Improving The Detection of Plagiarism in Scientific Articles Using Machine Learning Approaches

Muhammad Agreindra Helmiawan[1], Irfan Fadil[2], Dwi Yuniarto[3], Fathoni Mahardika[4], Fidi Supriadi[5]
{ agreindra@stmik-sumedang.ac.id[1], fadilirfan@stmik-sumedang.ac.id[2], duart0@stmik-sumedang.ac.id[3], fathoni@stmik-sumedang.ac.id[4], fsupriadi@stmik-sumedang.ac.id[5] }

STMIK Sumedang

Abstract. One of the modern problems that occur in the current research and publication process is the duplication of the results of other people's research that is presented again by other parties. With the ease of the resources obtained, the more open the opportunity to bring up a problem called Plagiarism. This is attempted to be completed by the computer system with new approaches to detect and predict the existence of plagiarism in research automatically. In this article, approaches and methods for detecting plagiarism use machine learning techniques, where machine learning is empowered to become an algorithm as construction and evaluation in detecting plagiarism. Technically, this algorithm will compare and analyze the compatibility of words and sentences in documents with other document databases so that the analysis becomes an evaluation material, prediction, and determination that the document is plagiarism or not. The purpose of this study is to protect intellectual property and ideas, as well as the results to improve better performance and level of accuracy in detecting plagiarism.

Keywords: Machine Learning, Detection, Plagiarism, Algorithm.

1 Introduction

Along with the rapid development of information technology (IT), the IT-based learning process is not a necessity but an obligation, especially in the field of education which should be the basis for developing knowledge, especially in the field of informatics engineering[1]. Technical developments in computer hardware and software now make it possible to introduce automation into almost all aspects of the system between humans and machines. With this technical capability, which system functions must be automated. This education sector is risky in duplicating scientific articles or research conducted by researchers, this will have a serious impact so that it is necessary to overcome and prevent plagiarism in scientific article writing [2]. We describe the model for the type and level of automation that provides a framework and objective basis for making choices using the machine learning method that identifies words and sentences[3] which indicate the writing of scientific articles there are duplications or similarities of articles[4]. Each research topic has the substance of the work of the author, the words and paragraphs of the research produced by the researcher have the scientific value of the author's thoughts, but do not rule out the words and paragraphs that have similarities that result in indicated plagiarism [5]. The instrument which is the basis that influences detection including the substance of scientific articles by automation using the system[6], reliability and consistency

ICONISTECH I. 2019, July 11-12, Bandung, Indonesia

DOI 10.4108/eai.11-7-2019.2297759

in system automation can be utilized to solve the problem of plagiarism. In this study, we propose a new system in detecting the similarity of scientific articles with the Machine Learning method approach to explore the level of similarity of sentences in paragraphs and build them into the form of a system of automation to detect the similarity of scientific articles.

2 Method

The research method that the researchers used used the Prototype approach. The prototype is one of the methods used in developing software through certain stages starting from analysis and gathering needs, rapid design, prototype formation, coding, and software testing. In this research method the author conducted several stages according to the following figure:

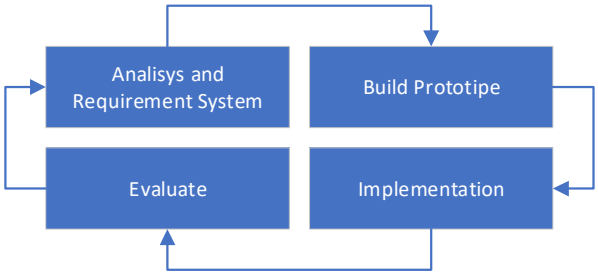

Fig. 1 Prototype Model

The prototype method can be divided into several steps in the research stages as follows :

a. System Requirements Analysis, researchers conducted interviews with expert experts who were competent in this field of research. The interview intends to obtain information regarding the current Named Entity Recognition system. In addition to being interviewed, in this case the researcher conducted a review of the results of other people's research as a comparative material and references relating to this research..
b. Building a prototype, researchers will design and model the system to be built that can be modeled in the form of UML, Interface Design, ERD. The purpose of this modeling is to oversee the system process so that it matches the problems studied before entering the implementation / coding stage
c. Implementation, research will translate prototypes that have been done and then modeled in the form of applications that can run and can be tested.
d. Testing, in this case the researcher conducted a test of the application to find out whether there were still errors or errors in the program code. The testing process is carried out from the beginning to the end.

