Classification of Criminal Crimes From Data Twitter Using Class Association Rules Mining

Husna Gemasih[1], Rayuwati[1], Azhari SN[2] and Mursalin[3]
{husna_gemasih@yahoo.com}

[1]Department of Informatics, Universitas Gajah Putih, Aceh, Indonesia
[2]Department of Computer Science and Electronics, Universitas Gadjah Mada, Yogyakarta, Indonesia
[3]Department of Mathematics Education, Universitas Malikussaleh, Aceh Utara, Indonesia

Abstract. The police obtain criminal crime data from the field based on reports from a person or group, from the data the police can evaluate the crimes that occurred. The police have no reports of crimes from other parties such as from social media. One such social networking media is Twitter. The information conveyed by Twitter users in a tweet usually contains something related to himself or his environment, including the occurrence of a crime. Such information will serve as data for classification as well as to know the trends of criminal crime. This research uses classification data mining technique that is Class Association Rules (CARs). CARs will find all frequent ruleitems through a series of stages and build rules using frequent ruleitems obtained, then rules will be obtained. The resulting rule will be evaluated to determine the strength of the rule using the Laplace Accuracy equation, which will produce the best rule. These rules will serve as models for the new data classification. The result of accuracy test of this method by using 100 test data is 96%.

Keywords: Twitter Data, Class Association Rules, Laplace Accuracy

1 Introduction

Criminality is all kinds of actions and actions that are economically and psychologically harmful in violation to the laws prevailing throughout the state of Indonesia as well as social and religious norms [1]. During the period 2010-2014 in Indonesia, the number of incidents of crimes against life (murder) and incidents of violations of rights/ property without the use of violence tended to decline. While for the incidence of crimes against physical / bodies (violence) and the incidence of crime-related narcotics fluctuates with the tendency to increase. Number of crime incidents against morality, crimes against rights / property for the use of violence and crime events related to fraud fluctuate. Then for the incidence of crimes against people's independence (abduction) tends to increase [2].

Police of the Republic of Indonesia or abbreviated with the Police in relation to the Government is to play a role in maintaining security and public order, and providing protection. The police only have data on crime reports from the field or in the field alone, have no reports of crimes from other parties such as from social media. Data from social media can be used to become one of the supporting reports on the crimes that occurred. Given the data

WMA-1 2018, January 20-21, Indonesia

DOI 10.4108/eai.20-1-2018.2281925

reports from the field and from social media, the police have the material consideration or evaluation of crimes that occur to be more accurate. Twitter is one of the social networking media has become part to the pattern of community communication.

Twitter data on crime can be used to generate new information such as to know the trend of crime. The amount of Twitter data allows to use data mining techniques. Minning data method chosen for this research is Class Association Rules Mining. This method has high accuracy and strong flexibility in handling textual data [3].

2 Literature Review

2.1 Classification

According to Prasetyo (2012) classification is a job of valuing data objects to include them in a certain class of available classes. In the classification, there is two main work done, namely the development on the model as a prototype to be stored in a memory and the use with the model to perform the introduction/ classification/ prediction on another data object to being known as the class where the data object in the model already saved.

2.2 Classification

According to Liu (2007) the importance of an association rule can be determined by two parameters, namely support and confidence. Support is the percentage of combinations of items in the database, with confidence being the strong relationship between items in association rules. Suppose $I = \{i_1, i_2, \dots, i_n\}$ is a set of items or objects, the database consists of a set of transactions. An association rule is expressed in the form of: $X \rightarrow Y$, where $X \subset I$, $Y \subset I$ and $X \cap Y = \emptyset$, X is called antencedent and Y is called consequent.

