Compilation of Malay Criminological Terms from Online News

Joanna Chiew Ling Lee[1], Phoey Lee Teh[2], Sian Lun Lau[3] and Irina Pak[4]
[1,2,3,4]*Department of Computing and Information Systems, Sunway University, Bandar Sunway, 46150 Petaling Jaya, Selangor, Malaysia*
14000426@imail.sunway.edu.my
phoeyleet@sunway.edu.my

***Abstract*—A Malay language corpus has been established by the Institute of Language and Literature (Dewan Bahasa dan Pustaka, DBP in Malaysia). Most of the past research on the Malay language corpus has focused on the description, lexicography and translation of the Malay language. However, in the existing literature, there is no list of Malay words that categorizes crime terminologies. This study aims to fill that linguistic gap. First, we aggregated the most frequently used crime terminology words from Malaysian online news sources. Five hundred crime-related words were compiled. No automatic machines were in the initial process, but they were subsequently used to verify the data. Four human coders were used to validate the data and ensure the originality of the semantic understanding of the Malay text. Finally, major crime terminologies were outlined from a set of keywords to serve as taggers in our solution. The ultimate goal of this study is to provide a corpus for forensic linguistics, police investigations, and general crime research. This study has established the first corpus of a criminological text in the Malay language.**

***Index Terms*—Criminological Text; Malay Language; Part-of-Speech; Semantic Tagging**

I. INTRODUCTION

Part-of-speech tagging (POS) refers to each word of one sentence assigned to an appropriate part-of-speech tagging [1]. That is the procedure to identify each noun, verb, adjective or other parts of speech, which is known as the POS tagging [1]. POS tagger has been gaining widespread attention in the field of linguistics. The use of POS tagger has been applied in lexical feature extraction for word clustering [2], Twitter [3], and medical blogs [4]. Compared to other languages [5] such as English [6], [7], the development of the Malay language corpora in Malaysia is still lagging behind. To the best of our knowledge, there is yet to be a Malay language corpora that compile a specific and detailed list of criminological terms in Malay.

Linguistics literature [8] has highlighted how the Malay language has many loanwords from others languages. Since then, large-scale linguistic works have been established. Tasks such as word tagging and tokenizing are done in many different languages, including Arabic [9], Hebrew [10], German [11], Urdu [12], Burmese [13], Russian [14], Chinese [15] and Swedish [16]. In other words, the process of text segmentation involved in these studies has been used in many different languages for text analysis [17].

It is unarguably true that English is one of the most usable and established compared to any other language. Although several Malay corpora analysis have been conducted, the development of the English language remains an example at all times, at least both of the information on newspapers (e.g. Utusan Online or Berita Harian) only have a general tag to search for all the crime news online, which is "jenayah" in Malay. The word "crime" is too abstract and broad term, and yet limited to be of any help to forensic linguistic users. In particular, professionals in crime-related fields such as police, lawyers and forensic scientists may find it is helpful to search for materials related to crime with such a list of terms in Malay. With such list of tags, the availability of more relevant information will be available for crime-related fields academic or research purposes. While the Malay language is a medium of instruction in education, the majority of online communication in Malaysia remains to be in English [18]. Furthermore, the Malay language has yet to have a specific list of crime-related terminologies developed for crime-related news or information search.

Thus, this study aims to look at the creation of crime-related words by identifying the most frequently used criminology terms from online news articles. The rest of this paper is organized as follows: Section II presents the literature review, and Section III describes the method of this study using human coder and sentiment tools' setups. Section IV contains an analysis of the survey results, followed by the evaluation of sentiment tools' testing. Section V then concludes the paper.

II. LITERATURE REVIEW

A. Crime

Crime has always been a big societal issue, regardless of whether it is a knife crime or a cybercrime. In 2018, the crime rate in Malaysia is still on the rise [19]. In the worldwide crime index, Malaysia is ranked at number 15 (63.05%), while United States is at number 35 (49.58%) and United Kingdom at number 62 (41.20%). Malaysia's crime index was rated at 70.88% in 2012, decreased to 67.50% in 2014

