Accredited by National Journal Accreditation (ARJUNA) Managed by
Ministry of Research, Technology, and Higher Education, Republic Indonesia with Second Grade (Peringkat 2, Sinta 2)
since year 2017 to 2021 according to the decree No. 10/E/KPT/2019.

Published online on the journal's webpage: **http://jurnal.iaii.or.id**

RESTI JOURNAL
(System Engineering and Information Technology)
Vol. 4 No. 4 (2020) 711-716 | ISSN Electronic Media: 2580-0760

Fake News (Hoax) Identification on Social Media Twitter using Decision Tree C4.5 Method

Brenda Irena[1], Erwin Budi Setiawan[2]
[12]School of Computing, Informatics, Telkom University
[1]brendairena@student.telkomuniversity.ac.id, [2]erwinbudisetiawan@telkomuniversity.ac.id

Abstract

Social media is a means to communicate and exchange information between people, and Twitter is one of them. But the information disseminated is not entirely true, but there is some news that is not in accordance with the truth or often called hoaxes. There have been many cases of spreading hoaxes that cause concern and often harm a particular individual or group. So in this research, the authors build a system to identify hoax news on social media Twitter using the Decision Tree C4.5 classification method to the 50,610 tweet data. What distinguishes this research from some researches before is the existence of several test scenarios, classification only, classification using weighting feature, and also classification using weighting feature and feature selection. The weighting method used is TF-IDF, and the feature selection uses Information Gain. The features used are also generated using n-grams consisting of unigram, bigram, and also trigrams. The final results show that the classification test that uses weighting feature and feature selection produces the best accuracy of 72.91% with a ratio of 90% training data and 10% test data (90:10) and the number of features used is 5000 in unigram features.

Keywords: hoax, twitter, decision tree C4.5, TF-IDF, information gain

1. Introduction

In this increasingly digital era, the ease of communicating and exchanging information is increasingly felt by the presence of social media such as Twitter, Facebook, Instagram, and the others. Social media is a means for exchanging messages, whether it text messages, pictures, audio or video through the internet network. Nowadays, social media has become a daily necessity, active users on social media in Indonesia increased by 8.1% from 2019 to 2020. Twitter became one of the most actively used social media with more than 300 million users in 2020 [1]. Everyone has the freedom to upload a tweet, starting from tweets that contain positive content or even the negative one. In fact, they can also misuse the information or spread information that is not in accordance with the truth (hoaxes) freely.

Hoaxes are an intentional fake news, inflammatory, and inaccurate news. In Indonesia, social media is the highest medium of hoax news distribution, at 87.5% compared to chat applications (67%), websites (28.2%), television/radio (8.7%), newspaper (6.4%) and e-mail (2.6%) [2]. The incident of spreading hoax will continue to unsettle the society, because there will be many parties who are harmed by these occasion such as defamation, fraud, or others.

The number of negative impacts felt cause thoughts to take precautions against the spread of hoax news on social media, one of them is by building a hoax detection system. Research related to hoax detection on Twitter has been done several times before, including the research conducted by Achmad Fauzi et al using the Support Vector Machine method to predict the likelihood of someone spreading hoax news. The test was conducted on tweets related to the 2019 presidential election. Twitter features such as retweet, URL and hoax support features such as provocation and hostility can produce the best accuracy of 78.33% [3].

Research related to hoax detection was also carried out by Mykhailo Granik and Volodymyr Mesyura on a Facebook news upload dataset and the method used is Naïve Bayes Classifier. In this research, each word in a news article is interpreted independently and able to produce an accuracy of 74% in the dataset tested [4].

In research conducted by Errissya Rasywir and Ayu Purwarianti, a combination of feature selection techniques was carried out, in which the best combination was mutual information and information gain. Then classification is done by comparing 3

Accepted by editor: 30-06-2020 | Final Revision: 20-07-2020 | Online Publication : 20-08-2020

methods, SVM, Naïve Bayes and C4.5. The result shows that Naïve Bayes has the best accuracy among the other two methods. But the weakness is, the system is difficult to recognize news with a particular topic so that makes the news wrong classified [5].

