International Journal of Engineering & Technology, 7 (4.40) (2018) 112-117

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Mapping Criminal Location Entity from Indonesian Online Newspapers

Neno Sulistyawan, Sari Widya Sihwi, Wiranto

Informatics Department, Sebelas Maret University
Jl.Ir Sutami No 36A,Jebres, Surakarta, Indonesia
**Corresponding author E-mail:* sariwidya@staff.uns.ac.id

Abstract

The aim of this study is to extract entity locations on crime news in Indonesian online newspapers and to tag the locations into a map. The methods used in this study are rule-based algorithm, for identifying and extracting entity location of the crime, and SVM (Support Vector Machine) algorithm, for classifying which sentence containing the location of the crime. Every sentence containing criminal location was included in geocoding process so it could be mapped into a digital map. The accuracy of identifying the entity location by using rule-based algorithm is 96.2%. SVM model that has the best accuracy in classifying sentences that contains entity scene of the crime is Radial kernel whose accuracy is 95.77%.

Keywords: *Crime; Geocoding; Information Extraction; Rule Based Algorithm; Support Vector Machine.*

1. Introduction

Based on data released in 2016 from the Central Bureau of Statistics, crime cases that occurred in Indonesia in 2015 increased to 352,936, which were the highest number in five years. In this year also, a number of types of crime have increased, such as murder, rape and others. This figure shows that of 100,000 people in Indonesia, 140 of them are at risk of becoming victims of crime [1]. Knowing the places that became the location of the crime, of course, will further increase the vigilance of a person to become victims of similar abuse. This information can be a consideration for someone to make decisions of many things. Some examples of decision-making that take this into consideration are to find a place to live, to choose a school or to determine the way to go when traveling alone at night. But unfortunately, there is currently no criminal information in Indonesia presented in the form of map.

In the other hand, the rapid growth of information in the internet, including criminal information is getting higher everyday. Unfortunately, it is not balanced with the human speed to process the information. An automatic process, that, is called Information Extraction (IE), is required to take the necessary parts of the information[2].

Information Extraction (IE) is a technology related to a way of making unstructured or semi-structured texts or documents with specific domain into a structure relevant information. Unlike the Information Retrieval (IR) which concentrates on how to identify the relevant document from the document collection, IE generates structured data to be ready for subsequent processing [3]. Usually an IE process is defined by the input and extraction targets. One of the sub-task of IE is to help the process in identifying and extracting information that is called Named Entity Recognition (NER). NER process helps users to produce more meaningful corpus by identifying the proper name for the corpus and classify them into groups, such as the person's name, organization, location and so on [4]. NER have been used in many domain for many purposes, such as [5] that used NER for identifying biomedical instances, [6] and [7] that used NER to search-query processing in the travel domain, [8] that used NER to give alternate routes to commuters if violence is detected and many others.

In the domain of crime and security, some works have been done, but none of them used Indonesian language in their NER processing. Shabat and Omar [9] in 2015 designed a model for extracting crime-specific information from Malaysian newspaper and social media. Jayaweera et al [10] proposed an intelligent crime analysis system to analyze the growing volumes of crime related data. Arulanandam, Savarimuthu usand Purvis [11] extracted crime information from three different countries, New Zealand , Australia and India.

NER researches have been done widely for English because it is the most widely used language in the world. In spite of that, there are only few researches in NER for Bahasa Indonesia [12]. Some of the researched are [7], and [13]. [7] used Naive Bayes Classifier and got 74% precision with 70% recall. [13] used rule-based algorithm to identify the locations where the tropical disease outbreaks occurred and to identify the sentences that contain the occurrence date and the number of victims and it got 99.8% as its accuracy.

This paper suggested to extract criminal information taken from online newspaper using NER and presented them into a digital map. The data source of this research was gained from Tribunnews.com, an Indonesian online newspaper, because this newspaper has a special section for criminal cases in Jabodetabek cities (Jakarta, Bogor, Depok, Tanggerang, and Bekasi) since 2012.

In this study, the location entity extraction process can be divided into three main processes, which are entity location identification process using Rule-Based algorithm, training and classification process using Support Vector Machines (SVM) models, and location entity mapping process into a digital map. Rule-based was

used in the identification process because of its accuracy reached 89,47 used in [14] and 96% in [15]. In this study, Rule-based algorithm empowered with morphological and contextual component and Gazetteer administrative regions in Jabodetabek, which was obtained from the web page http://www.kemendagri.go.id/pages/data-wilayah . SVM model is used to train each sentence having a location entity. The SVM algorithm has been selected because of its good performance for a variety of classification problems [16], and its ability to tolerance to irrelevant and redundant attributes [17]. At the end, each entity criminal news site that passes through the process of checking duplication of news going through the process of geocoding to be displayed on a digital map.

