ANALYSIS AND DETECTION OF HOAX CONTENTS IN INDONESIAN NEWS BASED ON MACHINE LEARNING

Tansa Trisna Astono Putri[1], Hendryx Warra S[2], Irma Yanti Sitepu[3], Marita Sihombing[4], Silvi[5]

[1,2,3,4,5] TeknikInformatika
Universitas Prima Indonesia, Jl. Sekip, Medan Petisah, Kota Medan, Sumatera Utara 20111

E-mail: tansatrisna@unprimdn.ac.id[1], hendryxputra91@gmail.com[2], yirma79@yahoo.com[3],maritasihombing@gmail.com[4], silvi.ong97@gmail.com[5]

Abstract

Hoax newsthat contain incorrect (false) information often become public consumption on social media today. This hoax phenomenon raises doubts about information and makes confusion in the community. In this study, experiments conducted aimed at selecting the best algorithm in classifying hoax and non-hoax news with the number of data in 251 articles in Indonesian language (100 hoax articles and 151 non-hoax articles) using text mining method and machine learning based approaches. This research undergoes the text preprocessing phase which consists of tokenizing, case folding, filtering, stopwords removing, stemming and TF-IDF weighting using unigram and bigram combine features before processing it into classification text. The results of this research is the Random Forest algorithm that gets the best accuracy in classifying hoax and non-hoax news compared to the Multilayer Perceptron algorithm, Naïve Bayes, Support Vector Machine, and Decision Tree with an accuracy value of 76.47%.

Keywords : Classification, News, Hoax, Machine learning, Text mining

1. INTRODUCTION

Information technology is a tools and infrastructure (hardware, software, useware) system and methods for acquiring, sending, processing, interpreting, storing, organizing, and using data meaningfully [17]. Technologyof information also interpreted as science in the field of computer-based information and its development is very rapid [13].

Progress in information technology has an impact on the increasingly widespread news network and accessible easily through online social media. As online news networks expand, the quality of the news disseminated is also less. The news that is disseminated is not all true news, including there are also fake news or hoaxes which result in detrimental to some parties.

Hoax is an information or news that contains uncertain things or which really are not facts that occur [6]. Hoax can also be identified with the following [9]: the news comes from unclear / untrusted sources. Images, photos or videos used are the result of engineering, using provocative sentences, and contain political and racial elements.

The spreading of hoaxes among the public can have negative effects, such as damage, losses, both material and psychological, public distrust, and so on. One example of a hoax is that in early January 2018, a group of people joined in the organization of the Islamic Defenders Front (FPI) attacked and burned down the headquarters of the Indonesian Lower Society Movement (GMBI) in Bogor. According to a number of eyewitnesses, the attack was carried out by FPI members, which numbered around 150 people. The reason for the attack was triggered by the hoax news obtained by FPI members from social media that one of the FPI members had been stabbed by GMBI members. As a result of the attack, a GMBI headquarters and a house were burned down. The losses incurred reached hundreds of millions of rupiah.

The impact of the spread of fake news (hoaxes) will have bad consequences and harm many parties, hoaxes can cause losses from various aspects, both time and economy, public panic, worsening social relations and so on. To avoid these adverse effects, this study will help us classify the news before we are affected by hoaxes or even spread the hoax. The technique used in this study is the text classification system using a machine learning based approach. The algorithms used are: Multilayer Perceptron (MLP), Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM) and Decision Tree (DT).

Similar research has been carried out by Rasywir and Purwarianti [14], the difference between this research and the previous research was on the use of algorithms. Rasywir and Purwarianti's research used three algorithms, namely Naïve Bayes algorithm, Support Vector

Machine, and Decision Tree, while in this study, we compared five algorithms namely Multilayer Perceptron, Naïve Bayes, Support Vector Machine, Random Forest, and Decision Tree. Another difference from Rasywir and Purwarianti's research with our research is the text preprocessing phase, where the phase in the previous research begin with case folding first, then followed by tokenizing, stopwords removing, stemming and ended with TF-IDF weighting with three features i.e. unigram, bigram and combines unigram and bigram, while in the research we did, the phases of text preprocessing process begin with tokenizing, then the case folding, filtering, stemming, and TF-IDF weighting using unigram and bigram combined features. [14]

2. LITERATURE REVIEW

2.1. News

News is a report of information about events and opinions that are actual, important and interesting to convey to the public in the form of newspapers, radio and online media. One condition of the news is that the news must be based on an accident or event that actually happened. But as technology advances increasingly rapidly, the spreading of the news is also increasingly unclear whether the truth is in accordance with the facts or just hoaxes.