In this research, the author identifies hoax news by using the Decision Tree C4.5 classification method because this method is a very strong and well-known classification and prediction method [6]. Decision Tree C4.5 is an extension of ID3 with several improvements such as possible to use continuous data, missing values, and pruning. C4.5 is also able to use attributes with different weights [7]. In addition, this research also performed a test scenario using the Term Frequency Inverse Document Frequency (TF-IDF) weighting feature to look up the value/weight representation of each document in the dataset and also the Information Gain feature selection to select the best features that affect on determining the class of a document in which unigrams, bigrams, and trigrams are the feature extractions used. Tests conducted on Indonesian language tweet data of 50.610 tweets taken from the crawling process based on keywords. The purpose of this research is to implement the method described above and find out the system's performance in identifying hoax news using accuracy calculation with reference to the confusion matrix table. From the system built, the author also finds out what features are influential in identifying hoax news, what features are most often used in spreading hoax news, and what gram is the best in determining hoax news.

2. Research Method

2.1. System Flowchart

In Figure 1, the author represents a flowchart of how the whole system works in this research. The system starts from data collection, continue to labeling, then the unstructured data will be changed to be more structured through preprocessing. Then each word in the dataset will be weighted using TF-IDF weighting and features will be selected using information gain feature selection. The system then divides the data into two parts, training data for learning and test data to test the learning outcomes of the system by using four different split data ratios (90:10, 80:20, 70:30, 60:40). The use of different ratios aims to find out which ratios produce the best performance. The performance of system will be measured using accuracy.

2.2. Data

The dataset used in this research was taken from the crawling process on Twitter using an API that has been provided by Twitter developers. Every one time crawling process will be obtained 100 of the latest data tweet. The data taken is as many as 50.610 data tweets using keywords taken from topics that are currently being discussed or being trending topics on Twitter and also estimated to contain hoaxes in the span of December 2019 - March 2020. Table 1 shows the keywords used in data retrieval.

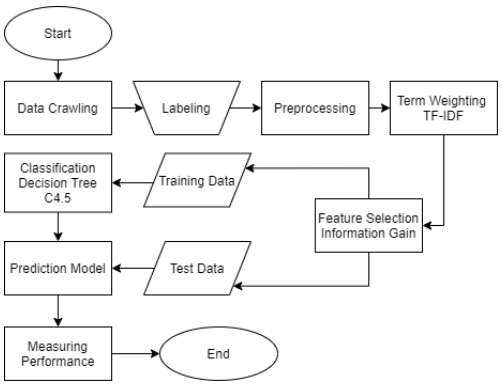

Figure 1. Flowchart

Table 1. Keywords

Keyword	Amount
#BanjirJakarta2020	5.468
#PecatAniesBaswedan	5.403
#gubernurterbodoh	6.492
#KebakaranHutan	8.627
#VirusCorona	8.719
Jokowi	15.901
Total	**50.610**

2.3. Labeling

Labeling is the process of determining the class of tweet data whether included in hoax news or not that is done manually by the author together with a final task group consisting of 3 people. In Table 2, it is explained about the features on Twitter that serve as a reference for the author in labeling.

Table 2. Feature that Identify Hoax.

Feature	Description
Tweet	whether the tweet contains elements of a hoax or not such as panic, provocation, hatred
Username	whether the username is set with the real name or pseudonym, contains numbers / symbols, contains elements of hatred or not
Location	location included in the tweet upload
URL	tweet include URL as a source
Following: Followers	the number of following is more than followers
Verified Account	whether the account has been verified or not

Data that has been collected is then labeled with a reference based on the Table 2. Comparison of data with hoax and non-hoax classes based on the results of labeling in the existing dataset is shown in Table 3.

Table 3. Labeling Result

Label	Amount	Percentage
HOAX	25.022	49.4%
NON-HOAX	25.588	50.6%

2.4. Preprocessing

In processing data, it takes several processes so that the raw data that has been obtained previously can be processed and managed easily. Preprocessing is a process of changing the form of unstructured text data into structured forms according to their needs [8]. In the system built, there are four stages of preprocessing, first case folding to eliminate characters other than a-z, then normalization to correct abbreviated words or words that are not clear to be match with those set in the normalization dictionary, then filtering to discard words that are considered not important, and the last one is stemming to change each word with a root word. Table 4 shows an example of the preprocessing flow that the system does with tweet data.

Table 4. Example of Preprocessing Flow

Process	Sentence
Initial	"@helmifelis Pilpresnya 2024 tp pembusukkan ke anies udh mulai.\n #LanjutkanRevitalisasiMonas"
Case Folding	pilpresnya tp pembusukkan ke anies udh mulai
Normaliza-tion	pilpresnya tapi pembusukkan ke anies udah mulai
Filtering	pilpres pembusukkan anies
Stemming	pilpres busuk anies

2.5. N-gram

N-gram is a substring along the n characters of a string. N-gram is a method used for word or character generation [9]. The n-gram model is formed based on the n-gram frequency that appears in the document. The document will be read word by word and will be made n-gram of those words. Each n-gram raised will be recorded in the table. Table 5 is an example of the n-grams formation in the sentence "pilpres busuk anies".

