OPEN ACCESS
Indo-JC Indonesia Journal of Computing
ISSN 2460-9056
socj.telkomuniversity.ac.id/indojc

IND. JOURNAL ON COMPUTING
VOL. 5, ISSUE. 2, SEPTEMBER 2020. PP. 107-118
DOI:10.21108/INDOJC.2020.5.2.499

The Effect of Information Gain Feature Selection for Hoax Identification in Twitter Using Classification Method Support Vector Machine

Isep Mumu Mubaroq[#1], Erwin Budi Setiawan[#2]

 # School of Computing, Telkom University
Bandung, West Java, Indonesia

[1]isepmumumubaroq@students.telkomuniversity.ac.id
[2]erwinbudisetiawan@telkomuniversity.ac.id (Corresponding Author)

Abstract

Nowadays social media twitter is popular media for news dissemination. News has elements that can be distinguished types of news, such as hoax that has elements of panic, worry, and anxiety that can have a significant impact in various fields of social, economic, educational, and political. Hoax prevention efforts need as possible before news viral, by to be developed method with functions to identify and hoax analyze. in this research we have proposed an approach Machine Learning with method Support Vector Machine (SVM) supported by feature selection Information Gain (IG) added Term Frequency–Inverse Document Frequency (TF-IDF) for word weighting system performance is very optimal in increasing accuracy by 37,51%, with accuracy reaching 96.55%.

Keywords: Hoax, Information Gain, Non-Hoax, Support Vector Machine, TF-IDF

Abstrak

Saat ini media sosial twitter merupakan media yang popular untuk penyebaran berita. Berita mempunyai unsur-unsur agar dapat dibedakan jenis beritanya, seperti *hoax* yang mempunyai unsur kepanikan, kecemasan dan kegelisahan yang dapat memberikan dampak signifikan diberbagai bidang sosial, ekonomi, pendidikan dan politik. Upaya pencegahan *hoax* perlu dilakukan agar berita tersebut tidak menjadi viral, yaitu dengan cara dikembangkan suatu metode yang berfungsi mengidentifikasi dan menganalisis *hoax*. Dalam penelian ini diusulkan suatu pendekatan *Machine Learning* dengan metode *Support Vector Machine (SVM)* untuk proses mengidentifikasi berita *hoax*. Penggunaan *Support Vector Machine (SVM)* didukung oleh fitur seleksi *Information Gain (IG)* ditambah *Term Frequency–Inverse Document Frequency (TF-IDF)* untuk pembobotan kata menghasilkan performansi sistem yang sangat optimal dalam peningkatan akurasi sebesar 37.51%, dengan akurasi mencapai 96.55%.

Kata Kunci: *Hoax, Information Gain, Non-Hoax, Support Vector Machine, TF-IDF.*

I. INTRODUCTION

Social media twitter is the media used to deliver messages and even irresponsible people also use twitter to spread hoax [1]. This tweet has a significant impact in various fields such as social, economic, education, and politic that cause anxiety in society [2]. Hoax is needed to do preprocessing data by crawling on twitter. Prevention of hoax needs to be developed with a method that functions to identify, analyze, and assess it as

RECEIVED ON JAN 2020. ACCEPTED ON AUG 2020

early as possible before the tweet becomes viral, thereby minimizing negative impacts, in addition to maintaining the creativity of social media as a place where hoax is disseminated. Hoax identification uses the machine learning approach, which is the classification method. The classification method used is Support Vector Machine which is a supervised learning method by contradicting a linear data classification process [3]. The use of the Support Vector Machine method has the advantage of being able to obtain hyperplane optimally by producing maximum margins between different classes [3] and being able to classify textual data that has a high dimension by giving the best result with several tests [4].

However, the Support Vector Machine has disadvantages in the selection of parameters of features that affect the results of the accuracy, therefore feature selection is needed to reduce the dimensions of the dataset so that time in the classification process will be reduced, by reducing features those are not relevant to obtain high classification [5]. Compare using method feature selection as mutual information, chi-square, and Information Gain for choosing feature selection with the best dimension from set feature [6]. Feature selection Information Gain gives the highest accuracy value [6]. The classification method has the highest accuracy and the optimal value is Support Vector Machine with feature selection Information Gain [7]. There is a selection of extraction features used, namely TF-IDF. TF is the number of words, while IDF shows the level used in the document and TF-IDF there is the formula for weight document is known to be efficient, easy and the result are more accurate. TF-IDF weighting is often used as a primary tool and ranking the relevance of documents [8].

In this research, determine the effect of feature selection Information Gain on identification hoax from twitter using method Support Vector Machine. The choosing of selected features can help to eliminate useless features to improve accuracy in the classification process. The Support Vector Machine method was chosen based on previous research which results in good accuracy in identification hoax on Twitter, as many 51.421 Indonesian-language tweets were taken by crawling periodically from August 2019 to December 2019, tweets were taken from trending that contained hoaxes. The data is labelled in January 2020 - March 2020 using excel with values 1 and 0, 1 for hoax value, and 0 for non-hoax data will be divided into training data and test data. Furthermore, the data used for the performance of the methods applied in the confusion matrix are accuracy, precision, recall, and f1-score. Training data and test data are run five times to get the average accuracy results and what will be taken is the highest accuracy value.

