Int. J. Advance Soft Compu. Appl, Vol. 10, No. 3, November 2018
ISSN 2074-8523

Hoax Classification with Term Frequency – Inverse Document Frequency Using Non-Linear SVM and Naïve Bayes

Ayundyah Kesumawati[1], Achmad Kurniansyah Thalib[2]

Department of Statistics,
Universitas Islam Indonesia, Jl. Kaliurang KM 14,5 Yogyakarta Indonesia
e-mail: [1]ayundyah.k@uii.ac.id, [2]14611134@uii.ac.id

Abstract

In recent years, there are crucial issues in the modern society that gain information on the internet. Spreading the news very easily but can lead to very difficult to filtering the information. The flow of information that provides broad benefits to society, can even enter into the psychology and social for the integrity of the Nation. Information that is easily obtained is extremely dangerous in terms of validity and is not uncommonly a hoax. The dataset that used in this research was gained from news website detik.com and turnbackhoax.id. in this research will provide the comparing of two methods there are Naïve Bayes Classifier (NBC) and Support Vector Machine (SVM) with Radial Basis Function. This research using the Term Frequency – Inverse Document Frequency Weighting (TF-IDFW) that separated each word to make it easy to analyze the text classification. The results obtained for accuracy NBC with training data of 1.480 and test data of 369 is 85.09% and for SVM obtained an accuracy of 83.74%. In addition, the merging of information with text mining, the keyword for the news category is "Price", followed by "KPK", "Stock", "Indonesia", "DPR", and "Police". For the hoax category, the most words are the word "Price", followed by "KPK", "Stock", "Indonesia", "DPR", and "Police".

***Keywords**: News, Hoax, TF-IDFW, NBC, Text Mining, SVM.*

1. Introduction

The development of information and communication technology especially the internet is very fast and make information and news very quickly and easily to obtained. Currently, in Indonesia as of August 2015, internet users reached 88.1

million users, of which 79 million users became active social media users [1]. That’s the huge number divided by it population.

The rapid spread of information flows generated by the development of information technology can lead to information in the form of rapidly circulating news. The rapid spread of news can lead to very difficult information filtering. Information that is easily obtained is very weak in terms of validity and sometimes it is classified as a hoax. Hoax is false news that is inventoried or twisted from the real reality. Hoax may aim to influence the reader with false information so that the reader takes action according to the contents of the hoax.

According to a survey conducted by "MASTEL" that 90.30% of respondents answered the news Hoax is a false news in deliberate, 61.60% argue inciting news, 59% argue accurate news, 14% opinion news forecast and science fiction, 12 % argue that the news cornered the government, 3% responded to the news that was not favored, and 0.60% argued did not know.

The minister of communications and informatics of Indonesia said that there are already nearly 800 thousand sites that spread Hoax on the internet. Of the large number of sites that spread Hoax on the internet it will be very difficult to be able grouping which information are Hoax and not on the internet. News and hoax data which are text data can be classified as unstructured data. In this research the news data will be analyzed into the original news class or hoax news using Naïve Bayes Classifier and Support Vector Machine. The classification process is performed using the Naive Bayes Classifier (NBC) and Non-Linear Support Vector Machine (SVM) algorithm. The NBC and SVM algorithms have a high degree of accuracy in terms of text classification [2].

The section of the rest of this paper as follows: Section 2 discusses the related works on the point SMS Classification using Naïve Bayes Classification, Hoax in Bahasa with Machine Learning, and Hoax Classification Using Linear Support Vector Machine. Section 3 describes the proposed method TF-IDF Weighting using Non-Linear Support Vector Machine. Section 4 provides the results and analysis for comparison between Non-Linear SVM and Naïve Bayes, and finally Section 6 provides discussions and conclusion of this research.

2. Related Work

Research on SMS Classification Using Naive Bayes Classification Method and Apriori Algorithm [3]. The purpose of this research is to get the classification model using Naïve Bayes and Apriori Algorithm to detect SMS which is spam or ham. The data is a collection of SMS collected from the system database. The results of this study were obtained from Apriori Algorithm of 98.7% and 94.7% using the Naive Bayes classification method.