2. Research Method

Data obtained from this research comes from the process of crawling online crime newspaper, Tribunnews.com. In the process of crawling, all html tags were removed, so that the documents produced were in the form of paragraphs. Furthermore, data from crawling process was filtered, so the news that is not a criminal case is removed. In this study, there are 1,010 criminal news selected, which consists of homicide, assault, robbery, theft, and rape / sexual abuse, and drug cases. All the news selected were pre-processed to be "clean" by doing sentence extraction, normalization punctuation, removal of multiple whitespace. After that, the "clean" data were processed in the stage of identification of the location of the entity, as can be seen in Figure 1.

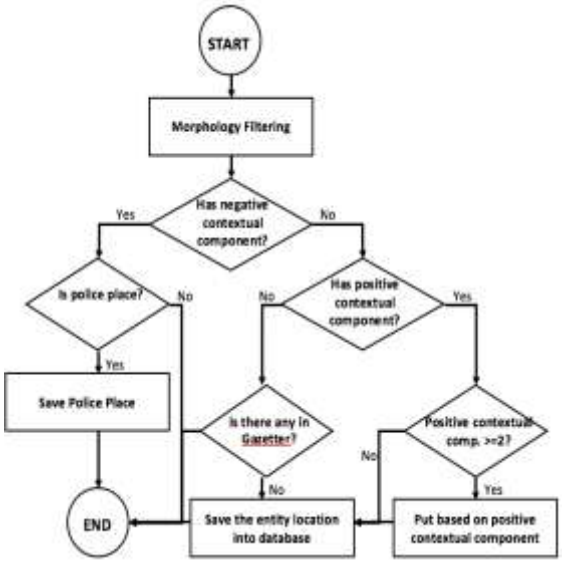

Fig. 1: Location Identification Flow

2.1. Filtering by The Morphology

Each sentence is filtered by using the morphology, which is a kind of formation and shape of token, in order to get any specific entities, such as location entity. Location entity is usually a single token that begins with a capital letter, or some tokens that are capitalized sequential. In addition, this process also handled entity extraction processes that has special characteristics, such as time entity, date entity, general entity and alias name of the person. Table 1 shows all of these filtering processes that can be described as regular expressions. These regular expressions were written using Java format.

Table 1: Regular Expression

No	Name	Regular Expression
1	General	\\s(([A-Z]{1}[a-z]+|[A-Z]{2,}(.)?)+(\\s+([A-Z]{1}[a-z]+|[A-Z]{1,}|\\d+)+)*)(\\s|,|.)")
2	Date	\\(\\d+{1,2}/\\d+{1,2}/?(\\d+{1,4})?\\)
3	Time	(\\s+)((((1|0)[0-9]|2[0-3])\\.([0-5][0-9]))(\\s+(WIB|wib))?)
4	Alias	(([A-Z]{1}[a-z]+|[A-Z]{2,})+(\\s+([A-Z]{1}[a-z]+))*(\\s+\\(\\d+\\s?(tahun)?\\))+)(\\s|,|.)?

2.2. Filtering by Contextual Components

Contextual component is a token that can be a reference for the establishment of the entity's location. A token that has a negative component would be blacklisted as an entity location. In the other hand, a sequence of tokens having contextual positive will be counted the number of positive contextual possessed by it. If there is only one contextual component, it will be directly stored in the database, while the sequence of tokens which has more than one component going through the process of separation. Figure 2 is the algorithm used for the separation of multiple contextual components. Table 2 is the example of contextual components used in this filtering.

