JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 1, Januari 2023, Page 214-222
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.v7i1.5384

Tagging Efficiency Analysis of Part of Speech Taggers on Indonesian News

Djatnika Widia Nugraha*, Donni Richasdy, Aditya Firman Ihsan

Faculty Of Informatics, Informatics Study Program, Telkom University, Bandung, Indonesia
Email: [1,*]djatnikawn@students.telkomuniversity.ac.id, [2]donnir@telkomuniversity.ac.id, [3] adityaihsan@telkomuniversity.ac.id
Correspondence Author Email: djatnikawn@student.telkomuniversity.ac.id

Abstract−Part of speech tagging (POS tagging) is a part of Natural Process Language (NLP). POS tagging is the process of automatic labeling of a word in a sentence according to the word class. There are various tagger methods in POS tagging, each tagger method has its own characteristics in its application. The research method used is Conditional Random Fields and Hidden Markov Model. The training of the two method models uses the Indonesian language corpus and Indonesian news texts as test data to determine which method is the most efficient based on the results of the accuracy and training time of each model. The method that has the best value is the CRF method with an accuracy value of 97.68 on the evaluation of the corpus test data and 90.02% for the sample Indonesian news dataset with a training time of 146.90 seconds, then there is the HMM method which has the highest accuracy value with a value of 94.25 % and shorter training time relatively shorter at 32.45 seconds and for the sample sentences containing 116 tokens, CRF method produces 90.05% accuracy which is higher than the HMM method which produces 79.31% accuracy.

Keyword: Part of speech Tagging; Natural Process Language; Efficient; Conditional Random Fields; Hidden Markov Model

1. INTRODUCTION

Part-of-speech Tagging (POS Tagging) is an automatic word tagging in which is the part is considered appropriate from a tag will be given to a word based on words that have been tagged before [1]. POS Tagging is also called as grammatical tagging or disambiguation word categories [2]. The use of POS tagging is important because it is used in several Natural Processing Language (NLP) applications such as word disambiguation, sentence parsing, questions answering, and translation machine [3,4]. Tagging words manually is a task that can save a lot of time, because you need to have special accuracy and skills in its application [5]. It is hoped that with POS tagging giving a tag will save a lot of time. There are 2 types of labeling rules in the POS tagging method, namely Rule-Based-Tagging and Stochastic Tagging. Rule-Based Tagging is the process of labeling according to dictionaries or rules that have been determined from training data sets. This type is also commonly used to solve problems in cases of unknown words or ambiguity in words, morphological and semantic information [1]. Then the second type is Stochastic Tagging which is done by using the corpus dataset as a data train to later determine the probability of a class of words. The use of this rule is also to determine the best tag from the model which has predicted which word class is considered appropriate for the word to be tagged [1].

Indonesia is rich of ethnics settled in different regions. Each of them has different local language for communication in their regions, meanwhile the most common language for communication is Indonesian. Indonesian is the national language used to communicate officially throughout Indonesia [6]. Indonesian is also used as journalism language either in electronic or digital media. As other languages Indonesian language has several grammatical categories, such as verbs, nouns, adjectives, adverbs, and so on which is usually found in online news articles widely available on social networks [3].

According to KBBI News is an incident report about what the organization has recently learned about important or interesting things [7]. News articles have distinctive characteristics like a being actual, factual, and interesting. The use of Indonesian words in the news is formal. The role of POS Tagging in Indonesian news is considered to be very helpful because with POS tagging we can decipher words to get information quite easily and required shorter time than manual labeling [5].

The research that will be carried out is regarding to efficiency of POS taggers in Indonesian news. As the previous studies that have been carried out by several researchers with various methods and resulted the quite good methods accuracy. One of them was the research conducted by Ritu Banga and Pulkit Mehndiratta on a comparison of the five tagger methods including Perceptron Taggers, TnT, Conditional Random Fields Tagger, Brill Tagger, and Classifier BasedPOS Tagger (CPOS). Perceptron taggers have the highest accuracy of 88.7% [1]. However, the dataset used in this research comes from the Twitter API and is in English. There is a difference with the research that will be conducted by text from Twitter, where the text is much shorter than news articles. As for other research regarding POS Tagging in Indonesian with each POS Tagger method [2,8,9,10,11,12].

It has been described above that in this research, the authors plan to find the most efficient POS tagger for Indonesian news by using the Conditional Random Fields (CRF) and Hidden Markov Models (HMM) taggers. Conditional Random Fields (CRF) is a probabilistic calculation method for determining the order of labels to be assigned to the sequence of observations [13]. And the Hidden Markov Model (HMM) method is a process of providing tags that can classify one series or sequence of tags for each word in one sentence. HMM also uses a probabilistic technique, in which the resulting sequence of two stochastic coefficients, one of the processes, is unobservable (hidden state) [9,13].

Submitted: **27/12/2022**; Accepted: **12/01/2023**; Published: **28/01/2023**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 1, Januari 2023, Page 214-222
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.v7i1.5384

The research conducted will use the HMM and CRF Tagger methods because both methods use probability techniques to determine a tag in a word. Then the dataset used is a dataset from online news in Indonesian. The corpus used is the corpus that comes from Indonesia Manually Tagged Corpus which contains the text of a news item that has been tagged manually with a total of 23 tag set [15]. The expected results in this research are that we can find out which POS taggers method between CRF and HMM taggers has the best efficiency. The method that is considered to have the best efficiency is the method that has a high accuracy value and a faster training time on the dataset.

2. RESEARCH METHODOLOGY

2.1 System Design

The system set up in the research is purposed to find the most efficient POS tagger on Indonesian news. In this process it is required the following steps.

