JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 4, Oktober 2022, Page 2187-2195
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i4.4566

Analysis of Telkom University News Subjects on Popular Indonesian News Portals Using a Combination of Hidden Markov Model (HMM) and Rule Based Methods

Rendhy Al-Farrel*, Donni Richasdy, Mahendra Dwifebri Purbolaksono

Faculty Of Informatics, Informatics Study Program, Telkom University, Bandung, Indonesia
Email: 1,*rendhyaf@students.telkomuniversity.ac.id, 2donnir@telkomuniversity.ac.id, 3mahendradp@telkomuniversity.ac.id
Correspondence Author Email: rendhyaf@students.telkomuniversity.ac.id

Abstract−News media are often found in everyday life as a means of information for the public about something that is happening. In news articles, it is common to see several sentences that support the object to increase its popularity by being promoted by the subject. Part of Speech Tagging can determine the class of words in the sentence according to Tagsets provided by the corpus. That way, the search for the subject in the news article can be found from the word class obtained from a corpus. This research was focused on finding the subject "who" repeatedly spreading the news about Telkom University by using Part of Speech Tagging with the Hidden Markov Model and Rule Based on a news dataset from popular news portals about Telkom University. The process is taking all news about Telkom University on popular news portals and classifying it using the Hidden Markov Model and Rule-Based. We conducted to enhance the research results by changing the probability estimator on Hidden Markov Model. After running some scenarios, the best results obtained by the Hidden Markov Model and Rule-Based are the Accuracy of 94.96%, the Precision of 94.99%, the Recall of 94.96%, and the F1-Score of 94.95%.

Keywords: Pos Tagger; Subject; Hidden Markov Model; Rule-Based

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

Submitted: **27/07/2022**; Accepted: **31/08/2022**; Published: **25/10/2022**

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.

2. RESEARCH METHODOLOGY

2.1 System Design

The research carried out will go through the stages that will be passed in building a POS Tagging. The following is a flow of steps that will be carried out.

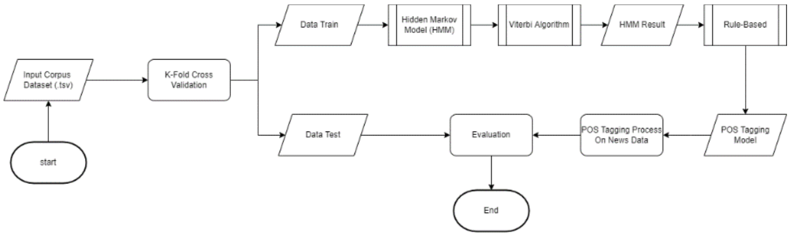

Figure 1. POS Tagging Flowchart

In Figure 1, the first step is to enter the corpus data into the system. Then the corpus data is entered into the K-Fold Cross Validation process to get the best train data and test data. Furthermore, the training data is entered into the Hidden Markov Model (HMM) process to train the model with the train data obtained from the previous process. After that, the process results of the prior process are decoded with the Viterbi Algorithm. Furthermore, the results from the previous process are entered into the Rule-Based process to validate the tags with the rules that have been created. Then, the final result of all the previous processes becomes a POS Tagging model, which will later be evaluated with test data. Before the POS Tagging model evaluates, there is a POS Tagging process that will be carried out on news data, and the following is the flow of the steps that will be carried out.

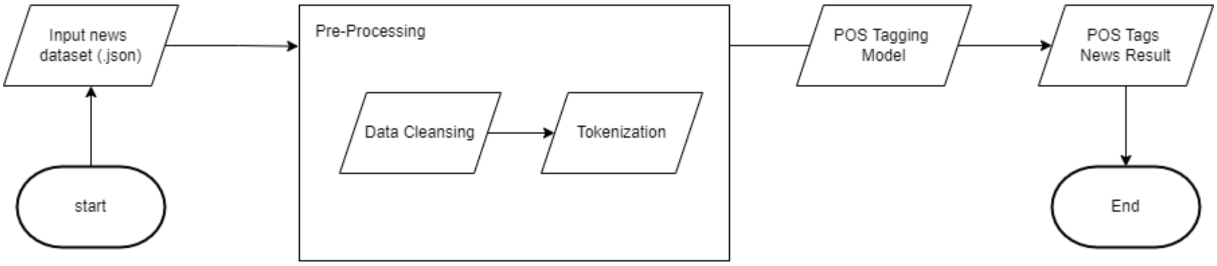

Figure 2. POS Tag process on news data

In Figure 2, the first step is to enter news data into the system. Then the news data will be entered into the Pre-Processing and performed two pre-processing, namely Data Cleansing and Tokenizing. Then, the POS Tagging that has been created in Figure 1 will be used to perform POS Tagging on the news data. Then, the results of the POS news tags will be used for subject analysis.