Research on Hoax in Bahasa with Machine Learning [4]. The goal of this study is to get the best model on machine learning that able to do classification on news hoax. The classification of news hoax or news with incorrect information is one of the text categorization applications. Like text-based categorization applications in general, this system consists of pre-processing, feature extraction, feature selection and execution of the classification model. In this study, experiments were conducted to select the best technique for each sub process using 220 Indonesian articles in 22 topics (89 hoax articles and 131 non-hoax articles). The result of this research is that Naive Bayes model show better accuracy than SVM and C4.5 with 91.36% accuracy.

Classification of Hoax Articles Using Linear Vector Machine Support with Term Frequency - Inverse Document Frequency [5]. The purpose of this study is to design a system that can classify information by passing through the field of science Mining. The results obtained that Accuracy of Linear SVM for hoax article classification and no hoax with 108 hoax data and 132 hoaxed articles are 95.8333%. Measurements were made by Cross Validation method with Fold 10. Also found that Support Vector Machine with Linear Kernel has an accuracy of 95.8333%.

There is a differences between this research with related works, where in this research only using news headlines that contained in news portal website detik.com and a collection of hoax on the site turnbackhoax.id. This research using Non-Linear Support Vector Machine with Radial Basis Function include comparison study between NBC and non-linear SVM.

Based on these study, related parties can classify text in the form of news well so that later information in it can be extracted well and presentation of information from observed data can provide useful information.

The purpose of this research is to classify news based on two classification are News category and Hoax category. Since the purpose of this research is to obtain the best method between Naïve Bayes Classifier and Non-Linear Support Vector Machine which have the highest degree of accuracy and get information from text mining in Hoax category, the problem formulation are get the best method to classify News and Hoax category with TF –IDF Weighting using Naïve Bayes Classifier and Non-Linear Support Vector Machine, and get information obtained from text mining results by Hoax data.

3. The Proposed Method

The Proposed Method of this paper is using the TF – IDF Weighting to combining the text classification to get quite smooth of class in hoax or news. The idea of this research is to split the sentence or text became a matrix and represented by weighting of each word. So, there are three theory that used in this research are

TF-IDF Weighting, Naïve Bayes Algorithm and Non-Linear Support Vector Machine. The algorithm of each method shown as below.

3.1. TF-IDF Weighting

In vector space model, TF-IDF is a widely used weighting method, which was firstly introduced from information retrieval. TF-IDF (Term Frequency-inverse Document Frequency), puts weighting to a term based on its inverse document frequency. It means that if the more documents a term appears, the less important that term will be, and the weighting will be less. It can be depicted as this:

$$a_{ij} = tf_{ij} * \log\left(\frac{N}{n_j}\right) \tag{1}$$

In formula (1), tf_{ij} represents the term frequency of term j in document i, N represents the total number of documents in the dataset, n_j represents the number of documents that term i appears. When N equals n_j, then a_{ij} becomes zero, this often appears in small dataset, so we need to apply some smoothing techniques to improve formula (1) as following [6]:

$$a_{ij} = \log(tf_{ij} + 1.0) * \log\left(\frac{N + 1.0}{n_j}\right) \tag{2}$$

Ilustration of TF-IDF Weighting

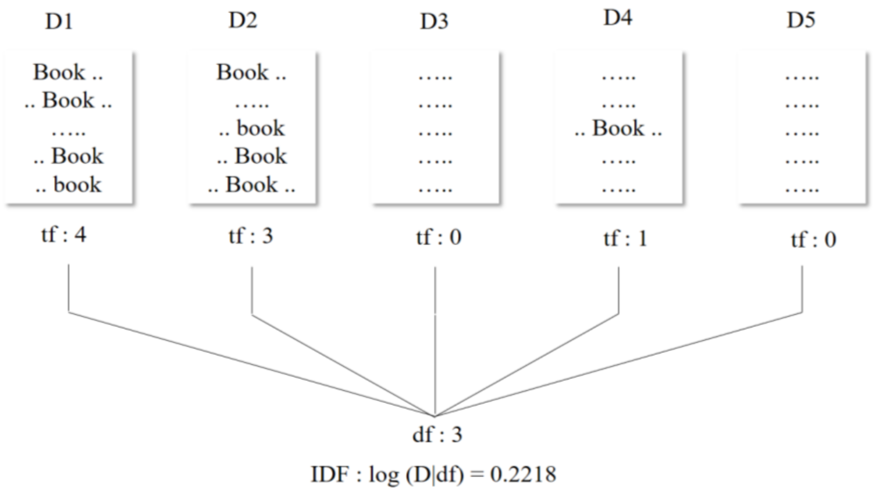


Fig.1 TF – IDF Algorithm

Figure 1 explained that the IDF will be the weight of each word and include as by input for the NBC and Non-Linear SVM.

3.2. Naïve Bayes Classifier

Naïve Bayes Classifier (NBC) or Bayesian Classification is a process of classification method used to determine the probability of a member of a class. NBC is a simple probabilistic classification technique. Bayes theorem has similar classification capabilities to both Decision Tree and Neural Network methods. Therefore, NBC is also effective, efficient, and reliable in handling large datasets and can handle irrelevant data. The symbol for X is the input vector containing the data and Y is the class label. The NBC formula for classification is as follows:

$$P(Y \mid X) = \frac{P(Y)\prod_{i=1}^{q} P(X_i \mid Y)}{P(X)} \tag{3}$$

Each X = $\{X_1, X_2, X_3, \ldots, X_q\}$ as many q attributes or q dimensions, where:

$P(Y\|X)$	= Probability of data with Vector *X* in *Y* Class
$P(Y)$	= Initial probability of *Y* Class (Prior Probability)
$P(X\|Y)$	= Final probability (*Posterior Probability*)
$\prod_{i=1}^{q} P(X_i\|Y)$	= Independent probability of *Y* Class in Vector *X*

For the value of P (X) will always remain so that the calculated only $P(Y)\prod_{i=1}^{q} P(X_i|Y)$ with choose the largest probability that will be used as a class selected on predicted results, this is the work system of NBC.

NBC can be used for both categorical and numerical data. For numerical data there is special treatment. The special treatment is to assume a certain form of probabilistic distribution and to estimate the distribution parameters with training data. Gaussian distributions are often chosen in presenting the conditional probabilities of continuous data in class $P(X_i, Y)$.. There are two parameters in Gaussian Distribution in example: mean, and variant [7]. Where the conditional probability of class Y_j for X_i data is:

$$P(X_i = x_i | Y = y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} exp^{-\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2}} \tag{4}$$

$i \geq \infty$ and $j \geq \infty$, where

X_i	= Attribute *i*
x_i	= Attribute value of *i*

Y = Class
y_j = Sub Class of Y
μ_{ij} = Sample *Mean*
σ_{ij}^2 = Varian sample

3.3. Non-Linear Support Vector Machine

In 1992 for the first time SVM was introduced by Vapnik as a series of eminent concepts in the field of pattern recognition. Today SVM is one of the fastest growing methods. SVM is a machine learning method that works on the principle of Structural Risk Minimization (SRM) which aims to find the best hyperplane that can separate classes in the input space. SVM is one of the classification methods in data mining. SVM can also predict both classification and regression [8]. Basically SVM has a linear principle, but now SVM has evolved to work on nonlinear problems. The way SVM works on nonlinear problems is by incorporating kernel concepts in a high-dimensional space. In this dimensionless space, will be sought separator or often called hyperplane. Hyperplane can maximize distance or margin between data classes. The best hyperplane between the two classes can be found by measuring the margins and then looking for the maximum point. The effort in finding the best hyperplane as class separator is the core of the process in the SVM method.