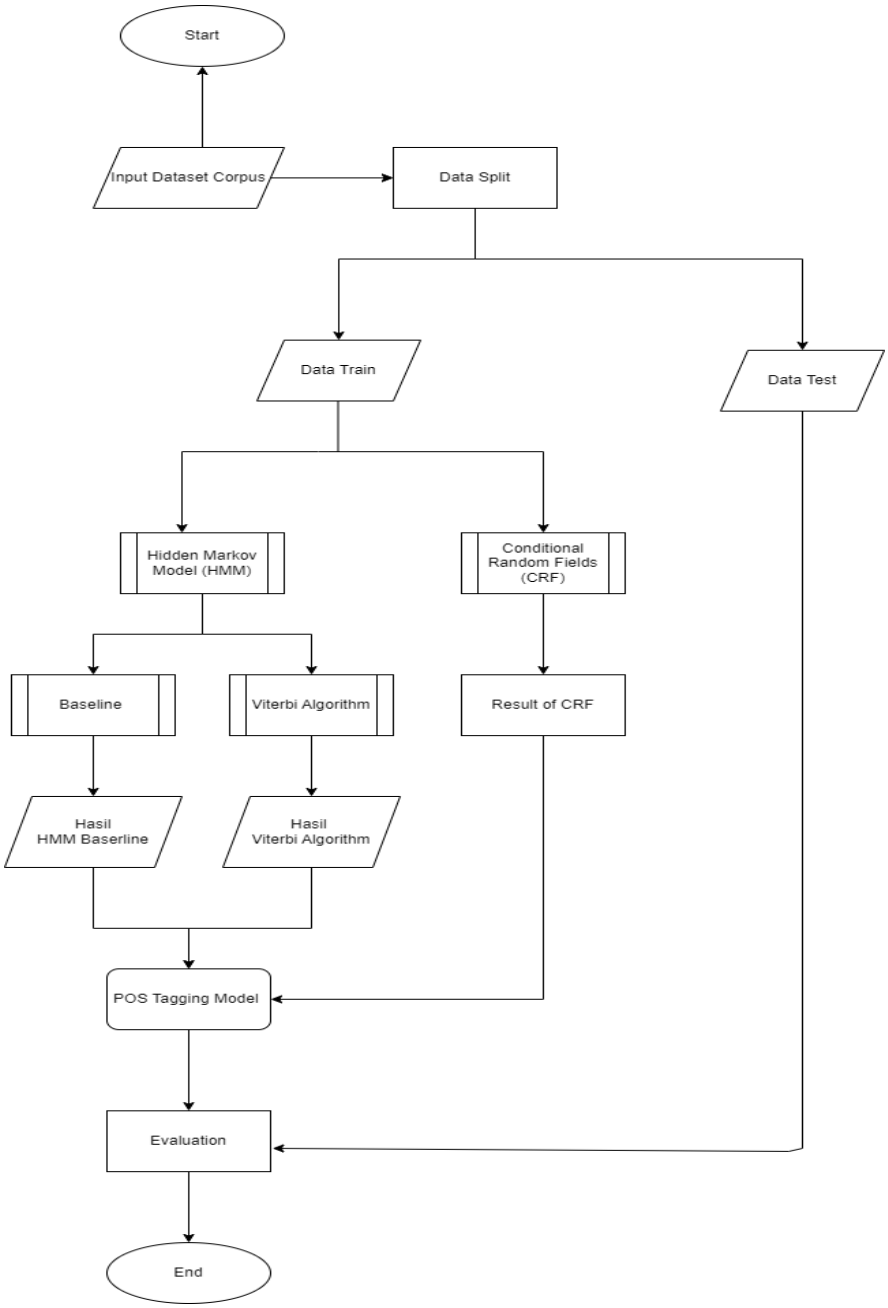

Figure 1. System Design

In Figure 1 is the system design carried out in this research. The steps taken include input corpus dataset, data splitting on train and test data, then the train data will be entered into the model HMM, HMM with Viterbi Algorithm and CRF method and for the final step is evaluate the post tagging method based on value of accuracy from the each of POS taggers model.

Submitted: **27/12/2022**; Accepted: **12/01/2023**; Published: **28/01/2023**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 1, Januari 2023, Page 214-222
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.v7i1.5384