2.2 Dataset

The dataset that will be used is derived from the results of Web Scraping using Scrapy and Selenium in the python programming language on popular news portal websites such as "detik.com" and from several categories such as detikNews, detikEdu, detikFinance, detikWolipop, detikInet, detikTravel, and detikHot in the period June to August. The elements taken are title, category, author, date, article, and scrape_time. After successfully doing Web Scraping, the program will generate a JavaScript Object Notation file (JSON). This study uses data obtained from news articles from popular news portals. The number of datasets collected to conduct this research is 445 news data. The following is the structure of the news dataset.

Table 1. The structure of news dataset

title	category	author	date	article	scrapetime
Tentang Telkom University, Universitas Swasta Terbaik Indonesia Versi Webometric	detikEdu	Novia Aisyah	Selasa, 29 Jun 2021 16:46 WIB	Telkom University menjadi universitas swasta terbaik di Indonesia versi Webometrics. Secara nasional, perguruan tinggi swasta (PTS) ini menempati peringkat ke-7 setelah Universitas Airlangga…	2021-07-14 18:57:08

Then, the corpus used in this study is the IDN tagged corpus, the manual annotated POS marking corpus for Indonesian. This corpus has 10,000 sentences and 250,000 more token tagsets. This corpus obtains data from the IDENTIC parallel corpus, where the corpus is a combined corpus of mixed sentence data from Indonesian and English from Penn Treebank Corpus, which is translated into Indonesian, international news portals and subtitles [13]. The following is the structure of the corpus data.

Table 2. The Structure Of Corpus Dataset

word	tag
Pendek	JJ
kata	NN
,	Z
kemiskinan	NN

2.3 Pre-processing Data

One of the first steps to classifying is to do pre-processing. The purpose of pre-processing is to make the data easier to process to provide quality data for the classification process. In this study, the author only provides pre-processing to news data and performs two pre-processing, namely Data Cleansing and Tokenizing. This is because if there is too much pre-processing, it is feared that it will change the sentence structure and affect the tagging later. The following are the stages of the pre-processing process carried out.

1. Data cleansing function is to clean the data until leaving only text data. Data cleansing is done to remove the Space.
2. Tokenizing is the process of breaking sentences on data into a collection of word pieces which are commonly called tokens.

Submitted: 27/07/2022; Accepted: 31/08/2022; Published: 25/10/2022

Table 3. Pre-Processing

Pre-processing	Before	After
Data Cleansing	Universitas Airlangga.\nTelkom University berlokasi di Bandung, Jawa Barat.	Universitas Airlangga. Telkom University berlokasi di Bandung, Jawa Barat.
Tokenization	Pemeringkatan ini berdasarkan sistem penilaian berbasis situs web masing-masing perguruan.	‘Pemeringkatan’, ‘ini’, ‘berdasarkan’, ‘sistem’, ‘penilaian’, ‘berbasis’, ‘situs’, ‘web', ‘masing’, ‘masing’, ‘perguruan’

2.4 Hidden Markov Model

Hidden Markov Model (HMM) is a statistical model with probability calculations and is modelled with unobservable states. HMM can also be referred to as a simple dynamic Bayesian network [19]. The probability calculation process is carried out when other events can be observed and seen directly. HMM consists of two states where the state is the hidden state and the observed state. The hidden state is the part that cannot be observed, while the observed state is the part that can be observed directly. HMM has five components, namely:

1. The set of the observed states

$$O = o_1, o_2, \dots, o_n \tag{1}$$

2. The set of the hidden state

$$Q = q_1, q_2, \dots, q_n \tag{2}$$

3. The transition probability is the probability for the state *i* move or move state *j*.

$$A = a_{01}, a_{02}, \dots, a_{n1}, \dots, a_{nm};\ a_{ij} \tag{3}$$

4. Emission probability is a process in which state *i* generates the probability of an observation o_t.

$$B = b_{i(o_t)} \tag{4}$$

5. Initial state and ending state that has nothing to do with observation.

$$q_0, q_{end} \tag{5}$$

The following is the equation (1) of the Hidden Markov Model in the case of Part of Speech Tagging.