The hoax is a deliberate manipulation of news and aims to provide false recognition or understanding [4]. Hoax is often found in the news, both through print and social media. The purpose of the hoax itself is very diverse, ranging from spreading hate speech, causing anxiety in society, influence the perceptions of the community, and so on.

The Constitution which regulates the Hoax is UUD Nomor 11 tahun 2008 concerning Article 28 of Electronic Information and Transactions, namely: (1) "*Setiap Orang dengan sengaja dan tanpa hak menyebarkan berita bohong dan menyesatkan yang mengakibatkan kerugian konsumen dalam Transaksi Elektronik. (2) Setiap Orang dengan sengaja dan tanpa hak menyebarkan informasi yang ditujukan untuk menimbulkan rasa kebencian atau permusuhan individu dan/atau kelompok masyarakat tertentu berdasarkan atas suku, agama, ras, dan antar golongan (SARA)*". Act No. 1 of 1946 concerning Regulations Criminal Law. Article 14, namely: (1) "*Barang siapa, dengan menyiarkan berita atau pemberitahuan bohong, dengan sengaja menerbitkan keonaran di kalangan rakyat, dihukum dengan hukuman penjara setinggi-tingginya sepuluh tahun. (2) Barang siapa menyiarkan suatu berita atau mengeluarkan pemberitahuan yang dapat menerbitkan keonaran dikalangan rakyat, sedangkan iapatut dapat menyangka bahwa berita atau pemberitahuan itu adalah bohong, dihukum dengan penjara setinggi-tingginya tigatahun*". Article 15 namely: "*Barang siapa menyiarkan kabar yang tidak pasti atau kabar yang berkelebihan atau yang tidak lengkap, sedangkan ia mengerti setidak-tidaknya patut dapat menduga bahwa kabar demiki anakan atau sudah dapat menerbitkan keonaran di kalangan rakyat, dihukum dengan hukuman penjara setinggi-tingginya dua tahun*."

2.2. Text Mining

Text mining is one branch of data mining that analyzes data in the form of text. Text mining itself is a process of mining data in the form of text with data sources that are usually obtained through documents, and aims to find words that can represent the contents of the document so that analysis connectivity between documents can be carried out [19]. And until now, text mining has been widely applied in fields such as: security, biomedicine, software and applications, online media, marketing, academics, and other fields (Saraswati, 2011).

2.3. Text Preprocessing

Text preprocessing is the initial stage of text mining. In this stage, the preparation process for documents and data is carried out so that the documents / data are ready to be processed and the classification process can be processed properly. There are also stages in the text preprocessing, namely:

a. Tokenizing

Tokenizing is the process of breaking or cutting sentences into words by using a space marker. For example: "Text preprocessing merupakantahapawaldaritext mining", it will be broken down into: "Text", "preprocessing", "merupakan", "tahap", "awal", "dari", "text", "mining".

b. Case folding

Case folding is the stage where all characters in the text are processed and converted into lower cases. For example: "TECHNOLOGY" will be changed to "technology", "Text" will be converted into the word "text", and so on.

c. Filtering

Filtering is the stage of removing punctuation from the results of tokens.

d. Stopwords removing

In removing stopwords stages, words that are not unique words or words that often appear and are not important will be removed or

discarded. Examples of words in question are prepositions, conjunctions, adverbs and substitute words, such as: "yang", "ke", "di", "sebuah", "pada", "oleh", "ini", "dari" , etc.

e. Stemming

Stemming process is a process carried out to obtain the basic words of a word by eliminating the suffixes of both prefixes, suffixes, infixes and combinations of prefixes and suffixes.

f. TF - IDF weighting

TF (Term Frequency) is the frequency of the appearance of a term in the document concerned. The greater the number of occurrences of a term (high TF) in the document, the greater the weight or will provide a greater value of conformity.

Whereas IDF (Inverse Document Frequency) is a calculation of how the terms are distributed widely in the collection of documents concerned. IDF shows the relationship between the availability of a term in all documents. The fewer the number of documents containing the term in question, the greater the IDF value.

