

Crime index based on text mining on social media using multi classifier neural-net algorithm

Teddy Mantoro¹, Muhammad Anton Permana², Media Anugerah Ayu¹

¹Departement of Computer Science, Faculty of Engineering and technology, Sampoerna University, Jakarta, Indonesia

²Departement of Computer Science, School of Computer Science, Nusa Putra University, Sukabumi, Indonesia

Article Info

Article history:

Received Sep 19, 2021

Revised Apr 14, 2022

Accepted Apr 22, 2022

Keywords:

Classification

Index crime

Patterns of crime

Social media

Text mining

ABSTRACT

Everyday criminal issues appear on social media, even some crime news is often disturbing to the public but it gives a warning to the public to remain careful and alert to the surrounding environment. However, following large amounts of crime information on social media is not effective, especially for busy people. Therefore, there is a need to efficiently and effectively summarize information in a way that is meaningful and easy to see, attracts people's attention, and can be used by law enforcement officials. The purpose of this study is to present the index crime based on social media by looking for patterns of crime. This study proposes the projected index crime based on crime trends by using text mining to classify tweet texts and post contents into 10 crime classes. The classification method uses the neural-net multi classifier algorithm which has several classifiers namely logistic regression, naïve bayes, support vector machine (SVM), and decision tree in parallel. In this approach, the classifier that provides the best accuracy will be the winning classifier and will be used in the next learning process. In this experiment, in using the multi classifier neural-net, the logistics regression classifier often provides the best accuracy.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Teddy Mantoro

Departement of Computer Science, Faculty of Engineering and technology, Sampoerna University

Jakarta, Indonesia

Email: teddy.mantoro@sampoernauniversty.ac.id

1. INTRODUCTION

Social media is widely used in the world, including in Indonesia, and its usage provides many opportunities to make decisions based on analytics and predictive support systems. Social media is also an ideal source of data to support many decisions [1], [2] and allows also to explain events, emotions, and other topics broadly [3]. Information in the current digital era can be in the form of multi-media, both text, images, and videos that can be accessed anytime and anywhere easily [4] through social media services. Of the various types of social networks, such as Twitter and Facebook [5], this is a social media site that is a mainstay for everyone who has the ability to share information very quickly [6]. Twitter has a text message interaction feature with a limit of one hundred and forty characters called tweets. On Twitter there are also trending topics termed topics, words, and topics [7], [8]. One of the topics that get the attention of social media users is the topic of crime.

Crime is a nuisance to people who want peace [9]. At first, crime was considered a phenomenon of spiritualism [10], where the crime was considered as an act committed by a person or group and in the background by several factors including differences in political ideology, population density and composition and political instability [11]. Crimes that occur in Indonesia are currently increasing with various types of crimes such as theft, violence, murder, and many more. Crimes that occur are very varied, even some time

duplication crimes happen, and also patterned so updated data support is needed in an effort to anticipate these crimes with the help of online media or social media [12]. A criminal act is the treatment of a person or group whose condition is detrimental to another person and constitutes a law that must be punished by law. Crime is currently increasing with various types of crime showing a pattern of crime in a place and a significant period of time [13]-[15].

Crimes or criminal acts that are currently happening in Indonesia are no longer a secret because of the proliferation of criminal acts everywhere. Sometimes the news of crime is disturbing for everyone. People are aware of the opportunities for crime in the surrounding environment. According to [16], [17] in every place, not only in Indonesia, crime increases in every type and aspect of crime, as the evidenced in present in their crime mapping studies.

In this case, a special way is needed to assist law enforcement officers or the police in the process of investigating criminal acts that occur every day. As we know, crime knows no age, gender, or place. Therefore, it is necessary to anticipate crimes against crimes that have been troubling residents in their daily lives. The data collection or the dataset that will be used for this research is taken from social media such as Twitter, Facebook [18]-[20]. The data collection process was carried out by crawling data using orange3, data miner, and beautiful soup.

The criminal dataset filtered from social media such as Twitter and Facebook are produced by a text mining process and then classified the crime using the neural-net multi classifier approach based on place and time. This approach has several classifiers namely logistic regression, naïve bayes, support vector machine (SVM), and decision tree in parallel. In this approach, the classifier that provides the best accuracy will be the winning classifier and will be used in the next learning process cycle to map the Index crime in a particular location.

2. RESEARCH METHOD

This session discusses the methods needed to achieve the research objectives. It provides details on the research procedure. Firstly, it discusses the research design that will be used to achieve the research objectives, the investigative process in reviewing the study information, and then the process in designing and developing a framework for identifying concerns based on the state of the art. In addition, it also describes an approach in evaluating the framework.

