Analysis and Implementation of Ontology Based Text Classification on Criminality Digital News

F Rahma1, D D Pangestuti2*, A Herdiani3, N Selviandro4
1,3 Informatics Engineering, School of Computing, Telkom University, Indonesia
2,4 Information System, Information Technology, Institut Teknologi Sepuluh Nopember, Indonesia

Email: dwitikapangestuti.18052@mhs.its.ac.id

Abstract. This research, we search the trend by utilize all information about the criminality type. When the trend has found, the possibility of crime in West Java will decrease. We need to analyse the data to get the information about the criminality trend by using Ontology based Text Classification Method. A news will be grouped into some criminality classes based on the relation around that groups so the process of classification can be done simply and specifically. The criminality trend of West Java is Property Crime with 47.5% of occurrences. We used several testing method such as F1 Score, Precision, Recall, and Accuracy. Based on the methods, the performance of system is running well or not in different point of view. The researcher obtained satisfactory result with F1 score 87.05%, accuracy 86.74%, recall 100% and precision 77.08% with using comparison of composition crime dataset and non-crime dataset is 40 : 40. It happens because testing needs to be done by considering the number of both datasets. The more balance the comparison while do pre-process, the higher accuracy that will get. In order to know whether the system is able to clarify the information accurately and can separate the non-criminal news dataset.

1. Introduction
Criminality rate in Indonesia always increase every year. Based on data from numbeo.com, in 2015, Indonesia criminality rate is rank at 58 from 120 countries. And in 2016 has increased to 51 from 117 countries [1]. While based on Crime Statistics data 2014, there are three provinces with the highest crime rate such as DKI Jakarta, North Sumatra, and West Java [2]. Among the three provinces, West Java is the province with the largest urbanization level in Indonesia. To increase the awareness, the public needs to know information about any criminal acts that often occur. Currently there is enough news about crime news in various online news portals. However, there are not many news portals that provide specific tags or special mark about criminality, so it is difficult for news readers to know what crimes are commonly happen in 2016. Therefore, it is necessary to analyze the crime news so the trend of criminality can be known. To analyze the news, we need to classify the news text so the information behind the news can we get. To classify the news text, we can implement ontology as classifier in the process classification.

Using ontology as classifier has been done in the previous research. In the e-Court European project, they use ontology to do information management of crime documents by using several ontology domain [3]. While in this research just use one specific domain. It is Criminality Domain. In the e-Court European Project, they focus on search storage the document in the ontology and classify the document not text. While in this research we focus on using ontology to classify many texts into several groups.

The technique to analyze the criminality trend is Ontology-based Text Classification technique. Ontology is use as classifier and it can be an effective classifier because it does not require a document to be tested and can see the semantic interconnection between groups [6]. Whereas if classification by
traditional method, the method requires many data sets to be tested and cannot visualize connectedness between groups. In addition to being an effective classifier, ontology can also handle cases of classification of text into certain groups by utilizing inter-class connectivity on the ontology [7]. While other classification techniques, cannot handle classification cases to more than one group. Another advantage is that text classification techniques based on Ontology classification can be done dynamically without having to re-test from classifier [4]. Using this technique, the classification of the news has done by looking at the linkage between a criminal group and another criminal group and possibly one having more than one criminal group, so the search for crime trends can be more specific. The results to be gleaned from this classification are the prevalent crime trend [5]. The specific classes based on data from Central Bureau of Statistics in West Java 2014. These classes are physical crime, property crime, life crime, crime of morality, narcotics crime, fraud crime, public order crime, crime against people's independence [2]. When all processes of classification have done by using ontology, the next process from this research to do is analysis the result. Based on the result, the purpose of this research to search the criminality trend can be done.

2. Method
The system to be built is a system that can classify the news text into several groups of criminal acts by utilizing inter-class connectivity and instances on the ontology. The input of this system is a dataset which is a collection of news texts from various online news portal of West Java. Then, the dataset of the news will be classified into several groups of criminal acts, so the trend of criminal acts can be known. The output of this system needs to be tested, to find out how accurately and successfully this system works. There are three parts of the work phase in doing this classification explained in Figure 1.