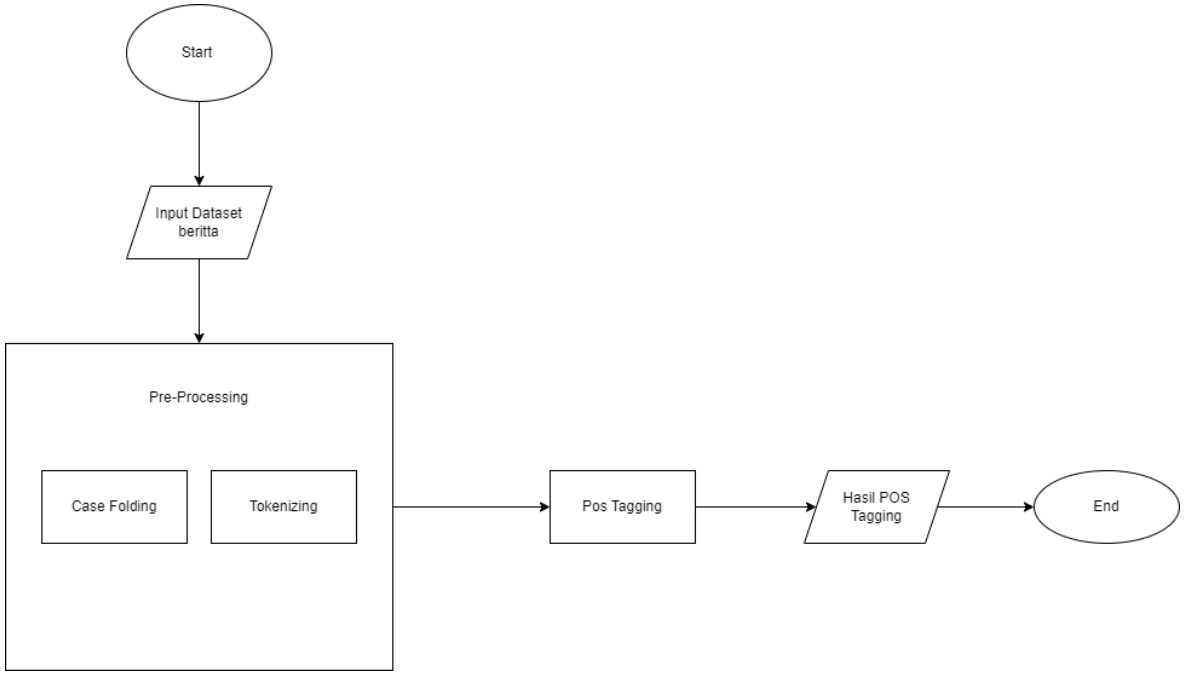

Figure 2. Process on Indonesian News Tagging

In figure 2 is a process on tagging Indonesian news, we can observe in the first steps is inputting Indonesian news data on the system, then pre-processing the data which consists of case folding and tokenizing, after that the next step is tagging the word already pre-processed by each post tagging method. Then in the last steps it produces words that have been labeled by the post tagging method.

2.2 Data Collection

The dataset that will be used is the result of Web Scraping using scrappy in programming in python. The data has been taken from online news articles in Indonesian language consisted of title, category, author, and scrape time. Files that have been scraped will be saved in JSON form to be used in this research. The corpus used also has been taken from the previous research, namely "Designing an Indonesian Part of speech Tagged and Manually Tagged Indonesian Corpus" with a TSV (Tab Separated Values) corpus file format [15]. In this corpus, there are 23 tag set used in this corpus. The following is an example of a web scrapping news dataset and a corpus tag set that will be used.

Tabel 1. Dataset Indonesia News

Title	Category	Author	Date	Article	Scrapetime
"Universitas Terbaik di Jawa Timur Versi Webometrics, Siapa yang Teratas?"	detikEdu	Novia Aisyah	Selasa, 29 Jun 2021 15:44 WIB	Universitas terbaik di Jawa Timur versi Webometrics - Webometrics beberapa waktu lalu merilis daftar universitas terbaik di Indonesia saat ini. Pemeringkatan ini berdasarkan sistem penilaian berbasis situs web masing-masing perguruan tinggi. Untuk para pelajar SMA kelas 12, informasi tersebut juga dapat dijadikan referensi memilih perguruan tinggi tahun depan.\nUrutan tiga besar secara berurutan ditempati oleh Universitas Indonesia, Universitas Gadjah Mada, dan IPB University.	2021-07-14 18:57:37

Table 2. Corpus Tagset

Tag	Description
CC	Coordinating Conjunction
CD	Cardinal Number
OD	Ordinal Number
DT	Article
FW	Foreign Word
IN	Preposition

Submitted: **27/12/2022**; Accepted: **12/01/2023**; Published: **28/01/2023**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 1, Januari 2023, Page 214-222
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.v7i1.5384

Tag	Description
JJ	Adjective
MD	Modal Verb
NEG	Negation
NN	Noun
NNP	Proper Noun
NND	Measurement Noun
PR	Demonstrative Pronoun
PRP	Personal Pronoun
RB	Adverb
RP	Particle
SC	Subordinating Conjunction
SYM	Symbol
UH	Interjection
VB	Verb
WH	Question
X	Unknown
Z	Punctuation

2.2 Pre-processing

The Pre-processing step is one of the first steps to carried out a classification. The purpose of this pre-processing is to facilitate the data processing so that the data can be used with good quality for later classification. In this study, there are 2 steps of Pre-Processing used, namely case folding white space and tokenization.

2.2.1 Case Folding

The case folding method used is to remove whitespace, whitespace is an empty character. The function of the method is to remove the space by calling the strip() function in python. The following are the results of using the folding whitespace case.

Table 3. Case Folding Whitespace

Before Case Folding Whitespace	After Case Folding Whitespace
Universitas terbaik di Jawa Timur versi Webometrics - Webometrics beberapa waktu lalu merilis daftar universitas terbaik di Indonesia saat ini. Pemeringkatan ini berdasarkan sistem penilaian berbasis situs web masing-masing perguruan tinggi. Untuk para pelajar SMA kelas 12, informasi tersebut juga dapat dijadikan referensi memilih perguruan tinggi tahun depan.\nUrutan tiga besar secara berurutan ditempati oleh Universitas Indonesia, Universitas Gadjah Mada, dan IPB University.	Universitas terbaik di Jawa Timur versi Webometrics - Webometrics beberapa waktu lalu merilis daftar universitas terbaik di Indonesia saat ini. Pemeringkatan ini berdasarkan sistem penilaian berbasis situs web masing-masing perguruan tinggi. Untuk para pelajar SMA kelas 12, informasi tersebut juga dapat dijadikan referensi memilih perguruan tinggi tahun depan. Urutan tiga besar secara berurutan ditempati oleh Universitas Indonesia, Universitas Gadjah Mada, dan IPB University.

2.2.2 Tokenization

Tokenization is a stage in the processing of text data. This stage is the stage of dividing the text into tokens or certain parts. The result of tokenization on this dataset as follows:

Table 4. Tokenization

Before Tokenization	After Tokenization
Universitas terbaik di Jawa Timur versi Webometrics -Webometrics beberapa waktu lalu merilis daftar universitas terbaik di Indonesia saat ini.	"Universitas" "terbaik" "di" "Jawa" "Timur" "versi" "Webometrics" "-" "Webometrics" "beberapa" "waktu" "lalu" "merilis" "daftar" "universitas" "terbaik" "di" "Indonesia" "saat" "ini" "."