2.1. Types and sources of data collection

The data used for this research is in the form of text obtained from social networking services like Twitter and Facebook [21], [22]. The data is a collection of comments or opinions regarding crimes that occurred in Indonesia. The author collects posting or tweets for specific criminals for classification. There are ten categories that will be used in this study which are corruption, domestic violence, fraud, kidnapping, molestation, murder, narcotics, persecution, rape, and theft. Those categories are taken based on the most common crime categories in Indonesia according to the Indonesian statistical bureau, i.e., (in Indonesian “*Badan Pusat Statistik*” (BPS)).

2.2. Text mining procedure

Text mining or better known as text analysis is the process of converting unstructured text into more structured data to facilitate analysis. In achieving this mining process, one must pass several stages such as data collection, preprocessing, tokenisation, remove stop word, stemming, transformation, feature extraction, and classification. All those step and techniques of text mining procedure can be described in Figure 1.

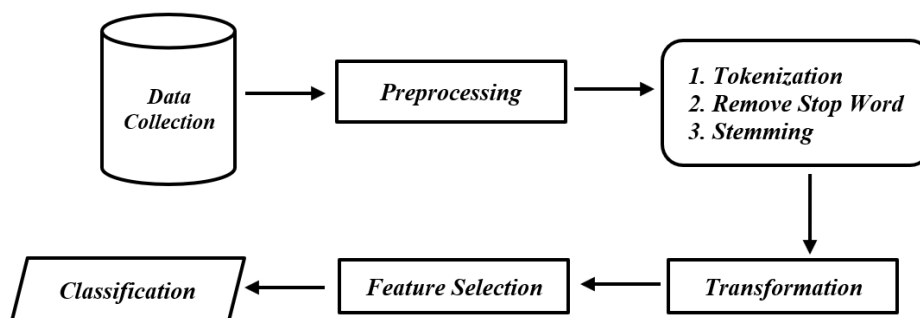


Figure 1. Step and techniques of text mining procedure

2.3. Pre-processing

Preprocessing is a process of preparing and cleaning data on a dataset or data collection for classification. The function of preprocessing is for reducing noise in text mining, so that it can help improve classification performance and speed up the classification process [23]. The preprocessing procedure is divided into three parts:

- 1) Tokenization, the process by which a set of text data is broken down into words, spaces, and after the text data is broken down, excess spaces and punctuation marks are removed, leaving only the letters characters left.
- 2) Remove stop words, stop words are words that have meaningless. The process worked by eliminating stop words.
- 3) Stemming, Stemming is the process of changing words that contain affixes into basic words.

2.4. Transformation

Each word has a weight in the corpus which is calculated using the term frequency-inverse document frequency (TF-IDF). This method is used to find the representation of the value obtained from the text data set that results will be formed or transformed into a vector between documents with terms or words.

$$W_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \quad (1)$$

w_i, j is the weight of term i in document j , tf_i , while j is the frequency of term i in document j , then N is the number of documents in the collection, so df_i is the frequency of term documents in the collection. The following is an example of the application of TF-IDF.

From Table 1 shows the weight of each data containing the word crime that leads to the crime keywords that has been determined. The sentences gets from the Indonesia language tweet and post Facebook. After the data is converted to vector form, then the classification begin by using neural-net multi classifier, which has several algorithms for classifying, such as naïve bayes, logistic regression, SVM, decision tree. These four classifiers algorithm, based on references, most of the time provide a considerable good accuracy. In the first cycle, these four classifiers algorithm were run in parallel, then the classifier that provides the best crime classification become the winning classifier and will be used in the next learning process cycle to map the index crime.

Table 1. Example of a term frequency (TF) matrix

No	Document		Tweet-1	Tweet-2	Tweet-3
1	In action	In Indonesian “beraksi”	0.000000	0.000000	0.408248
2	Fugitive	In Indonesian “buron”	0.000000	0.316228	0.000000
3	Checked	In Indonesian “diperiksa”	0.408248	0.000000	0.000000
4	Arrested	In Indonesian “ditangkap”	0.000000	0.316228	0.000000
5	Guess	In Indonesian “dugaan”	0.408248	0.000000	0.000000
6	Million	In Indonesian “juta”	0.000000	0.316228	0.000000
7	Village head	In Indonesian “kades”	0.408248	0.000000	0.000000
8	Kaligunting		0.408248	0.000000	0.000000
9	Corruption	In Indonesian “korupsi”	0.408248	0.000000	0.000000
10	Raw	In Indonesian “mentah”	0.000000	0.316228	0.000000

