Jurnal Mobile and Forensics (MF)
Vol. 3, No. 1, March 2021, pp. 6-16
P-ISSN: 2656-6257, E-ISSN: 2714-6685, DOI: http://dx.doi.org/10.12928/mf.v1i1.3763

NAÏVE BAYES FOR THESIS LABELING

[1]Fitria Nurhayati, [2*]Arfiani Nur Khusna, [3]Dimas Chaerul Ekty Saputra
[1,2]Department of Informatics, Faculty of Industrial Technology, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
[3]Department of Biomedical Engineering, The School Graduate, Universitas Gadjah Mada, Yogyakarta, Indonesia
e-mail: arfiani.khusna@tif.uad.ac.id *correspondence email

Abstrak

Penyusunan skripsi pada Program Studi Teknik Informatika UAD tebagi menjadi 2 bidang minat yaitu Sistem Cerdas (SC) dan Rekayasa Perangkat Lunak dan Data (relata). Data judul skripsi yang ada hanya digunakan sebagai arsip, dan belum pernah diolah maupun diklasifikasikan untuk mengetahui trend topik skripsi berdasarkan bidang minat mahasiswa setiap tahunnya. Sehingga belum terdapat rujukan untuk evaluasi kutikulum. Penelitian ini meliputi tahap koleksi data, pembagian data menjadi 2 bagian (data latih dan data uji), pemberian label manual data latih, text prepocessing, dan klasifikasi. Penelitian ini menggunakan 1290 data judul skripsi, menunjukkan hasil trend pengambilan judul skripsi pada tahun 2013 hingga tahn 2018 sebagian besar mahasiswa mengambil judul relata. Pengujian accuracy memiliki nilai 94.60%, precision 97.30% dan recall 85.70%.

The thesis preparation in the Department of Informatics Universitas Ahmad Dahlan is divided into two areas of interest, namely Intelligent Systems and Software and Data Engineering. Existing thesis title data is only used as an archive and has never been processed or classified to determine the trend of thesis topics based on student interest each year. The stages include data collection, the data is divided into two parts (training data and test data), manual labeling of training data, text preprocessing, and classification using Naive Bayes. The results show the trend of thesis title taking from 2013 to 2018 shows the thesis trend in the field of Intelligent Systems and Software. Accuracy testing uses Confusion Matrix and K-Fold Cross Validation with a k value is 10, has a value of 94.60%, precision of 97.30%, and a recall of 85.70%.

Kata Kunci: Confusion Matrix, Thesis Title, K-Fold Cross Validation, Classifitcation, Naïve Bayes.

INTRODUCTION

Education is the wise, hopeful and respectful cultivation of learning that undertaken in the belief that all should have the chance to share in life. Education can be known as a place of learning that can be done anywhere. Every human being must experience or run the education. The role of education is very important for humans today. Education is also useful for building human characteristics early on. To accomplish the basic education in Indonesia takes 12 years from Primary Intelligent Systemhool, Junior High Intelligent Systemhool and Senior High Intelligent Systemhool (Kurniawati, Suryadarma, Bima, & Yusrina, 2018).

Education taken after a high Intelligent Systemhool level of education is that include diploma, undergraduate, professional, masters, doctorate to specialist program. Such an education is organized by the College. Data from Ministry of Education and Culture in 2017, in Indonesia there are as many as 4.504 university units. Of the total tertiary institutions registered in the Ministry of Education and Culture, 3.136 units are private universities 122 units are state universities, the rest

Received January 28, 2021; Revised February 14, 2021; Accepted March 16, 2021

are universities under ministries or state institutions with service systems and religious colleges.

Universitas Ahmad Dahlan is one of the private collages in Indonesia. It was founded on November 18, 1960. Universitas Ahmad Dahlan develops fields of expertise or study programs in the fields of technology as well as in the social and humanities fields. Universitas Ahmad Dahlan which is has 11 faculties and 36 undergraduate programs. One of the steps that students must go through to get a bachelor's degree is making a research report or thesis as a final project. The thesis is a Intelligent Systemientific paper made by students and is prepared to fulfill some of the requirements for completing education in a bachelor's degree. Intelligent Systemientific work that is made can be in the form of research reports, such as library research, development research, field research, and laboratory research.