II. Literature Review

A. Hoax Identification

Research conducted by Achmad Fauzi related to Hoax Detection on Twitter with the Term Frequency Inverse Document Frequency Method and Support Vector Machine said that the accuracy results that showed the results of the Support Vector Machine method can determine the best margins and reach the maximum point, Support Vector Machine is also considered suitable to be used as a hoax detection method that produces an accuracy value of 78.33% and has features that are divided into 2 classes on the hoax that is provocation and hostility [8].

Based on Dina Maulina's research that Article Hoax Classification Using Support Vector Machine Linear with Term Frequency Inverse Document Frequency said that hoax intentionally 61.60% of inciting article hoax, 59% of the accurate article, and 14% of forecast article hoax. The research using Support Vector Machine with Term Frequency Inverse Document Frequency has value accuracy 95,833%. This method related used text classification hoax article using TF-IDF [9].

According to research by Ni Made Gita D.P related Identification Tweet Cyberbullying on Twitter Application using Method Support Vector Machine and Information Gain said that identification based best result on testing obtained from all testing parameters sequential training Support Vector Machine with accuracy 75% and based testing on threshold feature selection Information Gain. Because the Information Gain feature selection has a high value for classification representation, and this selection feature has the highest accuracy results compared to all the features in this research [7].

Depend of researching by Tansa Trisna about Analysis and Detection of Hoax Contains in Indonesian based on Machine Learning explains that the analysis process begins tokenizing, case folding, normalization, filtering, stopwords, stemming, weighting using TF-IDF and unigram and bigram features for combined into hoax and non-hoax classification text which has the highest accuracy value of 76.47% [10].

Previous study literature [7], [9], [8], [10], conclusions from previous study literature using the Support Vector Machine method have more value than other methods on paper [9] has the disadvantage that is not has a selection feature, so the verification results are less than optimal [7]. Then from the shortage of this paper will be added to the Information Gain selection feature on further research.

B. Preprocessing

The step of preprocessing is the process of raw data before to other processes by eliminating data that is not appropriate or changing the data into a younger form processed by the system. Preprocessing is very important especially the sentiment analysis on Twitter which contains mostly ambiguous or unstructured words or sentences and has a large noise [11].

C. TF-IDF

The TF-IDF method is a feature extraction method for calculating weights from documents that contain terms that often appear. Weight values indicate how important a term (term) is in a document [12]. There are two parameters in determining the word weight or *Wi* value in the TF-IDF method, namely the Term Frequency (TF) value and the Inverse Document Frequency (IDF) value.

TF is the frequency of occurrence of the term *t* in the document d_i, where the value of $f_t(x)$ is 1 if $x = t$ and 0 if $x \neq t$. To find the TF value it can be calculated by the formula equation (1), as follows [8] :

$$TF_{t,d} = \sum_{x \in d} f_t(x) \quad (1)$$

IDF is the number of documents containing the term *t.* Assume the total document in a collection is *D* and the documents included in collection *D* are assumed to be in the document d_i. IDF can be calculated from the number of documents $|D|$ divided by the number of documents in which contain the term t_i or called *df (t)*. To find the value of IDF can be formulated with equation (2) [8].

$$IDF_t = log\left(\frac{|D|}{df_t}\right) \quad (2)$$

Wi weight value is the product of TF value *(t, d)* with IDF value *(t)*. So the value of *Wi* weights can be calculated with the following Equation (3) [8].

$$W_i = TF_{t,d} \times IDF_t \quad (3)$$

D. N-Gram

N-Gram is a probabilistic model that is designed and developed to predict the next item and sequence of items. The items consist of letters, characters, and words according to the application will be made. The N-gram is racing against the next word in a certain order. N-gram is a container of a collection of words that has a length of n words [13]. N-Gram consists of substrings of length n characters from a string, or can be called N-grams is a piece of n characters from a string in continuity from the beginning of the word to the end of the word from the document [13]. Uni-gram, bi-gram, and tri-gram, the unique N-gram pattern of a language will appear more prominent [13] The N-gram method is also divided into several types. An example taken from the dataset is "jakarta banjir diawal tahun" can be seen in Table 1.