![Figure 1. General Framework System](image-url)

2.1 Preprocessing Module
2.1.1 Crawling Process
To obtain the data, it will do by doing process of crawling data. Crawling data has done on several news portals such as www.seputarjabar.com, www.jabar.tribunnews.com, www.fokusjabar.com, and www.lintasjabar.com. The crawling process has done on news that contains the tags of physical crime, property crime, life crime, crime of morality, narcotics crime, fraud crime, public order crime, crime against people's independence. Not just tags, but also terminology that fall into the category of criminal acts that exist

2.1.2 Preprocessing
In this process, there will be several tasks to do pre-processing like case folding, text clean up, tokenizing, stopword removal, phrase development, and synonym recognition [8].

2.2 Ontology
On this research, criminality ontology built based on Central Bureau of Statistics data. Before the development of ontology, writer do some research about the existing of other crime ontology. And there is ontology of criminal law in e-Court European project [3]. Based on that ontology, the structure of ontology is different with writer needed in this research. So, writer develop the ontology start from beginning. At this stage, there will be a grouping process words based on dictionary ontology data that
has been determined [9]. Ontology will map the result of classification of the category according to the class or instance. The process undertaken to find similarity of term on classes and instances in ontology, performed several stages start from make lexicon from ontology, make entity reference, and do summarization.

1. Make lexicon from ontology
   At this stage, a lexicon is created that holds the class and lexicon names to accommodate the instance names that exist on the ontology.

2. Make entity reference
   At this stage, the entity reference is created. Entity reference is a function to ensure or check whether the entity is in ontology.

3. Background knowledge compilation
   Based on the result of entity reference, it will be seen whether there is relevance of each entity with existing term.

4. Summarization
   Based on background knowledge compilation, it will be done a process summarization. In this process we will give weight to the occurrence of terms on the ontology and the process of adding to the weight of the occurrence of the term.

2.2.1 Classify Module
At this stage, when all the news datasets are categorize according to the similarity and connectedness of the ontology, any term that belongs to a particular class will be generalized into some form of its parent class.

3. Results and Discussion
After the system is finished, the next step is to test the system. The test is done by applying four evaluation methods, namely F1-Measure, Precision, Recall, and Accuracy. All of evaluation methods are using to measure the performance of built system [10]. While the dataset used in the test is the result of crawling from the system then labeled manually to see if the dataset is associated with ontology or not. This test is to measure the performance of the system that has been built. There are two types of datasets to use for testing. They are dataset of crawling results with criminal keywords amounted to 60 news and dataset of crawling results with non-criminal keywords amounted to 60 news. This dataset has collected with no keywords in its search only for the location of West Java.

Both of dataset will be used in three scenarios. The test scenarios performed as follows:
1. Do the test on comparative composition of criminal dataset to non-criminal dataset is 50:30 with using pre-processing data and non-preprocessing data.
2. Do the test on comparative composition of criminal dataset to non-criminal dataset is 40:40 with using pre-processing data and non-preprocessing data.
3. Do the test on comparative composition of criminal dataset to non-criminal dataset is 30:50 with using pre-processing data and non-preprocessing data.

When all of scenarios, the writer need to analyze the result of three scenarios. Here is the result of testing that showed on Table 1.
Table 1. Result of testing on ontology classification performance in F1 Score, Accuracy, Precision, and Recall

| Comparative | Sub-scenario | F1   | ACC   | PC    | RC    |
|-------------|--------------|------|-------|-------|-------|
| 50 : 30     | Un-preprocess| 57.50%| 62.22%| 40.35%| 100%  |
|             | Preprocess   | 82.22%| 83.15%| 69.81%| 100%  |
| 40 : 40     | Un-preprocess| 61.33%| 67.41%| 44.23%| 100%  |
|             | Preprocess   | 87.05%| 86.74%| 77.07%| 100%  |
| 30 : 50     | Un-preprocess| 59.37%| 70.45%| 42.22%| 100%  |
|             | Preprocess   | 86.56%| 87.14%| 76.31%| 100%  |

From the test results on Table 1, the best performance is on scenario 2 with same composition between criminal dataset and non-criminal dataset with doing preprocessing. If one of type of dataset has a bigger composition, the system still hard to classify the data into the correct classes. It happens because the system still hard to manage the ambiguity from the dataset.