2.4.1. Logistic regression

Logistic regression (logit) is a statistical analysis method that describes the relationship between the response variable (dependent variable) which is qualitative in nature having two categories or more explanatory variables (independent variable) sliding categories or intervals [24]. Description of the relationship between the response variables that have the trait qualitative or categories with explanatory variables that have been has two or more categories cannot be completed with the usual linear regression model using the method ordinary least squares (OLS) [25]. The algebraic equation model like OLS that we usually use: $Y = B_0 + B_1X + e$. Where e is the error variance or residual. With this regression model, it does not use the same interpretation as the OLS regression equation. Model the equation formed is different from the OLS equation. The following is the (2):

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = B_0 + B_1 X \quad (2)$$

\ln : natural logarithm. Where:

$B_0 + B_1X$: commonly known equation in OLS

While P accent is the logistic probability obtained by:

$$\hat{p} = \frac{\exp(B_0 + B_1 X)}{1 + \exp(B_0 + B_1 X)} = \frac{e^{B_0 + B_1 x}}{1 + e^{B_0 + B_1 x}} \quad (3)$$

2.4.2. Naïve bayes

The naïve bayes algorithm is one method that is often used in text mining research. A naïve bayes classifier is a probabilistic machine learning model that's used for classification task. The crux of the classifier is based on the bayes theorem. Naïve bayes is also a supervised learning with probabilistic elements, which assumes the terms appear independently, whereas a naïve bayes multinomial is a naïve bayes with a multinomial feature distribution.

2.4.3. SVM

SVM is a statistical method that can be applied to perform classification. SVM is a technique for finding hyperplanes that can separate two sets of data from two different classes. SVM has the basic principle of a linear classifier, namely a linear classification separable, but SVM has been developed to work on non-linear problem with including kernel concept in workspace high dimensions. In a sufficiently high-dimensional space, then look for a hyperplane maximizing the distance or margin between data classes [26].

2.4.4. Decision tree

Decision tree is a method that is able to make informal or simple decisions, but can also be used to predict results systematically, for example in biometrics or smart-card authentication [27] or more broader view in data analysis, especially text mining classification [28]. The decision tree algorithm belongs to the family of supervised machine learning algorithms. The purpose of this algorithm is to create a model that predicts the value of the target variable, where the decision tree uses a tree representation to solve the problem where leaf nodes correspond to class labels and attributes are represented on the tree's internal nodes.

2.5. Feature selection

Feature selection is done to make the classification process more efficient by reducing the amount of data to be analyzed. Feature selection identifies relevant features to be considered in the classification process. Ideally, the feature selection stage filters which features will enter the classification process or learning process through identifying the parts of the corpus that contribute to positive/negative sentiments, then combining the parts of the corpus with the aim of entering a document into one of the polar category.

2.6. Evaluation

In this study, the author uses several selected models that always show better accuracy, including: logistic regression (TF-IDF), naïve bayes, SVM and decision tree. The results of the classification are then evaluated using a confusion matrix which includes precision calculations, recall, and F1. Precision is the total of true positives (TP) divided by all positive predictions TP and false positives (FP):

$$Precision = TP / (TP + FP) \quad (4)$$

The recall is the total of TP divided by true positives TP and false negatives (FN):

$$Recall = TP / (TP + FN) \quad (5)$$

The F1 score serves to determine the balance between precision and recall. Here's how to calculate it:

$$F1 = 2x (precision \times recall) / (precision + recall) \quad (6)$$

2.7. Crime index measurement (CI_t)

The crime index is the percentage increase or decrease in a crime occurrence during a certain period of time compared to a certain time (which is used as the base time). The following formula for determining the crime index:

$$CI_t = \frac{\text{Number of crimes committed at time } t}{\text{Number of crimes committed at time } t_o} \times 100 \quad (7)$$

Where: t = time t and t_o = time base.

3. RESULTS AND DISCUSSION

3.1. Evaluation classification using logistic regression algorithm (TF-IDF)

The following is a presentation of the classification using logistic regression. The classification model shows precision, recall, F1-score and support. In this approach, the classifier that provides the best accuracy will be the winning classifier and will be used in the next learning process. Data in Table 2 describe that the accuracy of using logistic regression (TF-IDF) is 0.90. The Accuracy from several model among other are naïve bayes, SVM, and decision tree in parallel, logistic regression given accuracy much better than others model. The result from of the test will be used for next research or learning process to make sure this model more accurate.