Let T be a collection of transaction data consisting of transactions n. Each transaction is labeled with class y. Let I be the collection of all items in T, Y is the set of all class labels (or target items) and I ∩ Y = ∅. Class Association Rules (CARs) are the implications of form

$X \rightarrow y$, where $X \subseteq I$, and $y \in Y$

Support are defined in equation (1), whereas confidence is defined in equation (2). In general, CARs are different from normal association rules, namely:

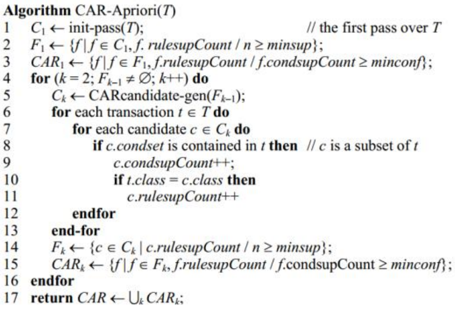
1. Consequent CARs have only one item, while consequent from normal association rules can have any number of items.
2. Consequent y from CARs can only be from label set Y class, that is, y ∈ Y. No item of I can appear as Consequent, and no class label can appear as condition rule. Conversely, normal association rules can have any item as condition or Consequent.

$$support = \frac{rulesupCount}{n} \qquad (1)$$

$$confidence = \frac{rulesupCount}{condsupCount} \qquad (2)$$

The main operation is to find all the rule items that have support above the minsup. A ruleitem is in the form: (condset, y), where the condset ⊆ I is an itemset, and y ∈ Y is the class label. The number of support from the condset (called condsupCount) is the number of transactions in T that contain the condset. The number of support of a ruleitem (called rulesupCount) is the number of transactions in T containing the condset and labeled with class y. Each ruleitem is basically a rule: condset → y, where n is the total number of transactions in T.

The rule generating algorithm, called CAR-Apriori, is shown in Figure 1.



```
Algorithm CAR-Apriori($T$)
$C_1 \leftarrow$ init-pass($T$);                    // the first pass over $T$
$F_1 \leftarrow \{f \mid f \in C_1, f.\,rulesupCount / n \geq minsup\}$;
$CAR_1 \leftarrow \{f \mid f \in F_1, f.rulesupCount / f.condsupCount \geq minconf\}$;
for ($k = 2$; $F_{k-1} \neq \emptyset$; $k$++) do
    $C_k \leftarrow$ CARcandidate-gen($F_{k-1}$);
    for each transaction $t \in T$ do
        for each candidate $c \in C_k$ do
            if $c.condset$ is contained in $t$ then  // $c$ is a subset of $t$
                $c.condsupCount$++;
                if $t.class = c.class$ then
                  $c.rulesupCount$++
        endfor
    end-for
    $F_k \leftarrow \{c \in C_k \mid c.rulesupCount / n \geq minsup\}$;
    $CAR_k \leftarrow \{f \mid f \in F_k, f.rulesupCount / f.$condsupCount $\geq minconf\}$;
endfor
return $CAR \leftarrow \bigcup_k CAR_k$;
```

Figure 1: CAR-Apriori Algorithm

2.3 Evaluation of Rules

Laplace Accuracy is used to estimate the accuracy of the resulting rule, defined in Eq. (3).

$$LaplaceAccuracy = \frac{(n_C + 1)}{n_{total} + k} \qquad (3)$$

k is the number of classes that exist, is the total number of examples that satisfy all the antecedent rules generated, of which there are examples of class c, which is the predicted class [6].

2.4 Trend

A trend is something that is popular at a certain point in time. While trends usually refer to a particular style in fashion or entertainment, there may be a tendency for warmer temperatures (if people follow trends associated with global warming). Trend can happen in any field and not just reflect fashion, culture and entertainment. Trend can also occur in the stock market, politics, and others [7].

2.5 Testing

Testing performance evaluation algorithm in doing categorization will be determined based on accuracy value, that is defined in equation (4) [8].

$$accuracy = \frac{correct\ number\ of\ categories}{number\ of\ documents\ test} \times 100\% \ (4)$$

3 Methods

The criminal classification system is a system that will categorize a text into one class. That is a class of crime against life, crime against morality, crime against people's freedom, property/goods crime, drug-related crime, fraud-related crime and physical/body. In outline can be seen in Figure 2.