Table 1
Malaysian Crime Categories

Crime Categories	Amount (the year 2016)
Acts of Violence	
Murder	456
Rape	1886
Robbery: Accomplices with Firearms	65
Robbery: Accomplices without Firearms	10,907
Robbery: Firearms	18
Robbery: Without Firearms	3463
Wounding	5531
Property Damage	
Theft	19894
Car Theft	10607
Motorcycle Theft	34754

This is accepted version of this manuscript, article is STILL IN PRESS, volume, issue and DOI is TO BE ASSIGNED

Heavy Vehicle Theft	3050
Snatch Theft	2963
Breaking, Entering and Stealing / Burglary	18760
Total Crime Index	112354

and rose again to 69.70% in 2015. Until 2017, the crime index had decreased to 63.05%. These numbers are still considered high, and Malaysia is still a country that is plagued by crimes. A recent open source statistic report of Malaysia has categorized crimes into two main categories [19]: 1) acts of violence, and 2) property damage. As shown in Table 1, these two categories can be separated into seven and six subcategories, respectively.

The subcategories in Table 1 show the various types and amount of crimes that are being committed in the country. It can therefore be deduced that it is a significant aspect that further analysis must be considered. However, the statistics only consider two different crime categories that are present in the country and do take into account other categories of crime that exist.

B. How Does Literature Categorize Crime Terminologies?

Many types of research have been done to categorize crime terminologies [20]–[22]. From mentioned literature, crime can be primarily categorized into the following seven categories: 1) property theft 2) violent crime 3) controlled substance/drug 4) terrorism 5) abuse 6) white collar crime, and 7) forced labour. As shown in Table 2, each of these broad categories of crime can then be broken down into different subcategories [20], [23]–[26].

Table 2
Categories of Crime

Major Categories (number of crimes)	Subcategories
Property Theft (6*)	Theft, Car Theft, Motorcycle Theft, Heavy Vehicle Theft, Snatch Theft, Breaking, Entering and Stealing /Burglary.
Violent Crimes (7*) *Jenayah Kekerasan*	Murder, Rape, Armed Robbery with Accomplices, Unarmed Robbery without Firearms, Armed Robbery, Unarmed Robbery, Wounding.
Controlled Substances/Drugs (7*) *Bahan-bahan Terkawal*	Trafficking, Drug Possession, Controlled Substance Violation and Other Crimes/Activity, Racketeering, Smuggling, Laundering Money from Controlled Substances, Tax Offenses.
Terrorism (8*) *Pengganasan*	Cyber Terrorism, State Terrorism, State Sponsored Terrorism, Nationalist Terrorism, Religious Terrorism, Left and Right Wing Terrorism, Anarchist Terrorism, Suicide Terrorism.
Abuse (7*) *Penderaan*	Child Abuse, Physical Abuse, Emotional Abuse, Sexual Abuse, Neglect, Bullying, Financial Exploitation.
White-Collar Crime (8*) *Jenayah Kolar Putih*	Antitrust, Securities Fraud, Mail Fraud, False Claims, Credit Fraud, Bribery, Tax Fraud, Bank Embezzlement.
Forced Labour (8*) *Buruh Paksa*	Forms of coercion, Prison Labour, Forced Overtime, Human Trafficking, Trafficking or Smuggling, Slavery, Child Labour, Bonded Labour.

*The number of subcategories in the category
Italic words are in the Malay language

In data classification, it is essential to group terms that share a common characteristic, meaning or quality. With the classification in Table 2, the process of categorizing crime terminologies becomes clearer.

A common way of categorizing a keyword is through keyword extraction [27]. This process is done based on the available list of keywords to accommodate the categorization of other keywords into those categories. However, an issue that may arise is that while there are subcategories that represent the general category of crime terms, there is no evidence or method to show that some types of crime belong in a particular subcategory, especially in the Malay language [27]. The use of keyword-based categorization to classify text into a corresponding category requires approximately 30 keywords to represent each category.

In this study, there are no keywords that are used to represent each category of crime. There are only a list of English words for crime and general terms without a source of references to their major categories [28]. Thus, making a list of words for crime is essential. As seen in Table 2, several major crime categories and their subcategories have been summarized and tabulated. This study aims to develop a list of crime-related Malay terminologies. However, it has also assisted us in producing a list of English terminologies. Until today, Malaysian police reports and documents are still written in the Malay language.