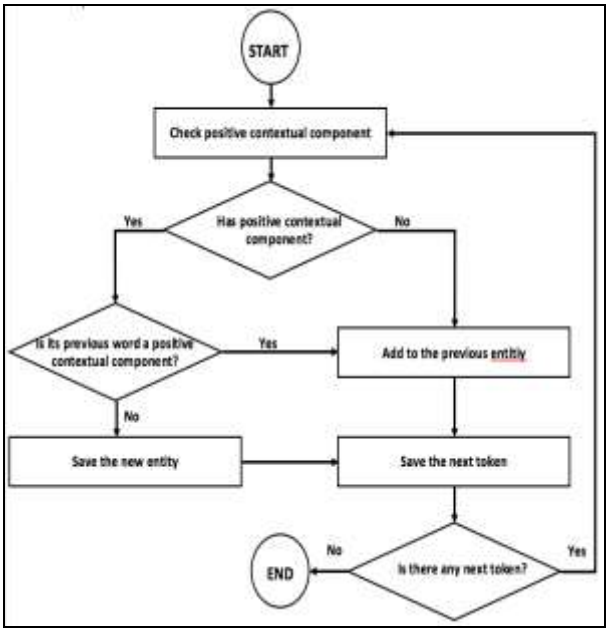

Fig. 2: Multiple Contextual Component Separation

Table 2: Contextual Component

Name	Description	Type	Example
Day	Name of the day	Negative	*Senin, Selasa, Rabu, Kamis*
Month	Name of the months	Negative	*Januari, Februari, Maret, April, Mei, Juni, Juli*
Administrative	Administrative District	Positive	*desa, kelurahan, kecamatan, kabupaten, kota*
Public Place	Name of public places	Positive	*pasar, terminal, masjid, bandara*

2.3. Matching the Gazetteer

Gazetteer is the list of administrative regions. This study used four Gazetteer, namely districts gazetteer, villages gazetteer, and district / city gazetteer, and gazetteer of the province, which were taken from the web page http://www.kemendagri.go.id/pages/data-wilayah. For each token that does not have a positive or negative contextual component then it will be matched with each gazetteer to decide whether the token is the location entity or not.

2.4. Extraction Feature Phase

At this stage, we will perform an extraction feature that can be used to determine the label on every sentence, whether the sentence is a sentence that contains the scene, or a sentence that does

not contain the scene. Table 3 contains the features that will be used for training.

Table 3: List of Features

Feature	Description
FirstSentence	Is it the first word?
ContainsRtRw	Does it contain an RT/RW?
ContainsStreet	Does it contain a street entity?
ContainsCommon	Does it contain a public place?
ContainsKelurahan	Does it contain a kelurahan?
ContainsKecamatan	Does it contain a subdistrict?
ContainsKabupaten	Does it contain a district?
ContainsDate	Does it contain a date?
ContainsDay	Does it contain name of the day?
ContainsTime	Does it contain a specific time?
ContainsAlias	Does it contain an alias?
ContainsCrimeTerm	Does it contain a crime term?
ContainsPolicePlace	Does it contain the place of any police office?
BeforeContainsDate	Does the previous word contain a date?
DuplicatePlaceScore	The simillarity score of an administrative district level in 1 sentence

2.5. Labelling Phase

At this stage, each sentence in all news will be labelled manually into a specific class. There are two classes, which are Sentences with Criminal Location (LCS), and Not a Location Criminal Sentence (NLCS). The results of this phase will serve as training data for SVM algorithm for the next stage.

2.6. Training Stage

In this research, the training will be conducted with SVM using a machine learning software called WEKA (Waikato Environment for Knowledge Analysis). Training data compiled by the format as shown in Table 4. F1 - F15 are all features of Table Feature (Table 3). The final classes consist of LCS (Location Criminal Sentences) and NLCS (Not a Location Criminal Sentence).

Table 4: Training Data Format

F1	F2	F3	F4	F5	…	F15	Class
0.	1	1	1	1	…	1	LCS
0	0	0	0	0	…	1	NLCS
0	0	0	0	0	…	1	NLCS

SVM can be explained simply as an attempt to find the best hyperplane which serves as a divider of two classes in the input space pattern which is a member of two classes: +1 and -1 and share alternative dividing lines (discrimination boundaries). Margin is the distance between the hyperplane to the nearest pattern of each class. The closest pattern is called a support vector. Attempts to locate the hyperplane is the core of the learning process on SVM [9]

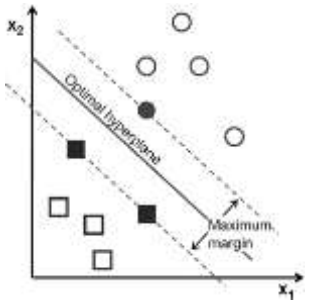

Fig. 3: Hyperplane in SVM [18]

