CHAPTER 1 INTRODUCTION

In this digital era, all types of events are reported openly and dynamically, including crime events. Many online news platforms provide crime news happening around the community in an up-to-date manner. The collections of news are compiled into relatively large crime data. However, the wealth of the data will become a pile of data only if it is not managed well.

In certain countries, the crime data are well managed to find detailed crime information, so the crime information is shared openly with the public. For example, the United Kingdom (U.K.) government has a crime map (U.K. Police 2013) available to the public. Also, in New Zealand (N.Z. Government 2013), the government provides crime information openly with coarse-grained details (e.g., the total number of thefts in a district or a province) [1]. However, Indonesiadoes not have much research focusing on crime information extraction. The Indonesians need to know the crime information openly based on the Crime Information Need Survey [2]. Around 78.3% of respondents are agreed that crime information needs to be shared openly with the public.[1].

For this reason, an exploration of the application of the Indonesian language crime information extraction is explored. Several studies on Crime Information Extraction have been carried out with several techniques and approaches. The results of this crime information extraction also need to be informative when it is visualized by presenting a crime trend in each of its province and can be accessed openly by Indonesian citizens.

This chapter includes the following subtopics, namely: (1) Rationale; (2) Theoretical Framework; (3) Conceptual Framework/Paradigm; (4) Statement of the problem; (5) Objective and Hypotheses; (6) Scope and Delimitation; and (7) Significance of the study.

1.1 Rationale

Indonesia is a country with the fourth largest population in the world [3]. With the high level of population density, an area's security level is an essential factor to consider. The level of security can be seen from many crimes occurring in areas. Based on data from numbeo.com

[1] The Author has surveyed a total of 54 respondents

[4], the crime index of Indonesia from 2017 to 2020 was fluctuated. However, the trend was increasing from 2018 to 2020, as in Figure 1 below.

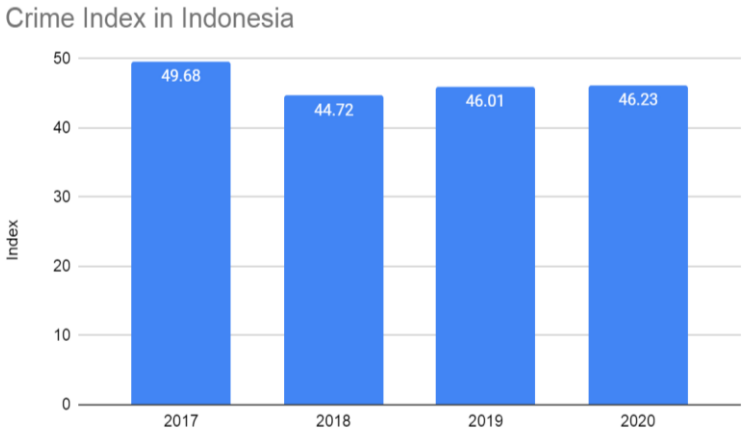

Figure 1. Indonesia Crime Index from 2017 to 2020 based on www.numbeo.com [4]

At the end of each year, the Indonesian police provide a report related to crime uploaded on the website of the Indonesian Statistics Agency (BPS) [5]. Although the report is uploaded online, this is less informative for the community based on Indonesian Criminal Information Needs [2]. However, with the rapid development of technology, many online news platforms present crime news around the community. The collections of news are compiled into relatively large crime data. However, the wealth of the data becomes a pile of data if they are not managed well. The method used to find important information relating to the crime news domain is the crime information extraction method. The results of the crime information extraction can be used for the public, such as providing information about the security of a region related to a crime, predictions of areas that need more protection, vulnerable times of increased crime rates, and many more.

In certain countries, crime information is available to the public on top of Google maps. For example, the United Kingdom (U.K.) government has a crime map (U.K. Police 2013) open to the public. However, many countries develop it only the coarse-grained crime information available to the public. For example, crime information in New Zealand (N.Z. Government 2013) contains coarse-grained information (e.g., the total number of thefts in a district or a province), which is not very useful for citizens [1]. For this reason, an exploration of the application of the Indonesian language crime information extraction is explored. The results of this crime information extraction can be informative if visualized by presenting a crime trend in each province.

In the process of crime information extraction, rules have an essential role in information retrieval and accuracy in extracting information of criminal entities that have been defined [6] [7]. The rules represent the entities according to the language structure. Based on [6] [7], If

the rules are not used in extraction, the misclassification of crime entities can increase. There are two main processes: the Named Entity Recognition (NER) process and the Dependency Parser process. In the Name Entity Recognition (NER) process, the pos-tagging process defines the nouns, verbs, and many more words. Meanwhile, in the Dependency Parser process, the rules identify the relation between words.

However, at this time, crime information extraction on Indonesian language news is rarely carried out?. Therefore, conducting crime information extraction in the Indonesian requires specific rules to extract crime entities. In general, to extract some information, entities usually use rules in the NER process. There have been several studies on NER in Indonesian. However, rules in that NER still have weaknesses. As a result, the entity extraction results are still ambiguous and less precise than that of I. Budi [8]. In that research, the Indonesian NER had a low performance of 67.37%.

