Effect of N-Gram on Document Classification on the Naïve Bayes Classifier Algorithm

Fitria Khoirunnisa [a,1], Novi Yusliani, M.T. [b,2], Desty Rodiah, M.T. [c,3]

[a] Research Bachelor, [b,c]Lecturer, Department Informatics Engineering, Faculty of Computer Science, Sriwijaya University,
Jalan Raya Palembang-Prabumulih KM.32 Indralaya, Kabupaten Ogan Ilir and 30662, Indonesia
[1] fitriakhoirunnisa120897@gmail.com; [2] noviyusliani@gmail.com; [3] destyrodiah@gmail.com

ARTICLE INFO

Article history
Received
Revised

Keywords
N-Gram
Naïve Bayes Classifier
Text Mining

ABSTRACT

News has become a major need for everyone, with news we can get the information needed. News can be distributed in the form of print mass media, electronic mass media and online media. The means of spreading the news now have grown very rapidly, making the amount of information being managed are bigger and word management classified also not small. Therefore, we need a system for classifying documents that are not structured. In this study, word processing in a document is done by N-Gram as a feature generation. The document classification process is carried out using the Naïve Bayes Classifier algorithm. This study examines the effect of N-Gram on document classification on the Naïve Bayes Classifier algorithm. The results of the classification accuracy of documents by applying N-Gram is 32.68% and without applying N-Gram is 84.97%. A decrease in the classification results occurs the number of features that result from solving N-Gram that is unique or dominant to another category. The accuracy of the results obtained shows that the application of N-Gram in the classification of documents using the Naïve Bayes Classifier algorithm gives a decreased effect on the performance of the classification.

1. Introduction

News has become a major need for everyone. This is because the news can produce the information that everyone needs. With the news, information about facts or is happening can be known. News is a form of reports about an event that is happening recently or the latest information from an event [1]. News can be distributed in the form of printed mass media such as newspapers, tabloids, newsletters, magazines, bulletins, and books. Electronic mass media such as radio, television and film. As well as online media, internet websites that contain actual information like print media.

The means of spreading the news have now been growing very rapidly, making the amount of information disseminated increased. Then the amount of information that is managed more and more and will make word management classified as not small. Thus, word management needed in categorizing news needs to know the essence of the document content. The process used in categorizing unstructured documents is the classification process of documents using the Naïve Bayes Classifier algorithm [2].

Naïve Bayes Classifier Algorithm is an algorithm used in text mining research. The Naïve Bayes Classifier algorithm is a classification method based on probability and the Bayesian theorem. The classification stage is determined based on the category value of a document that is the term that appears in the classified document [3].

One of the advantages of the Naïve Bayes Classifier algorithm compared to other algorithms [4] is to find out the category of a document by measuring the similarity of related documents through the process of recognizing text and documents. The Naïve Bayes Classifier algorithm is a statistical classification that can predict the probability of a class and a high degree of accuracy. Naïve Bayes Classifier algorithm can improve the results of classification accuracy and eliminate noise by choosing the right type of features and in accordance [5]. One of the right models in managing data in the form of text in data mining and word processing [6] to generate features is N-Gram [7].

N-Gram is a probabilistic model that can be used to generate characters and words and predict the next word in a particular word order. In character generation, N-Gram consists of substrings along the n characters of a string. N-Gram is used to take n-character fragments of a number of words that are continuously read from the source text to the end of the document [7]. The greater the nth value of a word is inversely proportional to the number of exit frequencies obtained, i.e. the smaller or less frequent exit. The use of the Bi-Gram and Tri-Gram models for language models is still possible, because the results of the number of frequency exits in the N-Gram term are still quite large and the data is still valid if it is processed further [7].

Research that applies the N-Gram process is [8] The N-Gram used is Uni-Gram, which is used in classifying sentiments for film reviews using various machine learning techniques, where the results of this study indicate that the N-Gram method as an approach to feature extraction. Where the results of the classification with Uni-Gram in this study resulted in an accuracy of 82.9%. Future studies applying word predictions to help speed up the typing process by applying N-Gram [7]. The results of this study indicate that N-Gram as a basic method in the prediction process is very helpful in sorting words. So that the prediction process becomes more effective, able to produce effective predictions above 20% of the total predictions that occur. Based on the results of this description, this study will examine the effect of N-Gram on the accuracy of the classification results of documents using the Naïve Bayes Classifier algorithm.