Table 2. The Classification model using logistic regression (TF-IDF)

Number of classification	Precision	Recall	F1-score	Support
0	0.89	1.00	0.94	55
1	0.80	0.80	0.80	15
2	0.00	0.00	0.00	3
3	1.00	0.90	0.95	10
4	1.00	0.50	0.67	2
5	1.00	0.60	0.75	5
7	0.83	1.00	0.91	5
8	0.86	0.75	0.80	8
9	0.00	0.00	0.00	1
10	0.95	1.00	0.98	20
Accuracy			0.90	124
Macro avg	0.73	0.66	0.68	124
Weighted avg	0.87	0.90	0.88	124

3.2. Evaluation using naïve bayes (TF-IDF)

The following is a presentation of the classification using naïve bayes. The classification model shows precision, recall, F1-Score and support. Mostly, other research used naïve bayes always gives good accuracy, moreover Probability calculation is the main reason this algorithm is a text classification friendly algorithm and a major favorite among many people. Data in Table 3 describe that the accuracy of using naïve bayes (TF-IDF) is 0.64. In this approach, the classifier shows better result but defently the result is lower than logistic regression. This model is also suitable for test classification.

Table 3. The Classification model using naïve bayes (TF-IDF)

Number of classification	Precision	Recall	F1-score	Support
0	0.58	1.00	0.73	55
1	0.62	0.73	0.67	11
2	0.00	0.00	0.00	2
3	1.00	0.20	0.33	10
4	0.00	0.00	0.00	1
5	0.00	0.00	0.00	2
7	1.00	0.07	0.13	14
8	0.00	0.00	0.00	12
10	1.00	0.76	0.87	17
Accuracy			0.64	124
Macro avg	0.47	0.31	0.30	124
Weighted avg	0.64	0.64	0.55	124
	Precision	Recall	F1-score	Support

3.3. Evaluation using SVM (TF-IDF)

The following is a presentation of the classification using SVM. The classification model show precision, recall, F1-Score and support. In this approach, SVM will be used for learning process and get the better result from the model. Table 4 presents that the accuracy of using SVM (TF-IDF) is 0.87. These result show that SVM can give better accuracy result, compate to logistic regression model, and the difference is only 0.03. In most cases. SVM gives more accurate result if it compared with naïve buyes, but in this learning logistic regression experiment, it gives better accuracy result.

3.4. Evaluation classification using decision tree (TF-IDF)

The following is a presentation of the classification using decision tree. These learning is the last model will be use to compare with others model. The process almost the same with other learning but definitely the result will see, how much accurate the process is work. Table 5 describes that the accuracy of

using decision tree (TF-IDF) is 0.80. After the text data has gone through all stages of the procedure, the text data were then divided into training data and testing data, with a percentage scale of 80% for training data and 20% for testing data. The following are the results of the evaluation of some of the algorithms that we apply to the classification text. From all the main steps for the text mining process used in this study, by following the pre-processing steps, apply the text mining, and actionable discoveries and discoveries information, the author may identify new information including patterns, issues and themes of the collected social media data. However, applying text to a data set requires continuous evaluation and improvement to achieve the best results.

Table 4. The classification model using support vector mechine (TF-IDF)

Number of classification	Precision	Recall	F1-score	Support
0	0.88	0.93	0.90	15
1	0.82	0.64	0.72	14
2	1.00	1.00	1.00	11
3	1.00	0.87	0.93	55
4	0.95	1.00	0.97	19
5	0.33	1.00	0.50	1
6	0.00	0.00	0.00	0
7	0.67	0.67	0.67	6
8	0.25	0.50	0.33	2
Accuracy			0.87	123
Macro avg	0.65	0.74	0.67	123

Table 5. The classification model using decission tree

Number of classification	Precision	Recall	F1-score	Support
0	0.58	0.70	0.64	10
1	0.75	0.80	0.77	15
2	0.50	0.50	0.50	4
3	0.98	0.89	0.93	61
4	0.95	0.82	0.88	22
5	1.00	0.40	0.57	5
6	0.00	0.00	0.00	1
7	0.40	1.00	0.57	4
8	0.00	0.00	0.00	1
Accuracy			0.80	123
Macro avg	0.57	0.57	0.54	123

