	Mauaji	Uhalifu kv	Ugaidi	Makosa ya	Madawa y	Mauaji ya Albi
2	Majira	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE
3	Mtanzania	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
4	Nipashe	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
5	Majira	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE
6	Mwanand	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE
7	Majira	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
8	Majira	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
9	Mtanzania	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	TRUE	FALSE
10	Mtanzania	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE
11	Mwanand	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE
12	Mtanzania	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
13	Nipashe	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE
14	Mtanzania	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
15	Mtanzania	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
16	Mwanand	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
17	Mwanand	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
18	Nipashe	TRUE	FALSE	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
19	Mtanzania	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
20	Mtanzania	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
21	Mwanand	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE

3.3 *Association rules generation*

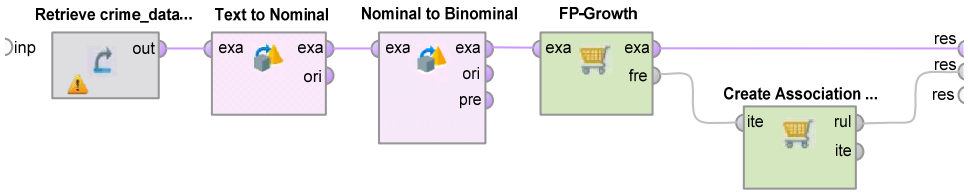
3.3.1 *Data preparations for rules generation*

After obtaining a set of mined crimes, we employed FP-growth algorithm to generate association rules between those crimes. Since the mining process was done from four different newspapers, frequency of occurrences of the mined crimes varied from one newspaper to another. To make FP-growth applicable in our case, we organised the mined dataset in tabular form in such a way that mined crimes were set as attributes of the table and the newspapers where the crimes were mined were set as rows, as shown in Figure 3. A Boolean value 'TRUE' or 'FALSE' was assigned in each attribute to indicate whether a particular crime was reported in a particular newspaper or not.

3.3.2 Rules generation process

The prepared dataset was then loaded to RapidMiner for association rules generation. To accomplish rules generation we used 'FP-growth' and 'generate association rules' operators. But due to the nature of our dataset the FP-growth could not be applied directly because it requires all attributes to be binominal. We then did some pre-processing to mould our dataset into the desired form. In fact, after retrieving the prepared crimes dataset, we used 'text to nominal' and 'nominal to binominal' operators to pre-process the data before FP-growth was used. This process is shown in Figure 4.

Figure 4 Operators involved in the association rules generation process (see online version for colours)



The text to nominal operator was applied to convert all text attributes in the dataset to nominal attributes. After that conversion, nominal to binominal operator was applied to change those nominal attributes to binominal. FP-growth operator was then applied to generate frequent itemsets. And, finally, to generate a set of association rules from the generated frequent itemsets we used the create association rules operator. Our aim was to generate only strong association rules between the mined crimes, so we set minimum support and confidence to be 0.95 and 1 respectively.

4 Results

4.1 Mined crimes and their frequencies of occurrence

Ten (10) crime incident types were extracted from the input files. Since the input files were in Swahili, the mined results were also in Swahili. The following are the results with their English translation in parentheses; *Mauaji* (Killings), *Unyama* (Brutality), *Milipuko* (Explosives), *Makosa ya kingono* (Sexual offenses), *Ujambazi/Uvamizi* (Invading/Gangs), *Uhalifu wa kutumia Bunduki* (Gunned crimes), *Uhalifu wa kutumia Silaha za jadi* (Traditional armed crimes), *Ugaidi* (Terrorism), *Madawa ya kulevya* (Drugs), and *Mauaji ya Albino* (Killing of people with albinism). Table 1 summarises this finding.

As shown in the presented results, *Mauaji* (Killings) occurred mostly in all of the four newspapers. Its frequency of occurrence was 60 in the four articles of the Majira newspaper, 59 in seven articles of the Mtanzania, 98 in eight articles of the Mwananchi, and 36 in six articles of the Nipashe. It can further be observed that other crime incident types with high frequency of occurrence were; *Unyama* (Brutality), *Milipuko* (Explosives) and *Ujambazi/Uvamizi* (Invading/Gangs).