In preparing the thesis, especially in the Department of Informatics at Universitas Ahmad Dahlan, it is divided into two areas of interest, namely the Intelligent Systems and Software and Data Engineering. Subjects from each area of interest also differ, some elective courses become part of Intelligent Systems family, and some become part of the Software and Data Engineering family.

Table 1 contains a grouping of elective courses that support each topic of diIntelligent Systemussion of Intelligent Systems and Software and Data Engineering

Table 1. Academic Interest

No	Elective Courses Software and Data Engineering (SDE)	Elective Courses Intelligent System (IS)
1.	Multimedia Introduction	Computer Vision
2.	Dynamic Website Programming	Machine Learning
3.	Data Mining	Decision Support System
4.	Robotics Informatics	Advanced Graphics
5.	Website Engineering	Information Retrieval System
6.	Cryptography	Pattern Recognition
7.	Digital Forensics	Natural Language Processing
8.	Parallel Programming	Game Development
9.	Software Quality Assurance	---

Based on Table 1 Academic Interest that each area of interest has a sub-specialization consisting of different elective courses. Transforming the data on the attributes of the research title to the research topic is still done manually in Microsoft Excel, so it is not very precise. Therefore, it is necessary to carry out the processing by performing automatic grouping (classification) of research titles using one of the techniques, namely text mining. So later it can be seen the trend of the topics that are developing in the implementation of the Department of Informatics at Universitas Ahmad Dahlan student thesis every year. Furthermore, it can be used as material for curriculum evaluation in the Informatics Engineering Study Program.

Rules on the curriculum contained in Law No. 12 Year 2012 Article 35, paragraph 2, it is written that the Higher Education Curriculum developed by each College with reference to the National Standards of Higher Education for each study program that includes the development of intelligence, character, and skills (DPR-RI & INDONESIA, 2012).

The curriculum is a set of plans and rules related to the achievement of learning targets, study materials, the learning process and the values used to guide the implementation of study programs. Curriculum arrangement and planning, consisting

of several stages, including the needs analysis stage, the development stage, the implementation stage, the evaluation stage and the follow-up stages for the good done by the study program (Alsubaei, 2016). Curriculum enhancement is carried out to produce graduates according to the target learning outcomes that have been implemented by the study program.

Curriculum development and institutional-Intelligent Systemale research roadmap planning require knowledge of the trend of thesis topics for students at both the university level and the level of the study program. Based on interviews conducted with Coordinator of Research Methodology and Thesis of the Department of Informatics, that the data that has been collected has never been processed or further classified to determine the trend of thesis topics based on the student's interest area that students take each year.

Classification is a categorization process carried out on a set of documents, classification is very important for the ease of users searching for documents, data classification from some document with specific is the title, is data classification in the form of text, so this type of classification can be done using the text mining method (Ogundare & Wiggins, 2018).

Text mining is an important stage in the Big Data analysis process, that is unstructured, such as a large number of text data (Xiang, Intelligent Systemhwartz, Gerdes Jr, & Uysal, 2015). Text mining is one part of data mining which is used to analyze and process data in the form of text which is semi-structured and unstructured. In contrast to data mining which is generally used to analyze data that is categorical, continuous, or ordinal (Suh, Park, & Jeon, 2010). In the text mining stage, several algorithms commonly used, including the C4.5 algorithm, the Naive Bayes algorithm, Cosine Similarity, TF-IDF weighting, Support Vector Machine (SVM), K-Nearest Neighbor (KNN).

Researches related to text mining have been done before, one of which is using TF-IDF weighting to make a web application about information retrieval on the site detik.com (Khusna & Agustina, 2018). The combination of the Naive Bayes Classification Algorithm and the Chi-Square method in text classification results in excellent performance (Alshalabi, Hamood, Tiun, Omar, & Albared, 2013).

Research conducted using the Naive Bayes Algorithm to analyze product review based on a specific aspect of the product. This review could be range from thousands and various opinions. This research has three phases, data processing using POS, feature selection using Chi-square, and classification using Naïve Bayes (Mubarok, Adiwijaya, & Aldhi, 2017).