Therefore, this thesis explores crime information extraction on Indonesian crime news by building rules suitable for extracting crime entities to carry out the criminal information extraction process correctly and correctly. The system extracts the types of crime, victim, perpetrator, place, and time of the incident [2]. If detailed criminal news information can be found, then the crime trend of each city can be identified.

1.2 Theoretical Framework

Information extraction is the task of extracting relevant information from unstructured data [9]. Along with that, Crime Information Extraction is an expanded field of information extraction that extracts unstructured data in the crime domain. In the field of Crime Information Extraction, several features are used to support the extraction process, such as Part of Speech Tagging (POS Tagging) features in Bahasa, Dependency Parser features in Bahasa, Crime Event Classification, and Rule-Based Crime Argument Extraction.

Rule-Based Crime Argument Extraction is a process to define other crime entities such as The Perpetrator, The victim, Location, and Time. This method has been applied in the previous study [6]. In addition, the previous research includs the rules in dependency parsing to know the dependency relations within the sentences. However, this rules applied in this study are based on POS Tagging, and Dependency Parsing Labels and combinations from both.

The feature in POS Tagging can be helpful to define the type of word because it obtains the POS tag from the text. At the same time, the dependency type can be beneficial to know the correlation between words. It identifies certain crime entities such as perpetrator, victim, location, and time entities easier.

This study attempts to find a specific crime event entity for each sentence used Crime Event Classification to classify the terms of words into some crime events. The classification uses ontology as a classifier. This method utilizes the structure of crime ontology built based on Police Crime Types on the Report of Crime Statistics 2018 [10].

1.3 Conceptual Framework/ Paradigm

Crime Information Extraction is a task to extract some entities in the crime domain. The examples of crime entities that can be extracted are *the crime type, location, time, victim, perpetrator, action taken against the criminal,* and many more. In previous studies [6], two main methods have a significant contribution, Natural Language Processing (NLP) unit and Entity Extraction, to conduct a good crime information extraction. First, there is a process named Dependency Parsing and Parts-of-Speech (POS) Tagging in the NLP unit. These methods are used in syntactic processing. Then in Entity Extraction, there is a process to extract potential named entities which fall under one of these classes: *Person, Location, Date, Time, Money, and Percent* by using the Standford Named Entity Recognizer.

By adapting these methods in Indonesian, it is expected to have a good result in extracting the information of Indonesian crime digital news. This is as a result of this process. Some changes are conducted in the NLP units and Entity Extraction. The Dependency Parsing and POS Tagging is applied in Indonesian using some Python libraries in the NLP unit. Then, the extraction is implemented by building some rules based on the result of NLP units. In general, this study focuses on extracting five types of crime entities - *Crime Type*, *Perpetrator Name, Victim Name, Location, and Time*- by using two main processes, namely Crime Type Classification and Rule-Based Crime Argument Extraction. First, the system classifies the sentence into more crime types in the Crime Type Classification using ontology as a classifier. This process determines the following process because if the sentence has the kind of crime, then the sentence is processed again to find any criminal arguments in the sentence. Finally, in the Rule-Based Crime Argument Extraction, the system extracts the sentence into some crime arguments using rules. The rules are built by combining some POS Tagging labels and Dependency Parser labels. The main focus of this study is to construct the relevant rules to extract the crime arguments correctly and nicely.

1.4 Statement of the Problem

The outline problem raised in this study is how to extract the crime entities using rules to extract the crime entities in Indonesian language.

The followings are sub-problems of this study:

- What is the process to define rules to extract crime argument entities based on POSTag and Dependency type features?
- What is the rule-based Indonesia crime information extraction system's performance?

1.5 Objective and Hypothesis

The objectives of this study are:

1. Building a system to extract crime entities in Indonesian.
2. Proposing a rule-based crime information extraction system

The benefit of this research is to provide information about crime trends so the vigilance of the Indonesian people towards crime in their area can be increased.

Based on the problems, objectives, and previous research findings, the hypothesis need to be considered is the similarity of syntactic patterns between Indonesian and English. The syntactic pattern between both is Subject-Predicate-Object. Here is the hypothesis as follows:

> "If the previous rules can be implemented in English, then it also can be implemented in Crime Information Extraction on Indonesian."

- Independent : The rules
- Dependent : Crime Information Extraction on Indonesian

The hypothesis can be measured by looking at the F-Measure score, which is influenced by True Positive, True Negative, False Positive, and False Negative generated by the system [8].

1.6 Scope and Delimitation

The scope of this study is Close-Domain Information Extraction. The domain is on Indonesian crime news only. The accumulated dataset was from one online portal news named detiknews.com from January to December 2020 in Indonesia.

1.7 Significance of the Study

This study defines rules for crime information extraction in Indonesian texts based on portal news in Indonesia. The system is going to identify the crime entities and measure the performance of the implemented system. This study groups the crime news into a crime dataset, and this dataset can be used on the overall process of the system. Based on some works that has been done, it is expected that these contributions may attract attention to further exploration. While for the rules, the rules are developed based on another research that creates rules to find the role on the sentence [11] and do some modifications to make it suitable for Indonesian characteristics.