In Malaysia's online news content (crime news) are generally tagged as 'crime' or 'jenayah'. No website is found to provide a list of tags that give further insight into the specific crime that the content belongs to. To fill the gap, the main aim of the study is to create a list of crime-related Malay terminologies.

III. Materials and Methods

A. Phase 1: Data Collection

The first stage of this study was to collect news from online newspapers in the Malay language, particularly news and articles that related to crime. Initially, 200 news articles were compiled. Manually, all words from the articles were recorded in a database, which separated the words by dates. For each year between 2014 and 2017, at least 50 articles were manually recorded. The number of selected online articles from Utusan Online was 71 (www.utusan.com.my), 60 from Berita Harian (www.bharian.com.my) and 69 from Harian Metro (www.hmetro.com.my). These websites generally feature newspaper articles for all categories and are written in the Malay language. The use of these newspaper articles makes it possible not only to obtain unique information of the way in which each newspaper reports or writes crime-related content, but also to consider the types of crime that have been, and are being, reported.

A random sampling method was used to select the articles and to ensure that the data collected was not biased [29]. The sampling method was carried out by using a random number generator, the maximum limit being the number of articles available on the newspaper webpage.

B. Phase 2: Pre-processing of data using Human Coder

Due to the lack of Malay sentiment tools, four human coders were used to read each newspaper article and verify the news content. Through the random sampling method, some collected articles were found to be irrelevant. For example, under the list of crime articles in Berita Harian, news on 'accidents' had been erroneously included. To overcome this issue, each news article was read through by human coders and would be removed from a list of top 500 crime-related keywords search if the article was unsuitable.

The second issue that had to be countered during the data pre-processing stage was the presence of duplicate news from different newspapers. Therefore, each article was regarded as

a distinct piece of news as the authors of the news article might have used different terms to write a similar story. This particular issue still has to be studied further. Figure 1

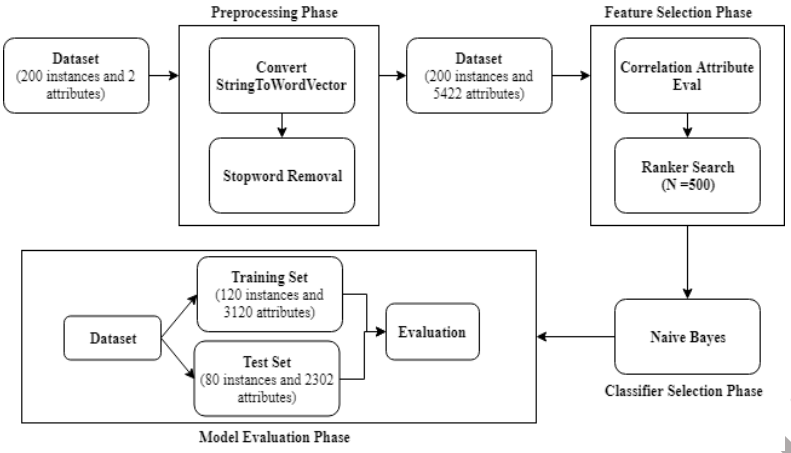


Figure 1: Flow of Data Pre-processing, Feature Selection to Evaluation

illustrates the flow of data pre-processing, and following phases (feature selection, evaluation etc.).

C. *Phase 3: Processing the Data*

From the literature review, a total of 52 subcategories of crime were identified and summarized (Table 2). The reduction process was done because there was an overlapping of attributes (words) that appeared in different categories. Furthermore, to ensure that the words in the final list have no similarity in meaning, a crime vocabulary in English was used to distinguish the semantic meaning of the words. This step was to benchmark all semantic meaning of each English word to a Malay meaning using four human coders.

Therefore, a list of wide-ranging crime vocabulary in English was obtained online from Cambridge Dictionary and Oxford Dictionary. The dictionaries were also used to translate English words to Malay, as sometimes one dictionary alone would not be able to provide the Malay equivalent of a word semantically. The human coder, therefore, had to determine the outcome. If no Malay translation of a word could be found in either dictionary, then the English–Malay Google Translate tool would be used to attain a rough translation.

D. *Phase 4: Processing the Training Set of Data*

This step was performed to create a list of categorized newspaper articles by comparing the list of words that appeared in the news with the list of Malay-translated words gathered from the previous step (Phase 3). When the text in the news article has a more frequent appearance of crime words in a specific category list (e.g. Murder), then the news will be categorized under that particular category.