In real world problem, generally the problem of data obtained is rarely linear. Many are nonlinear. In SVM itself there is a function called kernel. This kernel function is used to solve nonlinear problems. The Kernel function makes it possible to implement a model in a higher dimension space.

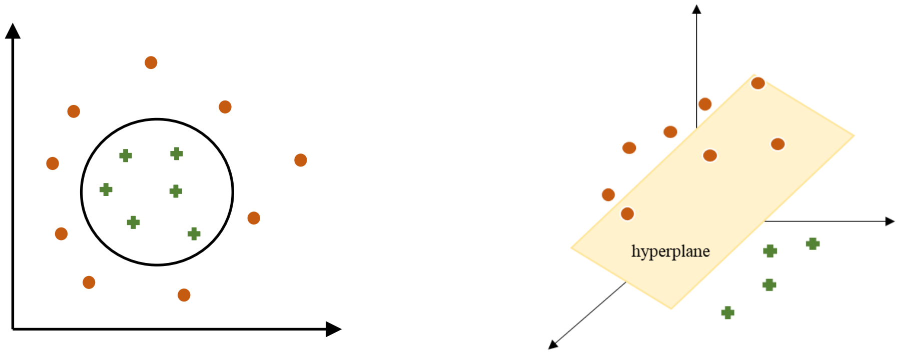


Fig. 2 Hyperplane

The following equation is common Kernel Function is Radial Basis Function

$$K(\vec{X}_i,\vec{X}_j)=\exp\left(\frac{\|\vec{x}_i\ -\vec{x}_j\|^2}{2\sigma^2}\right) \quad (5)$$

In the Radial Basis Function type of a support vector machine, the number of Radial Basis Function and their centers are determined by their number of support vector and their values.

4. Results, Analysis and Discussions

This section presents the results and analysis of the study together with the discussions. The analysis are given into four section: the Preprocessing data text using TF-IDFW algorithm, the analysis of hoax classification using Naïve Bayes Classifier and Support vector Machine and the performance comparison of two method NBC and Non-Linear SVM, and using wordcloud in hoax data.

4.1. Preprocessing Using TF-IDF Weighting Algorithm

This section discusses about preprocessing the headline news data using Term Frequency – Inverse Document Frequency. The process as a follow:

Table 1: Research Data

Category	News
Hoax	Hati-Hati Orang Iseng Tempel Sticker PKI di Belakang Mobil
News	5 Kasus Korupsi Kakap Eksploitasi Sumber Daya Alam yang Ditangani KPK
Hoax	Planet Nibiru Siap Menabrak Bumi
News	Investor Penerima Tax Holiday Tak Bisa Menerima Tax Allowance
Hoax	Jenazah Siyono utuh, berarti ybs Mati Syahid

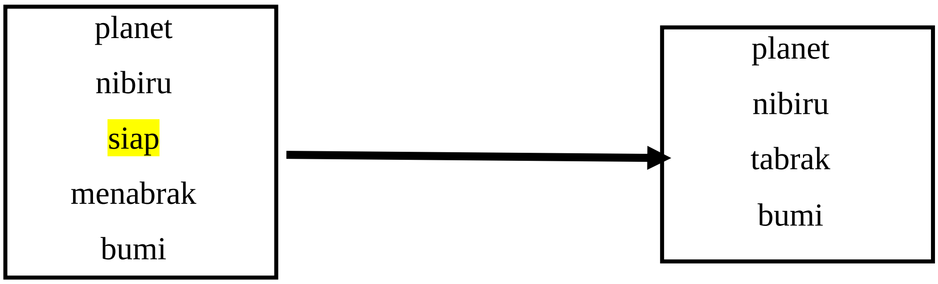
Table 1 shown the example headline data from turnbackhoax.id for the hoax category and detik.com for news category. the headline data using the Bahasa as the language of Indonesia. The next step is to case folding the data by reducing all letters to lower case on the other hand case folding can equate words that might better be kept apart. The result of case folding as follow.