TABLE 1 EXAMPLE NGRAM

Data	Jakarta banjir diawal tahun
Uni-Gram	|jakarta| |banjir| |diawal| |tahun|
Bi-Gram	|jakarta banjir| |diawal tahun|
Tri-Gram	|jakarta banjir diawal| |banjir diawal tahun|

E. Information Gain

Information Gain is used to calculating the effect of a feature on class similarity in a sentence. To get the Information Gain value entropy calculation is needed before the data is separated and after the data is separated [14]. A term is measured by counting the number of bits of information taken from the presence or absence of a term in a document. Information Gain can be calculated using equation (4) [15].

$$Entropy(S) = \sum_{j=1}^{k} - p_j \, log_2 \, p_j \quad (4)$$

Where *S* is a dataset, *k* is the number of *S*, p_j represents the number of samples for class *j*. The log function is taken based on 2 because the information is coded based on bits. Then the entropy value is calculated after the separator in equation **(2.5)** [5] [16].

$$Entropy(S, A) = \sum_{i=1}^{v} (\frac{|S_v|}{|S|} \times Entropy(S_v)) \quad (5)$$

With the value *v*, that is all values contained in attribute *A*, *Sv* is a subset of *S* where attribute *A* is valued at *v*. The information gain value is calculated from equation **(2.6)** [14]**.**

$$Gain(S, A) = Entropy(S) - Entropy(S, A)) \quad (6)$$

With the value of Gain *(S, A)* is the value of Information Gain. Entropy *(S)* where the value of the entropy is done before separation. Entropy *(S, A)* is the value of entropy performed after separation. The amount of Information Gain shows a large influence on an attribute on the classification of data [14].

F. Support Vector Machine

The next process is to classify input data that has been patterned and processed into a vector or binary using the Support Vector Machine method that can find the best hyperplane that functions as a separator of two classes in the input space. The pattern is a member of two classes +1 and -1 as an alternative to the dividing line. Margin is the distance between the hyperplane and the closest pattern of each class. The closest pattern is called the Support Vector Machine [9].

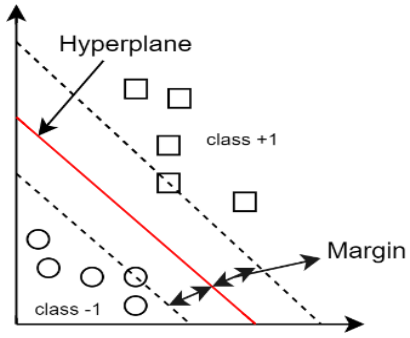


Figure 1 Best Hyperplane

Figure 1 is the best hyperplane search. The red line located in the middle is the hyperplane, then the dashed line that is above the hyperplane line is the margin where this line is the distance from the hyperplane in the nearest data [17].

G. Confusion Matrix

The next process is the classification process uses the Support Vector Machine method, the next steps are needed, namely evaluation. This evaluation stage uses the confusion matrix to calculate the value of accuracy, precision, and recall. A confusion matrix is a tool for conducting analysis that is usually used in Supervised Learning which is used to see the test results of a predicted model [18]. Here is a table from the confusion matrix:

TABLE 2 CONFUSION MATRIX

	Actual positive	**Actual negative**
Prediction Positive	TP	FN
Prediction Negative	FP	TN

Table 2 Confusion Matrix can be explained as True Positive (TP) the tweet is prediction positive hoax and actual the tweet true hoax. True Negative (TN) the tweet is prediction negative non-hoax and actual the tweet true non-hoax. False Positive (FP) the tweet is prediction positive hoax and actual the tweet false non-hoax, False Negative (FN) the tweet is prediction negative non-hoax, and actual the tweet true hoax. From the explain can be calculated the value of accuracy, precision, recall, and f1 score. Next to the formula equation.

$$Precision = \frac{TP}{TP+FP} \times 100 \quad (7)$$

$$Recall = \frac{TP}{TP+FN} \times 100 \quad (8)$$

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \times 100 \quad (9)$$

$$F1 - score = \frac{2 \times precision \times recall}{precision + recall} \quad (10)$$

Precision is used to measure positive patterns that are correctly predicted from the total predicted patterns in positive classes [19] (7), Recall is used to measure positive patterns that are correctly classified [19]. Next in the equation formula (8). To find the Accuracy value, the formula equation (9) will be used. F1 Score is a comparison of weighted average precision and recall, using the formula (10).

III. RESEARCH METHOD

A. Performance of System Support Vector Machine with Information Gain

The process carried out in the development of the system is data crawling carried out on social media twitter, after the data collected labelling process is carried out, the next stage is the preprocessing process where this process has several stages namely case folding, cleaning, normalization, and stopword removal. The next process is weighted using the TF-IDF of each word in the tweet data with N-gram supporters. Furthermore, to produce an optimal accuracy value, the selection feature is added using Information Gain. After all the weighting and selection features are given, the classification process is carried out using Support Vector Machine, which will produce value accuracy and value prediction. The performance of the build is shown in Figure 2.Fi