E. *Phase 5: Using the WEKA*

Using WEKA [30], the dataset which was originally a collection of text in String format was converted into each word or attribute using the StringToWordVector function. In this step, unnecessary attributes (for example, 'ada', 'akan', etc., in Malay) which may negatively affect the data due to an overlapping of words were filtered and removed using the keywords that could best help the classification prediction were obtained. Table 4 shows an example of a list of words (attributes) that were selected from the CorrelationAttributeEval feature selection.

F. *Phase 6: Feature Selection*

Phase 6 involved feature selection, also known as attribute selection, to remove noise feature. In this study, the GainRatioAttributeEval, InfoGainAttributeEval and CorrelationAttributeEval feature selection algorithms and applied rank search as algorithms were used. By using three different feature selection algorithms, the consistency of the study's evaluation could be proven. The best 500 extracted keywords that could best help the classification prediction were obtained. Table 4 shows an example of a list of words (attributes) that were selected from the CorrelationAttributeEval feature selection.

G. *Phase 7: Model Evaluation/Validation*

In Phase 7, the classified set of terms was evaluated. Naïve Bayes classifier was used to categorize the dataset as it is a simple probabilistic classifier which is effective in analyzing text in many domains. Particular classier was selected because it was successfully applied in text analysis in past study of [31]. Moreover, since there were seven different categories of crime to be classified, Naïve Bayes was chosen as it is known for multi-class prediction which could generate better output for text analysis.

The output model was evaluated through correctly classified instances, incorrectly classified instances, recall, precision, F-measure, and ROC Area.

IV. RESULTS

A. *Part 1: List of crime words according to the category*

List of words was gathered through the process of word searching related to each crime category. From the seven categories of crime, a total of 724 crime terminologies were collected. Following the conversion of words into Malay, a total of 521 crime words were left, as can be seen in Figure 2. Due to the nature of language, some Malay-translated words appeared to be similar. It follows that if similar words appeared within the same category, it would be eliminate thus reducing the redundancy.

Table 3 shows the number of words that represent each crime category. The words for each category were then utilized to categorize the training set.

B. *Part 2: News Categorization for Training Set*

Categorization process was done for news text. The frequency of the words and the category to which they belonged determined the category of the text as a whole. Thus, the training set containing the text and its corresponding crime category was developed.

508	Pengganasan	7	militan
509	Pengganasan	7	aktivis
510	Pengganasan	7	pengeboman
511	Pengganasan	7	osama
512	Pengganasan	7	laden
513	Pengganasan	7	gerila
514	Pengganasan	7	bunuh diri
515	Pengganasan	7	propaganda
516	Pengganasan	7	kempen
517	Pengganasan	7	al-qaeda
518	Pengganasan	7	bangkit
519	Pengganasan	7	is
520	Pengganasan	7	tentera
521	Pengganasan	7	taliban

Figure 2: Malay crime category and list of words

Table 3
Crime Categories and Number of Related Words

Category	Number of Words
Violent Crimes	129
Property Theft	72
Abuse	77
Forced Labour	47
White-collar Crime	74
Controlled Substances	77
Terrorism	52

C. *Part 3: Data Pre-processing*

In this process, the dataset pre-processing was applied to the original dataset. By applying an unsupervised method of filtering using StringToWordVector in WEKA, each word in the text was converted into its attribute. This led to an increase in the total number of attributes. The applied stop words then filtered the attributes by matching the same words to the existing attributes. This pre-processing phase helped obtain attributes in the training and test datasets.

D. *Part 4: Feature Selection*

A list of 500 most relevant attributes in the content of the output was finalized. Attributes with a low correlation were dropped from the list and thus improved the classifier's prediction, as they would no longer affect the output. CorrelationAttributeEval was applied as feature selection. Ranker Search method was applied as well.

E. *Part 5: Classification*

The results of the Naïve Bayes classifier using four different feature selections are shown in Table 4. This evaluation displays the accuracy of the model based on the datasets that were input into the WEKA Machine Learning tool.