In figure 3, there is a dividing plane that separates all objects based on the class. The first and second plane bordering the class into two, so as to obtain the equation. Provided data is denoted as $x \in \Re^d$ whereas each denoted by $y_i \in \{-1, +1\}$ for i = 1,2, ..., l, where l is the number of data. The second class is assumed to -1 and +1 can be separated completely by hyperplane dimension d, which is defined:

$$\vec{w}.\vec{x} + b = 0 \quad (1)$$

x = vector input
w = weight vector
b = bias

Pattern $\vec{x}$ which belongs to the class -1 (negative samples) can be formulated as a pattern that satisfies the inequality:

$$\vec{w}.\vec{x} + b \leq -1 \quad (2)$$

While the pattern that includes classes +1 (positive samples) meets inequality:

$$\vec{w}.\vec{x} + b \geq +1 \quad (3)$$

The margin can be found by maximizing the value of the distance between the hyperplane and the closest point, namely .This can be formulated as a Quadratic Programming (QP) problem, ie finding the minimum point of the equation (5) with respect constraint equation (8).

$$\min_{w} \tau(w) = \frac{1}{2}\|\vec{w}\|^2 \quad (4)$$

$$y_i(\overrightarrow{x_i}.\vec{w} + b)\text{-}1 \geq 0, \forall_i \quad (5)$$

$\min \tau(w)$= Quadratic problem
$\frac{1}{\|\vec{w}\|}$= Distance between the closest point hyperlane
This problem can be solved with a variety of computational techniques, such as by Lagrange Multiplier.

$$L(w, b, \alpha) = \frac{1}{2}\|\vec{w}\|^2 - \sum_{i=1}^{l} \alpha_i \left(y_i((x_i.w_i + b) - 1)\right)$$

$$\text{for } (i = 1, 2, 3, \ldots, l) \quad (6)$$

L = Lagrange Multiplier.
α_i = Weight value of each point
y_i = Value of output
The optimal value of Equation (18) can be calculated by minimizing L against $\vec{w}$ and $\vec{b}$, and maximizing L against $\propto_i$. Regarding to the nature that at the point of optimal gradient L = 0, equation (6) can be modified as the maximization problem that only contains αi, as equation (7)
with its constraint as equation (8)

$$\sum_{i=1}^{l} \propto_i - \frac{1}{2}\sum_{i,i=1}^{l} \propto_i \propto_i y_i\, y_i\, \overrightarrow{x_i}\,\overrightarrow{x_i} \quad (7)$$

$$\propto_i \geq 0\ (i = 1,2,3, \ldots, l) \sum_{i=1}^{l} \propto_i y_i = 0 \quad (8)$$

From the results of this calculation are mostly obtained αi as positive. Data correlated with positive αi is called a support vector. Sometimes there are two classes that can not be separated linearly. To resolve the problem of non-linear, SVM should be modified by incorporating kernel function. In non linear SVM, first mapped data x by the function x Φ to vector space into a higher dimension. In this new vector space, a hyperplane separating the two classes can be constructed. This condition is called as Kernel Trick that can be formulated as equation (9).

$$K(\overrightarrow{x_i}, \overrightarrow{x_i}) = \Phi(\vec{x_i}).\Phi \quad (9)$$

$K(\overrightarrow{x_i}, \overrightarrow{x_i})$ = kernel function
Φ = mapping function from input space to vector space

Kernel trick provides various facilities. This is because during the SVM learning process, to determine support vector, we just need to know the kernel function used, and do not need to know the form of a non-linear function of Φ. This study will use three common kernel types (Table 5) in the training process and also compare their results.

Table 5: Kernel Types in SVM

Kernel Types	Definition
Linear	$K(\overrightarrow{x_i}, \overrightarrow{x_i}) = (\overrightarrow{x_i}.\overrightarrow{x_i})$

Polynomial	$K(\vec{x_i}, \vec{x_i}) = (\vec{x_i}.\vec{x_i} + 1)^p$
Radial	$K(\vec{x_i}, \vec{x_i}) = exp(-\frac{\|x_i - x_j\|}{2\sigma^2})$

2.7. Classification Phase

The classification process is done with SVM algorithm based on the model of the training results, and the results of the classification and the location of the detected entity is stored into the database. Furthermore, the classification result from data $\vec{x}$ is the result of the equation (10).

$$f(\Phi(\vec{x})) = \vec{w}.\ \Phi(\vec{x}) + b \quad (10)$$

$$= \sum_{i=1 \cdot x_i \in SV}^{n} \Phi(\vec{x_i}).\Phi(\vec{x_i}) + b \quad (11)$$

$$= \sum_{i=1 \cdot x_i \in SV}^{n} \alpha_i\ y_i\ K(\vec{x_i}, \vec{x_i}) + b \quad (12)$$

Support vector of equation (10), (11) and (12) is subset from training set that is chosen as support vector, which is in another words data $\vec{x_i}$ correspond to $\alpha_i \geq 0$.