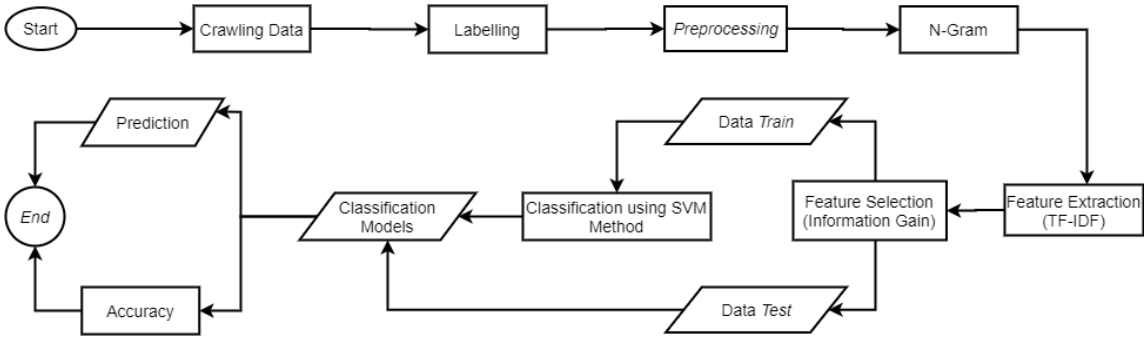


Figure 2 Performance System

B. Crawling

Crawling is a step to collect data or downloading data from the database. The data collected from the server on twitter in the form of user and twitter attribute [20]. Process crawling takes data large or small data from the web can be saved and found by using keyword [21]. This research final task collects data by crawling on API Twitter which was develop by Jaka Eka Sembodo et all using keyword, which can take maximal 200 tweets by once crawling [22].

C. Preprocessing

In preprocessing is the initial stage of preparing data for use to the next step. In this preprocessing reduces attributes that are not useful or relevant for the improvement of the classification process. The process of eliminating data that is not appropriate will change the form of unstructured data into structured data and can be used for the next process. The following of the preprocessing step is used in the hoax identification research. Case folding is a process that converts lowercase capital letters. Cleaning is a process carried out to eliminate characters, numbers, URLs, # hashtags. Normalization is the process of changing all abbreviated or incompatible words in the language dictionary of the document to be adjusted according to the writing rules and dictionary. Stopword Removal is the process of removing words that appear frequently and have no effect on the dataset.

TABLE 3 EXAMPLE OF PREPROCESSING

Preprocessing	Input	Output
Case Folding	Jangan nyampek Anies Baswedan menikmati hsl proyek d Monas yg penuh kejanggalannya	jangn nyampek anies baswedan menikmati hsl proyek d monas yg penuh kejanggalannya
Cleaning	jangn nyampek anies baswedan menikmati hsl proyek d monas yg penuh kejanggalannya	jangn nyampek anies baswedan menikmati proyek monas penuh kejanggalannya
Normalisasi	jangn nyampek anies baswedan menikmati proyek monas penuh kejanggalannya	jangan sampai anies baswedan menikmati proyek hasil monas penuh kejanggalanannya
Stopword Removal	jangan sampai anies baswedan menikmati proyek hasil monas penuh kejanggalanannya	anies baswedan nikmat proyek monas penuh janggal

D. Feature Extraction

Weighting uses TF-IDF, this weighting is used to gauge the importance of words in a document. TF is the frequency of occurrence of a word in one document, while IDF is a measure of the ability of words to make distinctions in categories. In this research, weighting uses uni-gram, bi-gram, tri-gram, then using weighting combinations such as uni-gram and bi-gram, bi-gram and tri-gram, and the third-weight combination between uni-gram, bi-gram and tri-gram.

E. Feature Selection

The step that needs to be considered in this selection feature is to measure a class on various data using entropy, by calculating the process of Information Gain values by the process of calculating the value of the gain and entropy, which are the determinants of the attribute to be discarded and the attributes to be used.

IV. RESULTS AND DISCUSSION

A. Dataset

In this research the data collection of the crawling of the keyboards were taken from trending hashtag which is estimated to contain hoax elements from August 2019 - December 2019 as many as 51,421 collected. After the data has been collected, the labelling process shows the amount of 25,329 data according to the class with a hoax label with a value of 1, the number of 26,029 non-hoax labels with a value of 0. Table 3 is an example of a dataset taken from the crawling process and collected according to its class. From result crawling will entry on the database using query SQL, then export format CSV.

TABLE 4 COLLECT DATASET WITH KEYWORD

Type	Keyword	Amount
Gubernur	#4niesGaBecusKerja	2221
	#PecatAniesBaswedan	1943
	#AniesDiserangBuzzerIstana	2006
	#GantiGubernurDKI	2120
	#gubernurterbodoh	2096
Presiden	#KasihanJokowi	2044
	#JokowiTakutKediri	2136
	#JokowiBasmiKorupsi	1867
	#JokowiKawalNatuna	1982
	#TanpaJKWJugaBisa	2194
Politik	#RezimBudakChina	1987
	#PilpresJiwasraya	2213
	#UlahRezimIndonesiaSuram	2302
	#PDIPMANTAPKOROPSINYA	1890
	#KPKdikebiriPDIP	1893
Kesehatan	#DukungJokowiBasmiCorona	2138
	#JokowiGagalTanganiCovid19	2086
	#LawanCoronaBersama	3012
	#WaspadaVirusCorona	1600
	#KamiTidakTakutVirusCorona	1515
Sosial	#NegeriIntoleran	1754
	#PapuaMajuPapuaMakmur	1967
	#NegeriSwasembadaUtang	2177
	#WajahBaruJakarta	2075
	#PapuaTermiskin	2203
Total Tweet		51.421

From dataset in table 4 to do process will classification with method support like extraction feature TF-IDF for to do weight of word support by N-gram with frequency word feature on tweet, show weight value frequency once used. Feature word from N-gram used can be seen in table 5.