Table 5. Example of N-gram Formation

N	N-gram produced
1	(pilpres), (busuk), (anies)
2	(pilpres busuk), (busuk anies)
3	(pilpres busuk anies)

Data from the preprocessing step will be generated by n-gram into unigram, bigram, trigram, and a combination of the three. The use of n-gram also aims to determine which is the best gram in determining hoax news. The number of features generated from n-gram breakdown in the dataset of 50.610 data tweets is shown in Table 6.

Table 6. Amount of N-gram Features

N-gram	Amount
Unigram	24.426
Bigram	156.491
Trigram	168.716
Unigram + Bigram	180.917
Bigram + Trigram	325.207
Unigram + Bigram + Trigram	349.633

2.6. TF-IDF

TF-IDF or Term Frequency-Inverse Document Frequency is one of the weighting techniques for a term or word in a document by calculating the weighting of the most commonly used terms. If TF weight terms in a document, IDF reduces the weight of a term if the appearance is widely spread throughout the document [10], with the equation:

$$W_{dt} = tf_{dt} \quad \text{x} \quad idf_t \tag{1}$$

where Wdt is the weight of term t to document d, $tfdt$ is the number of occurrences of term t in document d and $idft$ is the value of Inverse Document Frequency. The value $idft$ is obtained from:

$$idf_t = \log\left(\frac{D}{df}\right) \tag{2}$$

where D is the number of documents in the collection and df is the number of documents containing the term t.

2.7. Information Gain

Information Gain is used to select features that have the most significant information to the class in the data. [11]. To get the Information Gain value, entropy calculation is needed before the data is separated and after the data is separated, with the equation:

$$Entropy(S) = \sum_{i=1}^{k} pi \; log_2 \, pi \tag{3}$$

where Pi is the probability of S data in class i. K is the number of classes for S variable.

$$Entropy(S) = \sum_{i=1}^{v} \left(\frac{Sv}{S} \times \; Entropy(Sv)\right) \tag{4}$$

where v is all possible values of attribute A, Sv is a subset of S where attribute A is v. The information gain value is calculated from the following equation:

$$Gain\ (S, A) = Entropy\ (S) - Entropy\ (S, A) \tag{5}$$

where Gain (S, A) is the information gain value. Entropy (S) is the value of entropy before separation. Entropy (S, A) is the value of entropy after separation. The value of information gain indicates how much influence an attribute has on the data classification.

2.8. Decision Tree C4.5

Decision Tree is one type of supervised classification because the learning process on documents has a class label. A decision tree is a flow chart like a tree structure, where each internal node shows a test on an attribute, each branch shows the results of the test, and the leaf node shows the classes or class distribution [12]. Decision Tree used in this research is C4.5 because the characteristics of the data used are continuous.

C4.5 algorithm is an extension of the Decision Tree ID3 (Iterative Dychotomizer version 3) algorithm. The advantage of the C4.5 algorithm compared to ID3 is that

the C4.5 algorithm can handles both categorical and numeric value, can resolve missing data and pruning data [13]. In general, the process carried out by the C4.5 algorithm in building a decision tree is to choose attributes as the root based on the largest gain ratio, then determine the attributes that will be an internal nodes for each branch of the parent node, and make a decision node when attribute selection cannot be used anymore. Figure 2 Shows the Decision Tree C4.5 Calculation Flow.

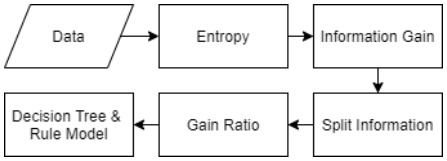

Figure 2. C4.5 Flow

Equations used in this algorithm include [14]:

a. Entropy

Entropy is used as a parameter to measure the heterogeneity (diversity) of a data sample. The smaller the value of Entropy, the better it is to use in extracting a class. The entropy value is defined by the formula:

$$Entropy(S) = \sum_{i=1}^{n} - pi\ log_2\ pi \qquad (6)$$

where S is the training data set, n is the number of partitions in S, and pi is the proportion of the sample in class i.