Based on the result of Criminality Law Ontology in e-Court European, it is different from this research. In Criminality Law Ontology in e-Court, the result is focus on how ontology can manage the information of criminal documents [3]. But in this research is focus on utilize the ontology to find the trend of criminality. By using the established system, the trend of crime that often reported by the West Java news portals is the "Crime of Property" with the percentage of occurrence is 45.7% of the 60 occurrences of all crimes on the news dataset. The following figure in Figure 2 show a result by system based on classification. It is classification which is describe all crime on a news dataset from obtained by the researchers (Figure 2).

![Rate of Criminality in West Java](image)

**Figure 2.** Graph of occurrence of all crime on a news dataset

Based on the research, ontology can be an effective classifier to classify with a good performance. The built system still has weakness to handle the ambiguity from the data, so on the ontology is not really good when the system handle much of variant data. The bigger data, the higher ambiguity words can be found. It means we need to handle the ambiguity words first and classify them into the right classes. In the future work, we focus on how to improve the performance of the system and be handle to resolve the ambiguity from the criminal dataset and non-criminal dataset.

4. Conclusion

Based on the tests conducted in this research, it can be conclude that ontology can be a classifier of the classification of text into several groups by utilizing the structure or connectedness of classes and instances in the design of ontology. The use of ontology in the classification of news can find the trend
of criminality with good performance with F1 87.05%, accuracy 86.74%, 100% recall and precision 77.08%. Trend of criminality that gained from the program is the Property Crime with the percentage of occurrence 47.5% of 58 occurrences of criminal acts on all data dataset. In the future work, we focus on how to improve the performance of the system and be handle to resolve the ambiguity from the data.

References
[1] Shevchenko, O., & Shevchenko, G. 2019. Powstanie rynku mediów w Gruzji. Wschód Europy. Studia humanistyczno społeczne, 4(1), pp. 105.
[2] Bandung, B. P. S. K. 2015. Kota Bandung dalam Angka. 2015. Bandung: BPS.
[3] Breuker, J. 2003. The construction and use of ontologies of criminal law in the ecourt european project. Proceedings of Means of electronic communication in court administration, pp. 15-40.
[4] Allahyari, M., Kochut, K. J., & Janik, M. 2014. Ontology-based text classification into dynamically defined topics. In 2014 IEEE International Conference on Semantic Computing pp. 273-278. IEEE.
[5] Silachan, K., & Tantatsanawong, P. 2011. Domain ontology health informatics service from text medical data classification. In 2011 Annual SRII Global Conference pp. 357-362. IEEE.
[6] Subhashini, R., & Akilandeswari, J. 2011. A survey on ontology construction methodologies. International Journal of Enterprise Computing and Business Systems, 1(1), pp. 60-72.
[7] Noy, N. F., & McGuinness, D. L. 2001. Ontology development 101: A guide to creating your first ontology.
[8] Liu, P., Yu, H., Xu, T., & Lan, C. 2017. Research on archives text classification based on Naive Bayes. In 2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). pp. 187-190. IEEE.
[9] Yang, X. Q., Sun, N., Zhang, Y., & Kong, D. R. 2008. General framework for text classification based on domain ontology. In 2008 Third International Workshop on Semantic Media Adaptation and Personalization. pp. 147-152. IEEE.
[10] Goutte, C., & Gaussier, E. 2005. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In European Conference on Information Retrieval. pp. 345-359. Springer, Berlin, Heidelberg.