Table 4
Results of Classifier Accuracy

Classifier	Feature Selection	Correctly Classified Instance (%)	Incorrectly Classified Instance (%)	Kappa Statistics
Naïve Bayes	None	82.50	17.50	0.7882
Naïve Bayes	GainRatioAttributeEval	78.75	21.25	0.7425
Naïve Bayes	InfoGainAttributeEval	78.75	21.25	0.7425
Naïve Bayes	CorrelationAttributeEval	83.75	16.25	0.8040
Average		80.94	19.06	0.7693

From Table 4, it can be seen that the correctly classified instance based on the weighted average of the four results is 80.94%. This not only shows the classification's high accuracy, but also signifies that of the 80 instances from the test dataset, the model managed to validate 80.94% of them. Kappa statistic represents agreement range between observers and perfect agreement is equal to a kappa of 1 [32]. Based on the kappa statistics, the average of 0.7693 suggests that the classification did not provide much room for "random guessing". To obtain a more comprehensive analysis of the results, the detailed analysis of WEKA outputs was studied. Table 6 shows the accuracy of the analysis based on each class from the CorrelationAttributeEval feature selection output.

From Table 5, based on the average of precision = 0.882, recall = 0.838 and f-measure = 0.839, the results suggest that the classification was reliable and accurate for most classes. The ROC [33] area also produced a high statistic (ROC Area = 0.980), reflecting high accuracy in the test. Accuracy is measured by the area under the ROC curve, whereby the closer the curve is to the Y-axis, the better the result will be.

Table 5
Detailed analysis based on the Naïve Bayes Classifier with CorrelationAttributeEval feature selection

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
Bahan-bahan Terkawal	0.778	0.000	1.000	0.778	0.875	0.855	0.972	0.955
Buruh Paksa	0.833	0.000	1.000	0.833	0.909	0.907	0.928	0.860
Jenayah Hartabenda	0.778	0.000	1.000	0.778	0.875	0.855	0.989	0.970
Jenayah Kekerasan	1.000	0.154	0.600	1.000	0.750	0.713	0.985	0.936
Jenayah Kolar Putih	1.000	0.029	0.833	1.000	0.909	0.900	0.990	0.906
Penderaan	0.429	0.014	0.750	0.429	0.545	0.538	0.975	0.800
Pengganasan	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000
Weighted Ave.	0.838	0.034	0.882	0.838	0.839	0.821	0.980	0.931

Figure 3 features the top 10 words from the seven crime categories. The classifier with the CorrelationAttributeEval feature selection with the highest accuracy is shown in Table 6. The attributes from the classifier were selected from the output of the feature selection process, and the words (attributes) that matched the list of crime words were selected to be in the top 10 words from the crime category. Figure 3 records the results where each category has its own set of top 10 words followed by the rank of each word, which affects the text classification.

While there are words that identify each category, there is the issue of overlapping words in more than one category. For instance, in the 'Jenayah Hartabenda' and 'Jenayah Kolar Putih' categories, the word 'curi' is evident in both. Classifier may manage to classify the text into its corresponding category due to other related words within a particular category.

Table 6 represents the category of crime and its respective texts. The frequency of the words in each text contributes to the words that describe the category. At least one of the top 10 words used in each category is present within the text. For example, the words 'mati', 'mayat' and 'cedera' are among the top 10 words, which describe the prevalence of violent

crimes in the representative text. Machine Learning tool on different feature selections, the

I. Conclusion and Future Work

Based on the validation of the classification from the

Jenayah Kekerasan	
membunuh	0.1914
mayat	0.1835
merogol	0.1522
api	0.1522
serbuan	0.1491
mangambil	0.1483
rekod	0.137
serangan	0.1346
bunuh	0.1264
mati	0.1245

Penderaan	
hospital	0.2348
keseksaan	0.1573
penderaan	0.1436
didera	0.1329
memukul	0.1168
anak	0.1065
buruk	0.1065
cedera	0.0967
menyerang	0.0961
mencederakan	0.0924

Bahan-bahan Terkawal	
dadah	0.222
narkotik	0.2115
pil	0.2052
erimin	0.1761
gram	0.1697
heroin	0.1601
ganja	0.1426
bahan	0.129
ekstasi	0.1229
syabu	0.1229

Jenayah Hartabenda	
mencuri	0.2069
merampas	0.2054
curi	0.193
emas	0.127
kenderaan	0.126
bunyi	0.1239
kecurian	0.1239
beg	0.1131
berat	0.1052
barang	0.103