b. Information Gain

The value obtained from the entropy calculation is still not original but the measurement of the attributes effectiveness in classifying the training data can be determined by the information that has been obtained, by the equation:

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} - \frac{si}{s} \times\ entropy(Si) \qquad (7)$$

where S is the training data set, A is the attribute, n is the number of partitions in attribute A, and Si is the i-th number of partitions.

c. Gain Ratio

Gain Ratio is a modification of information gain to reduce the bias attribute that has many branches. The equation is written as:

$$Gain\ Ratio = \frac{Gain(S,A)}{Split\ Information(S,A)} \qquad (8)$$

where split information is defined by the formula:

$$Split\ Information(S,A) = -- \sum_{i=1}^{C} \frac{si}{s} log_2 \frac{si}{s} \qquad (9)$$

where S1 through Sc is a subset created from splitting S using attribute A with variant C.

2.9. Measuring Performance

Measuring Performance is a step of analysis and evaluation of the system performance built. Performance is measured by the value of accuracy. Confusion matrix is a method used to calculate accuracy in the concept of data mining. Confusion matrix shown in Table 5.

Table 7. Confusion Matrix

	Predicted Class Yes	Predicted Class No
Actual Class Yes	True Positive (TP)	False Negative (FN)
Actual Class No	False Positive (FP)	True Negative (TN)

True Positive (TP) is when the actual and predicted class both is hoax. False Positive (FP) is when the actual class is not-hoax but the predicted class is hoax. True Negative (TN) is when the actual and predicted class both is not-hoax. False Negative (FN) is when the actual class is hoax but the predicted class is not-hoax. From the confusion matrix table, the accuracy value can be calculated. Accuracy is the level of closeness between the predicted value and the actual value. Accuracy is used to evaluate the number of prediction classes that correspond to the actual class [15]. The following is the equation of accuracy:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \qquad (10)$$

3. Result and Discussion

This section explain the test results of the system built. Testing in this research aims to determine the performance of system in classifying hoax news that is done with several test scenarios.

a. Testing of system performance using the Decision Tree C4.5 classification method.
b. Testing of system performance using the Decision Tree C4.5 classification method with the TF-IDF weighting feature.
c. Testing of system performance using the Decision Tree C4.5 classification method with TF-IDF weighting feature and information gain feature selection, where the feature selection is based on ranking on the limit of the number of features n=15,000, n=10,000, n=5,000, and n=3,000.

All three test scenarios are performed using all n-gram features (unigram, bigram, trigram, unigram + bigram, bigram + trigram, and unigram + bigram + trigram). In addition, each test scenario was also carried out 5 times with a ratio of training data and test data of 90:10, 80:20, 70:30, and 60:40.

3.1. Classification Test Results

Table 8 shows the results of system testing using the Decision Tree C4.5 classification method only.

From Table 8, the best accuracy is obtained with a percentage of 10% test data on the unigram + bigram feature that is equal to 61.13%. The use and

incorporation of n-gram is capable of producing good accuracy.

Table 8. Classification Result

N-gram	90:10	80:20	70:30	60:40
Unigram	60.93%	60.59%	59.70%	61.08%
Bigram	57.97%	58.71%	57.14%	57.85%
Trigram	56.38%	55.31%	56.12%	55.80%
Unigram+Bigram	61.13%	61.06%	60.46%	60.63%
Bigram+Trigram	60.81%	59.45%	58.98%	57.04%
Unigram+Bigram+ Trigram	57.73%	60.94%	59.83%	60.30%

3.2. Classification Test Results using Term Weighting

Table 9 shows the results of system testing using the Decision Tree C4.5 classification method with TF-IDF weighting feature.

Table 9. Classification using Weighting Result

N-gram	90:10	80:20	70:30	60:40
Unigram	62.96%	62.07%	62.02%	62.22%
Bigram	61.73%	62.60%	62.05%	61.78%
Trigram	62.47%	61.58%	62.12%	61.52%
Unigram+Bigram	62.79%	63.04%	62.44%	62.11%
Bigram+Trigram	62.93%	62.41%	61.94%	61.72%
Unigram+Bigram+ Trigram	63.05%	63.01%	62.95%	62.18%

From Table 9, the best test results obtained with 10% test data on the unigram + bigram + trigram feature that is equal to 63.05%. TF-IDF weighting also has an important influence on classification, because the TF-IDF weighting feature allows the system to not give too much weight to features that appear a lot in tweet documents so that the system will be more accurate in determining the topic discussed in a tweet, and therefore the accuracy produced in scenario 2 (classification using weighting feature) is better than scenario 1 (classification only).