Table 1 Mined crimes and their frequencies of occurrence

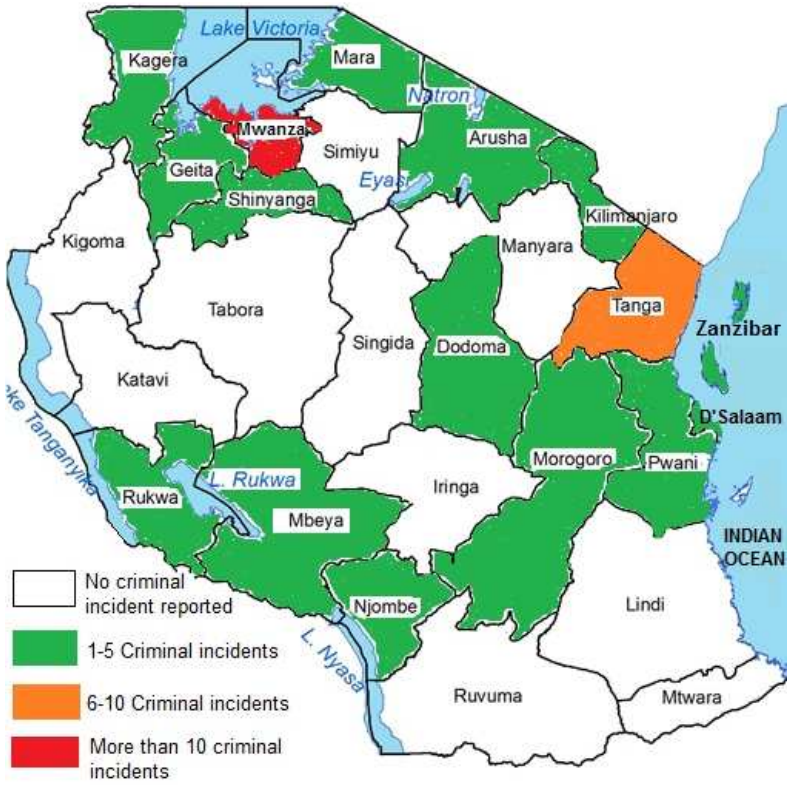
Types of crimes as mined from newspapers	*Majira*		*Mtanzania*		*Mwananchi*		*Nipashe*	
	Total freq.	*No. of articles*	*Total freq.*	*No. of articles*	*Total freq.*	*No. of articles*	*Total freq.*	*No. of articles*
Mauaji	60	4	59	7	98	8	36	6
Unyama	17	4	12	3	17	6	13	6
Milipuko	15	2	7	3	1	1	2	1
Makosa ya kingono	3	3	2	1	3	3	24	4
Ujambazi/Uvamizi	15	4	35	8	21	5	11	6
Uhalifu wa kutumia Bunduki	11	2	27	4	12	3	15	2
Uhalifu wa Silaha za jadi	10	3	12	5	18	5	13	5
Ugaidi	2	1	45	2	2	1	6	2
Madawa ya kulevya	0	0	15	1	2	2	1	1
Mauaji ya Albino	0	0	0	0	5	3	0	0

In the other hand, *Mauaji ya Albino* (Killing of people with albinism) were seldom reported. It appeared only in the Mwananchi newspaper, which tells that in May, 2016 there were very few incidents related to killings of people with albinism in Tanzania. The next least reported criminal incident was *Madawa ya kulevya* (Drugs), which was not reported at all in the Majira newspaper.

4.2 Crimes occurrence by regions

Investigating crimes distribution per regions was a second objective of this study. Results showed that out of 30 regions of Tanzania, newspapers reported crime occurrences in 16 regions. Zanzibar islands consist of five regions, but for the purpose of this study we treat the islands as a single region, that is, the region of Zanzibar. Figure 5 shows how crimes were distributed across the regions of Tanzania.

Figure 5 Map of Tanzania showing crimes distribution by region (see online version for colours)



Source: May, 2016

As shown in Figure 5, Mwanza region was leading by having an average criminal incidents of more than ten. Tanga followed with an average of 6 to 10 criminal incidents. Following these results, the two regions were categorised as relatively high crime zones in May, 2016. Consequently, the Tanzania Police Force and other stakeholders including

the general public could consider placing extra efforts to reduce the high crime rates in the regions.

