CHAPTER 1 INTRODUCTION

In this digital era, all kinds of events, including criminal events, are reported openly and dynamically. Criminal information is vital for society as environmental safety awareness [1]. There are many online news platforms that report the latest criminal news both domestically and internationally. Online news has various forms, including the form of articles with titles and bodies containing criminal information in the form of sentences and paragraphs. Digital data on online crime news can be collected and the amount is relatively large.

In certain countries, criminal data is managed to be extracted into criminal information that is needed by the public. The criminal information is also publicly accessible. For example, in the United States, the City of Phoenix, Arizona, has a Crime Statistics and Map on the www.phoenix.gov portal, and the City of Lee's Summit which has a Community Crime Map. United Kingdom (UK) also has crime information map accessible to the public. Meanwhile in Indonesia, the Central Statistics Agency (BPS) publishes the results of criminal statistics once a year, which is usually issued at the end of the year [13]. This makes people have to wait for the end of the year to see criminal data in the same year, when BPS releases statistical data. Meanwhile, based on the results of a preliminary survey, around 81.5% of respondents agree that crime information needs to be shared openly to the public [1]. The most significant informations needed by the Indonesian people are crime type, crime location, and crime date [1].

Based on the explanation above, it is necessary to explore and apply information extraction from online news articles in Indonesian language, especially information on crime type, crime location, and crime date. Several studies on the extraction of criminal information have been carried out using several techniques and approaches, and it is known that the results of this criminal information extraction will be informative if visualized by presenting crime trends in each province and publicly accessible by Indonesian citizens.

Chapter 1 discusses several subtopics: (1) Rationale; (2) Theoretical Framework; (3) Conceptual Framework/Paradigm; (4) Statement of the problem; (5) Objectives and Hypotheses; (6) Scope and Delimitation; and (7) Significance of the study.

1.1 Rationale

Every day criminal acts occur in Indonesia. Based on data from the Central Statistics Agency, a crime occurs every 1 minute 33 seconds [14]. Criminal behavior tends to be repeated in a certain location, especially a densely populated location with a low level of supervision [11] [16]. Based on data from numeo.com, Indonesia's crime index fluctuates from 2017 to 2021 [20] as illustrated in Figure 1 below:

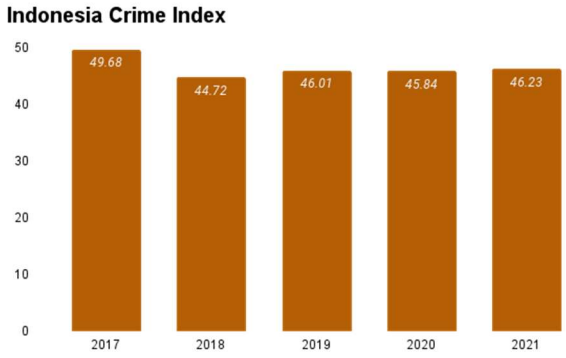

Figure 1. Indonesia Crime Index from 2017 to 2021 based on www.numbeo.com[20]

The Indonesian Central Statistics Agency (BPS) publishes the results of criminal statistics in Indonesia once a year. This data is usually released at the end of the year [13]. So to see criminal data this year, people need to wait until the end of the year when BPS has issued statistical data. Moreover, based on a survey of Indonesia's criminal needs, around 43.9% of respondents consider the information provided online by BPS to be less informative [1].

With the development of digital technology, many studies have been conducted to analyze criminal networks from unstructured digital documents. There is a lot of data that can be extracted in a digital document so that it can help investigations [2]. Another study in 2017 has started using online crime news sources to analyze text and extract location data using Named Entity Recognition (NER) [4]. Although in

this study, data extraction has not been carried out on the time of the incident, perpetrators, victims, and types of crimes [4]. In Indonesia, although there are still few researches on the extraction of criminal information in Indonesian news articles, researchers are starting to see crime news as a worthy source of data to study. Moreover, several well-known Indonesian online news companies already have good news accuracy with the correlation of the title, content and the facts [12]. Digital data on online crime news can be collected and the amount is relatively large.

As previously explained, in some other countries criminal data is managed to be extracted into publicly accessible criminal information. Likewise in Indonesia the Central Statistics Agency (BPS) releases criminal statistics data at the end of each year. This lack of criminal information in Indonesia requires improvement, one of which is through exploration and application of information extraction from Indonesian online news articles.

The process of extracting crime information in English texts has been carried out a lot. In their research in 2017, Mehedee Hassan and Mohammad Zahidur Rahman used two main processes for extracting criminal information, namely Named Entiry Recogniton (NER) and Support Vector Machine (SVM) as features [4]. The NER process is used to extract location and time terms from article content. While the SVM process is used as a classifier that will help classify whether the extracted terms are CRIME_SCENE or NOT_CRIME_SCENE based on previous research [4]. Then in another study in 2019 by Wijaya et al, a classification of types of crime with SVM was also carried out into 8 crime categories [6].

Appropriate techniques for extracting crime entities from Indonesian crime news are needed. Several prior studies on crime information extraction applied rules to perform crime information extraction. However, some misclassified instances could still be found. Thereby, the results of entity extraction are still ambiguous and inaccurate, especially in terms of location and time which could be found in more than one term in an article content. Budi [22] in his research revealed that NER in Indonesian has a low performance of 67.37%. In addition, the research conducted

by Fifinella [23] also obtained low performance 60.70% especially in terms of location and time.