TABLE 5 AMOUNT FEATURE NGRAM

Ngram	Amount
Unigram	27.820
Bigram	212.957
Trigram	201.459
Unigram + Bigram	240.777
Bigram + Trigram	414.416
Unigram + Bigram + Trigram	442.296

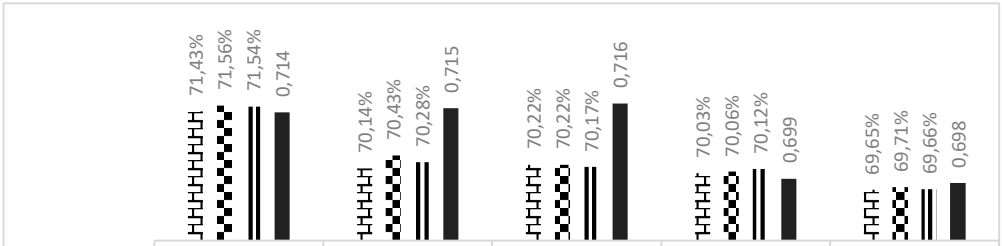
B. Result and Discussion

This test was conducted to determine the level of successfully the system performance that was made to see the performance of hoax identification on twitter social media and to know the effect of extraction features and selection features in this test. The following test was carried out on this final project research work.

1. Testing of system performance using the Support Vector Machine by selecting Information Gain features and without Information Gain. Tests carried out 5 times with 10%, 20%, 30%, 40%, 50% test data.
2. Testing the performance of the system performance using Support Vector Machine with TF-IDF weighting without using the Information Gain selection feature and using the Information Gain selection feature with a different percentage of test data, with threshold testing 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1. Testing consists of five times the test with the percentage of test data 10%, 20%, 30%, 40%, 50%.
3. Testing the system performance using the Support Vector Machine with N-gram without using the Information Gain selection feature and using the Information Gain selection feature with a different percentage of test data. Testing consists of five times with the percentage of test data 10%, 20%, 30%, 40%, 50%. To find out the effect of N-gram on accuracy.

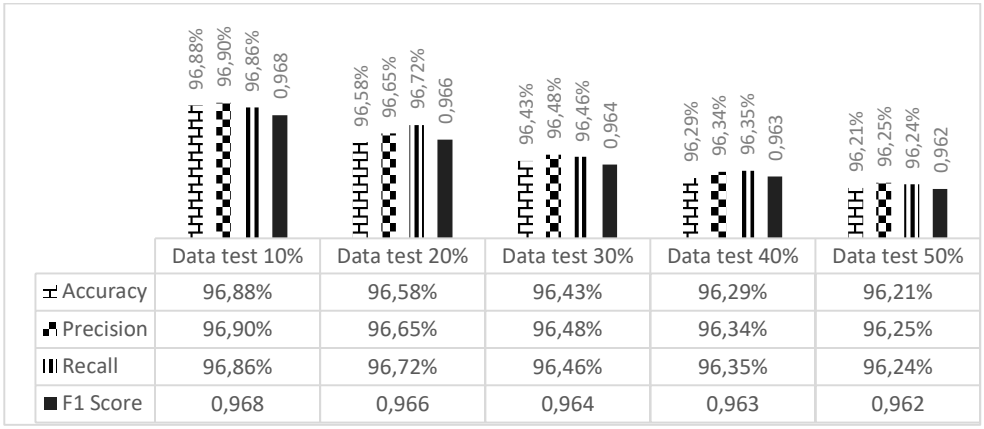
C. Results Analysis and Testing

Testing for classification with different percentage of test data can be seen in Figure 3 and Figure 4.



	Data test 10%	Data test 20%	Data test 30%	Data test 40%	Data test 50%
Accuracy	71,43%	70,14%	70,22%	70,03%	69,65%
Precision	71,56%	70,43%	70,22%	70,06%	69,71%
Recall	71,54%	70,28%	70,17%	70,12%	69,66%
F1 Score	0,714	0,715	0,716	0,699	0,698

Figure 3 Comparison of Confusion Matrix without Information Gain



	Data test 10%	Data test 20%	Data test 30%	Data test 40%	Data test 50%
Accuracy	96,88%	96,58%	96,43%	96,29%	96,21%
Precision	96,90%	96,65%	96,48%	96,34%	96,25%
Recall	96,86%	96,72%	96,46%	96,35%	96,24%
F1 Score	0,968	0,966	0,964	0,963	0,962

Figure 4 Comparison of Confusion Matrix with Information Gain

From the tests carried out it can be seen that the distribution of data test 10% has higher accuracy results in the amount of 71.43% without the Information Gain selection feature, and 96.88% with the Information Gain selection feature, it affects the amount of data 10% because with a smaller amount of test data will increase the value of accuracy in performance testing. So for all tests carried out using 10% test data.