Figure 2 shows that the topic tweet of corruption is the highest among others with the total of 267 tweets, in the second position is the topic of fraud and the third is murder. The Figure 3 shows a different pattern when compared to the trend topic on tweeters, where on Facebook the topic trend of theft is in the highest position with a total of 187, the second position is homicides with a total of 150 and the third is narcotics. The conclusion from the data we obtained from social media shows that DKI Jakarta dominates all crimes. The following is a graph of all types of crime by comparing from 3 provinces, namely: DKI Jakarta, West Java and Riau. Figure 4 shows that Jakarta's highest tweet trend crime is corruption, the second is murder and the third is narcotics. This study used the real data from social media and the result then was compared to the result from Central Bureau of Statistics. DKI Jakarta, West Java, and Riau were chosen based on the following reasons: DKI Jakarta and West Java has a very dense population so that crime is very likely happen there, and for Riau province, it is due to the location which is in the borders near other countries, which may have high vulnerability to crime.

3.5. Crime Index

In this study, crime index was divided by two i.e., the crime index based on type of crime and crime index based on time. The crime index based on type of crime, used the crime classifications. Table 6 shows the index of crime based on the type of crime. Table 7 shows that the index based on time and amount of crime, to count the index, N is defined as the biggest predicted crime in each month. It predicted may not more than 1000 crime in a month. The index was calculated by dividing it with 1000 and normalized to 100. Therefore, in calculating the index crime by predicting the number of crime trends on social media in the following months, which used the linear regression to estimate the index crime in the following months. Figure 5. shows that the projected index crime trends for the month July to December 2021 based on the previous calculation from February to June 2021.

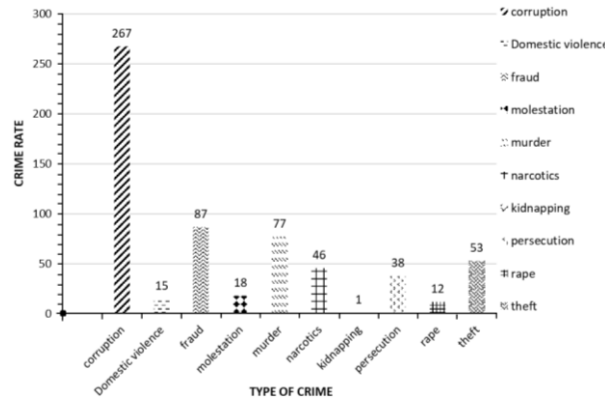


Figure 2. Trend tweet number of crime on Twitter

Figure 2 describing trend tweet from every single crime. There are 10 types of crime that are dominated by corruption as many as 267 tweets. Lately, there has been a lot of discussion about corruption which has become a trending topic, generally young people and intellectuals and influencers.

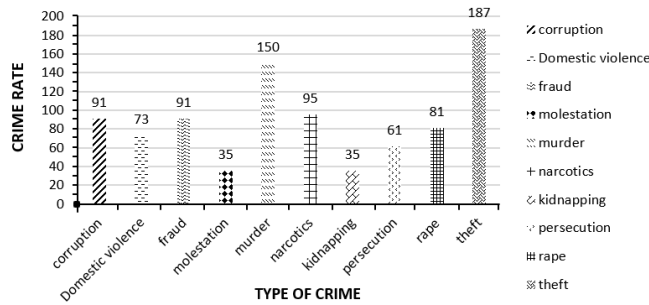


Figure 3. Trend tweet number crime index on Facebook

Figure 3 shows a different pattern when compared to the trend topic on tweeters, on Facebook the trend topic shows that theft is at the highest position with a total of 187, the second position is occupied by murder with a total of 150 and the third is narcotics. The conclusion from the data we obtained from social media shows that DKI Jakarta dominates all crimes. The following is a graph of all types of crime by comparing from 3 places, namely DKI Jakarta, West Java and Riau.

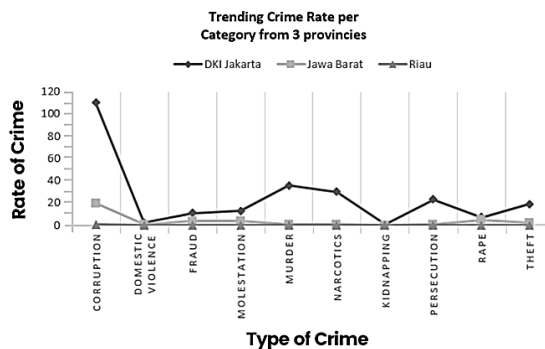


Figure 4. Type of crimes in DKI Jakarta, West Java, and Riau

Figure 4 comparison of 3 areas that are very vulnerable to crime, which are affected by population density and also borders with other countries. The result of the comparison is that DKI Jakarta has the highest crime rate dominated by corruption. To make sure our research is valid minimum the result is approach with the real data from, author will compare our data with Central Bureau of Statistics Indonesia.

