*Proceedings of the 2nd Adpebi International Conference on Management, Education, Social Science, Economics and Technology **(AICMEST**), Bandung, 15 December 2022*

ISBN 9798 88722 1274 *ISSN 2829 8438*

Designing Model System for Identification Violation of the Law Using Web Crawling and Text Mining

M Isnin Faried*, Dwi Atmodjo, Lely Priska
Faculty of Information Technology, Perbanas Institute
isninfaried@perbanas.id, dwi.atmodjo@perbanas.id, lely.priska@perbanas.id

Abstract

This study aims to create a model of an opinion detection system that will be used as a system to prevent violations of the ITE Law. Source of opinion from Indonesian-language twitter. The system was developed using web crawler technology and text mining with the implementation of one of the classification methods. This means that the research does not focus on sentiment analysis.
A web crawler with a focused crawling algorithm can make it easier to find legal sources that are used as guidelines for identifying violations from a user's opinion. The source of law in question is the ITE Law. The resulting convenience is being able to compile document sources on only specific topics and crawl relevant areas of the web. The resulting impact can reduce the amount of network traffic, resulting in significant savings in hardware and network resources.
Opinion Mining in this research uses the Naïve Bayes Multinominal Text (NBMT) algorithm. This algorithm is one of the algorithms in accordance with the opinion classification of twitter, capable of producing good accuracy. Another result obtained from the implementation of the NBMT algorithm is the speed of the process in the opinion classification process.
The resulting model is used as an alternative that can be used to prevent opinions that have potential violations, especially the ITE Law. To be more perfect, the author plans to develop research that focuses on opinion mining which has better accuracy, especially the detection of violations of the ITE Law.

Keywords : Model System, Violation the ITE Law, Focused Web Crawling, Text Mining

1. Introduction

The application of the Information and Electronic Transaction Law (UU ITE) in 2008 saw an increase in criminal cases against community activities in cyberspace. Based on the records of The Institute for Criminal Justice Reform (ICJR) quoted in the book (Arum et al. 2022) there was an increase in sentences of 96.8% in the 2016-2020 period. Cases that occur are generally due to the public's lack of understanding of this law (Kusumo, et al. 2021)
Based on these facts, this research was conducted in order to provide a solution in the form of a detection system that could prevent more victims from occurring.

1.1 Objectives

This research was conducted with the aim of creating a web technology-based system model that is used as an Early Warning System (EWS) for social media users in order to prevent criminal acts caused by postings in the form of opinions.
The proposed model uses the Text Mining method and the Web Crawler method. The Text Mining method used in opinion detection applies the Multinominal Naïve Bayes algorithm. The Web Crawler method with the Focus Crawler algorithm is used as an opinion search technique in social media. In this study, the data used came from Twitter, especially those containing opinions that violated the rules and were used as a corpus. The focus of research on violations of the ITE Law article 27 paragraph 3 is defamation.

ISBN 9798 88722 1274 *ISSN 2829 8438*

2. Literature Review

2.1 Text Mining

Text mining is used as text classification. By classifying the process of determining a text, it takes a long time and risks finding the meaning of the expanded data from the results of the topics needed. Naïve Bayes Classifier (NBC) is one of the algorithms used in text classification (Kalokasari et al. 2017). This section will discuss some of the results of research using NBC.

Zhirui and Chunyan (2020) conducted research on hotel reviews. The background of the problem is that previous studies have shown that there are many special high-frequency words but are not related to emotional tendencies and have the opposite effect, namely a decrease in classification efficiency and accuracy. The research proposal is to provide a naïve Bayes multinominal model combined with hotel comment characteristics to expand the list of stop words. Another target is that this method is also used for feature selection.

The results of the study show that with the above proposal, an increase in classification precision is obtained and after using an expanded list of stop words it has a small impact on the classification results but has increased operating efficiency to a certain extent. This research has shown that an extended list of stop words and this method can filter out features that have little effect on the classifier and reduce misclassification effectively.

Purwiantoro and Aditya (2020) conducted research on the classification of posts on social media. The targeted classification means SARA (Ethnicity, Religion, Race and Intergroup), radical and hoax. This study aims to apply the naïve Bayes multinominal algorithm in order to justify opinions originating from Twitter.

The results showed that the naïve Bayes multinominal algorithm is good for classification in the form of documents because it is capable of producing a high accuracy of 99.62%. The classification process does not take long, which is around 0.16 seconds.