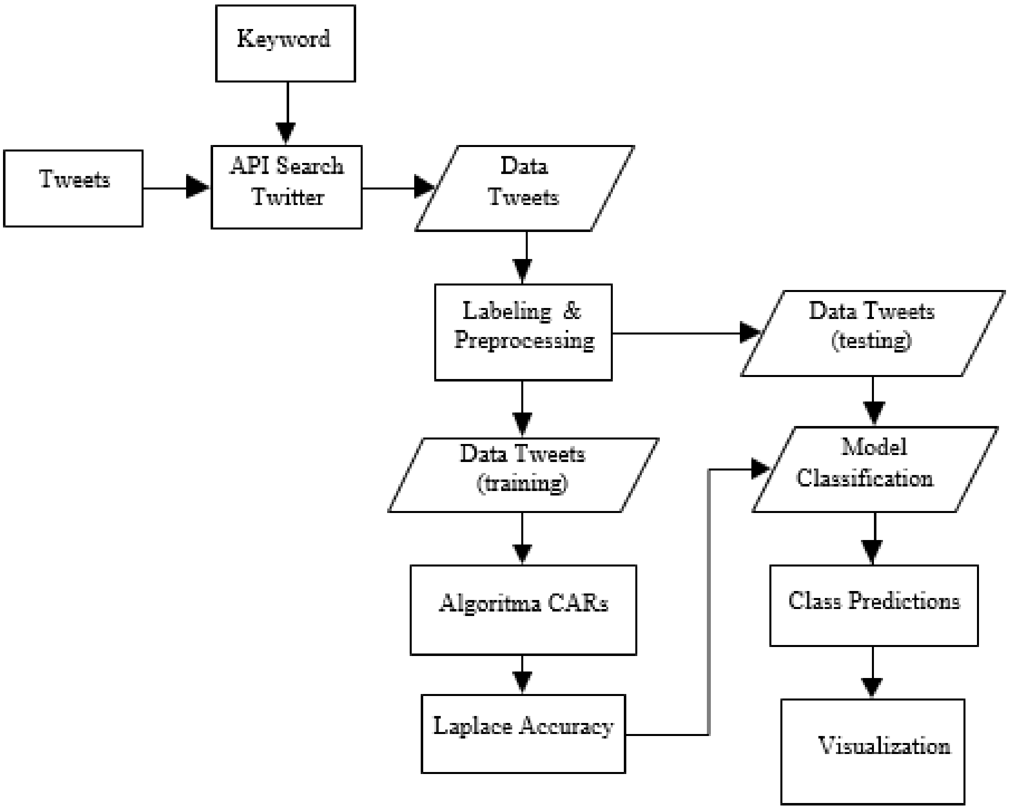


Figure 2: Overview of Crime Classification System

3.1 Tweet Collection

In outline can be seen in Figure 3.

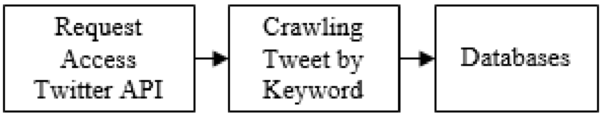


Figure 3. Tweet Collection

3.2 Labeling

The system will be created based on Twitter data, where the data will be done tweet labeling process. The data will be divided into training data and testing data. The training data will produce a model, which will be used to predict the testing data [9].

3.3 Process Preprocessing

Text mining is a step away from text analysis that performed automatically by a computer to extract a quality information from a series of texts summarized in a document [10].

After tweet data is obtained, the next step is to do preprocessing. In the process of preprocessing will do the cleaning of data tweets are downloaded *(dirty tweets)* so it will generate clean tweet data that will be used for the next process. In outline can be seen in Figure 4.

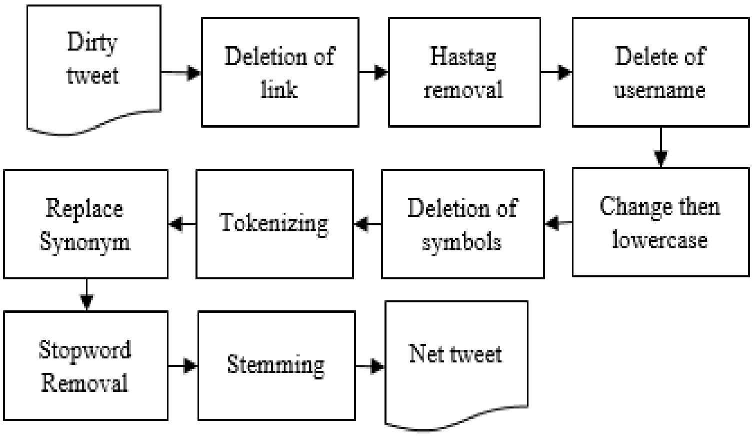


Figure 4: The Preprocessing Process

3.4 Development of Rules

The rule development process begins by finding all frequent ruleitems and constructing CARs, using the CAR-Apriori Algorithm in Figures 2[6], [11]–[15].