Table 6. Crime index based on type of crime

No	Type of crime	Crime index
1	Corruption	$\frac{267}{267} \times 100 = 100.00$
2	Domestic violence	$\frac{15}{267} \times 100 = 5.61$
3	Fraud	$\frac{87}{267} \times 100 = 32.58$
4	Kidnapping	$\frac{1}{267} \times 100 = 0.37$
5	Molestation	$\frac{18}{267} \times 100 = 6.74$
6	Murder	$\frac{77}{267} \times 100 = 28.8$
7	Narcotics	$\frac{46}{267} \times 100 = 17.22$
8	Persecution	$\frac{38}{267} \times 100 = 38.26$
9	Rape	$\frac{12}{267} \times 100 = 12.26$
10	Theft	$\frac{53}{267} \times 100 = 53.26$

Table 7. Crime index based on time (month)

No	Month	Crime index
1	December 2020	$\frac{2}{1000} \times 100 = 0.2$
2	January 2021	$\frac{19}{1000} \times 100 = 1.9$
3	February 2021	$\frac{11}{1000} \times 100 = 1.1$
4	March 2021	$\frac{51}{1000} \times 100 = 5.1$
5	April 2021	$\frac{17}{1000} \times 100 = 1.7$
6	Mei 2021	$\frac{126}{1000} \times 100 = 12.6$
7	June 2021	$\frac{536}{1000} \times 100 = 53.6$

Table 7 the index according to time its toward amount of crime, to count the index this research use N as the mount dividend 1000 as the highest amount in each month then times 100 to find the index. Therefore, in calculating the crime rate, this research tries to predict the number of crime trends on social media for the following months, in this process this research uses linear regression to find out the next index. Furthermore, in this research, try to predict the number of crime rates in the next month that we map for months July to December in 2021. The predictions we get index.

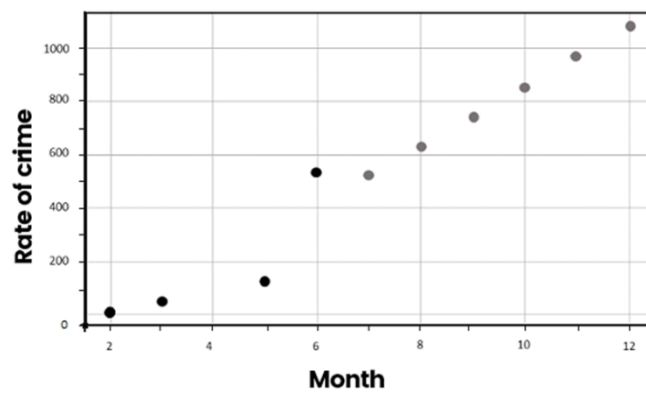


Figure 5. Projected number of crime trends for the month July to December (2021)

Figure 5 describes the projected number of crime trend for July till December. After June, the rate fell a little but the following month increased even more. So that it can be concluded to take the best decision.

4. CONCLUSION

This study proposes the projected index crime based on crime trends. After the crime data is collected, it is then converted to vector form. Followed by the classification which using neural-net multi classifier, which has several algorithms for classifying, such as naïve bayes, logistic regression, SVM, and decision tree. The classifier that provides the best accuracy will be the winning classifier and will be used in the next learning process.

Testing in classifying the results of tweets and opinions of users of social networks Twitter and Facebook against crime categories taken from each crime keyword that shows accuracy by using logistic regression algorithm of 0.90%, SVM gives accuracy of 0.87%, decision tree of 0.80%, and naïve bayes of 0.64%. The results obtained from the classification of types of crimes. Corruption is the highest number of tweets, which is 267. This is the first cycle, as these four classifiers algorithm will run in paralel, then the classifier that provides the best crime classification will be the winning classifier and will be used in the next learning process cycle to projected index crime based on crime trends.

Based on the results of the evaluation in this study by putting forward several suggestions for future work that can be used for further development. For instance, the success of a study is based on good data. Therefore, the timing of data collection is important because the opinions expressed by users of the Twitter social network service on the occurrence of crimes must be collected periodically due to the limitations of the limited Twitter data scrap so that it requires a fairly manual technique, namely periodic and scheduled data scrap. In the preprocessing part, this study have not done scrap data based on user posted location, the next study can considered this approach to improve the precision of this study.