2.4. Types of Algorithms

The algorithm used in this study is:

a. Multilayer Perceptron (MLP)

Multilayer Perceptron method is one of the most common topologies of artificial neural networks, where perceptrons are connected and form multiple layers. MLP consists of input layers, hidden layers, and output layers (Negnevitsky, 2005). Because MLP consists of several layers, the calculation process for this algorithm also goes through several stages, the first is calculating values from hidden layer (Ghosh et al, 2014).

$$hidden_k = \frac{1}{1+exp^{-hidden\_Net_j}} \quad (1)$$

Where the value of *hidden_Netj* can be calculated using equation (2).

$$hidden_{Net_j} = \sum_i w_{jk} \times mf_{ij}(y_{ij}) + bias_k \quad (2)$$

Where:

w_{jk} = Weight from neuron j on previous layer to neuron k on hidden layer.

$Mf_{ij}(y_{ij})$ = Member degree from attribute i to class j.

$bias_k$ = Bias from neuron k.

Then the result value from hidden layer will be used to calculate output layer used the this equation (Ghosh et al, 2014):

$$output_l = \frac{1}{1+exp^{-output\_Net_k}} \quad (3)$$

With the output value *output_Net*$_k$get from equation (4).

$$output\_Net_k = \sum_j w_{kl} \times hidden_k + bias_l \quad (4)$$

Where:

w_{kl} = Weight from neuron k on previous layer to neuron l on hidden layer.

$bias_l$ = Bias from neuron l.

b. Naïve Bayes

The algorithm developed by Thomas Bayes has been widely used as a method of classifying text that has a simple chance. This algorithm utilizes probabilities and statistics to predict future probabilities based on previous experience. One of his studies is a study that compares three algorithms namely Naïve Bayes algorithm, Support Vector Machine algorithm and Decision Tree algorithm in hoax news classification. And the result is that the Naive Bayes algorithm shows the best accuracy that is equal to 91.36% [14]. The equation of the Naïve Bayes theorem can be written as follows (Mitchell, 1997):

$$P(X|Y) = \frac{P(Y|X) \times P(X)}{P(Y)} \quad (5)$$

Where:

P(X|Y) = Probability of X based on the conditions of Y.

P(Y|X) = Probability of Y based on the hypothesis of X.

P(X) = Probability of X.

P(Y) = Probability of Y.

c. Support Vector Machine

The SVM method is a universal machine learning model that utilizes linear limiting functions as a basis (Joachims, 1998). And one of the interesting studies that drew our attention is the classification of hoax articles using SVM by TF-IDF weighting which shows an accuracy of 95.8333% (maulina & sagara). The equation for calculation of Support Vector Machine algorithm for data that may not be grouped correctly (Asiyah&Fithriasari, 2016) such:

$$min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{\lambda} \xi_i$$

Based on (6)

$y_i(wx_i + b) + t_i \geq 1$ dan $\xi_i \geq 0,1,\cdots,\lambda$

Where :
w = Hyperplane parameter searched.
C = Parameter that defines penalty due to error in classification of the data where the values are determined by the user.
λ = Number of partition/data.
i = Initial value
x = SVM input data
b = Alternative parameter towards the center of coordinates (bias).

d. Random Forest

The Random Forest method is a development of the CART method where the CART method applies the bootstrap aggregating (bagging) method and the random feature selection method [3] Random Forest is one method that can improve accuracy because it generates random nodes for each node [12]. One of study that uses the Random Forest method is the classification of new workforce research and using genetic algorithms as Random Forest optimization in the study, the highest accuracy results of the Random Forest were either optimized using genetic algorithms or did not reach 95% [2].

This method is used to build decision trees consisting of root nodes, internal nodes, and leaf nodes by randomly retrieving attributes and data according to the provisions in force.The decision tree starts by calculating the entropy value as a determinant of the level of impurity of the attribute and information gain value. To calculate the entropy value, use the formula as in equation (7), while the information gain value uses equation (8).

$$Entropy(Y) = -\sum_i P(C|Y)\, log_2\, P(C|Y) \quad (7)$$

Where:
Y = Set of cases.
P(C|Y) =Proportion of value Y towards class C.

$$IG(Y,a) = E(Y) - \sum_{veV(a)} \frac{|Y_v|}{|Y_a|} E(Y_v) \quad (8)$$

Where:
E = Entropy.
V(a) = All possible values in a set of cases a (Values a).
Yv = Subclass of Y with class v which is related to class a.
Ya = All of the values that corresponding to a.

e. Decision Tree

The advantages of the Decision Tree algorithm are this algorithm can handle continuous data and discrete data. This Decision Tree algorithm is the development of ID3 algorithm. As for previous studies that applied the Decision Tree algorithm, the classification of student graduation from the faculty of communication and informatics. The highest reached 73.91% [12]. The equation that can be used to calculate entropy is:

$$Entropy(S) = \sum_{j=1}^{k} -Pj\, log_2\, Pj \quad (9)$$

Where:
S = Set of cases.
k =Number of partition S.
Pj = Probability obtained from the number (yes/no) is divided by the total of the cases.