D. Analysis of the Effect of Information Gain Feature Selection on the Accuracy of Hoax

In Information Gain testing, you will choose features that meet the requirements. Following Information Gain testing using the threshold of the parameter.

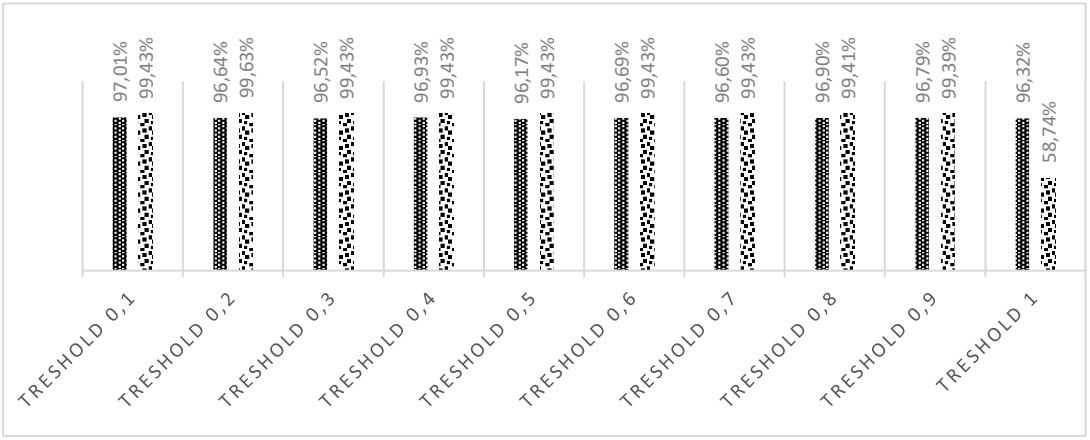
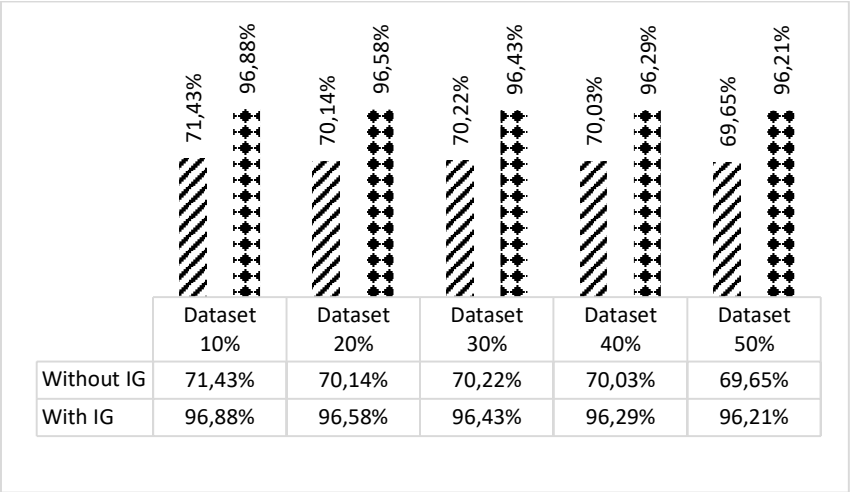


Figure 5 Threshold Parameters

Based on Figure 5, the best results of the Information Gain test on the 0.1 threshold parameter with an accuracy value of 97.01%. Then the next test is done by comparing the performance of Support Vector Machine system performance without using the Information Gain feature selection by using the Information Gain feature selection with a different percentage of test data. With the best use of 0.1 thresholds. Here are the results of tests that have been carried out showing the value 0.1 is the highest value for the selection of features used because Information Gain is influenced by the appearance of words on each feature, the more likely the value that appears the greater to give effect to accuracy.