2. Literature Study / Hypotheses Development

In this section contains the theoretical foundation and some research that has been done by previous researchers. This was made to strengthen the reasoning and rationality of the involvement of a number of variables in this study. It also functions as a scientific opinion that is integrated with the results of a literature review to build a researcher's mindset in relation to the problem being studied.

Research conducted by [5] said the Naïve Bayes Classifier algorithm can be improved classification accuracy for sentiment analysis by selecting the right type of features and eliminating noise by selecting the appropriate features. Where the research results show, the accuracy of the IMDB popular movie review dataset reached 88.80%. The selection of the right model can be generalized to a number of text categorization problems to increase speed and accuracy.

Research conducted by [7] implements N-Gram as the basic method of predictive processes for word prediction. The process begins by breaking down word by word and grouping them according to language models. Then, the scoring process is carried out to determine which words are suitable for the word prediction choice. The test results show that the N-Gram method as a basic method in the prediction process is very helpful in sorting words, so the prediction process becomes more effective, capable of producing effective predictions above 20% of the total predictions that occur. Keystroke saving that can be generated can reach 50% depending on the training data used. Apart from the N-Gram method itself, setting the weights for each word score also greatly influences the word prediction process.

Research conducted by [8] implements N-Gram as an approach to extracting features, namely Uni-Gram, Uni-Gram with Bi-Gram, Uni-Gram with Part of Speech (POS). , Adjective and N-Gram with position. This study classifies sentiments for film reviews using a variety of machine learning techniques. The machine learning techniques used are Naïve Bayes Classifier, Maximum Entropy, and Support Vector Machines (SVM). The results of this study found that Support Vector Machines (SVM) became the best method when combined with Uni-Gram with an accuracy of 82.9%.

Research conducted by [9] implements a language detection system in multi-language documents using N-Gram. The language variations used in his research are Indonesian and English. His research applies an approach in language detection using N-Gram. By determining the type of language automatically from a text or document based on certain criteria. Language detection aims to classify the language of a document based on training conducted using a collection of documents or corpus. Language detection is done as an initial filter on the corpus, so that it can produce quality data input.

His research uses 25 Indonesian languages and 25 languages English. The results of the study showed a very good performance in detecting language from documents with an average F-measure of 0.93.

3. Methodology

Text mining is a theory about processing large collections of documents that exist from time to time using several analyses [10]. Research in the field of text mining is handling problems related to the classification of documents. A document can be grouped into certain categories based on words or sentences in the document. Words or sentences in a document have a specific meaning and can be used to determine the category of a document.

The purpose of text processing is to know and extract useful information from data sources by identifying and exploring interesting patterns in the case of text mining. The data source used is a collection of documents that are not structured and obtained manually by downloading Indonesian text documents in text format (*.txt) on the sites http://news.kompas.com and http://www.republika.co .id.

Data in the form of text that will go through the training process and testing with the same process, namely through the process of text preprocessing, N-Gram, term frequency and Naïve Bayes Classifier algorithm in the classification of documents, all stages are illustrated in Figure 1 and Figure 2.