Table 2: Case Folding Result

Category	News	Case Folding Result
Hoax	Hati-Hati Orang Iseng Tempel Sticker PKI di Belakang Mobil	hatihati orang iseng tempel sticker pki di belakang mobil
News	5 Kasus Korupsi Kakap Eksploitasi Sumber Daya Alam yang Ditangani KPK	kasus korupsi kakap eksploitasi sumber daya alam ditangani kpk
Hoax	Planet Nibiru Siap Menabrak Bumi	planet nibiru siap menabrak bumi

Category	News	Case Folding Result
News	Investor Penerima Tax Holiday Tak Bisa Menerima Tax Allowance	investor penerima tax holiday menerima tax allowance
Hoax	Jenazah Siyono utuh, berarti ybs Mati Syahid	jenazah siyono utuh berarti ybs mati syahid

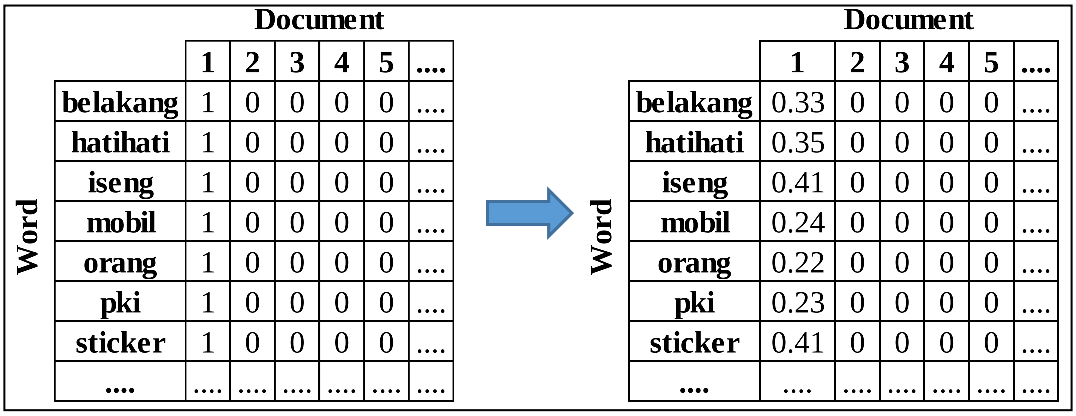
Table 2 shown the result of case folding, the uppercase became lowercase and there is no more punctuation in the sentences. The next step is by using the corpus of the Bahasa and by using tokenizing the sentences that contain any punctuation or the most common word such as yang, di, ke, siap, etc in Bahasa will be deleted, the result as follow:



(Tokenizing Result) (Filtering Result)

Fig. 3 Tokenizing Result

Fig. 3 shown the tokenizing result, the next step is to make a word matrix that contained the number of each word in sentences. Next, by using equation (2) the result of the TF – IDF Weighting matrix as follow:



Word	Document 1	2	3	4	5	....
belakang	1	0	0	0	0	....
hatihati	1	0	0	0	0	....
iseng	1	0	0	0	0	....
mobil	1	0	0	0	0	....
orang	1	0	0	0	0	....
pki	1	0	0	0	0	....
sticker	1	0	0	0	0	....
....	....	....	....	....	....	....

Word	Document 1	2	3	4	5	....
belakang	0.33	0	0	0	0	....
hatihati	0.35	0	0	0	0	....
iseng	0.41	0	0	0	0	....
mobil	0.24	0	0	0	0	....
orang	0.22	0	0	0	0	....
pki	0.23	0	0	0	0	....
sticker	0.41	0	0	0	0	....
....	....	....	....	....	....	....

Fig. 4 TF-IDF Weighting Process

Fig. 4 shown the result of the TF-IDF Weighting Process, the number of each word as variable and the number of document as an object that became input data for classification training and testing.