$$\mathrm{Tg_n} = \max(\mathcal{P}\left({}^{w_i}/_{tg_i}\right) \times \mathcal{P}\left({}^{\mathrm{tg_i}}/_{\mathrm{tg_{i-1}}}\right)) \tag{6}$$

$\mathrm{Tg_n}$: the search for word class

$\mathrm{tg_i}$: searches for the word class $\mathrm{w_i}$ from corpus

$\mathrm{w_i}$: searching word for word class

$\mathrm{tg_{i-1}}$: search word class as much as one before the word class of $\mathrm{w_i}$ in the corpus

$\mathcal{P}$: probability values

In Hidden Markov Model, there are four main components, namely:

2.4.1. States

In this study, states are POS tags (word classes) which will be used later in the study.

2.4.2. Initial Distribution

The initial Distribution of the probability observed states there are 23 POS tags (word class).

2.4.3. Emission Probability

This section represents the probability of the observed state with a tag that may be a word class. The following is the equation for Emission Probability

$$P(word|tag) = \frac{count(word|tag)}{count(tag)} \tag{7}$$

Word : observed word

Tag : word class

If using the Smoothing Technique, then the equation becomes as follows

$$P(word|tag) = \frac{count(word|tag)+1}{count(tag)+V} \tag{8}$$

V = Number of tags in POS Tags

2.4.4. Transition Probability

This section is the probability of comparison between observed and previous tags. Smoothing is also implemented in Transition Probability. Smoothing is the process of giving a probability distribution so that all word sequences can have several probabilities [20]. Here is the equation for Transition Probability

$$P(tag_i|tag_{i-1}) = \frac{count(tag_i|tag_{i-1})+1}{count(tag_{i-1})+V} \quad (9)$$

2.5 Viterbi Algorithm

Decoding for POS Tagging with Hidden Markov Model (HMM) is the Viterbi Algorithm. Viterbi algorithm is a process to get a series of words with predictable tags. The results from Viterbi are in the form of a series of words with predictable word classes, and then these results are used to calculate Accuracy, which is obtained from the comparison of predicted word classes generated Viterbi Algorithm with word classes from corpus data [18]. Viterbi algorithm is a process to find the most optimal path for a hidden state. With Viterbi Trellis, the algorithm will calculate recursively. Viterbi trellis is the probability value of HMM when it is in state *j* after passing through observations and passing the most probable order for each state at the time and value as in the following formula:

$$v_t(j) = \overset{max}{_{1\le i\le N-1}} v_{t-1}(i)a_{ij}b_j(O_t) \quad (10)$$

$v_{t-1}(i)$: t-1 for *Viterbi trellis*
a_{ij} : state q_i to state q_j for transition probability
$b_j(O_t)$: *observation state* o_t in *state j* for emission probability

2.6 Rule Based

Rule-based is a working method that uses language rules which are also known as grammar, to obtain the class of terms in a sentence. The method used is a method that uses the corpus to provide word classes in words. The rules are made only for Noun Phrase because in this study, the word class taken is the subject. The rules used are obtained from InaNLP, and rules are created based on the regular expression shown in the table below [21].

Table 4. Handwritten Rule Based

The rule for Rule-Based	
NP	{<NN|NNP>+<CC><NN|NNP|NNG>+}
NP	{<PRP><CC><NN|NNP|NNG>}
NP	{<NN|NNP><CC><PRP>}
NP	{<NN>+<JJ>}
NP	{<NN>+<DT>}
NP	{<NN>+<PRP>}
NP	{<NN>+<NNP>+}
NP	{<NNP>+<NN>+}
NP	{NN>+}
NP	{<NNP>+}
NP	{<NNG>}
NP	{<FW>+}
NP	{<PRP>}

2.6 Evaluation

The confusion matrix is widely used in machine learning to evaluate the classification model [22]. The confusion matrix consists of four essential components such as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). In addition, there is a performance evaluation with calculations such as Accuracy, Precision, Recall, and F1-Score. The following is the formula for calculating the performance evaluation.