Fourteen regions; four in the lake zone (Kagera, Geita, Shinyanga and Mara), two northern regions (Arusha and Kilimanjaro), two coast regions (Dar es Salaam and Pwani), two central regions (Dodoma and Morogoro), and three southern highlands regions (Rukwa, Mbeya and Njombe) and Zanzibar had an average of one to five criminal incidents. This made these regions to be categorised as low crime zones.

The four newspapers that were used in this study did not report any crime incident in southern regions (Ruvuma, Lindi and Mtwara), western regions (Kigoma, Katavi and Tabora) as well as Simiyu, Singida, Manyara and Iringa. Basing on this finding it can be fair to say that those regions were relatively safe in the time the news were reported.

4.3 Association rules generated

Six association rules were generated among the mined crimes. Figure 6 is a graphical visualisation of the generated rules while Figure 7 shows textual description of the rules.

Figure 6 Graphical visualisation of the generated association rules

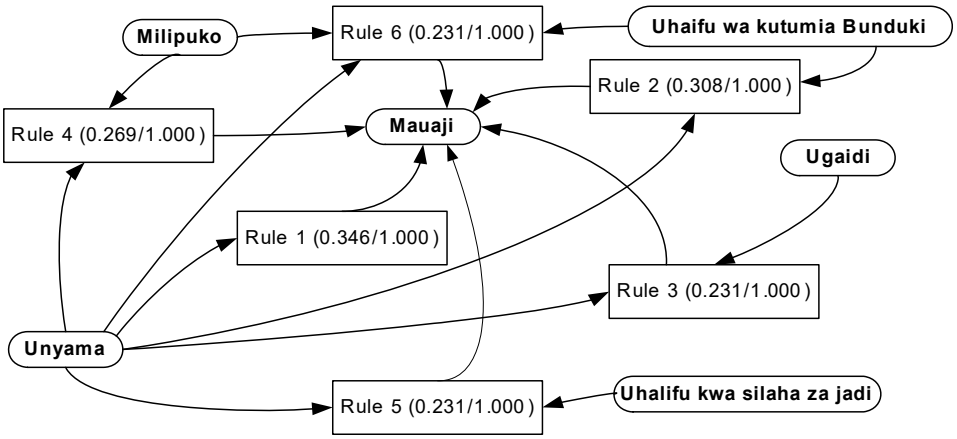


Figure 7 Textual description of the generated rules (see online version for colours)

[Unyama] → [Mauaji] (confidence: 1.000)
[Uhalifu wa kutumia Bunduki, Unyama] → [Mauaji] (confidence: 1.000)
[Ugaidi, Unyama] → [Mauaji] (confidence: 1.000)
[Milipuko, Unyama] → [Mauaji] (confidence: 1.000)
[Uhalifu kwa silaha za jadi, Unyama] → [Mauaji] (confidence: 1.000)
[Uhalifu kwa kutumia Bunduki, Milipuko, Unyama] → [Mauaji] (confidence: 1.000)

What can be inferred from the generated rules is that, since all the generated association rules concluded to *Mauaji* (Killings) then the premises (i.e., *Unyama* (Brutality), *Uhalifu wa kutumia Bunduki* (Gunned crimes), *Milipuko* (Explosives) and *Uhalifu kwa silaha za jadi* (Traditional armed crimes)) possibly resulted into killings. Therefore, it can be fair to

say; if brutality, gunned crimes, explosives and traditional armed crimes are contained, killings will also be contained.

Considering the patterns that we have mined it is not surprising to see what have been reported by other researchers about the rise of crime fear. In fact, this research confirms the existence of crime incidents which are the contributing factors to the reported fear. Police, and other law enforcement agencies, can use these mined patterns to boost their crime detection and prevention strategies.

5 Conclusions and future work

The main contribution of this work is the extraction of crime patterns from unstructured data in Swahili newspapers. By applying data mining techniques we were able to analyse and mine crime patterns and then generate association rules between the mined crimes. We were also able to show distribution of crime occurrences per regions of Tanzania. Mined patterns can help police officers and other law enforcement agencies to understand the crime situation from a different angle, and thus put in place more efficient proactive measures against future crimes.