Based on this background, this thesis will explore the extraction of crime type, crime location, and crime date in Indonesian online news articles. The technique used is to combine NER with SVM as a classifier that will help classify whether the extracted term is CRIME_SCENE or NOT_CRIME_SCENE, especially for the extraction of crime location and crime date in the article content section. To classify crime type, SVM will be used as claaification method.

1.2 Theoretical Framework

Information extraction is the process of determining structured information from unstructured text data [21]. Information extraction is performed to make information more accessible to end users and systems. Named Entity Recognition (NER) is one of the main components of information extraction, which aims to detect and classify named entities in the text [10]. NER is typically used to detect names, places, and document organization, and can be extended to identify genes, proteins, etc. as needed. Several previous studies used NER as a data extraction tool from criminal news texts in English such as unstructured digital investigative documents [2], online news newspaper articles [3], and online crime news [4]. In this study, NER will be used to extract criminal information such as crime type, crime location, and crime date, from Indonesian online news articles.

SVM is a machine learning method that uses the principle of Structural Risk Minimization (SRM) with the aim of finding the best hyperplane that can separate two classes in an input space [19]. The basic principle of SVM is a linear classifier which is usually used to solve linear problems [9] and further developed so that it can work on non-linear problems that can solve problems on non-linear problems such as Radial SVM [6]. The SVM model was first presented in 1992 at the Annual Workshop on Computational Learning Theory. This model was developed by Boser, Guyon, and Vapnik [19]. This development stimulates research interest in pattern recognition to investigate the potential capabilities of SVM both in theory and

practice. In this study, SVM will be used as a classifier that will help classify whether the term that has been extracted by NER is CRIME_SCENE or NOT_CRIME_SCENE, especially for the extraction of locations and times of criminal events in the article content section.

1.3 Conceptual Framework/ Paradigm

Extraction of crime information is a method for extracting multiple entities in the crime domain. Examples of some criminal entities that can be extracted are crime type, location, date, and many more. There are two main methods used in this research which refers to the previous research by Mehedee Hassan and Mohammad Zahidur Rahman in 2017 which used NER and SVM as features for crime information extraction in English texts, firstly named-entity recognition was performed on every word that was used. have the potential to become crime entities such as Person, Location, Date, and Organization, using the SpaCy Library by creating a model from the dataset in the form of a list of sentences that have been selected from crime news articles. SpaCy is an open source library that uses the Python programming language and is useful for efficiently handling NLP problems, one of which is NER. But unfortunately, until now SpaCy has not officially released the NER pre-train model for Indonesian [24]. Second, with SVM as a classifier that will help classify whether the extracted term is CRIME_SCENE or NOT_CRIME_SCENE, especially for the extraction of locations and times of criminal events which still have low accuracy [23].

By building a model from training data in Indonesian, it is hoped that it can give good results in extracting information on digital Indonesian criminal news. There are two processing blocks for different crime entities. First, the crime type classification process, where the dataset of each article already has a category or label *("culik", "narkoba", "begal", "rampok", "copet", "perkosa", "cabul",* and *"bunuh"*), so to determine the crime type instantly classification is done using the SVM algorithm. Second, the location and date extraction process, namely in this process the dataset used is a list of sentences that come from the contents of the article. By utilizing the SpaCy Library for NER, each sentence in the dataset is

extracted with information in the form of location and date. Then prelabelling is done on each sentence CRIME_SCENE or NOT_CRIME_SCENE for further classification with SVM to find out whether location and date are related to criminal events.

1.4 Statement of the Problem

Based on the background explanation, it is known that the prior studies of crime information extraction using rule-based extraction with combination of part-of-Speech tagging and Dependency Parsing still have low performance of 60.70% especially for crime location and date extraction. Therefore, the research question is "does adding named-entity recognition on crime information extraction from Indonesian crime news improve the crime type, crime location, and crime date extraction precision?"

1.5 Objective and Hypothesis

According to problem statement, the objective of this research is to improve news improve the crime type, crime location, and crime date extraction with more than 70% precision by proposing a combination of Named-Entity Recognition and Support Vector Machine.

Based on the problem and the objective, the hypothesis of this research is "If Named-entity recognition and support vector machine are used to extract crime type, crime location, and crime date from Indonesian crime news, crime type, the crime location, and crime date extraction precision will increase."

Independent variable: Named Entity Recognition and Support Vector Machine

Dependent variable: crime news, crime type, the crime location, and crime date extraction precision.

1.6 Scope and Delimitation

The scope of this research is Close-Domain Information Extraction. The domain is Indonesian crime news articles. The crime information that will be extracted are

crime type, crime location, and crime date. The dataset collected comes from an online news portal called detiknews.com from January 2020 to October 2021 in Indonesia. The fields used are title, content, and article category.

1.7 Significance of the Study

Considering the previous research by Fifinela in 2021 [23] which used a combination of post-tagging and dependency parser along with rules, this study aims to expand the extraction of crime information by combining the two methods, namely Named Entity Recognition (NER) as an information extraction method and Support Vector Machine (SVM) as an information extraction method. Classification technique to determine crime-scene related terms by modifying them so that they can be used in Indonesian. In this research, the dataset used is a collection of online crime news. This dataset will be used in the entire system process. Although there will be adjustments and modifications when the analysis process is at the sentence level because the initial dataset is a document level dataset. Based on several previous explorations, it is hoped that this contribution will be a trigger and attract attention for further exploration.