3.3. Classification Test Result using Term Weighting and Feature Selection

The feature selection testing technique is applied by selecting a number of features based on ranking limits for a number of specific features. In this test, the feature limit (n) selected is 15,000, 10,000, 5,000, and 3,000. The results of testing the feature limits shown in Table 10.

Table 10. Feature Limit Test Result

Amount of Feature	Accuracy
3.000	66.28%
5.000	68.40%
10.000	67.70%
15.000	65.79%

From Table 10, the best results of feature limit testing for information gain are obtained at 5,000 feature numbers with an accuracy of 68.40%. It proves that many features are irrelevant, and only 5000 features with the highest gain values that have the most important influence on the classification process. So, in the next test, 5,000 features will be used in determining the classification of hoax news.

Table 11 shows the results of system testing using the Decision Tree C4.5 classification method with TF-IDF weighting and information gain feature selection using 5000 features.

Table 11. Classification Test Result with Weighting and Feature Selection

N-gram	90:10	80:20	70:30	60:40
Unigram	72.91%	71.92%	71.50%	71.82%
Bigram	69.52%	68.60%	69.60%	68.86%
Trigram	69.38%	69.03%	67.71%	67.55%
Unigram+Bigram	70.94%	71.85%	69.69%	70.58%
Bigram+Trigram	69.46%	69.34%	69.14%	68.97%
Unigram+Bigram+ Trigram	71.60%	69.95%	70.03%	69.62%

From Table 11, the best accuracy results obtained with the percentage of 10% test data on the unigram feature that is equal to 72.91%.

In Figure 3, the author represents a chart which compares the test results of the three scenarios above.

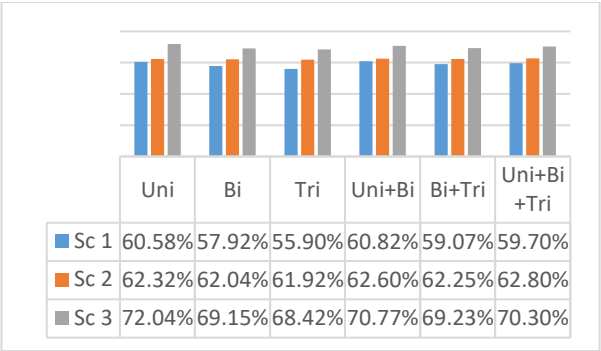

Figure 3. Result of Three Scenarios Comparison

When compared with scenario 1 (classification only) and scenario 2 (classification using weighting feature), this scenario (classification using weighting feature and feature selection) is able to produce the highest accuracy. From the average of all n-grams, scenario 1 produces an accuracy of 59.00%, scenario 2 is 62.32% and scenario 3 is 69.98%. The difference in the value of accuracy produced between scenario 1 and scenario 3 is quite significant at 10.98%. This is because the feature selection in scenario 3 is able to reduce irrelevant features and choose features that have more important information in identifying or classifying hoax news so as to improve the performance of predictive models.

From the system built, the most common features found in news hoaxes and the most widely used by Twitter users in creating or spreading hoax news are shown in Table 12. These features are based on the highest frequency of occurrence in the hoax news dataset.

The influential features in classifying hoax news are the relevant features obtained from the highest information gain (feature selection) value. The 10 most influential features are shown in Table 13.

Table 12. Most Features used in Hoaxes

Unigram	Bigram	Trigram
jokowi	presiden jokowi	jokowi puji anies
corona	anies baswedan	darurat nasional virus
anies	darurat nasional	sebar virus corona
banjir	bakar hutan	musuh jokowi serang
sebar	cari panggung	dukung lawan jokowi
panik	banjir jakarta	gubernur dki jakarta
buzzer	serang jokowi	gubernur rasa presiden
kadrun	lawan politik	bakar hutan australia
jakarta	sebar corona	bahaya virus corona
darurat	bikin panik	banjir di jakarta

Table 13. Most Influential Features

Unigram	Bigram	Trigram
luhut	bakar hutan	tunggu resmi pemkot
bakar	terima kasih	darurat nasional virus
hutan	resmi pemkot	anies baswedan nikmat
australia	darurat nasional	kadiv humas polri
kadrun	presiden jokowi	surati who ri
bacot	baswedan nikmat	bakar hutan australia
gabener	kadiv humas	kasih informasi terus
anies	wan aibon	stop hoax indonesia
goblok	info jual	jual hand sanitizer
becus	virus corona	forum anti fitnah

4. Conclusion

Based on the test results and analysis discussion that has been presented, it can be concluded that classification Decision Tree C4.5 coupled with the use of weighting feature and also feature selection of 5000 features with the highest gain value in the classification process resulting an increase in accuracy value of 10.98% by using various n-gram features. The use of TF-IDF weighting in the classification process can also improve the performance of the system. There are also some n-grams that do not affect the results, but some of the combinations are quite influential such as unigram + bigram and also unigram + bigram + trigram which results in quite high accuracy values in all test scenarios. But if on average of all tests, unigram is the best gram followed by unigram + bigram.