Based on pre-existing research, and the problems that occur in the Universitas Ahmad Dahlan Informatics Engineering study program, related to thesis title taking based on areas of interest, this study applies the Naive Bayes algorithm to classify the thesis titles of Informatics Engineering students and produce a trend for each thesis title. years so it is expected that the results will be more accurate.

METHODS

A. Text Mining

Text mining was developed in the 1980s and is becoming more and more effective with the increasing use of computerization. Text mining manages to find relationships and patterns that are not visible, for example measuring a person's level of happiness using his writing on Twitter and others (Allahyari et al., 2017).

Text mining is a part of text analysis that is done automatically and is carried out by a computer, the aim is to extract useful information from a collection of documents. The way this method works is to find words that represent the contents of the related document, then analyze the relationship between the documents using statistical calculations to determine group relationships, classifications and association patterns (Dreisbach, Koleck, Bourne, & Bakken, 2019).

The steps taken at the text mining stage are data cleaning and text preprocessing. Data cleaning aims to eliminate noise in a data or document, cleaning data on documents is useful for filtering invalid data (Katariya & Chaundri, 2015). While the stages in text preprocessing are shown in Figure 1 The following text preprocessing stages:

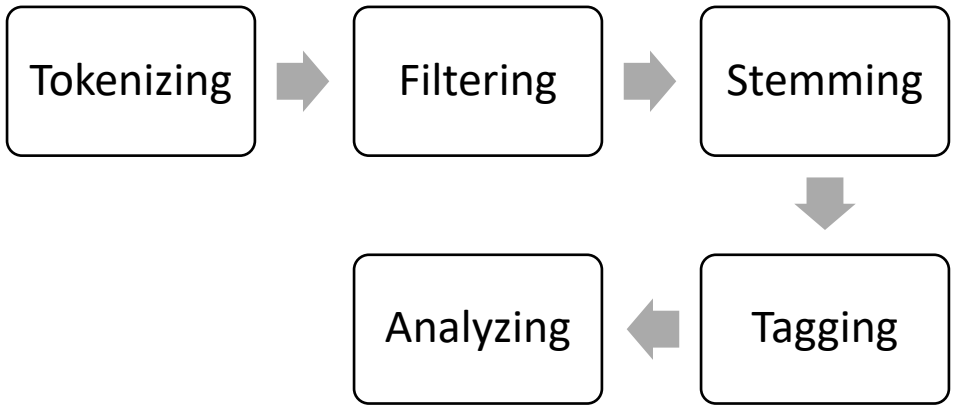

Figure 1. Preprocessing Text

a. ***Tokenizing***

The tokenizing process is done by cutting sentences into constituent words. In the tokenizing process, punctuation is also removed (Pinto, Goncalo Oliveira, & Oliveira Alves, 2016).

b. ***Filtering***

The filtering process is done by removing unnecessary words from the tokenizing process. This stage uses a stopword removal algorithm to remove unnecessary words. A stopword is a collection of words that have no meaning (Pinto et al., 2016).

c. ***Stemming***

The stemming process is done by looking for the root (root) of each word resulting from the filtering. At this stage, the process of returning various forms of words is carried out into the same representation. The stemming stage can also be interpreted as the process of removing affixes (Alami, Meknassi, Ouatik, & Ennahnahi, 2016).

d. ***Tagging***

The tagging process is the stage of finding the initial root form of each word then each word in the text will be categorized into grammatical functions such as nouns, pronouns, adjectives, verbs, adverbs, prepositions, determinants, and conjugations. Part Of Speech Tagging is important because several Natural Language Processing tasks, namely sentiment analysis, answering questions, and word

disambiguation need differentiation to overcome word ambiguity (Zampieri et al., 2018).

e. *Analyzing*

The analyzing process is the process of determining how far the relationship between words in an existing document is. This stage uses the calculation of Term Frequency (TF) and the Naive Bayes Algorithm (Wallace, Feng, Kandpal, Gardner, & Singh, 2019).

B. Term Frequency

The Term Frequency method is a way to give weight to the relationship of a word (term) to a document. The Term Frequency (TF) value is obtained based on the number of occurrences of a word in a particular document. For example, if a word appears 3 times in a document, then the TF value is three (Khusna & Agustina, 2018).