2.8. Testing Phase

In this study, the method used for testing is k-fold cross validation with k = 10. In this method, data is divided into k sections at random, then do k of experiments where each experiment using a piece of data to k as a data testing and utilizing the other part as training data. While the classification used to calculate the accuracy of precision, recall, and F-measure. Precision (P) is the number of correct classification of value divided by the total number of classification results (both positive and negative). Recall (R) is the number of correct classification results are worth divided by the number of correct value should be. While the F-measure (F) is a matrix that compute their accuracy value ratio of correct results and apply as average harmonious value of precision and recall.

P = (True Positive) / (True Positive + False Positive) (13)

R = (True Positive) / (True Positive + False Negative) (14)

F = 2PR / (P + R) (15)

2.9. Mapping Phase

Because a criminal case was usually reported in some articles in newspaper, so this condition should be checked to avoid duplication. In this study, if the case has the same date the entity, and the entity level location on the street / public place is the same, then the case is considered the same and are not included in the next process. For each location that is obtained from the location of the criminal sentence is going through the process of geocoding, namely the process of translation of the name of the location to the coordinates of latitude, longitude. In this study the process of geocoding done using the Google Map API. Query is sent comes from the name of the detected road or public location name, village, district, and city or county. The results of the query are stored as location coordinates, and displayed on a digital map.

3. Result and Analysis

3.1 Data Collection

In this study the source of the data used were taken from subsection *Kriminal* of section *Metropolitan* of online news www.Tribunnews.com. A news article in the form of web pages changed to remove all html tags it up into paragraphs paragraphs which will then be stored in the database. The data comes from the crawling crime news article in June 2012 - October 2015. The existing data are then selected manually to get the news article about the crime of robbery and theft, murder and persecution, drugs and rape. The data used in this study of 1,010 articles, which consists of 7,459 sentences containing entity location.

3.2 Sentence Pre Processing

For each paragraph saved is broken down into sentences. Sentences obtained from the separation mark '?', '!', And the '.' Is not a common abbreviation. It also becomes the removal of multiple whitespace and normalization of punctuation. Figure 4 is an example of extracting a sentence of each paragraph obtained from crawling process. After experiencing the process of extracting a sentence then obtained following sentences in figure 5.

Nasib malang menimpa Suryadi (50), warga Kampung Pesing. Ia menjadi korban pembacokan saat berniat melerai tawuran warga, Minggu (26/8/2012) pukul 02.00 WIB dini hari di Jl. Tubagus Angke Kel. Wijayakusuma Kec. Grogol Jakarta Barat. Kejadian berawal saat saksi yakni Achmad Joko (30) bersama dengan teman-temannya sedang makan mie rebus di sebuah warung. Tiba-tiba datang 10 orang tak dikenal menyerang sambil melempari batu, hingga saksi dan teman-temannya melarikan diri.

Fig. 4: Example of Sentences Before Extraction Process

Nasib malang menimpa Suryadi (50), warga Kampung Pesing.
Ia menjadi korban pembacokan saat berniat melerai tawuran warga, Minggu (26/8/2012) pukul 02.00 WIB dini hari di Jl. Tubagus Angke Kel. Wijayakusuma Kec. Grogol Jakarta Barat.
Kejadian berawal saat saksi yakni Achmad Joko (30) bersama dengan teman-temannya sedang makan mie rebus di sebuah warung.
Tiba-tiba datang 10 orang tak dikenal menyerang sambil melempari batu, hingga saksi dan teman-temannya melarikan diri.

Fig. 5: Example of Sentences After Extraction Process

3.3 Location Entity Identification Phase

3.3.1 Filtering Morphology Token

This section will do a filtering based on the pseudocodes in Fig. 2, 3, 4, and 5. The examples of entity results obtained can be seen in Table 5.