3.5 Classification Process

The classification process is done after obtaining the rules in the classification model. Classification is done by choosing the rules that match the test data, then calculate the average Laplace Accuracy value of the rules for each class. The highest average Laplace Accuracy value will be used as the prediction class [16].

4 Results

4.1 Testing Based on the Minimum Value of Confidence and Support

Testing using minimum value of confidence and support is test based on minimum value of confidence and support specified. The details can be seen in Table 1.

Table 1: Test Results Based on Confidence and Support

Support **Confidence**	**20%**	**40%**	**60%**	**80%**
2 %	58	52	48	45
4 %	22	17	16	16
6 %	17	15	15	15
8 %	3	2	2	2

Table 1 shows the number of rules generated using the minimum value of confidence and support. The higher the minimum value of confidence and support used, the resulting rule will be less. If the minimum value of confidence and support is lower than the number of rules generated will be more and more.

4.2 Testing of Laplace Accuracy Value

This test is used to get the correct Laplace Accuracy value. All rules obtained will be evaluated using Laplace Accuracy value based on Equation (3). The results from the evaluation of the rules can be seen in Table 2.

Table 2: Test Results Using Laplace Accuracy Value

Support Confidence	≥ 0,75				≥ 0,85			
	20%	40%	60%	80%	20%	40%	60%	80%
2 %	35	35	35	35	14	14	14	14
4 %	22	17	16	16	14	14	14	14
6 %	17	15	15	15	14	14	14	14
8 %	2	2	2	2	2	2	2	2

Table 2 shows that a rule with a Laplace Accuracy value of ≥ 0.75 produce a maximum number of rules of 35 rules compared with a Laplace Accuracy value of ≥ 0.85 only yielding as many as 14 rules.

4.3 Testing Accuracy

Testing accuracy using test data as many as 100 new data representing from seven classes. The results from the test can be seen in Table 3.

Table 3. Testing Results Accuracy

Support	Confidence	Accuracy	
		Laplace value ≥ 0,75	**Laplace value ≥0,85**
2 %	20%	88%	88%
4 %	20%	96%	88%
6 %	20%	96%	88%
8 %	20%	12%	12%
2 %	40%	88%	88%
4 %	40%	96%	88%
6 %	40%	96%	88%
8 %	40%	12%	12%
2 %	60%	88%	88%
4 %	60%	96%	88%
6 %	60%	96%	88%
8 %	60%	12%	12%
2 %	80%	88%	88%
4 %	80%	96%	88%
6 %	80%	96%	88%
8 %	80%	12%	12%

Table 3 shows the best accuracy obtained by using Laplace Accuracy value ≥ 0.75. From the test results above minimum confidence and support are taken the highest of the pair with minimum value of 80% confidence and support 6% with an accuracy of 96%.

4.4 Characteristics of Rules

Based on the minimum confidence value of 80% and 6% support and the result of evaluation of the rule using Laplace Accuracy value ≥ 0.75 obtained strong rule to be used as a classification model. The resulting rules can be seen in Table 4.

Table 4. Result Rules

No	**Rules**	**Confidence**	**Support**	**Laplace Accuracy**
1	Bunuh → terhadap nyawa	87.93	7.12	0.8
2	Aniaya → fisik/ badan	96.07	6.84	0.86
3	Keras → fisik/ badan	98.03	6.98	0.87
4	Perkosa → kesusilaan	94.54	7.26	0.85
5	Culik → kemerdekaan orang	97.95	6.70	0.87
6	Curi → hak milik/ barang	94.91	7.82	0.86
7	Maling → hak milik/ barang	94.33	6.98	0.85
8	Rampok → hak milik/ barang	94.33	6.98	0.85
9	Gelap → terkait penipuan	98.24	7.82	0.89
10	Korupsi → terkait penipuan	98.03	6.98	0.87
11	Tipu → terkapenipuan	98.27	7.96	0.89
12	Edar → terkait narkotika	100	8.10	0.90
13	Ganja → terkait narkotika	100	7.12	0.89
14	Narkoba → terkait narkotika	95.58	9.07	0.88
15	Sabu → terkait narkotika	100	7.40	0.9

4.5 Testing Accuracy

The test uses 13097 test data for 10 months collected from 1 January 2016 to 22 October 2016. Data is categorized into 7 classes. This test is done to find out the trend of crime.