4.2. Hoax Classification Using Naïve Bayes Classifier and Non-Linear Support Vector Machine

This section described the result of classification analysis and discusses about the result comparison between Naïve Bayes Classifier and Non-Linear Support Vector Machine using Radial Basis Function. The result as follow:

4.2.1. Naïve Bayes Classifier

Naïve Bayes method uses the probability value in determining the classification class. The main components of the Naïve Bayes concept are the prior and also the probability of each word. The analysis by using the Naive Bayes Classifier used 1.480 headline articles that classification in two class there are hoax or news. Table 3 shown the result between prediction and actual objects.

Table 3: *Confussion Matrix of Naive Bayes Classifier*

Prediction	**Actual**		**Total**
	News	**Hoax**	
News	139	32	171
Hoax	23	175	198
Total	162	207	369

To evaluate the model or machine learning that has been established, then made predictions on the data test after training data. In Table 3. shown that there are that the news supposed to be classified as news but it shown that there a misclassification. Total for predictive data obtained in news category is 171 objects. For hoax category, obtained total category for hoax prediction is 198 objects. If it compared with the actual data, total category of news is 162 objects, and for category hoax obtained 207 objects. When compared to these two results, it can be seen that there is a difference from the original data as a whole and the predictions data for each category. However, for the total of the total test data, we get 369 data. Based on Table 3 it can be calculated the value of accuracy and the number of error. The calculated as follow:

$$\text{Accuracy } = \frac{\sum(correct\ prediction)}{(total\ prediction)} = \frac{139+175}{369} = 0{,}8509$$

$$\text{Error } = \frac{\sum(wrong\ prediction)}{(total\ prediction)} = \frac{23+32}{369} = 0{,}14191$$

It can be concluded that using the Naïve Bayes Classifier predicted in a correctly class is 85,09%.

4.2.2. Non-Linear Support Vector Machine

The Non-Linear SVM method works by finding hyperplanes or inter-class dividing lines. The C and Gamma parameter values in the SVM method with the RBF Kernel will determine the accuracy of the classification. In this study, the C and Gamma values used are the default values assigned. For the value of C is a positive value that is C ≥ 0, and the default value of the model is 1, while for Gamma value is also $\gamma \geq 0$, with default value is 1 / data dim. The data used for machine learning or model is training data that has been made on preprocessing a, while to testing the accuracy of the machine or model used test data.

Table 4 : *Confussion Matrix of Non-Linear Support Vector Machine*

Prediction	**Actual**		**Total**
	News	**Hoax**	
News	139	37	176
Hoax	23	170	193
Total	162	207	369

To evaluate the model or machine learning that has been established, the accuracy of classification need to predictions on the test data after training on the train data. In Table 4 it can be conclude that there is misclassification. Total for predictive data obtained in news category obtained 176 objects. For hoax category in a prediction obtained 193 objects. When it compared in actual data, the total category of news is 162 objects, and for the category of hoax obtained 207 objects. Based on Table 4 it could also calculated the value of accuracy and its error. The calculated of the accuracy using Non-Linear SVM as follow:

$$\text{Accuracy } = \frac{\sum(correct\ prediction)}{(total\ prediction)} = \frac{139+170}{369} = 0{,}8374$$

$$\text{Error } = \frac{\sum(wrong\ \ prediction)}{(total\ prediction)} = \frac{23+37}{369} = 0{,}1626$$

It can be concluded that using the Non-Linear Support Vector machine predicted in a correctly class is 83,74%.

4.2.3. Comparison Naïve Bayes and Nonlinear Support Vector Machine

Comparison of accuracy of result classification method of Naïve Bayes Classifier and Non-Linear SVM is used to determine the best method in this research. The results of both methods obtained the value of accuracy of both methods as follows.

Table 5 : Comparison of Naïve Bayes Classifier and Non-linear SVM

	Naive Bayes	Non Linear SVM
Accuracy	85,09%	83,74%
Error	14,91%	16,26%

Based on Table 5, it can be concluded that the accuracy values for these two classification have quite little different. However, the accuracy value of the classification method for this case study is higher than the Non-Linear SVM RBF classification method. By looking at the comparison of accuracy and error, then with an accuracy of 85.09%, which means that 85.09% of the predictions of the Naive Bayes method fit exactly into the actual category, it can be concluded that the Naive Bayes classification method is better than the SVM RBF.