Xue et.al (2019) conducted research on analyzing Chinese comments. The aim of this study was to examine the effect of the sentimental classification of Chinese texts based on Naive Bayes applied to the Chinese language review of the film Douban. The stage of the research was to pre-process the text to compile training texts and test texts, as well as to analyze the emotional tendencies of the test texts with the sentimental classifiers that were made.

The results obtained show that the sentimental classifier based on the Naïve Bayes algorithm effectively embodies the emotional classification of Chinese reviews of Douban's films. However, there are drawbacks in this experiment. Differences in emotional words in different fields and differences in vocabulary that express emotions in different age groups can lead to inevitable mistakes.

Alsanad (2018) conducted research on the detection of Arabic topics using Discriminative Multinominal Naïve Bayes and Frequency Transforms. The research background because several studies on the classification of Arabic texts in recent years need improvement to improve accuracy and efficiency. This study proposes an effective approach in Arabic text classification and topic detection using discriminatory multi nominal naïve Bayes (DMNB) for the classifier process and frequency transformation. The proposed approach includes three main steps: Arabic text preprocessing, extraction and normalization of Arabic text features, and Arabic text classification.

The results obtained using the 10-fold cross-validation test technique show that the accuracy of the proposed approach is better than the state-of-the art approach. For development it is necessary to extend the approach with feature selection and reduction algorithms.

Adikara et al. (2020) conducted research to detect cyberbullying in Instagram comments. Detection targets will be divided into two classes, one classification is cyberbullying and the other is non-cyberbullying. The algorithm used is the Naïve Bayes Classifier with Bag of Words and Lexicon based features. Bag of Words features are extracted from the terms that appear in comments and Lexicon based features are extracted using a dictionary or what is commonly known as the sentiment lexicon.

The results of the study show that the algorithm can be used to detect comments that are classified as cyberbullying on Instagram through several stages. This study uses a small dataset, but using a combination of the Bag of Words feature together with Lexicon-Based Features obtains higher performance than using the features independently. Applying 5-fold cross-validation, the system produces good accuracy and precision for detecting cyberbullying on Instagram.

ISBN 9798 88722 1274 *ISSN 2829 8438*

2.2 Web Crawler

The presence of internet technology has the consequence of having a large amount of data which is known as big data. Data on the internet is not easily stored in a database locally because big data has various forms. Search engines like Google are only able to collect raw data but cannot provide accurate information. Web crawlers are an alternative solution to the shortcomings of search engines. (Yu et al. 2020).

Yu et al (2020) define a web crawler as a computer program that browses hyperlinks and indexes search results. The purpose of this technique is for crawlers to perform accurately and quickly in retrieving information on the internet. The same definition is given by Kausar et al (2013) which states that web crawlers are programs or software or programmed scripts that browse the world wide web systematically and automatically. Web crawlers will browse page to page by using the graphical structure of the web page. Based on these two definitions, it can be concluded that a web crawler is a computer program created with the aim of tracing hyperlinks using a directed graphic structure, indexing results so that processing can be fast and accurate.

There are several crawling techniques commonly used, namely :

A. General Purposed Crawling | Generic Web Crawlers
   General Purposed Crawling works by collecting as many search results web pages as possible from certain hyperlinks. Web pages are retrieved in bulk from several different hyperlinks. The results of the General Purposed Crawling process are a little slow because it browses search targets from various sources. (Kausar et al. 2013)
B. Focused Crawling | Focus Web Crawlers
   Focused web crawlers perform searches for specific web sites, thereby saving significant amounts of time, disk space and network resources. This is due to fewer pages being stored due to the focus on certain topics. (Yu et al. 2020)
C. Distributed Crawling
   The Distributed Crawling technique runs several processes together to browse and download pages from the Web. (Kausar et al. 2013)
D. Incremental web crawlers
   There are differences between Incremental Crawlers and General Purposed Crawling, especially in different search strategies. The Incremental Crawler implements a schedule that signals web pages and databases to crawl again at certain intervals based on some refresh policy. This technique focuses on the time interval between two changes to the same database, then searches data independently based on the change frequency of each web page and each deep web database. (Yu et al. 2020)

3. Methods

The Naïve Bayes Multinominal Text algorithm is used because it is suitable for the classification of opinion data from Twitter (Purwiantoro and Aditya 2020). In addition to having this suitability, this algorithm also has very good accuracy results, which is approximately 89.58%. This is reinforced by the number of studies conducted by previous researchers which has been described in the literature review section.