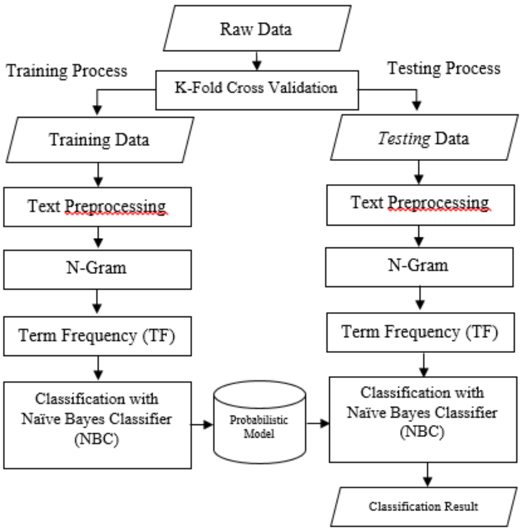

Fig. 1. Software Process Stages Diagram with N-Gram

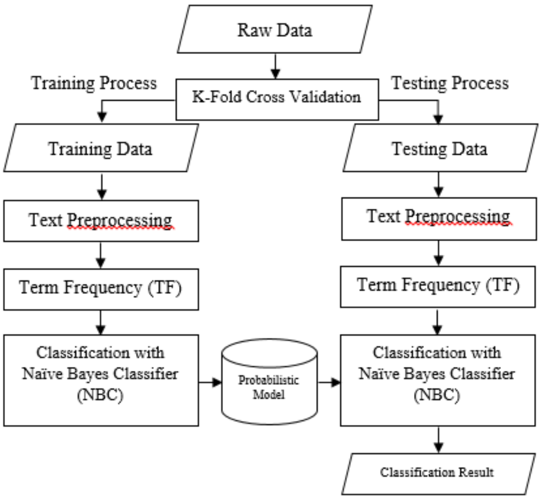

Fig. 2. Software Process Stages Diagram without N-Gram

a. Text Preprocessing

Before determining the features that are represented, extraction stages are needed to be carried out in general on the document, namely case folding, tokenizing, filtering and stemming. In this research the case folding process will change all the letters in the document into lowercase letters. Furthermore, the sentence will be broken up into pieces of words or known as tokenizing and through a filtering process that functions to take important words from the results of tokenizing. The last step is stemming the process of reducing words to basic forms after filtering, where the algorithm used is the Nazief and Andriani algorithm for Indonesian stemming.

b. N-Gram Model

After the text preprocessing is done, that is case folding, tokenizing, filtering and stemming, then the process of generating features. This stage aims to take the number n characters pieces from a word. Where the N-Gram stage is formed based on the frequency of the N-Gram that appears in the document. For each N-Gram raised, it will be recorded in a table. If the N-Gram has already appeared in the document, then the frequency for the N-Gram will be added by one, if not then the N-Gram will be added to the table with the number of occurrences of one. The number of N-Grams in a word is calculated by (1):

$$Ngrams_k = X - (N - 1) \tag{1}$$

Where to:

1. $Ngrams_k$ is a lot of N-Gram in a k word,
2. X is the number of letters in a word k, and
3. N is the type of N-Gram used, unigram, bigram, trigram, and so on.

The N-Gram model is distinguished based on the number of character chunks of n. To help in taking the pieces of words in the form of letter characters, padding is done with blanks at the beginning and end of a word.

c. Naïve Bayes Classifier Algorithm

The classification algorithm using probability and statistical methods proposed by the English scientist Thomas Bayes, which predicts future opportunities based on experience in the past so that it is known as the Bayes Theorem [11]. The basis of the Naive Bayes Classifier theorem used in this study is the following Bayes formula (2):

$$(A|B) = ((A|B) * (A))/(B) \tag{2}$$

The probability of event A as B is determined from opportunity B when A, opportunity A, and opportunity B. In its application (2) becomes (3):

$$(C_i|D) = ((D|C_i) * (C_i))/(D) \tag{3}$$

The Naive Bayes Classifier algorithm is a simplified model of the Bayes algorithm that is suitable for classifying text or documents. Equation (4) is:

$$V_{MAP} = \text{arg max } (v_j|a_{1,2},\ldots,a_n) \tag{4}$$

Based on (4), the Bayes formula can be written into (5) where:

$$V_{MAP} = \underset{v_j \in V}{\text{arg max}} \frac{P(a_1,a_2 \ldots a_n|v_j)P(v_j)}{P(a_1,a_2 \ldots a_n)} \tag{5}$$

$P(a_1, a_2 \ldots a_n)$ in (5) is a constant number, so it can be removed into (6) where:

$$V_{MAP} = \underset{v_j \in V}{\text{arg max}} P(a_1, a_2 \ldots a_n|v_j)P(v_j) \tag{6}$$

Because $P(a_1, a_2 \ldots a_n|v_j)$ in (6) it is difficult to calculate, it will be assumed that each word in the document has no connection, so the equation becomes:

$$V_{MAP} = \underset{v_j \in V}{\text{arg max}} P(v_j) \prod_i P(a_i|v_j) \tag{7}$$

If (7) is broken, then there are two more equations namely, (8) and (9) where:

$$P(v_j) = \frac{|\text{docs}_j|}{|examples|} \tag{8}$$

$$P(w_k|v_j) = \frac{\text{nk} + 1}{n+|vocabulary|} \tag{9}$$

Where to:

1. (v_j) is the probability of each document for a set of documents,

2. $(w_k|v_j)$ is the probability of the word w_k appearing in a document with the class category v_j,
3. $|docs|$ is the document frequency in each category,
4. |examples| is the number of documents available, and
5. n_k is the k-word frequency in each vocabulary category is the number of words in the test document.