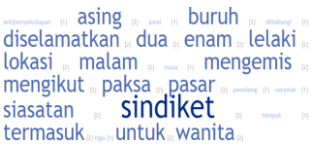
Jenayah Kolar Putih	
curi	0.193
rasuah	0.1819
suapan	0.1235
penyelewengan	0.1235
kenyataan	0.1188
bank	0.1077
palsu	0.1074
didakwa	0.0989
dana	0.0981
penipuan	0.0978

Pengganasan	
militan	0.1614
is	0.1512
risikan	0.1277
komputer	0.1258
keselamatan	0.1215
pengganas	0.1161
penipuan	0.0978
siber	0.0961
keganasan	0.0944
terrorism	0.0944

Buruh Paksa	
rasuah	0.1819
paksa	0.1756
sah	0.1317
buruh	0.1261
memukul	0.1168
penyeludupan	0.1147
buruk	0.1065
penjara	0.1037
kesakitan	0.0932
kerja	0.0912

Figure 3 Top 10 words from each crime category

Table 6
The category, representative sentences and words describing the category

Category	Representative sentence	Words describing category
Jenayah Kekerasan	seorang lelaki warga indonesia mati selepas terbabit dalam pergaduhan dengan rakan senegaranya di kediaman mereka di kampung buluh penyumpit, mukim kuah di sini, hari ini. ketua bahagian siasatan jenayah langkawi, asisten superintendan bee anak amba, berkata mayat lelaki berusia 38 tahun yang belum dikenali itu ditemui berlumuran darah di atas sofa dalam rumah terbabit pada jam 6. 15 pagi. siasatan awal mendapati mereka bergaduh sebelum maut manakala rakannya cedera.	
Jenayah Hartabenda	empat lelaki yang cuba merompak kedai emas di jalan besar sasaran, kuala selangor, pagi semalam, melarikan diri dengan tangan kosong selepas gagal memecahkan cermin pameran barang kemas. ketua polis daerah kuala selangor, superintendan ruslan abdullah berkata, kejadian berlaku pada 11.35 pagi dan tiada pelanggan ketika itu.	
Jenayah Kolar Putih	dua konstabel polis ditahan suruhanjaya pencegahan rasuah malaysia (sprm) petang tadi selepas disyaki meminta rasuah daripada ceti haram atau along di sungai petani. sumber berkata, kedua - dua anggota berusia 34 dan 37 tahun itu ditahan sprm cawangan sungai petani pada 2 petang tadi. anggota polis terbabit ditangkap kerana terbabit dalam permintaan wang rasuah berjumlah rm 10,000 daripada pengadu yang menjalankan kegiatan peminjaman wang haram.	
Penderaan	seorang wanita hong kong disabit kesalahan memukul, menyeksa, dan membiarkan pembantu rumahnya yang juga warga indonesia kelaparan, dalam kes yang mencetuskan kemarahan penduduk republik negara tersebut, tahun lalu. keputusan itu dibacakan di dalam kamar mahkamah, disambut sorakan penyokong erwiana sulistyaningsih yang merupakan bekas pembantu rumah, law wan - tung. wan-tung, 44, ibu kepada dua orang anak itu, ditangkap pada januari tahun lalu dan hukuman terhadapnya akan diputuskan pada 27 februari ini.	
Buruh Paksa	seramai 17 warga asing termasuk enam kanak - kanak berjaya diselamatkan oleh polis semalam, selepas dikesan menjadi buruh paksa satu sindiket untuk mengemis di beberapa lokasi pasar malam di puchong. ketua penolong pengarah bahagian kongsi gelap, judi dan maksiat (d7) bukit aman senior asisten komisioner rohaimi md isa berkata, semua pengemis berusia antara dua tahun hingga 50-an yang diselamatkan itu terdiri daripada dua lelaki, sembilan wanita dan enam kanak - kanak.	