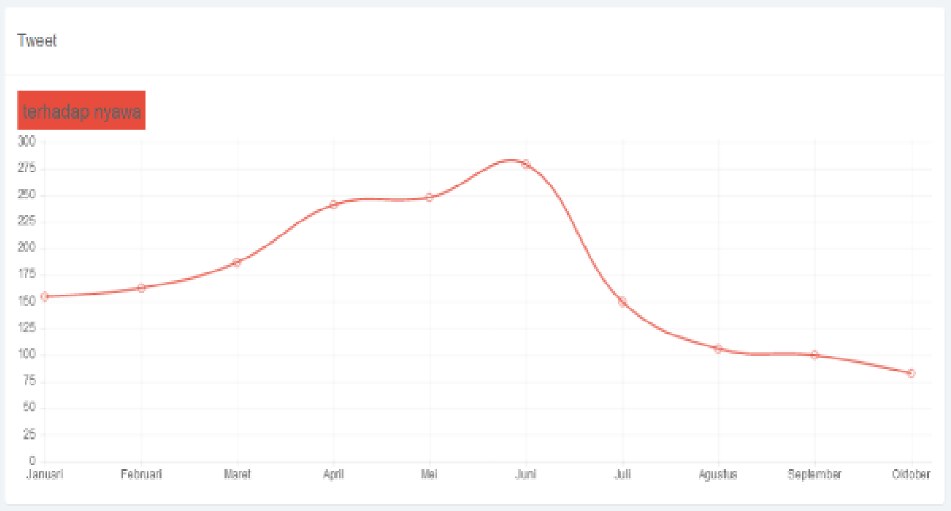


Figure 5. Visualization of Test Results of Crimes *Terhadap Nyawa*

Crimes *terhadap nyawa* have increased and decreased. As shown in Figure 5 crimes *terhadap nyawa*of the trend in June. Meanwhile, for crimes *terhadap fisik/badan* tend to decrease. Based on Figure 6 the crime *terhadap fisik/badan* experienced the trend in January.