C. Naïve Bayes Algorithm

Naive Bayes is a classification algorithm that is carried out using probability and statistical calculations, put forward by British Intelligent Systemientist Thomas Bayes, this method makes predictions based on the previous to find out future opportunities (Hosseinalizadeh et al., 2019). The advantage of using Naive Bayes is that this algorithm only uses a small amount of training data to determine the estimated parameters needed in the classification process (Pattekari & Parveen, 2012).

In this algorithm, each document is represented by attribute pairs a1, a2, a3, ..., n, where a1 is the first word, a2 is the second word and so on. Whereas V is the set of news categories. At the time of classification, the algorithm will look for the highest probability of all document categories tested (Vmap). The Vmap equation is as follows:

$$V_{map} = \text{argmax } P(v_j) \prod_i P(a_i|v_j) \quad (1)$$

V_{map} is the probability value calculated by Naive Bayes for the corresponding target function value. The frequency at which words occur is the basis for the value $P(v_j)$ and $P(a_i|v_j)$. The set of these probability values corresponds to the hypothesis to be studied. Hypotheses are then used to classify new data.

Value $P(v_j)$ calculated at the time of training data, obtained by the following formula:

$$P(v_j) = \frac{|doc\ j|}{|training|} \quad (2)$$

Where |doc j| is the number f document (thesis title) which has category j in training. Meanwhile | training | is the number of documents (thesis title) in the sample used for training. For the word probability a_i for each category $P(a_i|v_j)$, it is calculated at the time of training.

$$P(a_i|v_j) = \frac{|n_i+1|}{|n+kosakata|} \quad (3)$$

Where n_i is the number of occurrences of the word a_i in categorized documents v_j, while n is the number of all words in the document by category v_j and |vocabulary| is the number of words in the training example.

D. Confusion Matrix

Among the mechanisms that can be used to measure the validity of classification results, the ones that are often used are calculating the accuracy, precision and recall values (Caelen, 2017).

Confusion Matrix		True Values	
		True	False
Prediction	True	TP	FP
	False	FN	TN

Figure 2. Confusion Matrix (Mohajon, 2017)

Based on Figure 2 it can be explained that:

a. TP, namely True Positive, is the number of positive data classified correctly by the system.
b. TN, namely True Negative, is the amount of negative data classified correctly by the system.
c. FN, namely False Negative, is the amount of negative data but is classified incorrectly by the system.
d. FP, namely False Positive, is the number of positive data but is classified incorrectly by the system.

Accuracy is a calculation to get the result of the proportion of correct prediction. The calculation of accuracy value is shown in the following equation:

$$Accuracy = \frac{True\ prediction\ (positive\ and\ negative)}{Total\ item} \text{ x } 100\ \% \quad (4)$$

Precision (P) is a measure of the number of documents found to be relevant, the calculation of precision is shown in the following equation:

$$Precision = \frac{\#(Relevant\ item\ retrieved)}{\#(Retrieved\ item)} \quad (5)$$

Meanwhile, recall (R) is a measure of the number of relevant documents that can be recovered, as shown in the following equation:

$$Recall = \frac{\#(Relevant\ item\ retrieved)}{\#(Relevant\ item)} \quad (6)$$

E. K-Fold Cross Validation

K-fold cross validation is a method used to determine the average success rate of a system by looping using random attributes. This method can be used when the amount of data is limited. This test aims to determine the accuracy of the Naive Bayes

method which is applied to the classification of the thesis title when tested with different training data and testing data (Caelen, 2017).

Cross Validation is similar to the repeated random subsampling method, but the sampling is done in such a way that no 2 data tests overlap.