d. Classification with Naïve Bayes Classifier algorithm

The classification process stage with the Naïve Bayes Classifier is a stage for grouping documents that have been previously processed at the text preprocessing stage. Classification with the Naïve Bayes Classifier algorithm will go through the term frequency (tf) stage to calculate the frequency of occurrence of a term in the document, before the classification stage. Calculation of the term frequency (tf) produced will be recorded in a table. After that, the classification process is performed using the Naïve Bayes Classifier algorithm.

e. Classification with Naïve Bayes Classifier and N-Gram algorithm

The classification process stage with the Naïve Bayes Classifier algorithm is a stage for classifying documents that have been processed previously at the N-Gram stage. Classification with Naïve Bayes Classifier algorithm will go through stages using N-Gram. And then the classification process will be carried out with the Naïve Bayes Classifier algorithm.

In the initial testing phase of the study, the text data will first be classified through the term frequency (tf) using the Naïve Bayes Classifier algorithm. This stage aims to obtain classification results without applying the N-Gram process. The next process is the classification of text data by applying the N-Gram process that has been trained in advance, the results of the classification of the two tests will be compared. The test data was performed using the k-fold cross validation method with a k value of 10 fold, which means the experiment was carried out 10 times. This test aims to test the stability of accuracy if tested with training data and different test data. Then the classifier performance results from some test data will be summarized in the system evaluation table to calculate the accuracy of the classification results.

4. Results and Discussion

This research examines software using data in the form of online news documents in Indonesian. The document consists of five categories of online news, namely health news, political news, economic news, technology news, and sports news. Each category consists of 60 pieces of text type text documents (*.txt).

The experiments were carried out using the K-Fold Cross Validation method for data sharing. Each fold, testing the selected testing data, uses the Naïve Bayes Classifier algorithm without using N-Gram modelling and the Naïve Bayes Classifier algorithm using N-Gram modelling.

The testing process is carried out in accordance with the architecture of the software system. Where is the Naïve Bayes Classifier algorithm testing process without using N-Gram modelling, through data sharing using the K-Fold Cross Validation stage. Data processing using the text preprocessing process, term weighting is done with the term frequency, the formation of probabilistic models with prior formulas and the document classification process using the Naïve Bayes Classifier algorithm.

While the document testing process is on the Naïve Bayes Classifier algorithm using N-Gram modelling, through data sharing using the K-Fold Cross Validation stage. Data processing using the text preprocessing process, the data will be done in the process of generating features using N-Gram modelling, character weighting is done with term frequency, the formation of probabilistic models with prior formulas and the process of document classification using the Naïve Bayes Classifier algorithm. After the system classification results are obtained, the classification results will be evaluated using accuracy calculations. Table 1 below is the result of classification on the system using the Naïve Bayes Classifier algorithm without using N-Gram modelling and document classification with the Naïve Bayes Classifier algorithm using N-Gram modelling on the 1st fold consisting of 30 data testing.

Table 1. Document Classification Test Result Table

Fold To-	ID	Naïve Bayes Classifier		Conclusion	Naïve Bayes Classifier and N-Gram		Conclusion
		Manual Classification	System Classification		Manual Classification	System Classification	
1	149	Politik	Politik	True	Ekonomi	Kesehatan	False
1	210	Teknologi	Teknologi	True	Kesehatan	Kesehatan	True
1	169	Politik	Politik	True	Ekonomi	Kesehatan	False
…	…	…	…	…	…	…	…
1	7	Kesehatan	Kesehatan	True	Olahraga	Kesehatan	False

Furthermore, the process of calculating the percentage of accuracy shown in table 2 is the result of classification testing data using the Naïve Bayes Classifier algorithm without using N-Gram modelling and document classification with the Naïve Bayes Classifier algorithm using N-Gram modelling for each fold.

Table 2. Evaluation Table in the K-Fold Cross Validation Experiment

Fold To-	Naïve Bayes Classifier	Naïve Bayes Classifier and N-Gram
	Percentage of Accuracy	Percentage of Accuracy
Fold 1	83,3%	26,7%
Fold 2	83,3%	36,7%
Fold 3	93,3%	36,7%
Fold 4	73,3%	26,7%
Fold 5	96,6%	36,7%
Fold 6	80,0%	33,3%
Fold 7	83,3%	26,7%
Fold 8	80,0%	33,3%
Fold 9	93,3%	33,3%
Fold 10	83,3%	36,7%
Average	**84,97%**	**32,68%**

The percentage of the calculation results of the accuracy of the results of the classification of documents with Naïve Bayes Classifier algorithm without using N-Gram modelling and document classification with Naïve Bayes Classifier algorithm using N-Gram modelling.