We use methods with Machine learning and consider paraphrasing patterns. Polaini is the use of synonyms, changes between active and passive sounds, changing the form of words and parts of speech, breaking sentences, replacing words with definitions or meanings and various structures of sentences[7]. However, in the context of paraphrasing, it is important to define what is good and bad paraphrasing. That way, we can identify duplicate cases, change

translations, use additional words, unrelated text, blank samples, automatic word substitution and do something unrelated because they fail to follow instructions, such as trying to improve the quality of text instead of paraphrasing it. To improve the accuracy of paraphrasing similarities, we use Named-Entity Recognition (NER) or Named Entity Recognition and Classification (NERC). NER and NERC are one of the main components of information extraction tasks that aim to detect and categorize named entities in a text. NER is generally used to detect people's names, place names and organization of a document, but can also be extended to identify genes, proteins and others as needed. NER is useful in many NLP (Natural Language Processing) applications such as question-answering, summaries and dialog systems because it can reduce ambiguity. NER also deals with other information extraction tasks such as relation detection, event detection, and temporal analysis[8].

In building the system, the thing that is done is by planning and building a system or software in following user needs, this method is applied in the system design process that is built, meaning that the system can identify all of its needs from the start with general specifications. Stages in the process model represent the stages of developing new software designs to be built. From the process, details will be known that the Developer must develop or add a blueprint, or delete details that are not needed by the User [9].

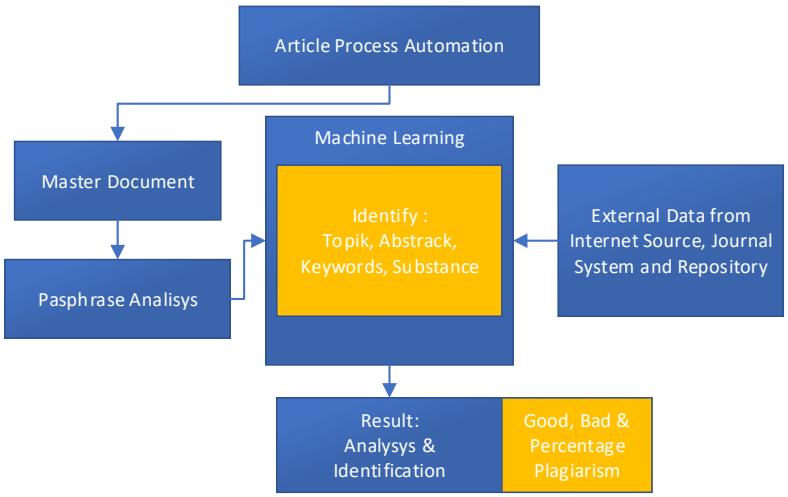

Fig. 2 System Procedure

3 Results

The automation system of the process of detecting the similarities of scientific articles is in the form of software that is useful for managing files of scientific articles starting from the topic of articles, abstracts, keywords and substance of scientific articles. So that this method and system can facilitate the process of plagiarism detection and can provide recommendations on the level of plagiarism of the article.

The method is used as a whole to generate prototypes in building machine learning to detect plagiarism in scientific articles. Overall model method to be used as can be seen in Figure 3

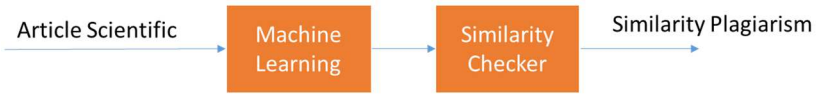

Fig. 3 Model Method

This Model method uses scientific article input to be processed using machine learning. So the expected result of the machine learning is a similarity plagiarism processed by algorithms similarity checker. So the next will be used for decision making to determine whether this standard article is acceptable or not.