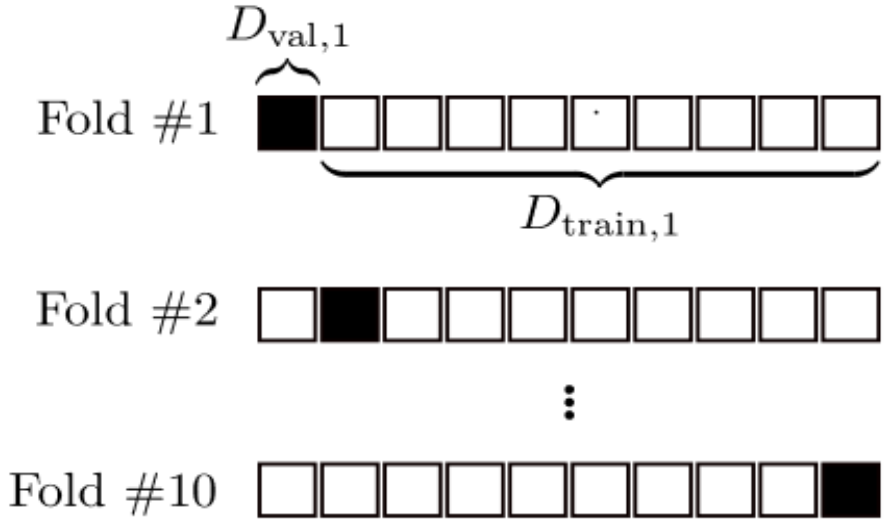

Figure 3. K-Fold Cross Validation

In Figure 3 K-fold Cross Validation it is visualized that the k-fold cross validation process divides the data randomly into k loose subsets, the model is trained in the training set and then applied in the validation set (Moayedi, Osouli, Nguyen, & Rashid, 2019).

RESULT AND DISCUSSION

The data used in this research uses thesis title data for the Universitas Ahmad Dahlan Informatics Engineering study program as many as 1290 thesis title data from 2013 to 2018 which will be grouped or classified into two areas of interest, namely Intelligent Systems (IS/SC) and Software and Data Engineering (SDE/Relata The first stage of the data that has been obtained is divided into two parts, namely training data and testing data, consisting of 500 data used as training data and labeled manually, then used to test 790 data which will be predicted periodically every year, and after the completion of the prediction, the 790 data will be updated into training data.

The test data is then carried out preprocessing with the tokenizing stage which aims to cut the sentence into words, followed by the filtering process which is the stage used to eliminate words that are considered unimportant. Furthermore, the stemming process is carried out aimed at converting each word into a root word. Then the weighting is carried out using term frequency, after the weighting results are obtained, classification is carried out using the Naive Bayes Algorithm.

The classification results are then tested using the Confusion Matrix and K-Fold Cross Validation values by dividing the 1290 thesis title data into 10 parts, this is done using the k = 10 value. The test is used to determine the level of accuracy, precision and recall in this study. Accuracy calculation is shown in table 3.1 Confusion Matrix value for 1290 data below:

Table 2. Confusion Matrix Value

Iteration	True SDE	False SDE	True IS	False IS	Total Data	Accuracy	Precision	Recall
1	83	4	30	12	129	0.875969	0.882352	0.714285
2	58	1	43	27	129	0.782946	0.977272	0.614285
3	86	0	33	10	129	0.922481	1.0	0.767441
4	108	0	15	6	129	0.953488	1.0	0.714285
5	96	0	30	3	129	0.976744	1.0	0.909090
6	94	0	34	1	129	0.992248	1.0	0.971428
7	111	1	15	2	129	0.976744	0.9375	0.882352
8	107	0	22	0	129	1.0	1.0	1.0
9	104	1	24	0	129	0.992248	0.96	1.0
10	86	1	42	0	129	0.992248	0.976744	1.0
Total	933	8	288	61				
Total Data			1290		1290	9.465116	9.76744	8.573166
Average						0.946512	0.973386	0.875317

Using the data calculation on the results of the first iteration obtained from the implementation of Jupyter Notebook using the Scikit-learn library, the accuracy value is obtained from equation (4) as follows:

$$\text{Accuracy} = \frac{(True\ relata + True\ SC)}{Total\ data\ pada\ iterasi\ 1} \text{ x } 100\ \%$$

$$= \frac{83+30}{129} \text{ x } 100\%$$

$$= \frac{113}{129} \text{ x } 100\%$$

$$= 87.5\%$$

So that the average accuracy value at 10 data iterations is as follows:

$$\text{Mean Accuracy} = \frac{\sum(accuracy1 + accuracy2 + \cdots + accuracy10)}{(10)}$$

$$= \frac{0{,}8759+0{,}7829+0{,}9224+0{,}9534+0{,}9767+0{,}9922+0{,}9767+1+0{,}9922+0{,}9922}{(10)}$$