Bahan-bahan Terkawal	polis menahan lima individu, termasuk tiga warga asing dan merampas pelbagai jenis dadah dianggarkan bernilai rm 6.7 juta sepanjang awal bulan ini sehingga kelmarin. pengarah jabatan siasatan jenayah narkotik (jsjn) bukit aman, datuk seri noor rashid ibrahim, berkata polis turut berjaya membongkar satu makmal dadah memproses dan membungkus pil ekstasi yang beroperasi di sebuah kondominium di jalan kuchai maju pada jumaat lalu.	
Pengganasan	pihak berkuasa turki telah membunuh hampir 900 orang yang didakwa anggota kumpulan militan negara islam (is) sejak januari lalu, kata agensi berita kerajaan, anatolia yang memetik sumber ketenteraan negara itu. menurut anatolia, daripada jumlah itu, seramai 492 ' pengganas ' telah dibunuh menerusi serangan udara manakala 370 lagi terbunuh dalam beberapa serangan meriam yang memusnahkan depot senjata mereka. bagaimanapun, angka kematian itu tidak dapat disahkan secara bebas setakat ini.	

results of recall = 0.838, precision = 0.882, f-measure = 0.839 and ROC Area = 0.980 proved that the determined results are accurate.

It can also be concluded that the word list used to categorize the text from the articles is accurate since the averaged correctly classified instance was recorded at 80.94%. Moreover, the built model was able to generate a high percentage of correctly classified instances. Therefore, the 521 words in the crime word list can be used in future work to assist in the tagging of crime in the Malay language.

Following the satisfactory results obtained in this study, it is suggested that in future research, a stemmer/lemmatizer could be applied to the dataset to acquire a cleaner dataset. Stemming is the process of reducing derived words, so that a general term could be generated. In this study, the attributes contained a multiple of the same words but with different prefixes onto it such as 'mem-', 'per-', '-an' etc. Due to these prefixes, the filtered dataset still carried attributes that represent the same words in different forms. Therefore, the application of a lemmatizer would be able to produce a more legitimate set of words.

One of the improvements for future study can be dealing with the multi-classification of the words. When text can exist in more than one category, known as multi-label classification. Therefore, in future work, the multi-label classification should be taken into consideration for instances where words may exist in more than one category.

REFERENCES

[1] H. Shi, W. Zhan, and X. Li, "A Supervised Fine-Grained Sentiment Analysis System for Online Reviews," Intell. Autom. Soft Comput., vol. 21, no. 4, pp. 589–605, 2015.

[2] A. Y. Nouira, Y. Jamoussi, and H. B. G. Hajjami, "Extracting Actions with Improved Part of Speech Tagging for Social Networking Texts," in 2016 IEEE International Conference on Computer and Information Technology (CIT), 2016, pp. 161–166.

[3] O. F. W. Onifade and M. A. Malik, "SASM: A tool for sentiment analysis on Twitter," in 2015 2nd World Symposium on Web Applications and Networking (WSWAN), 2015, pp. 1–5.

[4] Y. Batch, M. M. Yusofa, S. A. M. Noaha, and T. P. Lee, "MTag: A model to enable collaborative medical tagging in medical blogs," Procedia Comput. Sci., vol. 3, pp. 785–790, 2011.

[5] S. Piao et al., "Lexical Coverage Evaluation of Large-scale Multilingual Semantic Lexicons for Twelve Languages," in 10th edition of the Language Resources and Evaluation Conference, 2016, pp. 2615–2619.

[6] P. Rayson, "From Key Words to Key Semantic Domains," Int. J. Corpus Linguist., vol. 13, no. 4, pp. 519–549, 2008.

[7] DawnArcher, "Slurs, insults, (backhanded) compliments and other strategic facework moves," Lang. Sci., vol. 52, pp. 82–97, 2015.

[8] A. Teeuw, "The History of the Malay Language," in Modern Indonesian literature. Koninklijk Instituut voor Taal-, Land- En Volkenkunde, 1967, pp. 4–7.

[9] S. Kübler and E. Mohamed, "Part of speech tagging for arabic," Nat. Lang. Eng., vol. 18, no. 4, pp. 521–548, 2011.

[10] R. Bar-haim, K. Sima'an, and Y. Winter, "Part-of-speech tagging of modern hebrew text," Nat. Lang. Eng., vol. 14, no. 2, pp. 223–251, 2008.

[11] M. Koleva, M. Farasyn, B. Desmet, A. Breitbarth, and V. Hoste, "An automatic part-of-speech tagger for Middle Low German," Int. J. CORPUS Linguist., vol. 22, no. 1, pp. 108–141, 2017.