Table 3. Comparison of Document Classification Test Results

Algorithm	**Percentage of Accuracy**
Naïve Bayes Classifier	84,97%
Naïve Bayes Classifier + N-Gram	32,68%

Analysis of the results of the study explained the comparison of the results of the document classification test with the Naïve Bayes Classifier algorithm without using N-Gram modelling and document classification with the Naïve Bayes Classifier algorithm using N-Gram modelling, which is shown in Figure 3 of the system classification comparison graph.

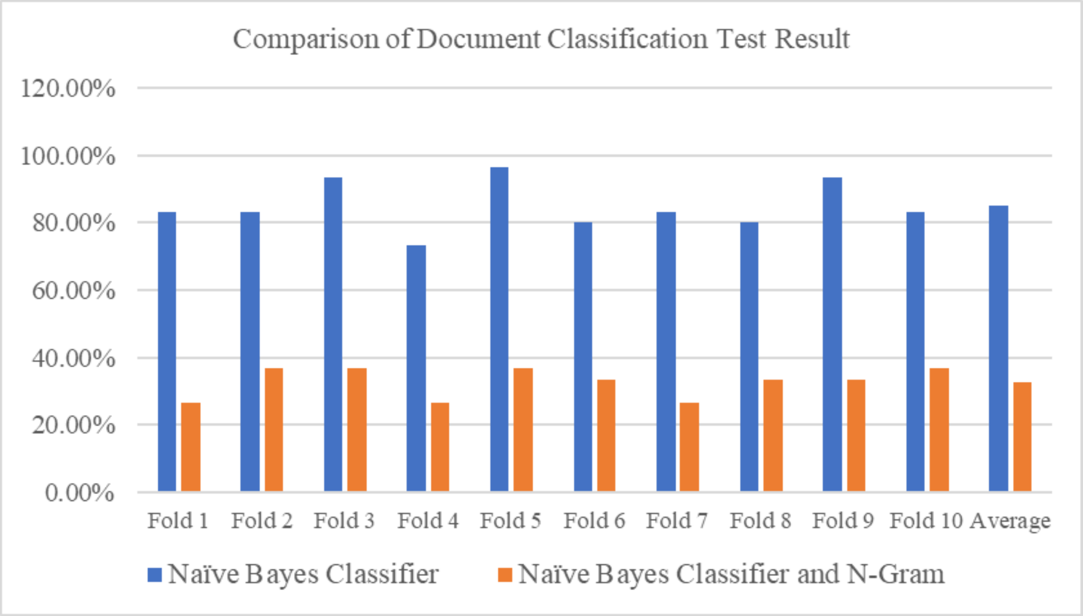

Fig. 1. Comparison of Document Classification Test Results

Figure 3 shows that the document classification results on the Naïve Bayes Classifier algorithm using N-Gram modelling results in lower accuracy, when compared to document classification with the Naïve Bayes Classifier algorithm without using N-Gram modelling. This shows that the use of the N-Gram model can affect the classification results in online news documents. The effect is because the system can be more accurate in classifying using words than the characters produced from the N-Gram modelling process.

The difference in the way of work produced between the two classification processes is in the classification process using the N-Gram model that is the result of the preprocessing text process that is produced in the form of terms broken down in the form of Bi-Gram. The number of characters generated requires sufficient space to accommodate the characters from the N-Gram solving process in the training data. That makes the number of features that result from solving N-Gram unique or dominant to other categories. So, the opportunities generated from the term frequency are so poor to be applied to the Naïve Bayes Classifier algorithm process.

The time generated in the classification process using N-Gram modelling is much longer compared to without using N-Gram modelling. The process of breaking the term into characters will influence the timing of the document classification process. Overall classification results of the two classification processes using the Naïve Bayes Classifier algorithm are still relatively low. This is because the process of training data in the formation of probabilistic model data affects the results of classification.

5. Conclusion

Based on the research results described above, there are some conclusions, namely the N-Gram modelling process can give effect to the results of the classification of online news documents by using the Naïve Bayes Classifier algorithm, where the accuracy is 32.68% and the accuracy results without using N-Gram modelling namely 84.97% and in this study, the effect of the N-Gram modelling process on the results of the classification of online news documents lies in the documents of health category. The number of characters generated from the N-Gram modelling process is dominant towards the health category.