Future work is to collect data for longer period of time, mine and check if there will be seasonality of patterns. Also extract crime patterns from structured data in crime databases and establish correlation between patterns of crimes from what is being reported in police and what is reported in the media.

References

Elyezjy, N.T. and Elhaless, A.M. (2015) 'Investigating crimes using text mining and network analysis', *International Journal of Computer Applications*, September, Vol. 126, No.8, pp.0975–8887.

Gaddis, I. Morisset, J. and Wane, W. (2013) *Law and Order: Countering the threat of crime in Tanzania* [online] http://blogs.worldbank.org/africacan/law-and-order-countering-the-threat-of-crime-in-tanzania (accessed July 2016).

Gangavane, H.N. and Nikose, M.C. (2015) 'A survey on document clustering for identifying criminal', *International Journal on Recent and Innovation Trends in Computing and Communication*, Vol. 3, No. 2, pp.2459–463, ISSN: 2321-8169

Hale, C. (1996) 'Fear of crime: a review of the literature', *International Review of Victimology*, Vol. 4, No. 2, pp.79–150, ISSN: 0269–7580.

Hipp, J., Güntzer, U. and Nakhaeizadeh, G. (2002) *Data Mining of Association Rules and the Process of Knowledge Discovery in Databases*, Springer-Verlag Berlin Heidelberg.

Isafiade, O. and Bagula, A. (2013) 'Citisafe: adaptive spatial pattern knowledge using FP-growth algorithm for crime situation recognition', in *Proc. IEEE International Conference on Ubiquitous Intelligence and Computing*, IEEE, pp.551–556.

Jackson, J. (2009) 'A psychological perspective on vulnerability in the fear of crime', *Psychology, Crime and Law*, Vol. 15, No. 4, pp.365–390, ISSN 1068-316X.

Jani, V.H. (2014) 'Survey of identifying criminal pattern using data mining algorithm', *International Journal of Innovative Research in Technology*, Vol. 1, No. 7, pp.5–7, ISSN: 2349-6002.

The United Republic of Tanzania (2013) *Crime Statistics Report: January–December, 2012*, Tanzania Police Force & National Bureau of Statistics.

The United Republic of Tanzania (2014) *Annual Crime Report 2013*, Tanzania Police Force & National Bureau of Statistics.

The United Republic of Tanzania (2015) *Crime Statistics Report: January–December, 2014*, Tanzania Police Force & National Bureau of Statistics.

The United Republic of Tanzania (2016) *Crime and Traffic Incidents Statistics Report: January–December, 2015*, Tanzania Police Force & National Bureau of Statistics.

Twaweza (2014) *Are We Safe? Citizens Report on the Country's State of Security*, Sauti za Wananchi, Brief No. 9.

Usher, D. and Rameshkumar, K. (2014) 'A complete Survey on application of frequent pattern mining and association rule mining on crime pattern mining', *International Journal of Advances in Computer Science and Technology*, Vol. 3, No. 4, pp.264–275, ISSN 2320-2602.

Varghese, B.V., Unnikrishnan, A., Jacob, P. and Jacob, J. (2010) 'Correlation clustering model for crime pattern detection', *International Journal of Advancements in Computing Technology*, Vol. 2, No. 5, pp.125–128.

Wambura, P.M. (2015a) *Crime and Security in East Africa: Burundians Feel Most Secure*, Afrobarometer Dispatch, No. 10.

Wambura, P.M. (2015b) *Police Corruption in Africa Undermines Trust, But Support for Law Enforcement Remains Strong*, Afrobarometer Dispatch, No. 56.

Wang, T., Rudin, C., Wagner, D. and Sevieri, R. (2012) *Learning to Detect Patterns of Crime*, Massachusetts Institute of Technology.

Zaman (2013) *Big Data and UK Policing – White Paper*, daywatcher.com.

Zubi, Z.S. and Mahmmud, A.A. (2014) 'Crime data analysis using data mining techniques to improve crimes prevention', *International Journal of Computers*, Vol. 8, No. 1, pp.39–45, ISSN: 1998-4308.