The focus web crawler was chosen in this research because it will browse pages with a certain topic, namely UU ITE. Consider using this technique because it can save a lot of time, disk space, and network resources. Implementation of a focused web crawler by implementing a module to filter web links, namely the web page rating module.

This web page assessment module works to obtain the content of the ITE Law, the web page relevance evaluator will compare the relevance between the content on the web page and the topic.

4. Data Collection (12 font)

This study uses data sourced from Twitter in Indonesian-language with the specific topic of opinion that violates the ITE Law. Data is taken without the span of the year of occurrence because of the difficulty in obtaining the appropriate sources. This data is used as a reference for the opinion mining process.

ISBN 9798 88722 1274 *ISSN 2829 8438*

5. Results and Discussion (12 font)

In accordance with the research title, the focus of the research carried out by researchers is the creation of a system model that will be used as a tool for detecting potential violations of laws. The violations that were used as objects were violations of the 2008 ITE Law.

The model design is made into 2 parts, namely part A and part B. Part A seen in figure 1 is the initial part of the model where process number 1 is designed in the form of a user text writing feature on social media. This feature in the study was prepared with web technology. The written text will be stored as a text block (number 2) and will be processed by the "classification" module to be processed using the text mining method in the image as number 3. The processed "classification" results will be stored and using the web crawling method will be adjusted to the ITE Law which located on the Ministry of Communication and Information website. This crawling stage uses the Focused Crawling algorithm.

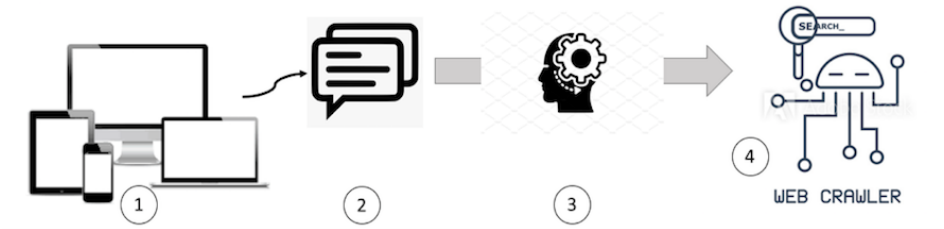

Figure 1. System Model Design part A.

The next process, namely part B, is shown in Figure 2. This section will carry out the process of adjusting the results of "classification" at the target site, namely the Ministry of Communication and Information which is described as number 5. The adjustments made to the results of "classification" are matched with Article 27 paragraph 3 concerning defamation (number 6). The research object is limited to defamation cases in order to focus more on design with the aim of preventing violations. The result of the adjustment is in the form of a user entry classification which is commonly known as an opinion. If the classification process is fulfilled, then this system model will display a message "potential" violation at the user layer (number 7).

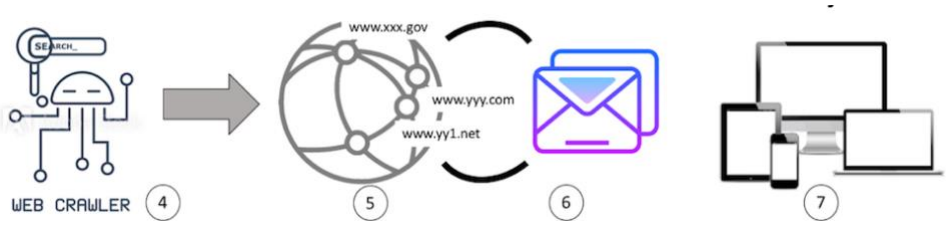

Figure 2. System Model Design part B.

To test this model, an application prototype is made by implementing several algorithms to support the model. The text classification section uses the Naïve Bayes Multinominal Text algorithm. In the prototype the researcher does not code but only utilizes existing libraries and mixes and matches them with python. The web crawler section uses a focused crawler algorithm. In order for the model implemented in the form of a prototype to be seen, the results need to be tested with the prepared sampling data.

Testing carried out on the prototype is black box testing. Based on Setiawan (2021) on the decoding site Black Box Testing is testing of software that is carried out to observe input and output results without knowing the code structure of the software. This test is generally carried out at the end of development with the aim of knowing whether the software can function properly. Prototype testing, especially the text classification section.