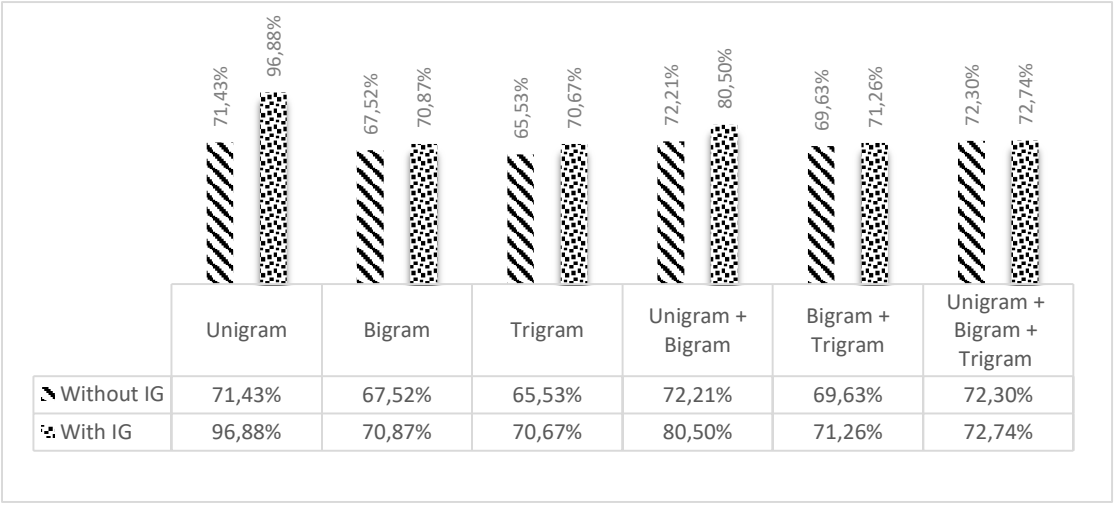
	Dataset 10%	Dataset 20%	Dataset 30%	Dataset 40%	Dataset 50%
Without IG	71,43%	70,14%	70,22%	70,03%	69,65%
With IG	96,88%	96,58%	96,43%	96,29%	96,21%

Figure 6 Effect of Information Gain on Accuracy

Based on Figure 6, the tests performed, a comparison of the results of good system performance using Information Gain to proves the percentage of test data influences the accuracy of data performed by the system, the tests performed are influenced by many data features and large gain values will affect the process of system performance. The number of features acquired was 191.247 and only 190.165 were used because the number of features was affected by the 0.1 threshold value limit. So that the features used to be used in the process of system performance, so the value of accuracy is better or increased.

Information Gain as a feature selection in this research implements the best value of the 0.1 threshold parameter. The results obtained with an average performance of 70.21% without Information Gain and with Information Gain 96.55%, this shows a very significant comparison of 37,51% due to the number of features used in the Information Gain limitation.

E. Analysis of the Effect of N-gram on Accuracy



	Unigram	Bigram	Trigram	Unigram + Bigram	Bigram + Trigram	Unigram + Bigram + Trigram
Without IG	71,43%	67,52%	65,53%	72,21%	69,63%	72,30%
With IG	96,88%	70,87%	70,67%	80,50%	71,26%	72,74%

Figure 7 Effect of N-gram on Accuracy

Based on Figure 7, the effect of N-gram on classification has a comparison of the average results of N-gram without Information Gain getting a value of 69.77%, while the results of the average value of N-gram with Information Gain getting 77.15%, the significant difference in accuracy is 10.57% for overall N-gram. For the overall N-gram comparison, unigram has the highest accuracy value without Information Gain of 71.43%, while with Information Gain 96.88%. That is because the word or term feature appears more frequently on unigrams compared to other N-gram, so unigrams have a major influence on increasing the value of accuracy in the performance process of the hoax identification system from Twitter.

Tests conducted showed the results of the analysis of system performance with an accuracy value of 96.55%. This test was less than the maximum because the influence of the labelling process contained the same words and in the prediction, the process produced an output of ambiguous feature words contained on the hoax and non-hoax word labels, so the system testing cannot predict the features of the same ambiguous word so it lowers the value of accuracy.

TABLE 6 FEATURE ON N-GRAM

Label	Feature Unigram	Feature Bigram	Feature Trigram
Hoax	"panik"	"bakar hutan"	"proyek monas penuh"
	"lawan"	"penuh janggal"	"kabur ruang isolasi"
	"tolak"	"timbul panik"	"akibat bakar hutan"
Non-Hoax	"kerja"	"terima kasih"	"batas sosial skala"
	"bantu"	"batas sosial"	"periksa fakta tingkat"
	"dukung"	"mohon info"	"peran aktif wujud"

Table 6 shows examples of word features most often appearing on uni-grams, bi-grams, tri-grams in each of the hoax and non-hoax classifications that have been determined by the system obtained from the TF-IDF weighting results.

V. CONCLUSION

We have proposed an approach effect of Information Gain feature selection on the identification of hoax from twitter using the Support Vector Machine, with functions to identify and hoax analyze, thereby minimizing negative impacts. The conclusions obtained from the results of this research indicate the effect of Information Gain feature selection on the classification process using the Support Vector Machine method with the highest value of 10%, with an accurate value without using the Information Gain feature selection of 70.21%, test data showing a significant reaching of 37,51%, for the system testing uses the Information Gain feature selection of 95.66% with the use of 0.1 thresholds the tests performed are influenced by many data features and large gain values. The number of feature acquisition is 191.247 and only 190.165 are used. So there are features that are utilized to be used in the system performance process which results in better or improved accuracy. The results of the accuracy value are not optimal because the effect of labelling process has the same word and in the prediction process, it produces an ambiguous feature word output contained on the hoax and non-hoax word labels.