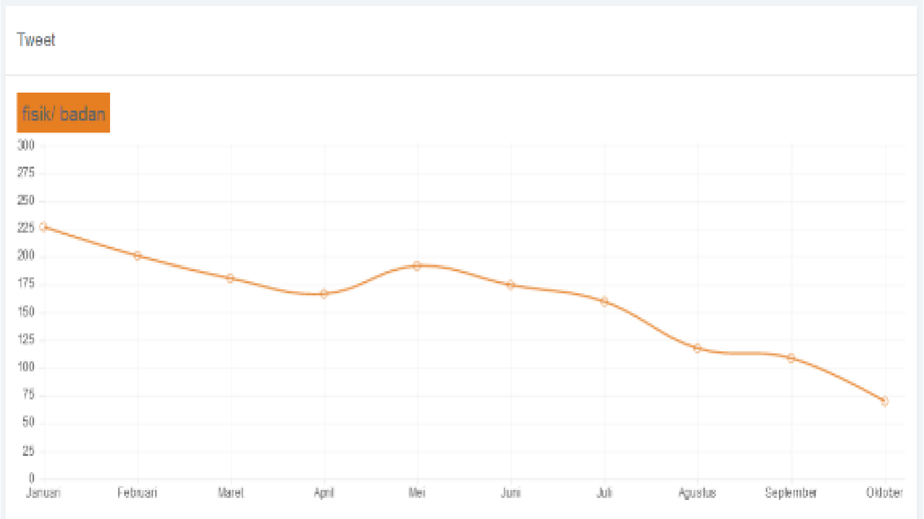


Figure 6. Visualization of Test Results of Crimes *Terhadap Fisik/Badan*

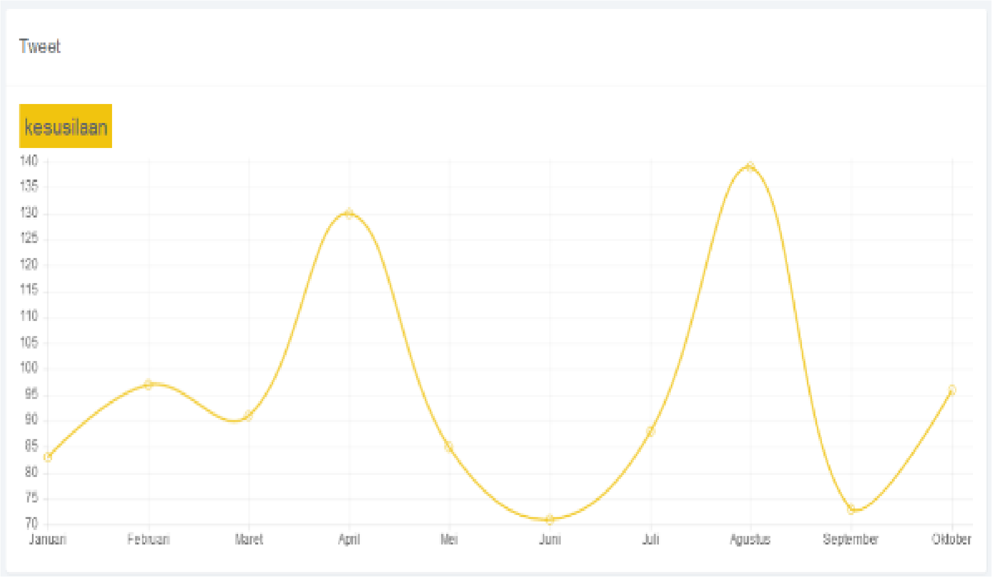


Figure 7. Visualization of Test Results of Crimes *Terhadap Kesusilaan*

Crimes *terhadap kesusilaan* fluctuate. As shown in Figure 7 crimes *terhadap kesusilaan* of the trend in August. Then, for crimes *terhadap kemerdekaan orang* also fluctuate. Based on Figure 8 crimes *terhadap kemerdekaan orang*of the trend in June.