ISBN 9798 88722 1274 *ISSN 2829 8438*

5.1 Results

Prototyping the Web Crawler section as described in the design model was coded with the target website of the Ministry of Communication and Information and the text of the law on "pencemaran nama baik". Figure 3 below is part of the web crawler coding.

```
import nest_asyncio
nest_asyncio.apply()
import twint
c = twint.Config()
c.Search = 'pencemaran nama baik'
c.Limit = 10
c.Pandas = True
twint.run.Search(c)
Tweets_df = twint.storage.panda.Tweets_df
Tweets_df.head()
Tweets_df.to_csv("data.csv", index=False)
```

Figure 3. Example of Coding Focused Crawler.

The results obtained from Web Crawling are good, which means that the selection of Focused Crawling for research needs is appropriate because it also proves that the process is carried out quickly and saves disk space.

In the text user classification section, it was found that the recognition accuracy was not good. For some data entries in accordance with the applied data sampling algorithm can determine the classification that has the potential for violations. There are results that cannot be recognized as having potential violations. The source of the data used was user text (opinion) entries which were factually disputed but in this study gave the opposite result. The results of the classification test can be seen in Figure 4.

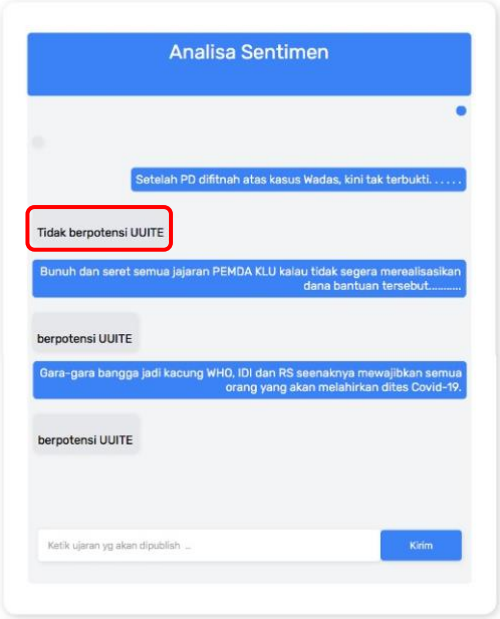

Figure 4. Text Classification Test Results

ISBN 9798 88722 1274 *ISSN 2829 8438*

Based on the black box testing method used which cannot see the structure of the code created, the analysis that can be done from these results is:

1. The NBMT algorithm is not suitable for analyzing text contained in laws in Indonesian
2. Incomplete corpus owned in research. The concept of making a corpus is done by collecting several opinions on the case that are found on the internet. But in fact the opinions found are not comparable to the existing cases.
3. There are not many studies that discuss specifically violations of the ITE Law

5.2 Proposed Improvements

Based on the results presented, it is necessary to develop further research with special discussion topics or research that focuses on existing sub-sections. The sub-section referred to here is the Classification section that needs to make a special research topic for text analysis contained in the law. There is still an opinion that legal sentences have multiple interpretations. This opinion was quoted from the online legal site page dated 15 July 2014 regarding Guidelines for understanding legal norms. Based on this fact, it is very interesting to conduct research in the field of text mining or natural language processing (NLP). Another interesting topic is the analysis of public opinion which has the potential to violate laws not only in the ITE Law but other legal products. The goal is to see patterns of opinion that tend to potentially violate the law.

6. Conclusion

The model created has good potential to be implemented considering the results found in the prototype trials are quite good. The application of a focused web crawler algorithm contributes to a short process. The algorithm used to carry out text classification gives results that are not completely failures. The resulting model is used as an alternative that can be used to prevent opinions that have potential violations, especially the ITE Law.

Specifically for the text classification section, it is necessary to develop research that focuses on this topic by using other algorithms or analyzing specifically the topic of texts in Indonesian laws and opinions in languages that have potential violations.

In the future works, the author plans to develop research that focuses on opinion mining which has better accuracy, especially the detection of violations of the ITE Law.

Acknowledgements

The authors would like to thank the Ministry of Education, Culture, Research and Technology especially Directorate General Higher Education of Research and Technology which has supported this research with research funding and PT Tri Dokumindo, especially for supporting the research. Finally, we would also like to thank Perbanas Institute, especially the Research and Community Service that always supports all the research activities.