$$= \frac{9{,}465116}{10}$$

$$= 0.9465116$$

The precision value with equation (5) uses the results of the first iteration data which is implemented in the Jupiter notebook using the Scikit-learn library as follows: *Precision* Software and data engineering (Relata)

$$\text{Precision} = \frac{(True\ relata)}{(Total\ data\ yang\ diprediksi\ relata)}$$

$$= \frac{83}{95}$$

$$= 0.87$$

Precision Intelligent System (SC)

$$\text{Precision} = \frac{(True\ sistem\ cerdas)}{(Total\ data\ yang\ diprediksi\ sistem\ cerdas)}$$

$$= \frac{30}{34}$$
$$= 0.88$$

So that the average value of precision in the first iteration is obtained $\frac{0,87+0,88}{2}$ = 0,88, while the average value of precision obtained from 10 iterations is

$$\text{Mean Precision} = \frac{\sum(precesion1+\ precision2+\cdots+precison10)}{(10)}$$
$$= \frac{0,8823+0,9772+1+1+1+1+0,9375+1+0,96+0,9767}{(10)}$$
$$= \frac{9,73386}{10}$$
$$= 0.973386$$

And the recall value with equation (6), using data from the results of the first iteration implemented on the Jupiter notebook, obtained the following calculation results:
Recall Software and data engineering (Relata).

$$\text{Recall} = \frac{(True\ relata)}{(Total\ data\ asli\ relata\ )}$$
$$= \frac{83}{87}$$
$$= 0.95$$

Recall Intelligent System (SC)

$$\text{Recall} = \frac{(True\ sistem\ cerdas)}{(Total\ data\ asli\ sistem\ cerdas\ )}$$
$$= \frac{30}{42}$$
$$= 0.71$$

So that the average recall value in the first iteration is obtained $\frac{0,95+0,71}{2}$ = 0,83. Meanwhile, the average recall value from 10 iterations is as follows:

$$\text{Mean Recall} = \frac{\sum(recall1+\ recall2+\cdots+recall10)}{(10)}$$
$$= \frac{0,7142+0,6142+0,7674+0,7142+0,9090+0,9714+0,8823+1+1+1}{(10)}$$
$$= \frac{8,573166}{10}$$
$$= 0.857317$$

Based on the results of the calculation of the thesis title classification system test using 1290 data with a total of 10 iterations using the Confusion Matrix and K-Fold Cross Validation, the accuracy value is 94.65%, the precision value is 97.33%, and the recall value is 85.73%.

CONCLUSION

Based on the results of the research that has been done, it can be concluded that a thesis title classification system has been created which can provide very useful information to determine the number of students who take thesis titles based on their fields of interest each year. So that it can help the thesis coordinator to classify the thesis title automatically and can be used as material for evaluation of study programs related to the fields of interest that exist in the Department of Informatics Universitas Ahmad Dahlan.

The trend of taking the thesis from 2013 to 2018 in Department of Informatics Universitas Ahmad Dahlan in the field of relative interest. The accuracy performance obtained from the evaluation method uses the K-Fold Validation using the value of k = 10 on 1290 data, resulting in an accuracy value of 94.65%, a precision value of 97.33% and a recall value of 85.73%. The Naive Bayes method in this study produces a good classification level with high accuracy values. So, it is suitable for predicting the class of the thesis title every year.

This research still has some shortcomings, it is hoped that there will be a further development of this research. Suggestions that can be used as a reference for further development, such as the upload process are still done manually using files with the xlsx extension. So, it is hoped that development will be carried out which can retrieve data every semester from the existing thesis title database.

The application of word weighting still uses the Term Frequency (TF) because in its application, it only calculates the occurrence rate of words. So, it is necessary to do further research related to the application of TF-IDF in the thesis title classification system. It needs to be considered for further research into the classification of the title of the thesis of the Universitas Ahmad Dahlan Informatics Engineering study program based on the scientific sub-group.

REFERENCE

Alami, N., Meknassi, M., Ouatik, S. A., & Ennahnahi, N. (2016). Impact of stemming on Arabic text summarization. *4th IEEE International Colloquium on Information Science and Technology (CiSt)*, 338–343.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. *ArXiv*.

Alshalabi, Al., Hamood, H., Tiun, S., Omar, N., & Albared, M. (2013). Experiments on the Use of Features election and Machine Learning Methods in Automatic Malay Text Categorization. *ICEEI*.

