1. Introduction

Background

Nowadays, science is developing very fast, as evidenced by the abundant number of scientific literature publications. Data from the Scimago Journal & Country Rank (ScimagoJR) website shows that in 2022, Indonesia managed to publish 43,300 scientific literature indexed by Scopus. Each of these scientific literature contains important information about the advancement of technology or science. However, scientific literature has several differences, when compared to other written works such as news, social media posts, and so on, in terms of structure and content, so that information processing in scientific literature also requires special techniques. Therefore, it is very important to develop methods and automate the processing of scientific literature [1].

Information extraction is the process of retrieving useful and structured information from a set of unstructured data. The results of information extraction can be used for various purposes in the data analysis process [2]. One application of its use is the analysis of the relationship between methods and research results, which are then used to develop model quality. Information extraction consists of various tasks, among which is entity extraction. Entity extraction, or what can also be referred to as named entity recognition (NER), is the procedure of identifying and classifying members of a group of words, such as location, organization, individual, and others, in a sentence or document [3]. In scientific literature, entity extraction can be performed to identify dataset, task, method, and evaluation metric. Illustration of the task can be seen in Fig. 1. In the figure, several words are highlighted in color to facilitate visualization. The word ”iris” is highlighted in green because Iris is identified as a dataset. ”Mengklasifikasikan” and ”klasifikasi” are highlighted in yellow because they are categorized as task entities in the given scientific study. The term ”K-Nearest Neighbors” and ”k-NN” can be classified as a method, thus highlighted in purple. Meanwhile, accuracy, precision, recall, and F1-score are highlighted in blue because they are identified as evaluation metric entities.

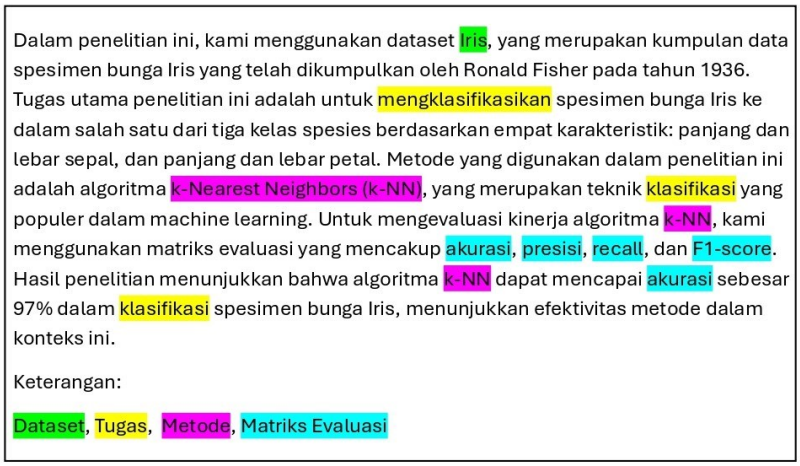

Fig.1. Illustration of Entity Extraction

Technology in information extraction and entity extraction continues to evolve, in line with the needs of data analysis. To improve the performance of the algorithms, datasets that support the research are also developed. One of the latest datasets that supports research in the field of information extraction is SciREX. SciREX is specifically designed for information extraction analysis in scientific literature and allows analysis of the entire document, not just paragraphs like previous datasets. The model built in SciREX research has been able to perform several tasks, including entity extraction, relation extraction, and importance detection using Conditional Random Fields (CRF) as one of its algorithms [4].

Inefficiencies in information extraction can hinder scientific progress and cause delays in solving problems that arise in society. Therefore, it is important to continuously review the development of information extraction models by conducting careful and in-depth research. Several methods are developed to accomplish the task of information extraction, including CRF and HMM. CRF and HMM have been used in several different studies, both of which are considered capable of providing high accuracy. Recent research [5] shows that HMM is capable of performing the entity extraction task well and produces a notable accuracy of 0.851. Meanwhile, another study [6] utilized CRF along with other models to solve a traditional Chinese medicine NER task and was able to produce an F1 value of 0.862. In this study, a comparison is made on the accuracy of using Conditional Random Fields (CRF) and Hidden Markov Models (HMM) in solving the task of entity extraction in scientific literature. The lack of research on entity extraction in Indonesian scientific literature makes this research important for the development of information extraction methods. To fulfill the research

purposes, a dataset was developed from Indonesian journals on the theme of informatics.

Topics and Limitations

The problem solved in this research is how to implement entity extraction in Indonesian scientific articles using the Hidden Markov Model (HMM) and Conditional Random Fields (CRF) methods. To address the problem, certain limitations were established:

1. The dataset comprises 100 scientific research papers in Indonesian.
2. These papers were selected based on their publication within the last decade and their relevance to informatics topics, specifically sentiment analysis, question-answering systems, classification, and recommendation systems.
3. From each research paper, key entities were extracted, including the dataset used, the task addressed, the method employed, and the metric for evaluation.

Objective

Based on the topics and limitations, the purpose of this research is to implement entity extraction in Indonesian scientific articles using the Hidden Markov Models (HMM) and Conditional Random Fields (CRF) methods and analyze the results.