a. Accuracy is a calculation to measure how well the value of a model is so that the model can classify data correctly and accurately. Accuracy is a benchmark tool to find the predicted value with the actual value. The following equation (11) is the formula for calculating Accuracy.

$$accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (11)$$

b. Precision is a calculation to find the value of performance accuracy between the requested data and the results of the prediction data. The formula (12) is the Precision calculation equation.

$$precision = \frac{TP}{TP+FP} \quad (12)$$

c. Recall is a calculation for the success rate of a model in determining information. The following equation (13) is the formula for calculating Recall.

$$recall = \frac{TP}{TP+FN} \quad (13)$$

Submitted: **27/07/2022**; Accepted: **31/08/2022**; Published: **25/10/2022**

d. F1-Score is a performance matrix that considers the calculation of Recall and Precision. The following equation (14) is the formula for calculating F1-Score.

$$F1\ Score = 2 \times \frac{precision \times recall}{precision + recall} \tag{14}$$

3. RESULT AND DISCUSSION

In the evaluation stage of this study, there are 3 test scenarios to evaluate the system that has been built. Scenario 1 is a model performance test based on K values from 3 to 10 for K-Fold Cross Validation. Scenario 2 aims to find the best corpus for the system that has been created. Then, scenario 3 aims to compare the estimators for the system created. An estimator is an optional function or class that maps the frequency distribution of a condition to a probability distribution for the Hidden Markov Model (HMM).

3.1 Scenario 1 Effect of K-Fold Cross Validation on Corpus Dataset

This scenario aims to determine the effect of the K-Fold Cross Validation for the original corpus dataset on the POS Tagging. In this test scenario, the researcher uses the K-Fold Cross Validation provided by the scikit-learn library to determine the train and test for the model. The following are the results of the comparison of the original corpus dataset after K-Fold Cross Validation.

Table 5. Comparison of K-Fold Cross Validation on Corpus

K-Fold	Accuracy (Train)	Accuracy	Precision	Recall	F1-Score
3	81.66	81.51	86.85	81.52	82.93
4	90.26	90.08	91.9	90.08	90.61
5	91.54	91.42	92.78	91.42	91.82
6	92.2	92.12	93.29	92.12	92.47
7	92.46	92.36	93.47	92.36	92.69
8	92.67	92.56	93.6	92.56	92.87
9	92.89	92.77	93.73	92.77	93.06
10	**92.91**	**92.76**	**93.69**	**92.76**	**93.04**

Based on the results obtained from this test scenario, the POS Tagging was created using the Hidden Markov Model and Rule-Based. The model has much better performance results with the K-Fold Cross Validation, where the K value used is a value of 10. Although the performance results given the value of K = 9 is greater, the train of the value of K = 10 is greater than the value of K = 9 with an accuracy value of 92.91%. Therefore, the results obtained from this test scenario with a value of K = 10 are the Accuracy of 92.76%, the Precision of 93.69%, the Recall of 92.76%, and the F1-Score of 93.04%.

3.2 Scenario 2 The Effect of Modification on Corpus Dataset

From the results of the previous test scenario, the dataset that has been carried out with the K-Fold Cross Validation will be used as the dataset used for the following test scenario. For test scenario 2, we will test some corpus data. The test to be carried out is to compare the results of the original corpus data with the modified corpus data. Where the modification made is to add word data such as the names of individuals whose first names are taken only and the word class "NNP" to the corpus. The following is the result of the comparison of the model's performance against the differences in the corpus.

Table 6. The Comparison of Modification on Corpus

Dataset	K-Fold	Accuracy (Train)	Accuracy	Precision	Recall	F1-Score
Original Corpus	3	81.66	81.51	86.85	81.52	82.93
	4	90.26	90.08	91.9	90.08	90.61
	5	91.54	91.42	92.78	91.42	91.82
	6	92.2	92.12	93.29	92.12	92.47
	7	92.46	92.36	93.47	92.36	92.69
	8	92.67	92.56	93.6	92.56	92.87
	9	92.89	92.77	93.73	92.77	93.06
	10	92.91	92.76	93.69	92.76	93.04
Modified Corpus	3	81.61	81.51	86.87	81.52	82.94
	4	90.25	90.07	91.89	90.07	90.6
	5	91.57	91.45	92.81	91.45	91.84
	6	92.22	92.12	93.31	92.12	92.48
	7	92.47	92.37	93.49	92.37	92.71

Dataset	K-Fold	Accuracy (Train)	Accuracy	Precision	Recall	F1-Score
	8	92.66	92.57	93.62	92.57	92.89
	9	92.9	92.78	93.75	92.78	93.08
	10	**92.94**	**92.82**	**93.78**	**92.82**	**93.11**

From the results of test scenario 2, it is evident that modifying the corpus dataset can affect performance and provide good performance. Therefore, it is better to use a corpus that has been modified. The results obtained from this test scenario are an Accuracy of 92.82%, a Precision of 93.78%, a Recall of 92.82%, and an F1-Score of 93.11%.