Table 6: Morphology Filtering Results

No	Sentences	Entities
1	*Aksi pembegalan menggunakan senjata api terjadi siang tadi, Selasa (21/4/2015) pukul 13.30 WIB di Jl Raya Cipayung, Kelurahan Cipayung, Kecamatan Pancoran Mas, Depok, Jawa Barat*	*Aksi, Selasa, WIB, Jl Raya Cipayung, Kelurahan Cipayung, Kecamatan Pancoran Mas, Depok, Jakarta Barat, (21/4/2015), 13.30 WIB*
2	*Korban diketahui bernama Endang Suhendar, warga Babakan Dangdeur, Kelurahan Pasar Ribu, Kecamatan Cibiru, Bandung, Jawa Barat.*	*Korban, Endang Suhendar, Babakan Dangdeur, Kelurahan Pasar Ribu, Kecamatan Kalibiru, Bandung, Jawa Barat*

3.3.2 Filtering Contextual Component

At this stage, each word will be matched to one or two tokens sequentially in the previous process. For every one or more token sequence containing negative contextual will not be included in the next process.

Seorang juru parkir di sebuah apotek di Jalan Raya Pondok Kelapa, Duren Sawit, Jakarta Timur menjadi korban penembakan oleh kawanan perampok pada Jumat (22/8/2014) pukul 20.30 WIB.

Fig. 6: Example 1 of Filtering Contextual Component

In the above sentence in figure 6, "Jumat" fulfils morphology rules, but because the word is in the negative contextual of days, then "Jumat" will not be included in the process, as well as the name of the month. For some tokens containing a contextual component will be included as a positive entity location. In the above

sentence says "*Jalan Raya Pondok Kelapa*" inserted into the location because the entity meets the rules of morphology and has a contextual component is positive, that said "*jalan*".
While for some token sequence that contains more than one positive contextual component will be separated in accordance with its positive contextual component. As in the following sentence (Fig 7), it cuts token sequence on fragment in "*Jalan Ciherang Sukatani*" on a fragment of the above meet the rules of morphology, but because the token sequence has a positive contextual component more than one, then the sequence of tokens must be broken down into "*Jalan Ciherang*", "Sukatani".

Pasalnya, korban saat dirampok di Jalan Ciherang Kelurahan Sukatani, Tapos, Depok, Rabu (8/4/2015) sekitar pukul 14.50, tidak membawa barang berharga.

Fig. 7: Example 2 of Filtering Contextual Component

3.3.3 Matching the Gazetteer

For a token word that does not have contextual components will be matched with a database region administrative location. At this stage, there is often founded cases with sequenced tokens that has the same location name, such as "Beji,Beji" in the sentence in Figure 8. In the gazetteer, "Beji" was discovered as a village, but due to the location of the previous entity has been found as a village, then the matching Gazetteer was raised to the subdistrict level. So,first "Beji" was identified as a village, and second Beji was identified as a district.

Dengan berbagai kunci duplikat, SS hampir saja berhasil menggasak uang Rp 2,1 Juta dari kotak amal yang ada di Indomaret di Jalan Bambon Raya, RT 2 RW 17, Beji, Beji, Depok, Sabtu

Fig. 8: Example of matching the gazetteer

Based on identification entity location with a combination of filtering morphology, filtering contextual component, and Gazetteer, from 2,851 entity existing location in 1010 news, there are 108 entities location that is not detected, or in other words the accuracy of the rule-based algorithm is 96.2%. Some errors occur due to the identification of the data at the time of the administrative area written are different from the data administration area when the study was conducted. Most of the errors also occur due to a typing error in the case of location name entity that did not use capital letters. In addition, there were also some error detection due to the similarity of the names of people who were detected as the name of administrative areas

3.4 Extraction Feature Phase

At this stage, the existence of the entity location, either entity street, village, district, county/city will be converted into binary values. This stage will also count duplicatePlaceScore value. For example, the sentences with duplicatePlaceScore can be seen in Figure 9.

Berdasarkan pengakuan kedua orang saksi yang merupakan pembantu pada rumah tersebut, Agus Prasetyo (31) warga Jalan Muhajir 2/28 RT 05/04 , Bambu Apus, Pamulang, Tangerang Selatan dan Lasmiati (36) warga Bukit Indah blok D 13/9 RT 06/06, Serua, Ciputat, Tangerang Selatan peristiwa terjadi berlangsung sangat cepat.

Fig. 9: Example of a sentence with dupicatePlaceScore

In the above sentence in Figure 9, "Bambu Apus" and "Serua" were detected as administrative entity, whereas Pamulang and Ciputat were detected as districts entities. Although this phrase has a location entity, the sentence is not a criminal location sentence because it tends to show the origin of the victim. The same pattern was also shown in a sentence that shows the location of the origin of the perpetrators. The greater the value of the duplication level entity with different locations, it will increase possibility that the sentence is not a criminal location sentence.

