1. INTRODUCTION

Natural Language Processing is a tool for processing text, where one of the tools is Parts of Speech (POS). Parts of Speech Tagging are used to annotate tags directly on a word in a sentence [1]. POS tagger is designed to analyze corpus data to determine the word class used in the data, where this tool will accept raw text as input and provide word class labels such as nouns, verbs, adjectives, adjectives, pronouns, conjunctions and other word classes [2]. POS Tagger is divided into two parts, namely rule-based and stochastic-based. Rule-based tagging is done by adjusting the rules that have been made. Meanwhile, Stochastic based tagging is done by using the corpus as training data for the model to determine the probability of a word class so that the word class is the best word class for the word [1].

News media are often found in everyday life as a means of information for the public about something that is happening. News is a fact or opinion that makes people interested in recognizing current events [3]. News articles generally have distinctive characteristics such as Factual, Actual and Unique and have many word classes through a series of words strung together by a particular subject. In news articles, it is not uncommon to find several supporting sentences to increase the popularity of the object promoted by the subject. This subject can be a big influence on the object, so many people start to look at the object described by the subject. Thus, this shows that the subject in branding is crucial so that it can influence the public to glance and be interested in the object referred to by the subject. Therefore, POS Tagging can be used to search for subject word classes from a series of sentences in news articles to find out which subject has the most significant role in influencing society.

Branding is the process of giving a company the possibility to tell the story of the company from a communicative aspect. This is done as a promise to meet customer expectations of the company and allow companies to position their products differently from their competitors [4]. In recent years there has been a rapid growth of internationalization in several industries. This makes the market for Higher Education unaffected because students are now more accessible and willing to move further away from home to study at the desired university [4]. Therefore, branding can be the main tool for universities to differentiate themselves and improve their reputation from other universities. Through branding, universities are able to attract the attention of students with the reputation results given in the branding. The role of the subject in branding also serves to enhance brand of the university with the large number of news articles created by the subject. That way, it can attract people's attention to consider the best university for them.

Telkom University is one of the private universities in Indonesia that has a vision to become a World Class University where Telkom University is actively involved in technological development [5]. Telkom University offers 31 study programs owned by seven faculties, one of which is the Informatics study program at the Faculty of Informatics [6]. To improve the reputation of Telkom University, media that can promote Telkom University is needed. One of the media that has a large audience is the news media. Therefore, the subject can do branding using news media to promote the brand.

Various news media have various information and different interests, such as the news portal located on "detik.com" with the information provided, which makes people want to know the phenomena that occur in the world [7]. Like the information presented by Agrakom, namely detik, the news media became a source of research on the subject. Detik was born on July 9, 1998, with his first story written by Budi D. Currently, detik is a popular Indonesian news site that delivers news in various categories [8]. Detik is a popular news portal with a high level of popularity. With this popularity, detik.com has become one of the news media as the main reference source for presenting information, especially among internet users in Indonesia [9].

In this study, the author will analyze the subject on popular news portals using news data that has been collected through the news portal "detik.com" and using the Hidden Markov Model (HMM) and Rule-based to perform POS Tagging. Hidden Markov Model (HMM) is a statistical model where the system being modelled is processed with hidden parameters [10]. The Hidden Markov Model (HMM) is chosen because the Hidden Markov Model is a corpus-based method [11]. Then the Rule-Based was chosen because of the direct writing of the rules. This Rule Based allows writers to make sentence and phrase rules [12]. In addition, the corpus used in this study is the IDN Tagged Corpus, a manual annotated POS tagging corpus for Indonesian [13]. In addition, the combination of these classification methods has a better accuracy performance in making predictions than the Hidden Markov Model alone. The combination of the Hidden Markov Model and Rule-Based can produce an accuracy of 86.62%, and the Hidden Markov Model alone produces an accuracy of 84.59% [14]. Based on the exposure of the research, this study will analyze the subject using a combination of Hidden Markov Model and Rule-Based due to high accuracy results so that it can provide an accurate model. In this study, we conducted to enhance the research results by changing the probability estimator on Hidden Markov Model. We also implement K-Fold Cross Validation on the corpus dataset to take the best data training and test. Finally, we modified the corpus by adding 1003 tokens with the label “NNP” to the corpus to enhance the performance of the model.

In research that was conducted by Yaroslav M, et al. in 2018 [15]. This study aims to test the Hidden Markov Model (HMM) to get predictions for Asynchronous Ventilator Patients. The results of this study indicate that the K-Fold Cross Validation can affect the results of the Hidden Markov Model (HMM), which managed to get the best results compared to the Hidden Markov Model without K-Fold Cross Validation.

Further research has been conducted by Muljono, et al. in 2017 [16]. This study aims to perform a morphological analysis for Part Of Speech (POS) Tagging Indonesia. This study examines and compares the performance of the Hidden Markov Model (HMM) on the original corpus with the modified corpus. The results of this study prove that the modification of the corpus can affect and have a good impact on the performance of the model made.

Other research was conducted by Muhammad Ridho A, et al. in 2021 [14]. This study aims to build a Hybrid POS Tagger using a combination of Rule-Based and Hidden Markov Model (HMM) to solve ambiguity problems using HMM and Viterbi Algorithm. The results of this study indicate that the results of the model performance from the combination of the Hidden Markov Model (HMM) and Rule-based are very good, with an accuracy of 86.62% compared to the Hidden Markov Model which produces an accuracy of 84.59%.

In another study conducted by Sindhya KN, et al. in 2019 [17]. The purpose of this study is to examine the Malayalam language because the existing research so far is the only research on common languages such as English. Then, POS Tagger for Malayalam is quite rare because Malayalam has a complicated structure compared to other foreign languages. This study shows that the Hidden Markov Model provides the best Accuracy in Malayalam so that the Hidden Markov Model can be used for other languages such as Indonesian.

Other research that has been conducted by Nitin S, et al. in 2017 [18]. Researchers have a goal to do POS Tagging stochastic based on the Viterbi Algorithm in Indonesian. This research was conducted because there are many studies on POS Tagging in English which have model performance with good accuracy values. This is because Indonesian has a more complicated sentence structure than other foreign languages. The results of the study used 10-Fold Cross Validation and gave 93.23% accuracy values, 94.55% recall values and 93.23% precision values.