ACKNOWLEDGEMENTS

The authors would like to thank to Directorate General of Higher Education, Ministry of Research and Technology of Indonesia (DGHE-MRTI). This research work was funded by DGHE-MRTI under Grant No. 10/E1/KPT/2021.





REFERENCES

- [1] J. P. Singh, Y. K. Dwivedi, N. P. Rana, A. Kumar, and K. K. Kapoor, "Event classification and location prediction from tweets during disasters," *Annals of Operations Research*, 2017, doi: 10.1007/s10479-017-2522-3.
- [2] S. Asur and B. A. Huberman, "Predicting the Future with Social Media," *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 2010, pp. 492-499, doi: 10.1109/WI-IAT.2010.63.
- [3] M. S. Gerber, "Predicting crime using Twitter and kernel density estimation," *Decision Support Systems*, vol. 61, pp. 115-125, 2014, doi: 10.1016/j.dss.2014.02.003.
- [4] A. Ristea, M. A. Boni, B. Resch, M. S. Gerber, and M. Leitner, "Spatial crime distribution and prediction for sporting events using social media," *International Journal of Geographical Information Science*, vol. 34, no. 9, pp. 1708-1739, 2020, doi: 10.1080/13658816.2020.1719495.
- [5] S. Kaur, P. Kaul, and P. M. Zadeh, "Monitoring the dynamics of emotions during Covid-19 using twitter data," *Procedia Computer Science*, vol. 177, pp. 423-430, 2020, doi: 10.1016/j.procs.2020.10.056.
- [6] W. He, S. Zha, and L. Li, "Social media competitive analysis and text mining: A case study in the pizza industry," *International Journal of Information Management*, vol. 33, no. 3, pp. 464-472, 2013, doi: 10.1016/j.ijinfomgt.2013.01.001.
- [7] J. C. Campbell, A. Hindle, and E. Stroulia, "Chapter 6 - Latent Dirichlet Allocation: Extracting Topics from Software Engineering Data," *The Art and Science of Analyzing Software Data*, pp. 139-159, 2015, doi: 10.1016/B978-0-12-411519-4.00006-9.
- [8] T. Vo *et al.*, "Crime rate detection using social media of different crime locations and Twitter part-of-speech tagger with Brown clustering," *Journal of Intelligent & Fuzzy Systems*, vol. 38, no. 4, pp. 4287-4299, 2020, doi: 10.3233/JIFS-190870.
- [9] S. S. Wijaya, M. Anugerah Ayu and T. Mantoro, "Providing Real-time Crime Statistics in Indonesia Using Data Mining Approach," *2019 5th International Conference on Computing Engineering and Design (ICCED)*, 2019, pp. 1-5, doi: 10.1109/ICCED46541.2019.9161096.
- [10] D. K. Tayal, A. Jain, S. Arora, S. Agarwal, T. Gupta, and N. Tyagi, "Crime detection and criminal identification in India using data mining techniques," *AI and Society*, vol. 30, pp. 117-127, 2014, doi: 10.1007/s00146-014-0539-6.
- [11] S. Mansour, "Social media analysis of user's responses to terrorism using sentiment analysis and text mining," *Procedia Computer Science*, vol. 140, pp. 95-103, 2018, doi: 10.1016/j.procs.2018.10.297.
- [12] L. Wang, L. Zhang, and J. Jiang, "Duplicate Question Detection With Deep Learning in Stack Overflow," in *IEEE Access*, vol. 8, pp. 25964-25975, 2020, doi: 10.1109/ACCESS.2020.2968391.
- [13] G. Hajela, M. Chawla, and A. Rasool, "A Clustering Based Hotspot Identification Approach for Crime Prediction," *Procedia Computer Science*, vol. 167, pp. 1462-1470, 2020, doi: 10.1016/j.procs.2020.03.357.
- [14] G. G. Monkman, M. J. Kaiser, and K. Hyder, "Text and data mining of social media to map wildlife recreation activity," *Biological Conservation*, vol. 228, pp. 89-99, 2018, doi: 10.1016/j.biocon.2018.10.010.
- [15] S. Yadav, M. Timbadia, A. Yadav, R. Vishwakarma, and N. Yadav, "Crime pattern detection, analysis & prediction," *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, 2017, pp. 225-230, doi: 10.1109/ICECA.2017.8203676.
- [16] B. Jefferson, *Crime Mapping*, Second Edition, vol. 3. Elsevier, 2020.
- [17] C. Vandeviver and W. Bernasco, "The geography of crime and crime control," *Applied Geography*, vol. 86, pp. 220-225, 2017, doi: 10.1016/j.apgeog.2017.08.012.
- [18] H. Achrekar, A. Gandhe, R. Lazarus, Ssu-Hsin Yu, and B. Liu, "Predicting Flu Trends using Twitter data," *2011 IEEE*





- Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2011, pp. 702-707, doi: 10.1109/INFOCOMW.2011.5928903.
- [19] F. Franch, "(Wisdom of the Crowds)²: 2010 UK Election Prediction with Social Media," *Journal of Information Technology & Politics*, vol. 10, no. 1, pp. 57-71, 2013, doi: 10.1080/19331681.2012.705080.
- [20] S. Chainey, L. Tompson, and S. Uhlig, "The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime," *Security Journal*, vol. 21, pp. 4-28, 2008, doi: 10.1057/palgrave.sj.8350066.
- [21] L. Vomfell, W. K. Härdle, and S. Lessmann, "Improving crime count forecasts using Twitter and taxi data," *Decision Support Systems*, vol. 113, pp. 73-85, 2018, doi: 10.1016/j.dss.2018.07.003.
- [22] A. Bermingham and A. F. Smeaton, "On Using Twitter to Monitor Political Sentiment and Predict Election Results," *Psychology*, pp. 2-10, 2011. [Online]. Available: <https://doras.dcu.ie/16670/1/saaip2011.pdf>
- [23] L. W. Kennedy, J. M. Caplan, and E. Piza, "Risk Clusters, Hotspots, and Spatial Intelligence: Risk Terrain Modeling as an Algorithm for Police Resource Allocation Strategies," *Journal of Quantitative Criminology*, vol. 27, pp. 339-362, 2011, doi: 10.1007/s10940-010-9126-2.
- [24] V. Ingilevich and S. Ivanov, "Crime rate prediction in the urban environment using social factors," *Procedia Computer Science*, vol. 136, pp. 472-478, 2018, doi: 10.1016/j.procs.2018.08.261.
- [25] G. C. Oatley and B. W. Ewart, "Crimes analysis software: 'Pins in maps', clustering and Bayes net prediction," *Expert Systems with Applications*, vol. 25, no. 4, pp. 569-588, 2003, doi: 10.1016/S0957-4174(03)00097-6.
- [26] V. K. Jain and S. Kumar, "An Effective Approach to Track Levels of Influenza-A (H1N1) Pandemic in India Using Twitter," *Procedia Computer Science*, vol. 70, pp. 801-807, 2015, doi: 10.1016/j.procs.2015.10.120.
- [27] T. Mantoro and A. Milišić, "Smart card authentication for Internet applications using NFC enabled phone," *Proceeding of the 3rd International Conference on Information and Communication Technology for the Moslem World (ICT4M) 2010*, 2010, pp. D13-D18, doi: 10.1109/ICT4M.2010.5971895.
- [28] H. W. Kang and H. B. Kang, "Prediction of crime occurrence from multimodal data using deep learning," *PLOS ONE*, vol. 12, no. 4, pp. 1-19, 2017, doi: 10.1371/journal.pone.0176244.

BIOGRAPHIES OF AUTHORS







Teddy Mantoro     is a Professor in Computer Science at Sampoerna University, Jakarta, Indonesia. His research interest is in the area of Artificial Intelligence, Mobile Computing, Information Security, and Intelligent Environment. He obtained a PhD, an MSc and a BSc, all in Computer Science, and his PhD was from School of Computer Science, the Australian National University (ANU), Canberra, Australia. He is a Senior Member of IEEE (Institute of Electrical and Electronics Engineers) and a Professional Member of ACM (Association for Computing Machinery). He can be contacted at email: teddy.mantoro@sampoernauniversity.ac.id.



Muhammad Anton Permana     is a researcher at postgraduate program, School of Computer Science, Nusa Putra University West Java, Indonesia. His research interests include the applications of Artificial Intelligence, and Mechine Learning. He can be contacted at email: anton.permana@nusaputra.ac.id.



Media Anugerah Ayu     is a Professor in Department of Computer Science, Sampoerna University, Indonesia. She earned a PhD degree in Information Science and Engineering from the Australian National Univeristy (ANU), Australia. Her research interests include machine learning, sentiment analysis, ubiquitous computing, Internet of Things (IoT), intelligent systems, activity recognition, user centered application development, and social computing. She is also an active member of professional organizations in computing and IT, including a Senior Member of IEEE and a Professional Member of ACM. She can be contacted at email: media.ayu@sampoernauniversity.ac.id.