Alsubaei, M. A. (2016). Curriculum Development: Teacher Involvement in Curriculum Development. *Journal of Education and Practice*, *7*(9), 106–107.

Caelen, O. (2017). A Bayesian interpretation of the confusion matrix. *Annals of Mathematics and Artificial Intelligence*, *81*(3–4), 429–450. https://doi.org/10.1007/s10472-017-9564-8

DPR-RI, & INDONESIA, P. R. (2012). *UNDANG-UNDANG REPUBLIK INDONESIA NOMOR 12 TAHUN 2012 TENTANG PENDIDIKAN TINGGI. UU RI*.

Dreisbach, C., Koleck, T. A., Bourne, P. E., & Bakken, S. (2019). A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *International Journal of Medical Informatics*, *125*(December 2018), 37–46. https://doi.org/10.1016/j.ijmedinf.2019.02.008

Hosseinalizadeh, M., Kariminejad, N., Chen, W., Pourghasemi, H. R., Alinejad, M., Mohammadian Behbahani, A., & Tiefenbacher, J. P. (2019). Gully headcut susceptibility modeling using functional trees, naïve Bayes tree, and random forest models. *Geoderma*, *342*(October 2018), 1–11. https://doi.org/10.1016/j.geoderma.2019.01.050

Katariya, N. P., & Chaundri, M. S. (2015). Text Preprocessing for Text Mining Using Side

Information, *3*, 3–7.

Khusna, A. N., & Agustina, I. (2018). Implementation of Information Retrieval Using TF-IDF Weighting Method On Detik.Com's Website. *TSSA-IEEE*.

Kurniawati, S., Suryadarma, D., Bima, L., & Yusrina, A. (2018). Education in Indonesia: A white elephant? *Journal of Southeast Asian Economies*, *35*(2), 185–199. https://doi.org/10.1355/ae35-2e

Moayedi, H., Osouli, A., Nguyen, H., & Rashid, A. S. A. (2019). A novel Harris hawks' optimization and k-fold cross-validation predicting slope stability. *Engineering with Computers*, 1–11.

Mohajon, J. (2017). Confusion Matrix for Your Multi-Class Machine Learning Model | by Joydwip Mohajon | Towards Data Science. Retrieved November 21, 2020, from https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826

Mubarok, M. S., Adiwijaya, A., & Aldhi, M. D. (2017). Aspect-based sentiment analysis to review products using Naïve Bayes. *AIP Conference Proceedings*, *1867*(August). https://doi.org/10.1063/1.4994463

Ogundare, O., & Wiggins, N. (2018). Identifying Sub-documents in a Composite Scanned Document Using Naive Bayes, Levenshtein Distance and Domain Driven Knowledge Base. *5th International Conference on Soft Computing and Machine Intelligence, ISCMI 2018*, 84–87. https://doi.org/10.1109/ISCMI.2018.8703245

Pattekari, S. A., & Parveen, A. (2012). Prediction System for Heart Disease Using Naive Bayes. *International Journal of Advanced Computer and Mathematical Sciences*, *3*(3), 290–294.

Pinto, A., Goncalo Oliveira, H., & Oliveira Alves, A. (2016). Comparing the performance of different NLP toolkits in formal and social media text. *In 5th Symposium on Languages, Applications and Technologies (SLATE'16)*.

Suh, J. H., Park, C. H., & Jeon, S. H. (2010). Applying text and data mining techniques to forecasting the trend of petitions filed to e-people. *Expert Systems with Applications 37*, 7255–7268.

Wallace, E., Feng, S., Kandpal, N., Gardner, M., & Singh, S. (2019). niversal adversarial triggers for attacking and analyzing NLP. *ArXiv*.

Xiang, Z., Schwartz, Z., Gerdes Jr, J. H., & Uysal, M. (2015). What can big data and text analytics tell us about hotel guestexperience and satisfaction? *International Journal of Hospitality Management*, *44*, 120–130.

Zampieri, M., Malmasi, S., Nakov, P., Ali, A., Shon, S., Glass, J., & Lee, C. (2018). Language Identification and Morphosyntactic Tagging. *The Secind Vardial Evaluation Campaign*.