This research is expected to be able to apply other methods of text mining in document categorization in addition to applying the Naïve Bayes Classifier algorithm in combination with N-

Gram modelling and applying methods that can determine the position of characters in a term in the classification process in N-Gram modelling.

Acknowledgment

This research was completed because of the support of various parties. Thank you to the online kompas and republika for presenting news to be studied as material for this research. And all parties who support the completion of this research.

References

[1] C. Juditha, “News Accuracy in Online Journalism (News of Alleged Corruption The Constitutional Court in Detiknews),” J. Pekommas, pp.145-154, 2013.

[2] A. Indriani, “Klasifikasi Data Forum dengan menggunakan Metode Naive Bayes Classifier,” Semin. Nas. Apl. Teknol. Inf., pp.28-33, 2014, https://doi.org/10.1155/2011/172853

[3] A. Hamzah, “Klasifikasi Teks Dengan Naïve Bayes Classifier (NBC) Untuk Pengelompokan Teks Berita Dan Abstract Akademis,” Pros. Semin. Nas. Apl. Sains Teknol. Periode III, pp.269-277, 2012, doi: 1979-911X.

[4] I. Setiawan and D. Nursantika, “Klasifikasi Artikel Berita Menggunakan Metode Text Mining Dan Naive Bayes Classifier,” Pros. SENIATI, pp.1-6, 2017.

[5] V. Narayanan, I. Arora, and A. Bhatia, “Fast and accurate sentiment classification using an enhanced Naive Bayes model,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 194–201, 2013, doi: 10.1007/978-3-642-41278-3_24.

[6] S. S. Kumar and A. Rajini, “An Efficent Sentimental Analysis for Twitter Using Neural Network based on Rmsprop,” *IOSR J. Eng.*, pp. 17–25, 2018.

[7] S. A. Sugianto, L. Liliana, and S. Rostianingsih, “Pembuatan Aplikasi Predictive Text Menggunakan Metode N-gram-based,” J. Infra, 2013.

[8] K. Saranya and S. Jayanthy, “Onto-based sentiment classification using machine learning techniques,” Proc. 2017 Int. Conf. Innov. Information, Embed. Commun. Syst. ICIIECS 2017, vol. 2018-January, pp. 1–5, 2018, doi: 10.1109/ICIIECS.2017.8276047.

[9] B. Zaman, E. Hariyanti, and E. Purwanti, “Sistem Deteksi Bahasa pada Dokumen menggunakan N-Gram,” MULTINETICS, 2015, doi: 10.32722/vol1.no2.2015.pp21-26.

[10] A. P. Wijaya, “Klasifikasi Dokumen dengan Naive Bayes Classification (NBC) Untuk Mengatahui Konten,” J. Datamining Indones., pp.1-6, 2012.

[11] S. L. B. Ginting and R. P. Trinanda, “Teknik Data Mining Menggunkan Metode Bayes Classifier Untuk Optimalisasi Pencarian Pada Aplikasi Perpusatakan,” Univ. Pas., 2013.

[12] A. Fathan Hidayatullah, M. Rifqi Ma, and arif Program Studi Manajemen Informatika STMIK Jenderal Achmad Yani Yogyakarta Jl Ringroad Barat, “Penerapan Text Mining dalam Klasifikasi Judul Skripsi,” Semin. Nas. Apl. Teknol. Inf. Agustus, pp. 1907–5022, 2016.

[13] T. Maghfira, I. Cholissodin, and A. Widodo, “Deteksi Kesalahan Ejaan dan Penentuan Rekomendasi Koreksi Kata yang Tepat Pada Dokumen Jurnal JTIIK Menggunakan Dictionary Lookup dan Damerau-Levenshtein Distance,” J. Pengemb. Teknol. Inf. dan Ilmu Komput., pp.498-506, 2017.

[14] J. Mooney, Machine Learning Text Categorization. Austin: University of Texas, 2006.

[15] A. Musthafa, “Klasifikasi Otomatis Dokumen Berita Kejadian Berbahasa Indonesia,” Fakultas Sains dan Teknologi universitas Islam Negeri Maulana Malik Ibrahim Malang, 2009.

[16] C. Sammut and G. Webb, *Encyclopedia of machine learning and data mining*. Springer Publishing Company, Incorporated. 2016.

[17] Zulkifli, T. A. Wibowo, and G. Septiana, “Pembobotan Fitur Ekstraksi pada Peringkasan Teks Bahasa Indonesia Menggunakan Algoritma Genetika,” eProceeding Eng. 2.2, pp.6481–6489, 2015.