2.3 Data Split

The purpose of data split here is to divide train data and test data. This data sharing has an influence on the results of each taggers method to be tested. In this study, the data will be divided into several ratios for each train data and test data.

2.4 Hidden Markov Model

Hidden Markov Model is a statistical model of a system that performs a calculation of an event that cannot be observed based on the observed event [16]. There are five components in the HMM method for POS tagging, namely the hidden

Submitted: **27/12/2022**; Accepted: **12/01/2023**; Published: **28/01/2023**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 1, Januari 2023, Page 214-222
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.v7i1.5384

state which represents the order of tags in the training corpus, the observed state which represents the sequence of data in the training corpus, and the transitional probability which represents possible tags [17]. In the POS tagging process in the HMM method, the observed data is a collection of sentences which will then be determined what class of word or tag is appropriate [16]. The process in the HMM method can be written in the following equation:

$$p(t_i|w_i) = \frac{p(t_i)}{n} x \frac{p(t_i|t_{i-1})}{p(t_{i-1})} x \frac{p(w_i|t_i)}{p(t_i)} \quad (1)$$

Where t_i is the word class of w_i in the corpus, n is the number of words in the corpus, and p is the probability or probability. In the HMM method there are 4 main components, namely states which are labels or word classes, the initial distribution is the probability of the observerd states, the emission probability is the probability of the observed state with a tag that may be a label, and the transition probability which is a comparison between the observed tags. with the previous tag.

2.4 Algorithm Viterbi

The Viterbi algorithm is a process for obtaining a series of words with a predictable word class or tag. The results of this algorithm are a series of words with predicted word classes, then these results will be used to calculate the accuracy value obtained from the comparison of predicted word classes produced by the Viterbi Algorithm with word classes from corpus data [10]. The Viterbi algorithm is a process for finding the most optimal path for the hidden state. The following is the formula of the Viterbi algorithm.

$$v_t(j) = \max_{1 \le i \le N-1} v_{t-1}(i) a_{ij} b_j(O_t) \quad (2)$$

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

The process starts with a dynamic algorithm to find the Hidden State Path. In the case of POS Tagging, the hidden state is a sequence of word classes (tags). After getting the results of the Emission Probability and Transition Probability from the previous process, the Viterbi Algorithm will determine the POS tag that is considered the most appropriate for the word.

3.5 Conditional Random Fields

Conditional random fields is a framework for building probabilistic models for sequential segmentation and labeling data [3,8]. CRF is a type of undirected probabilistic graphical model that is commonly used to compare a relationship between observations and build a consistent interpretation [5,18,19]. A representation of a model of CRF can be seen in Figure 1. Where x is a word and y is a tag.

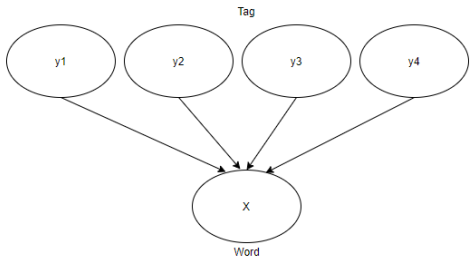

Figure 3. Representation of CRF Graph

CRF is a probabilistic approach whose main process is to calculate the conditional probability between a random variable and adjacent variables [8]. It can be seen in Figure 1 that x is a random variable for input and y is a random variable for sorting the appropriate labels. In this study the variable x is a data to be observed, x is a word in online news and y is a tag or label. The conditional probability from y to x can be written in the equation:

$$p(y|x\lambda) = \frac{1}{z(x)} \exp\left(\sum_j \lambda_j \sum_{i=1}^{n} f_j(y_{i-1}, y_i, x, i)\right) \quad (3)$$

Where λ_j is a value to be observed from the training data and z(x) is a normalization function.

2.5 Confusion Matrix

Confusion Matrix is a performance measurement for classification problems in machine learning [20]. Confusion Matrix will display and compare an actual value with the predicted value results, where the results of the comparison will produce an evaluation metric such as accuracy, precision, recall and f1-score. There are four values that can be generated by the confusion matrix, including True Positive (TP), False Positive (FP), False Negative and True Negative (TN). [21]. Confusion Matrix is used to calculate the values of accuracy, precision, recall, and f1-score. But in this research is uses only accuracy method. The calculation formula of accuracy as follows The accuracy value is

Submitted: **27/12/2022**; Accepted: **12/01/2023**; Published: **28/01/2023**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 1, Januari 2023, Page 214-222
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.v7i1.5384

obtained from the number of positive data that is predicted to be positive and negative data that is predicted to have a negative value and will be divided by the total number of data in the dataset.

$$\text{Acc} = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}} \quad (4)$$

3. RESULT AND DISCUSSION

The evaluation stage in this study there are 4 test scenarios to evaluate the system that has been built. Scenario 1 is a test of the data splitting effect on processing time and accuracy. Scenario 2 aims to increase the accuracy of the HMM method by using the Viterbi algorithm. Scenario 3 aims to optimize the POS taggers method. Then the 4th scenario carries out the process of labeling news sentences by each POS taggers method.