3.3 Scenario 3 Effect of Estimator on POS Tagging Model

Based on the results of test scenario 2, the comparison of the dataset that has been tested will be used as the dataset used is the modified corpus dataset for test scenario 3. In test scenario 3, a trial will be carried out to find out how much influence changes the estimator on Hidden Markov Model (HMM). The estimator of the HMM is LidstoneProbDist. Estimators to be tested against the HMM are SimpleGoodTuringProbDist and LaplaceProbDist. Of the three estimators, we will compare which estimator is the best for POS Tagging. The following are the results of the comparison of the estimators on the model.

Table 7. Estimator Comparison

Estimator	K-Fold	Accuracy (Train)	Accuracy	Precision	Recall	F1-Score
LidstoneProbDist	3	81.61	81.51	86.87	81.52	82.94
	4	90.25	90.07	91.89	90.07	90.6
	5	91.57	91.45	92.81	91.45	91.84
	6	92.22	92.12	93.31	92.12	92.48
	7	92.47	92.37	93.49	92.37	92.71
	8	92.66	92.57	93.62	92.57	92.89
	9	92.9	92.78	93.75	92.78	93.08
	10	92.94	92.82	93.78	92.82	93.11
SimpleGoodTuring ProbDist	3	86.42	86.27	88.3	86.27	86.83
	4	93.27	93.13	93.26	93.13	93.12
	5	94.15	94.03	94.07	94.03	94.01
	6	94.58	94.45	94.5	94.45	94.44
	7	94.83	94.71	94.75	94.71	94.7
	8	94.92	94.81	94.88	94.81	94.81
	9	95.02	94.92	94.96	94.92	94.92
	10	**95.06**	**94.96**	**94.99**	**94.96**	**94.95**
LaplaceProbDist	3	80.46	80.41	82.12	80.41	80.54
	4	88.07	88.16	88.64	88.16	88.12
	5	91.57	91.45	92.81	91.45	91.84
	6	90.57	90.46	90.72	90.46	90.39
	7	90.82	90.67	90.94	90.67	90.59
	8	91.07	90.91	91.19	90.91	90.83
	9	91.26	91.07	91.3	91.07	90.99
	10	91.35	91.19	91.41	91.19	91.1

To see the comparison, the following is a visualization of the results of the comparison of LidstoneProbDist, SimpleGoodTuringProbDist, and LaplaceProbDist estimators on the POS Tagging using HMM using K-Fold Cross Validation where the K value used is 10, because the data series displays the best results from each comparison.

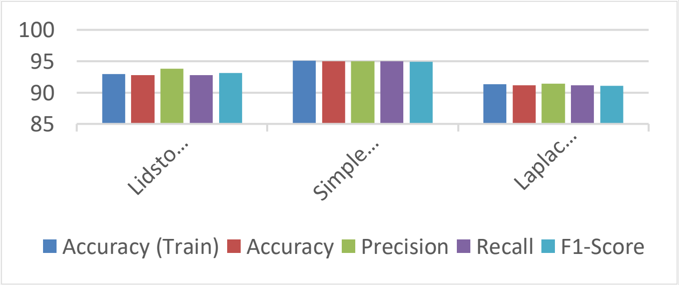

Figure 3. Visualization of Estimator Comparison Results

The results of this test scenario 3 show that the estimator SimpleGoodTuringProbDist has a better effect on the POS Tagging than the estimator LidstoneProbDist and LaplaceProbDist. The results obtained by the estimator SimpleGoodTuringProbDist on the POS Tagging with an Accuracy of 94.96%, a Precision of 94.99%, a Recall of 94.96%, and an F1-Score of 94.95%.