This machine learning Model uses the Named Entity Recognition (NER) approach. In that approach, the scientific article will be detected based on the number of words that NER can recognize based on the word Tag that can be seen in the research [10]. The number of entities to be recognized in this research include: Person, Norp, Facility, ORG, GPE, LOC, Product, Work Of Art, and LAW. This entity is used based on observations from some scientific articles used in this study.

With this Machine Learning Model, the method done to build the approach from this Machine Learning can be seen as shown in Figure 4.

Fig. 4. Machine Learning Approach

Using this model, there is a lot of scientific article to build training. The more training, the greater possible accuracy can be achieved. At the input of this scientific article, will be taken part consisting of the title of the research, the name of the researcher, abstract research, and keyword research. These parts will be carried out training process using the help of generators in the research [10]. The result of processing the training will then be built model using Anago library. The result of this model, will be used to check the documents of the new scientific article to be checked similarity the word with existing database in the application Similarity checker.

In the similarity checker method, all parts of the scientific articles have been detected by the learning engine. Will be used as a comparator for new scientific article data. On this method it is necessary an algorithm to be able to compute similarity words that can be detected by machine learning. The Similarity Checker method can be seen in Figure 5.

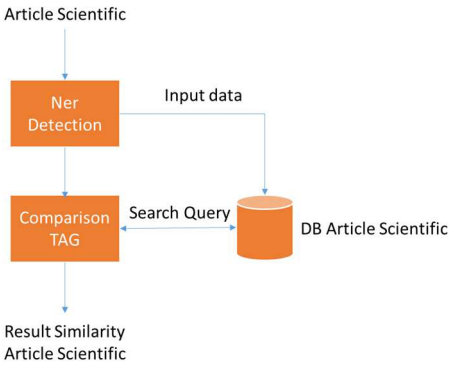

Fig. 5 Similarity Checker Method

The main method of the Similarity checker is the comparison TAG. In this process any word TAGS that have been processed by NER detection will be done in the search query process to the database article Scientific. Each of the same TAGS will be raised in result similarity article Scientific. If in one document has a TAG in common more than 5. Then it is certain that the article has something in common.

Implementations planning in this machine learning approach will be made in the from of system. This system is based on a web that will be built using CodeIgniter, a framework of the PHP programming language and MYSQL DBMS with a CodeIgniter (CI) framework that has the advantages of being open-source with a very small and MVC footprint..

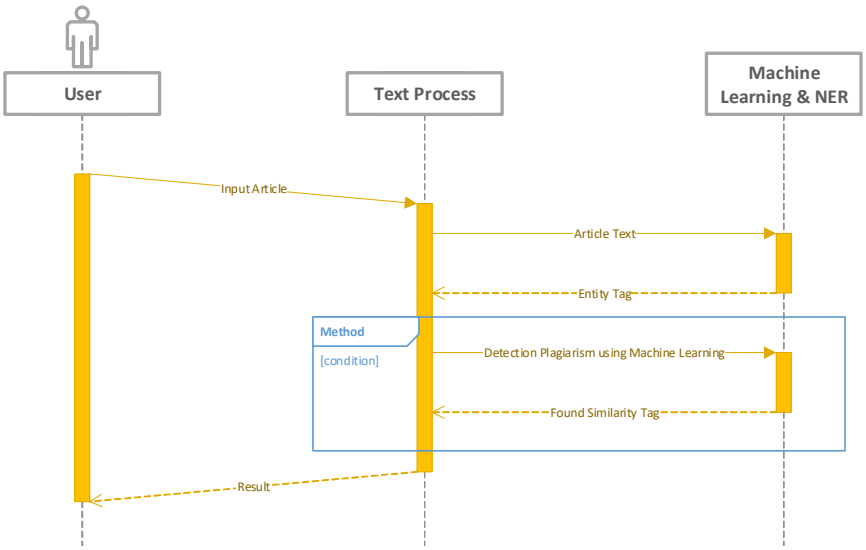

Fig. 6 Sequence Diagram Scanning Process

Based on the results of the model formulation that the author has completed, in the analysis phase, this model will be discussed about the implementation that the author has done. The following results from the implementation have been carried out