3.5 Labelling Phase

In the Labelling phase, all selected data were labelled manually. In this research, there should be only one sentence that contained criminal location. The example can be seen in Table 7.

Table 7: Examples of Morphology Filtering Results

No	Sentences	Label
1	*Satinah (23) pembantu rumah tangga (PRT) yang menyayat Jason Matthiew Simanjuntak (3,5 tahun) di rumahnya, Jl Bintara VI Gang Sawo, RT 03 RW06, Kelurahan Bintara, Kecamatan Bekasi Barat, akan menjalani tes kejiwaan, Rabu (19/11/2014), besok.*	LCS
2	*Tes kejiwaan harus dilakukan terhadap Satinah karena ia diduga mengidap gangguan jiwa.*	NLCS
3	*Kabid Humas Polda Metro Jaya, Kombes Pol Rikwanto mengatakan pasca ditangkap, Satinah langsung ditahan di Rutan Pondok Bambu, Jakarta Timur.*	NLCS

3.6. Classification Results and Testing

From 1,010 criminal news articles, 7,459 sentences contain location entities. Once they extracted in numeric form, the data would be processed using SVM algorithm with experiments on some kernels to get the best results. The results can be seen in Table 8.

Table 8: Training Results on three kernel types

Kernel Type	Original Class	Classification Results		Precision	Recall	F-Measure (%)
		LCS	NLCS			
Linear	LCS	6262	187	0.959	0.958	95.76
	NLCS	129	881			
Polynomial	LCS	6393	56	0.947	0.948	94.4
	NLCS	334	676			
Radial	LCS	6267	182	0.959	0.958	95.77
	NLCS	133	877			

3.7. Mapping Phase

After classifying the location entities on the news have been done, then they are stored in the criminal locations table and check for entity duplications. If there are similarities, the entity will not go through the next process. For any unique news, its location query, which is combination of street name or public place name with village name, district name and township, are used as a query in geocoding. Example of submitted queries for geocoding process can be seen in Table 9.

Table 9: Geocoding Results

No	Query	Coordinate (latitude, longitude)
1	Jalan Daan Mogot Cengkareng	(-6.1568568, 106.7203384)
2	Jalan Raya Cipayung Cipayung	(-6.4358935, 106.7967598)

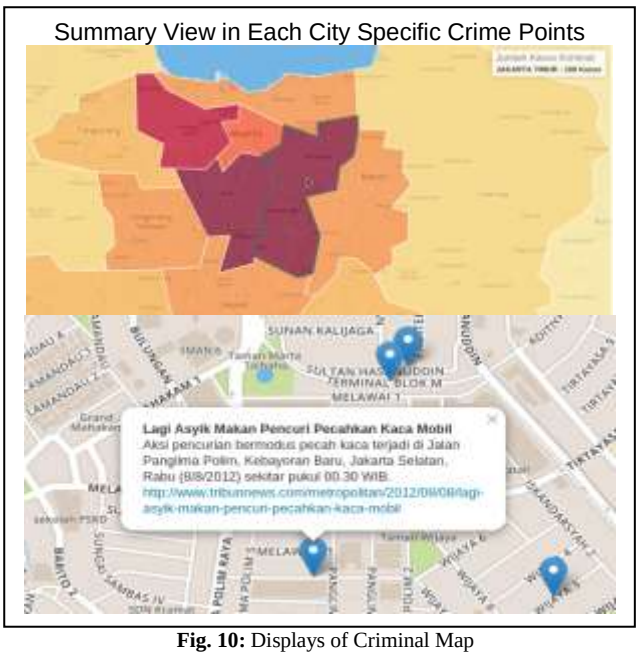

Fig. 10: Displays of Criminal Map

The results of this process will be stored on a database of criminal location. Implementation of the digital map application was built using Google Maps API. This Application has two views, as can be seen in Figure 10. The up side of Figure 10 is a summary display, which shows the number of points recorded crimes in any city in Jabodetabek area. The areas with darker colors denote the number of criminal cases is higher. The bottom side of Figure 10 is Specific display, which shows the placement of marker point in accordance with location entities detected as criminal location sentences.

4. Conclusion

The conclusion of this study is that the study has produced an application to display the locations of crime on a digital map. The level of accuracy of the rule-based algorithm that is used to perform the extraction location entity is 96.2%, while the level of the best SVM model accuracy for classifying sentences containing entity scene of the crime is 95.77% by using kernel radial.
Based on some mistaken identification, it is suggested the possible use of statistical methods, such as SVM algorithm combined with morphology token, contextual component, bag-of-word and gazeteer for location identification process entity in future studies. To overcome the problem of duplication of criminal cases that exist in the mapping stage, Future studies are suggested to use Levenshtein Distance algorithm to assess the similarity of the name of the entity street or public place for any criminal cases that have been saved.