4. CONCLUSION

After conducting subject analysis research on popular news portals with news data of 445 and using a combination of the Hidden Markov Model (HMM) and Rule-based in POS Tagging, it can be concluded that making a classification model with the K-Fold Cross Validation where the best K value is 10, can produce a good model because train and test data are the best data so that the model can do training with the train that makes the model predict accurately. Modifying the corpus data also affects the POS Tagging. Because the data is reproduced, the train used for model training has more capital information to make predictions. Changes in the estimator in Hidden Markov Model (HMM) also produce the best performance for the model because it can maximize the performance of the model and each word class so that the model can predict with maximum and accurate performance. The best results obtained by the POS Tagging are an Accuracy of 94.96%, a Precision of 94.99%, a Recall of 94.96%, and an F1-Score of 94.95%. Suggestions for further research are to try to use a corpus dataset with an enormous amount of data to find out whether the amount of data owned can affect the model using the Hidden Markov Model (HMM) and Rule-Based.

REFERENCES

[1] D. E. Cahyani and M. J. Vindiyanto, “Indonesian part of speech tagging using hidden markov model - Ngram viterbi,” *2019 4th Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng. ICITISEE 2019*, pp. 353–358, 2019, doi: 10.1109/ICITISEE48480.2019.9003989.

[2] D. N. Prabhu Khorjuvenkar, M. Ainapurkar, and S. Chagas, “Parts of speech tagging for Konkani language,” *Proc. 2nd Int. Conf. Comput. Methodol. Commun. ICCMC 2018*, no. ICCMC, pp. 605–607, 2018, doi: 10.1109/ICCMC.2018.8487620.

[3] A. Y. Rofiqi, “Clustering Berita Olahraga Berbahasa Indonesia Menggunakan Metode K-Medoid Bersyarat,” *J. Simantec*, vol. 6, no. 1, pp. 25–32, 2017.

[4] B. T. Within, “Branding for Universities A qualitative case study on Jönköping University BACHELOR THESIS WITHIN : Business Administration NUMBER OF CREDITS : 15 ECTS PROGRAMME OF STUDY : Marketing Management,” no. May, 2019.

[5] Q. Setyani, R. Andreswari, and M. A. Hasibuan, “Target Analysis of Students Based on Academic Data Record Using Method Fuzzy Analytical Hierarchy Process (FAHP) Case Study: Study Program Information Systems Telkom University,” *2018 6th Int. Conf. Cyber IT Serv. Manag. CITSM 2018*, no. Citsm, pp. 1–6, 2019, doi: 10.1109/CITSM.2018.8674334.

[6] D. Y. Putri, R. Andreswari, and M. A. Hasibuan, “Analysis of Students Graduation Target Based on Academic Data Record Using C4.5 Algorithm Case Study: Information Systems Students of Telkom University,” *2018 6th Int. Conf. Cyber IT Serv. Manag. CITSM 2018*, no. Citsm, pp. 1–6, 2019, doi: 10.1109/CITSM.2018.8674366.

[7] Y. P. D. Sasongko, “Pertarungan wacana dalam pemberitaan revisi undang undang Komisi Pemberantasan Korupsi di Kompas.com dan Detiknews.com,” *J. Signal*, vol. 8, no. Vol 8, No 1 (2020): JURNAL SIGNAL, pp. 36–48, 2020, [Online]. Available: http://jurnal.unswagati.ac.id/index.php/Signal/article/view/3011.

[8] C. R. Yulianti and H. Setiawan, “Analisis Framing dan Diksi Berita pada Media Online Detik Travel dan CNN Indonesia Sebagai Bahan Ajar Teks Berita,” *Edukatif J. Ilmu Pendidik.*, vol. 4, no. 1, pp. 803–814, 2022, doi: 10.31004/edukatif.v4i1.1895.

[9] D. Yulistiani and A. Parmawati, “an Analysis of Deictic Expression in the Article Selected From Detiknews About Krakatoa’S Mount Disaster 2018,” *Proj. (Professional J. English Educ.*, vol. 3, no. 6, p. 751, 2020, doi: 10.22460/project.v3i6.p751-756.

[10] H. Z. Muhammad, M. Nasrun, C. Setianingsih, and M. A. Murti, “Speech recognition for English to Indonesian translator using hidden Markov model,” *2018 Int. Conf. Signals Syst. ICSigSys 2018 - Proc.*, pp. 255–260, 2018, doi: 10.1109/ICSIGSYS.2018.8372768.