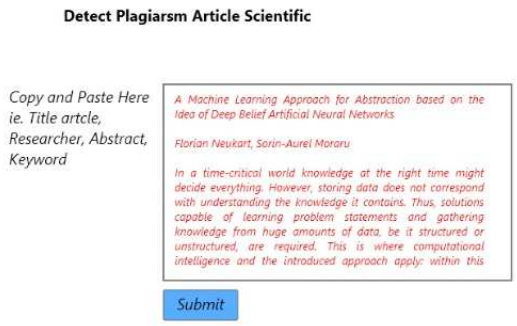

Fig. 7 Interface Design Plagiarism System

Fig. 8 Interface Design Result Similarity

4 Conclusions

In this study, we explored prototypes of word and sentence processing which included identifying and improving the accuracy of plagiarism in scientific articles. Thus, we introduce a model to assist in resolving the problems of originality and novelty in scientific article writing. We explore effectiveness, modeling, and efficiency and focus on plagiarism detection features. We conclude that the approach we take will be feasible for use by experts. In the future, we hope to integrate more methods and other algorithms that can be combined with Machine learning.

References

[1] I. K. A. E. Nugraha, K. Agustini, S. Si, M. Si, and I. G. P. Sindu, "Analisis Pemanfaatan E-Learning Sebagai Knowledge Management Dalam Mendukung Proses Pembelajaran Di Jurusan Pendidikan Teknik Informatika Undiksha," *Kumpul. Artik. Mhs. Pendidik. Tek. Inform. (ISSN 2252-9063)*, vol. 6, no. 1, 2017.

[2] H. Santoso and S. Sos, "Pencegahan dan Penanggulangan Plagiarisme dalam Penulisan Karya Ilmiah di Lingkungan Perpustakaan Perguruan Tinggi," *Hari Santoso, S. Sos*, vol. 1, pp. 1–23, 2015.

[3] M. Tschuggnall and G. Specht, "From plagiarism detection to bible analysis: The potential of machine learning for grammar-based text analysis," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2016, pp. 245–248.

[4] E. Gharavi, K. Bijari, K. Zahirnia, and H. Veisi, "A Deep Learning Approach to Persian Plagiarism Detection.," in *FIRE (Working Notes)*, 2016, pp. 154–159.

[5]A. Rosyadi, A. Z. Arifin, and D. Purwitasari, "Pengklasteran Berbasis Segmen Menggunakan Paragraf Untuk Identifikasi Topik Pada Deteksi IndikasiPlagiarisme," *J. Inspir.*, vol. 6, no. 2, 2016.

[6] M. A. Helmiawan, "Thesis Process Automation System," *PPMRTI*, vol. 12, no. 1, 2018.

[7] S. Burrows, M. Potthast, and B. Stein, "Paraphrase acquisition via crowdsourcing and machine learning," *ACM Trans. Intell. Syst. Technol.*, vol. 4, no. 3, p. 43, 2013.

[8] J. Foley, S. M. Sarwar, and J. Allan, "Named Entity Recognition with Extremely Limited Data," *arXiv Prepr. arXiv1806.04411*, 2018.

[9] K. K. Mohan, A. Srividya, and A. K. Verma, "Prototype dependability model in software: an application using BOCR models," *Int. J. Syst. Assur. Eng. Manag.*, vol. 7, no. 2, pp. 167–182, 2016.

[10] I. Fadil, D. Yuniarto, E. Firmansyah3, D. Herdiana, F. Supriadi, A. Rahman, "File Training Generator For Indonesian Language In Named Entity Recognition Using Anago Library"