3.1 Scenario 1 Effect of Data spilt on processing time and accuracy

Splitting data is done to divide the ratio of train data and test data on each tagger. Splitting data affects the accuracy of each taggers result and also on the processing time required to train the data for each POS taggers method. These results can be observed in the following table:

Table 6. Result of Data Splitting

Methods	Ratio	Time	Accuracy (Train)	Accuracy (Test)
CRF	80:20	146.90s	98.77%	97.68%
	60:40	97.10s	98.14%	96.85%
	50:50	72.41s	98.10%	96.64%
HMM	80:20	32.45s	97.39%	94.52%
	60:40	26.12s	97.35%	93.88%
	50:50	19.27s	97.26%	93.55%

Based on the results obtained from this test scenario, the CRF method has a higher accuracy value than the others, namely 97.18% at a ratio of 80% train data and 20% test data, but the time required for the data train process has the longest time range. from the HMM method where the result is 154.06 seconds, while the HMM method has a lower accuracy than the CRF method with 94.77% of 80% train data and 20% test data, but the advantage of this method is that the train time process is relatively fast compared to the HMM method. The CRF is 36.02 seconds.

3.2 Scenario 2 Viterbi Algorithm

Based on the results of the previous test scenario, it can be observed in Table 6 where the HMM has a lower accuracy on the test data than the CRF, so in this scenario, the HMM method will be tested using the Viterbi algorithm. The results of this test also involve the addition of the HMM method with the Viterbi algorithm to see what effect it has on the resulting accuracy value. The test results can be observed through the following table:

Table 7. Result of Algorithm Viterbi

Method	Time Processing	Accuracy (Test)
HMM Viterbi Algorithm	431.56s	96.25%

From the results of test scenario 2, it turns out that the HMM with the Viterbi algorithm has an influence on the processing time on the model and the accuracy value on the test. The accuracy value before using the Viterbi algorithm, the HMM method has the highest accuracy value of 94.64%, with the Viterbi algorithm the accuracy value of the HMM method has increased to 96.25%, while the processing time required by this algorithm has also increased to be longer, namely 431.56 seconds in the ratio.

3.3 Scenario 3 Optimization

The next test is to optimize the taggers method with the aim of getting the best results from parameter testing, to determine the best parameters for the HMM method using K-Fold Cross-Validation and CRF method using Grid Search Validation. The test results of this scenario produce the following final results:

Table 7. Result of K-Fold Cross Validation

K-Fold	Accuracy	Precision	Recall	F1-Score
3	93.85	94.20	93.86	93.89
4	94.06	94.39	94.06	94.10
5	94.13	94.44	94.13	94.17
6	94.15	94.44	94.15	94.17
7	94.31	94.57	94.32	94.33
8	94.36	94.63	94.36	94.38

Submitted: **27/12/2022**; Accepted: **12/01/2023**; Published: **28/01/2023**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 1, Januari 2023, Page 214-222
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.v7i1.5384

K-Fold	Accuracy	Precision	Recall	F1-Score
9	94.36	94.61	94.37	94.37
10	94.44	94.65	94.44	94.44

Table 8. Result Of Grid Search Cross Validation

Parameter C1	Parameter C2	Mean fit Time	Mean Test Score	Mean Train Score
0.01	0.01	256.01	97.29	98.73
0.01	0.1	272.35	97.29	98.73
0.01	1	274.10	97.29	98.73
0.1	0.1	276.73	97.29	98.73
0.1	1	274.53	97.29	98.73
1	0.01	258.08	97.29	98.73
1	0.1	270.10	97.29	98.73
1	1	199.97	97.29	98.73

The results of this test scenario 3 on the Table 7 we can observe that HMM method by applying 10 K-Fold with different accuracy, precision, recall, f1-socre results in each fold. The fold that has the lowest evaluation value is in fold-10 with an accuracy value of 94.44%, precision 94.65%, recall 94.44%, fi-score 93.44%. Meanwhile, the fold that has the lowest evaluation value is fold-3 with an accuracy value of 93.85%, precision 94.20%, recall 93.86%, F1-Score 93.89%. Then based on Table 8 the results for the grid search cross-validation test on the CRF method showed a mean test score of 97.29% and a mean train score of 98.73 for each parameter c1 and c2, but there were differences in the mean fit time for each parameter. In table 8 it can be observed that the best mean fit time is at parameters c1=1 and c2=1 resulting in 199.97 seconds, while parameters c1 = 0.01 and c2 = 1 have the longest time with a mean fit time of 276.73 seconds.

3.4 Scenario 4 POS Tag on Indonesian News

After going through the previous scenario stages, the next scenario will label the news articles that have been obtained and data processing has been carried out by each tagger method. It can be seen in table 1 that the news data taken is only the article column with a total of 445 columns. The labels or tags on the news generated by each method will be compared with the results. The comparison of tags will be taken from the sample sentences contained in the news article column and the results are as follows:

Table 9. Sample of Sentence on Indonesian News

Sentence with actual tag
Telkom/NNP University/NNP menjadi/VB universitas/NN swasta/JJ terbaik/JJ di/IN Indonesia/NNP versi/NN Webometrics/NNP ./Z Secara/IN nasional/NN ,/Z perguruan/NN tinggi/JJ swasta/JJ (/Z PTS/NNP)/Z ini/PR menempati/VB peringkat/NN ke-7/OD setelah/SC Universitas/NNP Airlangga/NNP ./Z Telkom/NNP University/NNP berlokasi/VB di/IN Bandung/NNP ,/Z Jawa/NNP Barat/NNP ./Z Perguruan/NN Tinggi/NN yang/SC dibuka/VB sejak/IN 2013/CD tersebut/PR terakreditasi/NN A/SYM oleh/IN Badan/NNP Akreditasi/NNP Nasional/NNP Perguruan/NNP Tinggi/NNP (/Z BAN-PT/NNP)/Z ./Z Telkom/NNP University/NNP juga/RB menjadi/VB Perguruan/NN Tinggi/NN Swasta/NN (/Z PTS/NNP)/Z pertama/OD di/IN wilayah/NN Jawa/NNP Barat/NNP dan/CC Banten/NNP yang/SC pertama/OD kali/NND mengantongi/VB akreditasi/NN A/SYM dari/IN Badan/NNP Akreditasi/NNP Nasional/NNP Perguruan/NNP Tinggi/NNP (/Z BAN-PT/NNP)/Z pada/IN tahun/NN 2016/CD ./Z Hingga/IN saat/IN ini/PR tercatat/VB hanya/RB 104/CD Perguruan/NN Tinggi/JJ yang/SC terakreditasi/NN A/SYM dari/IN Total/NN Perguruan/NNP Tinggi/NNP di/IN Indonesia/NNP (/Z 4.603/CD Perguruan/NN Tinggi/NN)/Z ./Z
Sentence with HMM tag
('Telkom', 'NNP'), ('University', 'NNP'), ('menjadi', 'VB'), ('universitas', 'NN'), ('swasta', 'JJ'), ('terbaik', 'JJ'), ('di', 'IN'), ('Indonesia', 'NNP'), ('versi', 'FW'), ('Webometrics.', 'FW'), ('Secara', 'FW'), ('nasional,', 'FW'), ('perguruan', 'FW'), ('tinggi', 'JJ'), ('swasta', 'JJ'), ('(', 'Z'), ('PTS', 'NNP'), (')', 'Z'), ('ini', 'PR'), ('menempati', 'VB'), ('peringkat', 'NN'), ('ke-7', 'JJ'), ('setelah', 'SC'), ('Universitas', 'NNP'), ('Airlangga', 'NNP'), ('.', 'Z'), ('Telkom', 'NNP'), ('University', 'NNP'), ('berlokasi', 'VB'), ('di', 'IN'), ('Bandung', 'NNP'), (',', 'Z'), ('Jawa', 'NNP'), ('Barat', 'NNP'), ('.', 'Z'), ('Perguruan', 'SC'), ('Tinggi', 'JJ'), ('yang', 'SC'), ('dibuka', 'VB'), ('sejak', 'IN'), ('2013', 'CD'), ('tersebut', 'PR'), ('terakreditasi', 'Z'), ('A', 'NNP'), ('oleh', 'IN'), ('Badan', 'NNP'), ('Akreditasi', 'NNP'), ('Nasional', 'NNP'), ('Perguruan', 'NNP'), ('Tinggi', 'NNP'), ('(', 'Z'), ('BAN-PT', 'NNP'), (')', 'Z'), ('.', 'Z'), ('Telkom', 'NNP'), ('University', 'NNP'), ('juga', 'RB'), ('menjadi', 'VB'), ('Perguruan', 'IN'), ('Tinggi', 'NN'), ('Swasta', 'JJ'), ('(', 'Z'), ('PTS', 'NNP'), (')', 'Z'), ('pertama', 'OD'), ('di', 'IN'), ('wilayah', 'NN'), ('Jawa', 'NNP'), ('Barat', 'NNP'), ('dan', 'CC'), ('Banten', 'NNP'), ('yang', 'SC'), ('pertama', 'OD'), ('kali', 'NND'), ('mengantongi', 'VB'), ('akreditasi', 'IN'), ('A', 'NNP'), ('dari', 'IN'), ('Badan', 'NNP'), ('Akreditasi', 'NNP'), ('Nasional', 'NNP'), ('Perguruan', 'NNP'), ('Tinggi', 'NNP'), ('(', 'Z'), ('BAN-PT', 'NNP'), (')', 'Z'), ('pada', 'IN'), ('tahun', 'NN'), ('2016', 'PR'), ('.', 'Z'), ('Hingga', 'IN'), ('saat', 'NN'), ('ini', 'PR'), ('tercatat', 'VB'), ('hanya', 'RB'), ('104', 'CD'), ('Perguruan', 'NND'), ('Tinggi', 'NN'), ('yang', 'SC'),

Submitted: **27/12/2022**; Accepted: **12/01/2023**; Published: **28/01/2023**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 1, Januari 2023, Page 214-222
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.v7i1.5384

('terakreditasi', 'VB'), ('A', 'NNP'), ('dari', 'IN'), ('Total', 'NN'), ('Perguruan', 'RB'), ('Tinggi', 'JJ'), ('di', 'IN'), ('Indonesia', 'NNP'), ('(', 'Z'), ('4.603', 'NNP'), ('Perguruan', 'NNP'), ('Tinggi', 'NNP'), (')', 'Z'), ('.', 'Z')