Acknowledgement

This research was supported by Hibah Peneliti Utama Sebagai Rujukan Hibah MRG-UNS 2016 from Universitas Sebelas Maret, Indonesia.

References

[1] S. D. of P. and S. Statistics, *Statistik Kriminal 2016*. Jakarta: Central Bureau of Statistics, 2016.
[2] R. Feldman and J. Sanger, *The Text Mining Handbook*. New York, NY, USA: Cambridge University Press, 2006.
[3] B. Liu, *Web Data Mining Exploring Hyperlinks, Contents, and Usage Data*, 2nd ed. Springer-Verlag Berlin Heidelberg, 2011.
[4] S.-M.-R. Beheshti, S. Venugopal, S. H. Ryu, B. Benatallah, and W. Wang, "Big Data and Cross-Document Coreference Resolution: Current State and Future Opportunities," no. November, 2013.
[5] B. Alshaikhdeeb and K. Ahmad, "Biomedical Named Entity Recognition : A Review," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 6, no. 6, pp. 889–895, 2016.
[6] B. Cowan, S. Zethelius, B. Luk, T. Baras, P. Ukarde, and D. Zhang, "Named Entity Recognition in Travel-Related Search Queries," *Proceeding AAAI'15 Proc. Twenty-Ninth AAAI Conf. Artif. Intell. Pages*, pp. 3935–3941, 2015.
[7] K. E. Saputro, S. S. Kusumawardani, and S. Fauziati, "Development of semi-supervised named entity recognition to discover new tourism places," in *2016 2nd International Conference on Science and Technology-Computer (ICST)*, 2016, pp. 124–128.
[8] T. Mahmood, G. Mujtaba, L. Shuib, N. Z. Ali, A. Bawa, and S. Karim, "Public bus commuter assistance through the named entity recognition of twitter feeds and intelligent route finding," *IET Intell. Transp. Syst.*, vol. 11, no. 8, pp. 521–529, 2017.
[9] H. Shabat, "Named Entity Recognition in Crime News Documents Using Classifiers Combination," vol. 23, no. 6, pp. 1215–1222, 2015.
[10] I. Jayaweera and C. Sajeewa, "Crime Analytics : Analysis of Crimes Through Newspaper Articles Crime Analytics : Analysis of Crimes Through Newspaper Articles," no. April, 2015.
[11] R. Arulanandam, B. T. R. Savarimuthu, and M. A. Purvis, "Extracting Crime Information from Online Newspaper Articles," *Proc. Second Australas. Web Conf. (AWC 2014), Auckland, New Zeal.*, no. Awc, pp. 31–38, 2014.
[12] A. S. Wibawa and A. Purwarianti, "Indonesian Named-entity Recognition for 15 Classes Using Ensemble Supervised Learning," *Procedia Comput. Sci.*, vol. 81, no. May, pp. 221–228, 2016.
[13] T. F. Abidin, R. Ferdhiana, and H. Kamil, "Automatic Extraction of Place Entities and Sentences Containing the Date and Number of Victims of Tropical Disease Incidence from the Web," *J. Emerg. Technol. Web Intell.*, vol. 5, no. 3, pp. 302–309, 2013.
[14] R. Alfred, L. C. Leong, C. K. On, and P. Anthony, "Malay Named Entity Recognition Based on Rule-Based Approach," vol. 4, no. 3, 2014.
[15] G. A. Leroy, "Crime Information Extraction from Police and Witness Narrative Reports," 2008.
[16] Y. Li, K. Bontcheva, and H. Cunningham, "SVM based learning system for information extraction," in *Deterministic and statistical methods in machine learning*, J. Winkler, M. Niranjan, and N. Lawrence, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 319–339.
[17] H. Shabat, N. Omar, and K. Rahem, "Named Entity Recognition in Crime Using Machine Learning Approach," in *Information Retrieval Technology*, A. Jaafar, N. Mohamad Ali, S. A. Mohd Noah, A. F. Smeaton, P. Bruza, Z. A. Bakar, N. Jamil, and T. M. T. Sembok, Eds. Cham: Springer International Publishing, 2014, pp. 280–288.
[18] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. New York, NY, USA: Cambridge University Press, 2000.