Suggestions for further research, it is highly recommended to maximize the work of labeling, and also preprocessing, especially in making stopwords and normalization dictionary in order to produce relevant features to optimize the value of accuracy. In addition, it is also recommended to test more various of feature limits for information gain to find the most optimal feature limits.

References

[1] Data Reportal, 2020. *Digital 2020: Indonesia*. [Online] (Updated 18 Feb 2020).
Available at: https://datareportal.com/reports/digital-2020-indonesia [Accessed 2 June 2020]

[2] Mastel, 2019. *Hasil Survey Wabah HOAX Nasional 2019*. [Online] (Updated 10 Apr 2019)
Available at: https://mastel.id/hasil-survey-wabah-hoax-nasional-2019. [Accessed 20 September 2019]

[3] Fauzi, A., Setiawan, E.B., Baizal, Z.K.A., 2018. Hoax News Detection on Twitter using Term Frequency Inverse Document Frequency and Support Vector Machine Method. *The 2nd International Conference on Data Information Science*. Bandung, Indonesia 15-16 Nov 2018. IOP Publishing.

[4] Granik, M., Mesyura, V., 2017. Fake News Detection using Naïve Bayes Classifier. *IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*. Kiev, Ukraine, 29 May-2 Jun 2017, IEEE.

[5] Rasywir, E., Purwarianti, A., 2015. Jurnal Cybermatika. *Eksperimen pada Sistem Klasifikasi Berita Hoax Berbahasa Indonesia Berbasis Pembelajaran Mesin*, 3(2), pp.1-8.

[6] Quinlan, J.R., 2014. *C4.5: Programs for Machine Learning*. California: Morgan Kaufmann Publishers.

[7] Hssina, B., dkk., 2014. International Journal of Advanced Computer Science and Applications (IJACSA). *A comparative study of decision tree ID3 and C4.5*. 4(2), pp.11-19.

[8] Siregar, Z.U., Siregar, R.R., Arianto, R., 2019. Jurnal Kilat. *Klasifikasi Sentiment Analysis pada Komentar Peserta Diklat menggunakan Metode K-Nearest Neighbor*, 8 (1), pp.81-92.

[9] Chandra, D.N., Indrawan, G., Sukajaya, I.N., 2016. Jurnal Ilmiah Teknologi dan Informasi ASIA. *Klasifikasi Berita Lokal Radar Malang Menggunakan Metode Naïve Bayes Dengan FItur N-gram*. 10 (1), pp.11-19.

[10] Chelliah, C.D., Gowri, M., Subramanian, B., S. M. A. Kalaiarasi, and Ramaraj, N., 2010. Journal of Engineering Science and Technology. *A novel term weighting scheme MIDF for text categorization*. 5 (1), pp.94-107.

[11] Maulida, I., Suyatno, A., Hatta, H.R., 2016. Jurnal SIFO Mikroskil. *Seleksi Fitur Pada Dokumen Abstrak Teks Bahasa Indonesia Menggunakan Metode Information Gain*. 17(2), pp.249-258.

[12] Sunjana, 2010. Jurnal Fakultas Hukum UII. *Aplikasi Mining Data Mahasiswa Dengan Metode Klasifikasi*, pp.24-29.

[13] Singh, S., Giri, M., 2014. International Journal of Advanced Information Science and Technology (IJAIST). *Comparative Study Id3, Cart and C4.5 Decision Tree Algorithm: A Survey*. 3(1), pp.47-52.

[14] Han, J., Kamber, M., Pei, J., 2012. *Data Mining Concepts and Techniques*. 3rd ed, San Fransisco: Morgan Kauffman Publishers.

[15] Galdi, P., Tagliaferri, R., 2018. Data mining: Accuracy and error measures for classification and prediction. *Encyclopedia of Bioinformatics and Computational Biology*, pp.431-436.