[12] W. Anwar, X. Wang, L. Li, and X.-L. Wang, "A Statistical Based Part of Speech Tagger for Urdu Language," in A Statistical Based Part of Speech Tagger for Urdu Language, 2007, pp. 3418–3424.

[13] C. Myint, "A Hybrid Approach for Part-of-Speech Tagging of Burmese Texts," in 2011 International Conference on Computer and Management (CAMAN), 2011, pp. 1–4.

[14] V. V. Petrochenkov and A. O. Kazennikov, "A statistical tagger for morphological tagging of Russian language texts," Autom. Remote Control, vol. 74, no. 10, pp. 1724–1732, 2013.

[15] T.-H. Chang, F.-Y. Hsu, C.-H. Lee, and H.-M. Lee, "Part-of-speech tagging for Chinese unknown words in a domain-specific small corpus using morphological and contextual rules," in Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering(NLPKE-2010), 2010, pp. 1–6.

[16] K. Smith, B. Megyesi, S. Velupillai, and M. Kvist, "Professional language in Swedish clinical text: Linguistic characterization and comparative studies," Nord. J. Linguist., vol. 37, no. 2, pp. 297–323, 2014.

[17] I. Pak and P. L. Teh, "Text Segmentation Techniques: A Critical Review," in Studies in Computational Intelligence, I. Zelinka et al., Ed. Springer International Publishing AG 2018, 2018, pp. 167–181.

[18] S. M. Ali and P. Krish, "Gender-Specific English Language Use of Malaysian Blog Authors," J. Lang. Stud., vol. 16, no. 3, pp. 21–35, 2016.

[19] M. O. D. Portal, "Jenayah indek Seluruh Malaysia," Data for Citizen Wellbeing, 2018. .

[20] M. Mahmood, "TERRORISM- DEFINITION AND TYPES," 2016. [Online]. Available: https://chainsoff.files.wordpress.com/2016/11/terrorism-definition-and-types.pdf. [Accessed: 02-Jun-2017].

[21] S. Wheeler, "White Collar Crimes and Criminals," Faculty Scholarship Series, 1988. [Online]. Available: http://digitalcommons.law.yale.edu/fss_papers/4127/. [Accessed: 02-Jun-2017].

[22] B. T. Yeh, "Drug Offenses: Maximum Fines and Terms of Imprisonment for Violation of the Federal Controlled Substances Act and Related Laws," in Congressional Research Service, 2015, pp. 1–14.

[23] NSPCC, "The definitions and signs of child abuse (NSPCC child protection fact sheet)," online Child Prot. Resour., no. April, pp. 1–9, 2009.

[24] ITUC, "Forced Labour Guide," 2008.

[25] S. W. Buell, "' White Collar ' Crimes," pp. 837–861, 2014.

[26] B. Yeh, "Drug offenses: Maximum fines and terms of imprisonment for violation of the federal controlled substances act and related laws," Congr. Res. Serv., p. 8, 2015.

[27] J. An and Y. P. Chen, "Keyword extraction for text categorization," Proc. 2005 Int. Conf. Act. Media Technol. 2005. (AMT 2005)., pp. 556–561.

[28] MyVocabulary.com, "Crime vocabulary, Crime word list - www.myvocabulary.com," MyVocabulary.com, 2018. .

[29] C. Teddlie and F. Yu, "Mixed Methods Sampling," J. Mix. Methods Res., vol. 1, no. 1, pp. 77–100, 2007.

[30] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software," ACM SIGKDD Explor., vol. 11, no. 1, pp. 10–18, 2009.

[31] S. H. Lu, D. a. Chiang, H. C. Keh, and H. H. Huang, "Chinese text classification by the Naive Bayes Classifier and the associative

classifier with multiple confidence threshold values," Knowledge-Based Syst., vol. 23, no. 6, pp. 598–604, 2010.
[32] I. Pak and P. L. Teh, "Machine Learning Classifiers: Evaluation of the Performance in Online Reviews," Indian J. Sci. Technol. ISSN, vol. 9, no. 945, pp. 974–6846, 2016.
[33] P. Martínez-Camblor, C. Carleos, and N. Corral, "General nonparametric ROC curve comparison," J. Korean Stat. Soc., vol. 42, no. 1, pp. 71–81, 2013.