Sentence with CRF Tag

('Telkom', 'NNP'), ('University', 'NNP'), ('menjadi', 'VB'), ('universitas', 'NN'), ('swasta', 'JJ'), ('terbaik', 'JJ'), ('di', 'IN'), ('Indonesia', 'NNP'), ('versi', 'NNP'), ('Webometrics', 'NNP'), ('.', 'Z'), ('Secara', 'IN'), ('nasional', 'JJ'), (',', 'Z'), ('perguruan', 'NN'), ('tinggi', 'JJ'), ('swasta', 'JJ'), ('(', 'Z'), ('PTS', 'NNP'), (')', 'Z'), ('ini', 'PR'), ('menempati', 'VB'), ('peringkat', 'NN'), ('ke-7', 'CD'), ('setelah', 'SC'), ('Universitas', 'NNP'), ('Airlangga', 'NNP'), ('.', 'Z'), ('Telkom', 'NNP'), ('University', 'NNP'), ('berlokasi', 'VB'), ('di', 'IN'), ('Bandung', 'NNP'), (',Jawa', 'NNP'), ('Barat', 'NNP'), ('.', 'Z'), ('Perguruan', 'NNP'), ('Tinggi', 'NNP'), ('yang', 'SC'), ('dibuka', 'VB'), ('sejak', 'IN'), ('2013', 'CD'), ('tersebut', 'PR'), ('terakreditasi', 'NN'), ('A', 'NNP'), ('oleh', 'IN'), ('Badan', 'NNP'), ('Akreditasi', 'NNP'), ('Nasional', 'NNP'), ('Perguruan', 'NNP'), ('Tinggi', 'NNP'), ('(', 'Z'), ('BAN-PT', 'NNP'), (')', 'Z'), ('Telkom', 'NN'), ('University', 'NNP'), ('juga', 'RB'), ('menjadi', 'VB'), ('Perguruan', 'NNP'), (‘Tinggi’, ‘NNP’),('Swasta', 'NNP'), ('(', 'Z'), ('PTS', 'NNP'), (')', 'Z'), ('pertama', 'OD'), ('di', 'IN'), ('wilayah', 'NN'), ('Jawa’,’NNP’) (’Barat', 'NNP'), ('dan', 'CC'), ('Banten', 'NNP'), ('yang', 'SC'), ('pertama', 'OD'), ('kali', 'NND'), ('mengantongi', 'VB'), ('akreditasi', 'NN'), ('A', 'NNP'), ('dari', 'IN'), ('Badan', 'NNP'), ('Akreditasi', 'NNP'), ('Nasional', 'NNP'), ('Perguruan', 'NNP'), ('Tinggi', 'NNP'), (' (', 'NNP'), ('BAN-PT', 'NNP'), (')', 'Z'), ('pada', 'IN'), ('tahun', 'NN'), ('2016', 'CD'), ('.', 'Z'), ('Hingga', 'IN'), ('saat', 'NN'), ('ini', 'PR'), ('tercatat', 'VB'), ('hanya', 'RB'), ('104', 'CD'), ('Perguruan', 'NN'), ('Tinggi', 'JJ'), ('yang', 'SC'), ('terakreditasi', 'NN'), ('A', 'NNP'), ('dari', 'IN'), ('Total', 'NNP'), ('Perguruan', 'NNP'), ('Tinggi', 'NNP'), ('di', 'IN'), ('Indonesia', 'NNP'), ('(', 'Z'), ('4.603', 'CD'), ('Perguruan', 'NN'), ('Tinggi', 'JJ'), (')', 'Z'), ('.', 'Z')]]

Table 10. Sample sentence accuracy

Token	HMM Tag (Correct)	HMM Tag (Wrong)	Accuracy
116	92	24	79,31%
Token	**CRF Tag (Correct)**	**CRF Tag (Wrong)**	**Accuracy**
116	105	15	90.05%

It can be observed in the sample sentence testing with 116 tokens carried out for both methods, both of which have different tags in the sentences for each method. In Table 9 and Table 10, we can observe the differences in tags in the HMM and CRF methods. The CRF method has 105 correct tags and 15 wrong tags with an accuracy value of 90.05%, this is quite different from the HMM method, the number of correct tags is only 92 and 24 incorrect, so it produces an accuracy value of 79.31%.

3.5 Analysis of Test Result

After carrying out 4 test scenarios, it can be seen that each test can affect the result value of the POS taggers method. So to get optimal results it needs some testing. In scenario 1, splitting data also affects the accuracy value of each POS tagger method, where the ratios tested are 80:20, 60:40, and 50:50. The three ratios result in a ratio of 80:20 getting the best accuracy value between the two POS taggers methods. This happens because the train data has an influence on the resulting value and on the process of training the model. In test scenario 1 it can be concluded that the more data that is trained, the better the accuracy results obtained. From each decrease in the data splitting ratio, the results obtained are also decreasing, but the less data that is trained, the shorter the processing time on the data train, and the more ratio of data that is trained, the longer the processing time for data train training.

For scenario 2, it is a test of the HMM method by adding the Viterbi algorithm, the results obtained for this addition have increased accuracy results from before. Where the Viterbi algorithm is able to increase the accuracy value from 94.52% to 96.25% but these results are also obtained by increasing the processing time of the Viterbi algorithm. Then scenario 3 tests with grid search cross-validation for CRF taggers and also k-fold cross-validation for HMM taggers. This test was carried out at a value of K=1 to K=10. The results of the k-fold cross-validation show that the value of K = 9 is a better value than the other k values, this test also shows that the comparison between folds is not too significant for the performance measurement value, but the value of K = 9 has a better result value.

Optimization of hyperparameter tuning is carried out for the CRF method by trying to find which parameters are the most optimal for this method. The parameter values to be searched for are parameters c1 and c2. The search for the best parameter is in the range c1 = [0.01,01, 1] and c2 = [0.01,01,1] with a value of cv = 3 and a total of 27 fits, the total time to do the fitting is 39min with the results of all parameters having an average value -Train and test averages at the same number in this corpus dataset. This happens because of the consistency of the data and the absence of imbalanced data, so that the process carried out by the CRF and Grid Search Cross Validation methods produces consistent results for each parameter.

In the next test, namely doing POS tagging on Indonesian news by each method. News data that has been pre-processed before will be tagged by the previously processed corpus. In the news tagging process, the CRF method has a higher accuracy value than HMM. This happens because of the different tags in the words of the two methods.