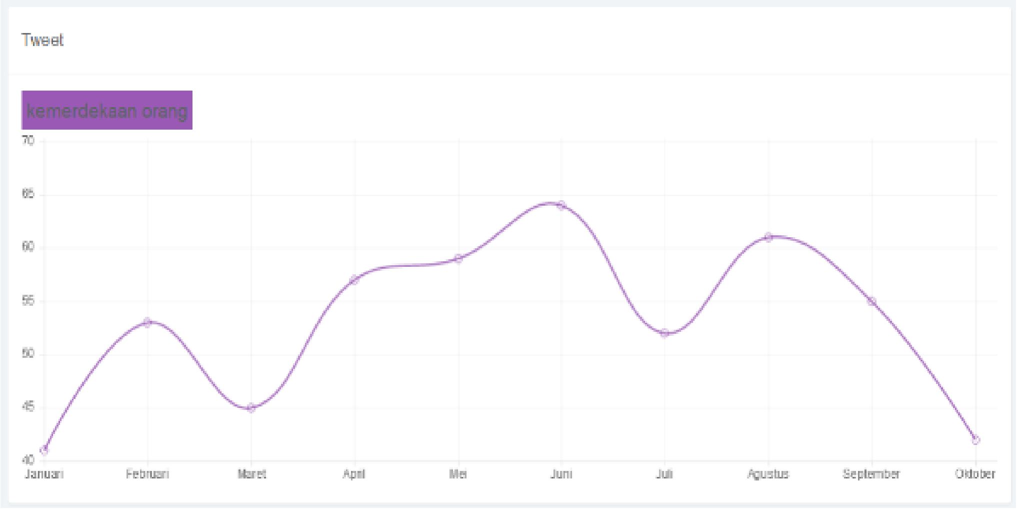


Figure 8. Visualization of Test Results of Crimes *Terhadap Kemerdekaan Orang*

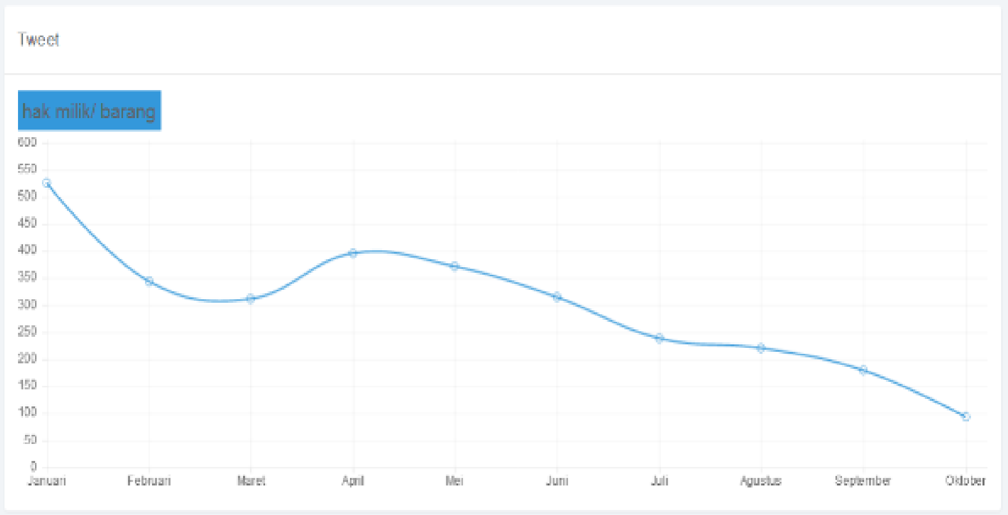


Figure 9. Visualization of Test Results of Crimes *Terhadap Hak Milik/ Barang*

Crime *terhadap hak milik barang* tends to decline. As shown in Figure 9 shows the trend of crime *terhadap hak milik barang* occurred in January. Based on Figure 10, crimes *terkait penipuan* fluctuates tends to decrease. The picture shows crime-terkait penipuan being a trend in February.