Suggestions that can be considered for further research are tests conducted to be improved in the preprocessing process, especially in the normalization dictionary and stopword to produce standard words, so that not many features are wasted and minimize ambiguous words to obtain optimal accuracy values.

REFERENCES

[1] B. Mansyah, "Fenomena Berita Hoax Media Sosial (Facebook) dalam Menghadapi Pemilihan Umum Gubernur DKI Jakarta Tahun 2017," Bandung, 2017.

[2] R. I. D. E. Agung B. Prasetijo, "Hoax Detection System on Indonesia News Sites Based on Text Classification using SVM dan SGD," *ICITACEE*, p. 45, 2017.

[3] M. S. M. A. Muhammad Hilman Aprilian Nurjaman, "Analisis sentimen pada ulasan buku berbahasa inggris menggunakan Information Gain dan Support Vector Machine," *e-Proceeding of Engineering*, vol. 4, no. 3, p. 4900, 2017.

[4] S. A. F. A. Irene Mathilda Yulietha, "Klasifikasi Sentimen Review Film Menggunakan Algoritma Support Vector Machine," *e-Proceeding of Engineering*, vol. 4, no. 3, p. 4740, 2017.

[5] S. G. Alper Kursat Uysal, "A novel probabilistic feature selection method for text classification," *Knowledge-Based System*, vol. 36, pp. 226-235, 2017.

[6] B. O. Shahana P.H, "Evaluation of Features on Sentimental Analysis," *ICICT*, pp. 1585-1592, 2015.

[7] M. A. F. I. L. S. D. Ni Made Gita Dwi Purnamasari, "Identifikasi Tweet Cyberbullying pada Aplikasi Twitter menggunakan Metode Support Vector Machine (SVM) dan Information Gain (IG) sebagai Seleksi Fitur," *e-ISSN*, vol. 2, no. 11, p. 5328, 2018.

[8] E. B. S. Z. K. A. B. Achmad Fauzi, "Deteksi Berita Hoax di Twitter dengan Metode Term Frequency Inverse Document Frequency dan Support Vector Machine," *Universitas Telkom*, p. 2, 2019.

[9] R. S. Dina Maulina, "Klasifikasi Artikel Hoax Menggunakan Suupport Vector Machine Linier dengan Pembobotan Term Frequency – Inverse Document Frequency," *Jurnal Mantik Penusa*, vol. 2, no. 1, p. 35, 2018.

[10] H. W. S. I. Y. S. M. S. Tansa Trisna Astono Putri, "Analysis and Detection of Hoax Content in Indonesian News Based on Machine Learning," *JIPN*, vol. 4, no. 1, pp. 19-26, 2019.

[11] S. Mujilahwati, "Pre-processing Text Mining pada Data Twitter," *SENTIKA*, pp. 49-56, 2016.

[12] H. H. a. H. L.P. Jing, "Improved Feature Selection Approach TFIDF in text mining," *IEEE Trans Knowl Data Eng*, pp. 944-946, 2017.

[13] L. S. R. Sendy Andrian Sugianto, "Pembuatan Aplikasi Predictive Text Menggunakan Metode N-Gram-Based," *Semantic Scholar*, 2018.

[14] A. H. I. L. S. Azizah Zain, "Identifikasi Tweet Hoax yang Berhubungan dengan Pemilihan Presiden 2019 Menggunakan Naive Bayes Classifier," *Open Library Telkom*, pp. 4-14, 2019.

[15] R. S. W. A. S. Abdul Razak Naufal, "Penerapan Bootstrapping untuk Ketidakseimbangan Kelas dan Weighted Information Gain untuk Feature Selection pada Algoritma Support Vector Machine untuk Prediksi Loyalitas Pelanggan," *ISSN*, vol. 1, no. 2, p. 103, 2015.

[16] informatikalogi, "Entropy & Information Gain," informatikalogi.com/201019, 2019.

[17] S. A. F. A. Irene Mathilda Yulietha, "KIasifikasi Sentimen Review Film Menggunakan Algoritma," *Telkom University*, 2017.

[18] C. O. a. P. Ti, "Performance Evaluation of the Data Mining Classification Method," pp. 249-253, 2014.

[19] I. T. a. I.Technology, "Review on Evaluation Metrics For Data Classification Evaluations," *Int J. Data Min. Knowl. Manag*, vol. 5, no. 2, pp. 1-11, 2015.

[20] E. B. S. Z. A. B. Jaka Eka Sembodo, "Data Crawling Otomatis pada Twitter," *Researchgate*, p. 12, 2016.

[21] E. B. S. Eias Raihandtsa Mamuri, "Mendeteksi Pesan Berita Palsu (Hoax) pada Twitter dengan Algoritma AdaBoost dan ANP," *Universitas Telkom*, p. 1, 2019.

[22] E. B. S. A. B. Jaka Eka Sembodo, "Data Crawling Otomatis pada Twitter," *Ind. Symposium on Computing*, no. doi :10.21108/INDOSC.2016.111, pp. 11-16, 2016.