[11] Ankita and K. A. Abdul Nazeer, “Part-of-speech tagging and named entity recognition using improved hidden markov model and bloom filter,” *2018 Int. Conf. Comput. Power Commun. Technol. GUCON 2018*, pp. 1072–1077, 2019, doi: 10.1109/GUCON.2018.8674901.

[12] A. N. M. Fahim Faisal, M. A. Rahman, and T. Farah, “A rule-based bengali grammar checker,” *Proc. 2021 5th World Conf. Smart Trends Syst. Secur. Sustain. WorldS4 2021*, pp. 113–117, 2021, doi: 10.1109/WorldS451998.2021.9514031.

[13] K. Kurniawan and A. F. Aji, “Toward a Standardized and More Accurate Indonesian Part-of-Speech Tagging,” *Proc. 2018 Int. Conf. Asian Lang. Process. IALP 2018*, pp. 303–307, 2019, doi: 10.1109/IALP.2018.8629236.

[14] M. Ridho Ananda, M. Yudistira Hanifmuti, and I. Alfina, “A Hybrid of Rule-based and HMM-based Part-of-Speech Tagger for Indonesian,” *2021 Int. Conf. Asian Lang. Process. IALP 2021*, pp. 280–285, 2021, doi: 10.1109/IALP54817.2021.9675180.

[15] Y. Marchuk *et al.*, “Predicting Patient-ventilator Asynchronies with Hidden Markov Models,” *Sci. Rep.*, vol. 8, no. 1, pp. 1–7, 2018, doi: 10.1038/s41598-018-36011-0.

[16] Muljono, U. Afini, and C. Supriyanto, “Morphology analysis for Hidden Markov Model based Indonesian part-of-speech tagger,” *Proc. - 2017 1st Int. Conf. Informatics Comput. Sci. ICICoS 2017*, vol. 2018-Janua, no. 0, pp. 237–240, 2017,

Submitted: 27/07/2022; Accepted: 31/08/2022; Published: 25/10/2022

doi: 10.1109/ICICOS.2017.8276368.

[17] S. K. Nambiar, A. Leons, S. Jose, and Arunsree, "POS Tagger for Malayalam using Hidden Markov Model," *Proc. 2nd Int. Conf. Smart Syst. Inven. Technol. ICSSIT 2019*, no. Icssit, pp. 957–960, 2019, doi: 10.1109/ICSSIT46314.2019.8987786.

[18] N. Sabloak, B. A. Hardono, and Deri Alamsyah, "Part-of-Speech (POS) Tagging Bahasa Indonesia Menggunakan Algoritma Viterbi," *Progr. Stud. Tek. Inform. STIMIK GI MDP Palembang*, no. x, pp. 1–11, 2017.

[19] Y. A. Rohman and R. Kusumaningrum, "Twitter Storytelling Generator Using Latent Dirichlet Allocation and Hidden Markov Model POS-TAG (Part-of-Speech Tagging)," *ICICOS 2019 - 3rd Int. Conf. Informatics Comput. Sci. Accel. Informatics Comput. Res. Smarter Soc. Era Ind. 4.0, Proc.*, pp. 0–5, 2019, doi: 10.1109/ICICoS48119.2019.8982411.

[20] M. D. Drovo, M. Chowdhury, S. I. Uday, and A. K. Das, "Named Entity Recognition in Bengali Text Using Merged Hidden Markov Model and Rule Base Approach," *2019 7th Int. Conf. Smart Comput. Commun.*, pp. 7–11, 2019.

[21] A. Purwarianti, A. Andhika, A. F. Wicaksono, I. Afif, and F. Ferdian, "InaNLP: Indonesia natural language processing toolkit, case study: Complaint tweet classification," *4th IGNITE Conf. 2016 Int. Conf. Adv. Informatics Concepts, Theory Appl. ICAICTA 2016*, pp. 5–9, 2016, doi: 10.1109/ICAICTA.2016.7803103.

[22] M. Hasnain, M. F. Pasha, I. Ghani, M. Imran, M. Y. Alzahrani, and R. Budiarto, "Evaluating Trust Prediction and Confusion Matrix Measures for Web Services Ranking," *IEEE Access*, vol. 8, pp. 90847–90861, 2020, doi: 10.1109/ACCESS.2020.2994222.

Submitted: **27/07/2022**; Accepted: **31/08/2022**; Published: **25/10/2022**