4. CONCLUSION

After conducting research on the efficiency analysis of POS taggers on Indonesian news, it can be concluded that the CRF method has a higher accuracy value than the HMM method. However, each of these methods has its advantages

Submitted: **27/12/2022**; Accepted: **12/01/2023**; Published: **28/01/2023**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor 1, Januari 2023, Page 214-222
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.v7i1.5384

and disadvantages. In the CRF method, the highest accuracy value is 97.68% for the test corpus data with a training time of 146.90s from a data sharing ratio of 80:20 for train and test data. While the Accuracy value of the HMM method is at 94.52% with a relatively short training time of 32.45s. The Viterbi algorithm for the HMM method can increase the Accuracy value to 96.25% but the time required for processing this algorithm becomes longer, namely 431.56s. Optimization was also carried out on each POS Taggers method with a different optimization method but still the results of the optimization of the CRF method still had an average value higher than the HMM with a value of 97.29%. Testing on the news sample sentences also shows that CRF has an Accuracy value that is higher than HMM with a value of 90.05%, while the HMM method only gets an Accuracy value of 79.31%. From these results indicate that the CRF method has a higher Accuracy value than the HMM method, it can be concluded that the CRF method is better than the HMM method because for each test the value of the Accuracy CRF is always above the HMM, but for time efficiency HMM is better because the training process on data is relatively faster than the CRF method.

REFERENCES

[1] R. Banga and P. Mehndiratta, "Tagging Efficiency Analysis on Part of Speech Taggers," Proc. - 2017 Int. Conf. Inf. Technol. ICIT 2017, pp. 264–267, 2018, doi: 10.1109/ICIT.2017.57.
[2] A. Z. Amrullah, R. Hartanto, and I. W. Mustika, "A comparison of different part-of-speech tagging technique for text in Bahasa Indonesia," Proc. - 2017 7th Int. Annu. Eng. Semin. Ina. 2017, 2017, doi: 10.1109/INAES.2017.8068538.
[3] A. Zilziana, A. A. Suryani, and I. Asror, "Part of Speech Tagging Menggunakan Bahasa Jawa Dengan Metode Condition Random Fields," e-Proceeding Eng., vol. 7, no. 2, pp. 8103–8111, 2020.
[4] A. Chiche and B. Yitagesu, "Part of speech tagging: a systematic review of deep learning and machine learning approaches," J. Big Data, vol. 9, no. 1, 2022, doi: 10.1186/s40537-022-00561-y.
[5] D. Kun Indarta and A. Romadhony, "Aspect and Opinion Extraction of Indonesian Lipsticks Product Reviews using Conditional Random Field (CRF)," KST 2021 - 2021 13th Int. Conf. Knowl. Smart Technol., pp. 113–117, 2021, doi: 10.1109/KST51265.2021.9415829.
[6] A. S. Nasution et al., "Sejarah Perkembangan Bahasa Indonesia," J. Multidisiplin Dehasen, vol. 1, no. 3, pp. 197–202, 2022.
[7] "KBBI." https://kbbi.web.id/berita
[8] F. Pisceldo, M. Adriani, and R. Manurung, "Probabilistic Part of Speech Tagging for Bahasa Indonesia," Proc. 3rd Int. MALINDO Work. Coloca. event ACL-IJCNLP, no. May, 2009.
[9] S. Briandoko, A. R. Dewi, and M. A. Setiawan, "Perbandingan Algoritma Conditional Random Field dan Hidden Markov Model pada Pos Tagging Bahasa Indonesia," vol. 2, no. 2, 2018.
[10] N. Sabloak, "Part-of-Speech (POS) Tagging Bahasa Indonesia Menggunakan Algoritma Viterbi," no. x, pp. 1–11, 2016.
[11] M. Kamayani, "Perkembangan Part-of-Speech Tagger Bahasa Indonesia," J. Linguist. Komputasional, vol. 2, no. 2, p. 34, 2019, doi: 10.26418/jlk.v2i2.20.
[12] 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.
[13] W. AlKhwiter and N. Al-Twairesh, "Part-of-speech tagging for Arabic tweets using CRF and Bi-LSTM," Comput. Speech Lang., vol. 65, 2021, doi: 10.1016/j.csl.2020.101138.
[14] M. Franzese and A. Iuliano, "Hidden markov models," Encycl. Bioinforma. Comput. Biol. ABC Bioinforma., vol. 1–3, pp. 753–762, 2018, doi: 10.1016/B978-0-12-809633-8.20488-3.
[15] V. Krishnapriya, P. Sreesha, T. R. Harithalakshmi, T. C. Archana, and J. N. Vettath, "Design of a POS tagger using conditional random fields for Malayalam," 2014 1st Int. Conf. Comput. Syst. Commun. ICCSC 2014, no. December, pp. 370–373, 2003, doi: 10.1109/COMPSC.2014.7032680.
[16] A. Mulyanto, Y. A. Nurhuda, and N. Wiyanto, "Penyelesaian Kata Ambigu Pada Proses POS Tagging Menggunakan Algoritma Hidden markov Model (HMM)," Pros. Semin. Nas. Metod. Kuantitatif, no. 978, pp. 347–358, 2017.
[17] 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, doi: 10.1109/ICICOS.2017.8276368.
[18] W. Khan et al., "Part of Speech Tagging in Urdu: Comparison of Machine and Deep Learning Approaches," IEEE Access, vol. 7, pp. 38918–38936, 2019, doi: 10.1109/ACCESS.2019.2897327.
[19] S. Fu, N. Lin, G. Zhu, and S. Jiang, "Towards Indonesian Part-of-Speech Tagging : Corpus and Models," Proc. Lr. 2018 Work. Belt Road Lr., vol. 1, pp. 2–7, 2018, [Online]. Available: http://universaldependencies.org/
[20] J. Xu, Y. Zhang, and D. Miao, "Three-way confusion matrix for classification: A measure driven view," Inf. Sci. (Ny)., vol. 507, pp. 772–794, 2020, doi: 10.1016/j.ins.2019.06.064.
[21] Sarang Narkhede, "Confunsion Matrix," Towards Data Science. https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62