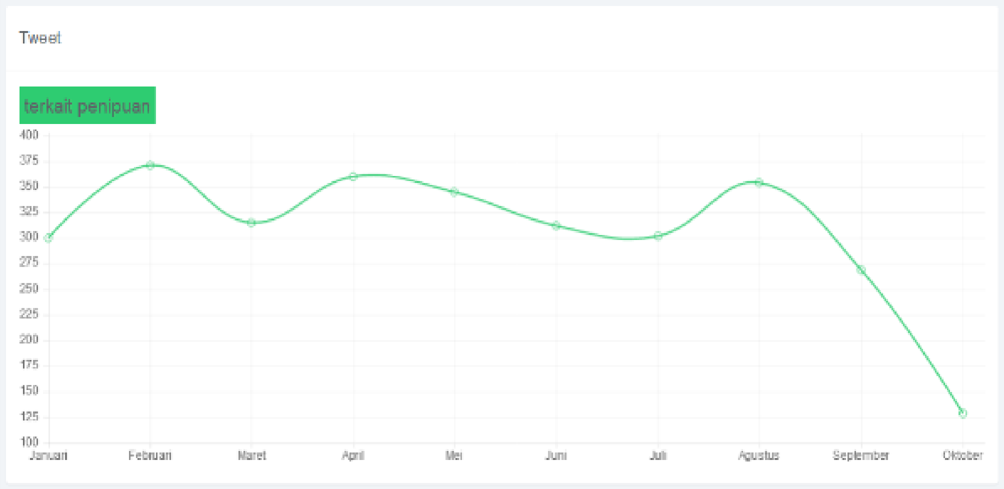


Figure 10: Visualization of Test Results of Crimes *Terkait Penipuan*

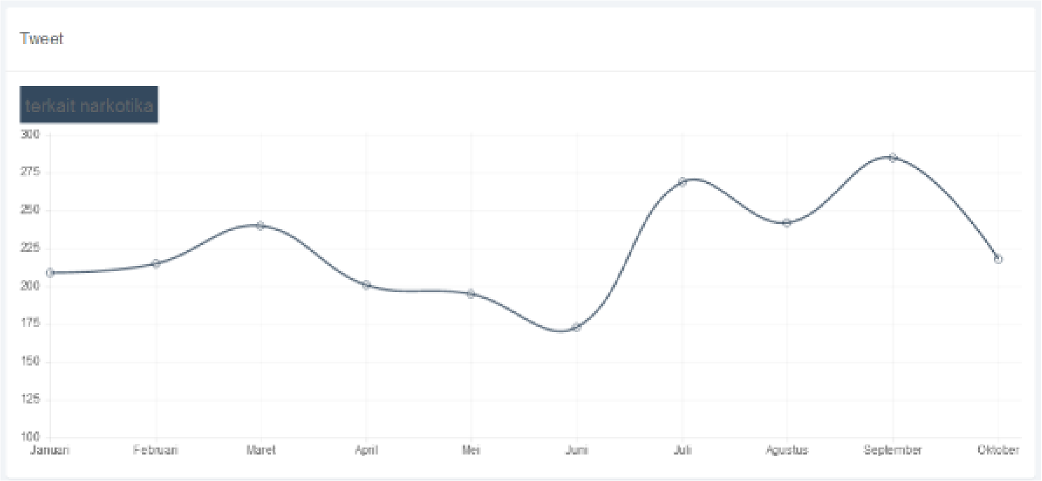


Figure 11. Visualization of Test Results of Crimes *Terkait Narkotika*

Figure 11 shows the crimes *terkait narkotika* being a trend in September.

5 Conclucion

Conclusions Based on the research that has been done then it can be concluded that:

1. Based on the accuracy test using the minimum confidence value of 80% and the minimum support of 6% which has the highest accuracy.
2. From the data as much as 13.097 for 10 months found that, crime *terhadap nyawa*of the trend in June, the crime *terhadap fisik/ badan*of the trend in January. Crime *terhadap kesusilaan* ofthe trend in August. Crime *terhadap kemerdekaan orang*of the trend in June. Crimes *terhadap hak milik/ barang*of the trend in January. Crimes *terkait penipuan*of the trend in February, and crimes*terkait narkotika* became a trend in September.
3. Accuracy test results using 716 training data and 100 test data indicate that the Class Association Rule Mining method has an accuracy of 96%.

References

[1] K. Kartono, *Patologi sosial*, I. Jakarta: rajawali pers, 2009.
[2] B. P. Statistik, “Statistik Kriminal 2015,” 2015. .
[3] Nidhi and V. Gupta, “Recent Trends in Text Classification Techniques,” vol. 35, no. 6, pp. 45–51, 2011.
[4] E. Prasetyo, *Data Mining Konsep dan Aplikasi menggunakan MATLAB*. 2012.
[5] B. Liu, *Web Data Mining*. 2007.
[6] J. Yin, X. & Han, “CPAR : Classification based on Predictive Association Rules,” *Proc. SIAM Int. Conf. Data Mining. San Fr. CA SIAM Press*, pp. 369–376, 2003.
[7] Vocabulary, “Trend,” 2016. .
[8] A. Hamzah, “Klasifikasi Teks Dengan Naïve Bayes Classifier (NBC) Untuk Pengelompokan

Teks Berita Dan Abstract Akademis," *Pros. Semin. Nas. Apl. Sains Teknol. Periode III*, no. 2011, pp. 269–277, 2012.

[9] S. Rodiyansyah and E. Winarko, "Klasifikasi Posting Twitter Kemacetan Lalu Lintas Kota Bandung Menggunakan Naive Bayesian Classification," vol. 6, no. 1, pp. 91–100, 2012.

[10] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, vol. 12. 2011.

[11] R. Kartika and I. Mukhlash, "Penerapan Algoritma Klasifikasi Berbasis Aturan Asosiasi untuk Data Meteorologi," vol. 1, no. 1, pp. 1–6, 2013.

[12] R. Amseke; and E. Winarko, "Aplikasi Algoritma CBA untuk Klasifikasi Resiko Pemberian Kredit," vol. 8, no. 2, pp. 121–132, 2014.

[13] N. Ransi and E. Winarko, "Algoritma CPAR untuk Analisa Data Kecelakaan," vol. 8, no. 2, 2014.

[14] C. M. Rahman, "Text Classification using the Concept of Association Rule of Data Mining," 2000.

[15] F. Thabtah, "A review of associative classification mining," 2007.

[16] M. Nandhini and S. N. Sivanandam, "An improved predictive association rule based classifier using gain ratio and T-test for health care data diagnosis," *Sadhana - Acad. Proc. Eng. Sci.*, vol. 40, no. 6, pp. 1683–1699, 2015.