ISBN 9798 88722 1274 *ISSN 2829 8438*

References

Adikara, P.P., Adinugroho, S. and Insani, S., Detection of Cyber Harassment (Cyberbullying) on Instagram Using Naïve Bayes Classifier with Bag of Words and Lexicon Based Features, *Proceeding of The 7th International Conference on Sustainable Information Engineering and Technology (SIET'20)*, pp. 64-68, Malang, Indonesia, November 16-17, 2020

Alsanad, A., Arabic Topic Detection Using Discriminative Multinominal Naïve Bayes and Frequency Transforms, *Proceedings of the 2018 International Conference on Signal Processing and Machine Learning (SPML'18)*, pp. 17-21, Shanghai, China, November 28-30, 2018

Arum, N.S., Wahyudin, A. and Kardina, A., Panduan Penanganan Perkara Pelanggaran Kebebasan Ekspresi Daring bagi Pendamping Hukum, SAFEnet, 2022

Kalokasari, D.H., Shofi, I.M. and Setyaningrum A.H., Implementasi Algoritma Multinominal Naïve Bayes Classifier Pada Sistem Klasifikasi Surat Keluar, *Jurnal Teknik Informatika*, vol. 10, no. 2, 2017

Kausar, M.A., Dhaka, V.S. and Singh, S. K., Web Crawler : A Review, *International Journal of Computer Applications*, vol. 63, no. 2, pp. 31-36, 2013

Kusumo, V.K., Junia, I.L..R., Prianto, Y. and Ruchimat, T., Pengaruh UU ITE Terhadap Kebebasan Berkespresi Di Media Sosial, *Proceeding of Seminar Nasional Hasil Penelitian dan Pengabdian Masyarakat 2021*, pp. 1069-1078, Jakarta, Indonesia, October 21st , 2021.

Panduan Memahami Larsa Bahasa Hukum, Available : https://www.hukumonline.com/berita/a/panduan-memahami-laras-bahasa-hukum-lt53c489209fd8e/ , Last Accessed on December 10, 2020

Purwiantono, F.E. and Aditya, A., Klasifikasi Sentimen Sara, Hoaks Dan Radikal Pada Postingan Media Sosial Menggunakan Algoritma Naïve Bayes Multinominal Text, *Jurnal TEKNOKOMPAKI*, vol. 14, no. 2, pp. 68-73, 2020

Republik Indonesia, "Undang-undang tentang Informasi dan Transaksi Elektronik (ITE)." 2016.

Setiawan, R., Black Box Testing untuk Menguji Perangkat Lunak, Available : https://www.dicoding.com/blog/black-box-testing/ , Last Accessed on December 10, 2020

Vaseeharan, T. and Aponso, A., Review On Sentiment Analysis of Twitter Posts About News Headlines Using Machine Learning Approaches and Naïve Bayes Classifier, *Proceeding of the 12th International Conference on Computer and Automation Engineering (ICCAE 2020)*, pp. 33-37, Sydney, NSW, Australia, February 14-16, 2020

Yu, L., Li, Y., Zeng, Q., Sun, Y., Bian, Y. and He, W., Summary of web crawler technology research, *Journal of Physics : Conference Series (ISPECE)*, 2020

Xue, J., Liu, K., Lu, Z. and Lu, H., Analysis of Chinese Comments on Douban Based on Naïve Bayes, *Proceeding of the International Conference on Big Data Technology (ICBDT 2019)*, pp. 121-124, Jinan, China, August 28-30, 2019

Zhirui, Y. and Chunyan, L., Analysis of Sentiment Classification of Hotel Reviews Based on Multinominal Naïve Bayes, *Proceeding of the 11th International Conference on E-business, Management and Economic (ICEME'20)*, pp.11-14, Beijing, China, July 15-17, 2020

ISBN 9798 88722 1274 *ISSN 2829 8438*

Biographies

Isnin Faried is a lecturer in the informatics department of the Faculty of Information Technology, Perbanas Institute-Jakarta, Indonesia. Research interests in web technology, machine learning, text mining, and security system.

Dwi Atmodjo, W.P. is a lecturer in the informatics department of the Faculty of Information Technology, Perbanas Institute-Jakarta, Indonesia. Research interests in internet of things (IoT), multiplatform programming, machine learning, and web technology.

Lely Priska, D.T. is a lecturer in the informatics department of the Faculty of Information Technology, Perbanas Institute-Jakarta, Indonesia. Research interests in e-Government, Human Computer Interaction (HCI), and smart city.