4.3. Text Mining of Hoax

This section discusses text mining about the headline in hoax datasets. The result as follow

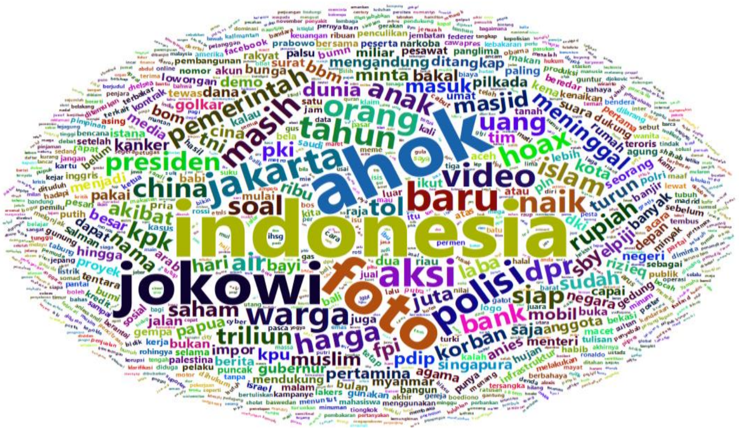


Fig. 5 Word Cloud of Hoax

It can be concluded that in Fig. 5 there are 4 word had quite big size of word in word cloud hoax category there is Indonesia, Ahok, Jokowi and Foto. It means that the frequency of that word is the most dominant word frequency if compared with other words. The word order of the word cloud represented by figure 6 as follow:

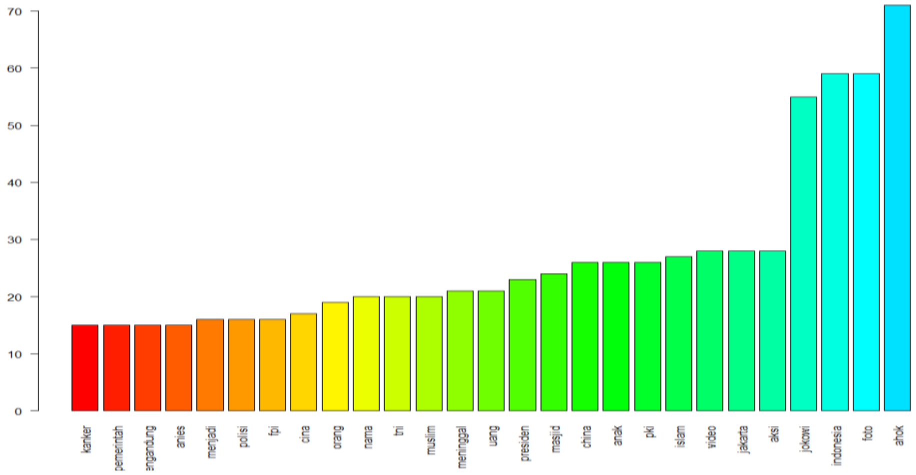


Fig. 6 The Frequncies of Word in Hoax Datasets

Fig.6 represent that within the range of 2015-2018, there is one thing that can be extracted. In 2017, there are politics issues about the election of the Governor of DKI Jakarta, when it was widely circulated hoax news about the candidate governor who was caught in the denigration of religion at the time. For that reason, the word "Ahok" becomes one of the dominant that appears among other words. Because of words with high frequency are clearly visible on the word cloud, it can be assumed that these words are considered important for this news category. Table 6. Shown the association each word of three highest frequencies word, the result as follow:

Table 5 : Word Association in Hoax Datasets

Indonesia		Jokowi		Ahok	
Kata	**Asosiasi**	**Kata**	**Asosiasi**	**Kata**	**Asosiasi**
kirim	0.27	bloomberg	0.25	teman	0.23
terulang	0.22	dibiarkan	0.25	parade	0.16
tragedi	0.22	memuji	0.25	pro	0.16
cina	0.16	presiden	0.22	ruang	0.16
agp	0.15	senen	0.18	menjamur	0.16
memasukkan	0.15	rezim	0.18	miliarbulan	0.16
makar	0.15	menjamur	0.18	jakarta	0.15
dibodohkan	0.15	miliarbulan	0.18	kampanye	0.15
peradilan	0.15	moral	0.18	djarot	0.15
wibawa	0.15	media	0.16	sumbangan	0.13
lte	0.15	pemberitaan	0.14	bersalaman	0.13
penemu	0.15	bodoh	0.14	kaos	0.13
peringatkan	0.15	makin	0.14	diri	0.13
tenaga	0.14	psk	0.14	alami	0.13
kerja	0.13	dituduh	0.13	kabur	0.13
lengkap	0.12	kebakaran	0.13	dibuat	0.13
tunggal	0.12			makin	0.13
komunis	0.12			psk	0.13
mineral	0.12			menuntut	0.12
tdl	0.12			poster	0.12
				monas	0.12

In Table 5. Could be conclude that the association of the 3 words that most often appear in the category hoax. The result of the association indicates that when there is the word "Indonesia", "Ahok", and "Jokowi", the words in figure 4 above are the most frequently followed. The word "Indonesia" is most authentic with "send", "repetitive", "tragedy", "china". Similarly, the word "Indonesia", for the words "Ahok" and "Jokowi", the words contained in Figure 4 are the words that most often follow the words when the words "Ahok" and "Jokowi" appear.

6 Conclusion

Based on the results of the analysis and discussion that has been done in this research, can be obtained conclusion as follow:

1. Based on the results of news and hoax classification analysis, accuracy was obtained for NBC method with training data of 1,480 and 369 test data obtained by accuracy of 85.09% and for SVM RBF, with 83.74%.
2. From the results of digging information with text mining, the most common word for the hoax category are the word “Indonesia", followed by “Ahok“ and “Jokowi".

ACKNOWLEDGEMENTS

This research is supported by Academic Development Board of Universitas Islam Indonesia, Statistics Department of Universitas Islam Indonesia.

References

[1] Social, We Are. (2015). *Digital, Social & Mobile in Southeast Asia in Q4 2015*. http:// wearesocial.com / sg / special-reports / digital-southeast-asia-q4-2015. (15 April 2018)

[2] Naradhipa, A. R., & Purwarianti A. (2012). Sentiment classification for Indonesian message in social media. *Cloud Computing and Social Networking* (ICCCSN), International Conference (pp: 1-5).

[3] Ahmed, I., Guan, D. (2014). SMS Classification on Naive Bayes Classifier and Apriori Algorthm Frequent Dataset. *International Journal of Machine Learning and Computing*, 4(2).

[4] Rasyiwir, E., Purwarianti, A. (2015). Experiments on the Hoax Language Classification System Indonesian Based Machine Learning. *Cybermatika Journal*. 3(2).

[5] Sagara, R. (2017). Hoax Articles Classification Using Linear Vector Machine Support by Weighting Term Frequency - Inverse *Document Frequency*. *Thesis*. *Computer Science Faculty Universitas Amikom*. Yogyakarta.

[6] Liu, M., & Yang, J. (2012). An Improvement of TFIDF Weighting in Text Categorization. *In International Conference on Computer Technology and Science*. ICCTS 2012 (Vol 47).

[7] Subhan, A. & Ahmad, Z., F. (2017). *Implementation of Data Mining To Determine Daily Rain Potential By Using Naïve Bayes Algorithm*. Thesis. Informatic Engineering Faculty of Universitas Dian Nuswantoro. Semarang.

[8] Larose, D., T. & Larose, C., D. (2014). Discovering Knowledge in Data: An Introduction to Data Mining Second Edition, New Jersey: John Wiley & Sons Inc.