From the Entropy calculation in equation to (9), the Gain value can be calculated using the following equation (10):

$$Gain(S,A) = E(S) - \sum_{i=1}^{n} -\frac{|Si|}{|S|} \times E(Si) \quad (10)$$

Where:
S =Set of cases.
A = Attribute.
E = Entropy.
n = Number of partition attributes A.
|*Si*| =Number of cases on the partition i.
|*S*| = Number of cases in S.

3. RESEARCH METHODS

3.1. Data source

The data sources that we used in this study are taken from online news articles that have been labeled with hoaxes and non-hoax labels. With the number of hoax articles as many as 100 articles and 151 articles labeled non-hoax. So the overall article data for this study amounts to 251 articles.

3.2. Step Analysis

In this study, there are several steps for analyzing data, as follows:

a. Collect the data on news articles from various online news sites, namely: kompas.com, viva.co.id, detik.com, liputan6.com, timesdindonesia.co.id, turnbackhoax.id, wartaekonomi.co.id, analisadaily.com, metrotvnews.com, beritasatu.com, cnnindonesia.com, idntimes.com, republika.co.id, prokal.co, cekfakta.com, jpnn.com, okezone.com, sindonews.com, solopos.com, tempo.co,

merdeka.com, gridoto.com, tribunnews.com, malangtimes.com, jawapos.com, melekpolitik.com, kumparan.com, cumicumi.com, klikdokter.com, klikpositif.com.

b. Then the news data that has been collected and stored in CSV format will go through the stages of text preprocessing, and the first stage is the tokenizing stage, where all sentences from the news data that have been collected will be separated by syllables based on spaces.
c. Then in the folding case process, all the letters in the separated word are changed to lowercase letters so that the same words but have different upper and lower case writing are not processed as different words by the program. And then it will be continued with the normalization of words, where different words that have the same meaning will be changed to one word in common.
d. After the case folding and normalization process, the filtering process is carried out to remove characters other than letters and numbers, such as punctuation (.), Comma (,) or URL addresses, and will be considered asdelimeter.
e. Then for words that are not unique words or words that often appear in sentences, will be removed in the stopwords removing stage. This process aims to reduce the number of words that are processed by deleting words that have no effect or do not have positive or negative (net) characteristics of the sentence.
f. Then it proceed with the stemming process, where the affixes to the words will be removed to get the basic words. This process aims to make the words that have the same basic word as the same word and not the different words just because there is a prefix / suffix / suffix on the base word.
g. Then the last step in the text preprocessing is to calculate the weight of each word in the news data to represent the importance of the role of the word in the data with TF-IDF weighting using unigram and bigram combined features. The greater the weight value, the more important is the role of the word in forming a data (Weddiningrum, 2018).
h. Data is then divided into 2, namely: training data and test data with the percentage of training data is 80% and the test data is 20%. With a total data of news articles as many as 251, so the total number of training data is 201 articles and test data as many as 50 articles. Then selecting training data and test data will be done randomly by the program.
i. Conduct text classifications using machine learning algorithms, namely Multilayer Perceptron, Support Vector Machine, Naïve Bayes, Random Forest, and Decision Tree algorithms.
j. Compare the performance between the Multilayer Perceptron algorithm, Naïve Bayes, Support Vector Machine, Random Forest, and Decision Tree based on the level of accuracy, precision, recall, and F-1 score from each algorithm.Examples of the application of the text preprocessing process to classify news data can be seen in table (1) below, namely as follows:

Table 1.
Text preprocessing process

Classification process	News data
Preliminary Data	*Di sosial media beredar foto ini yang menyebutkan sebagai pesawat Lion Air JT-610.*
Tokenizing	*Di* *sosial* *media* *beredar* *foto* *ini* *yg* *menyebutkan* *sebagai* *pesawat* *Lion* *Air* *JT-610.*
Case folding and normalization	*di* *sosial* *media* *beredar* *foto* *ini* *yang* *menyebutkan* *sebagai* *pesawat* *lion* *air* *jt-610.*
Filtering	*di* *sosial* *media* *beredar* *foto* *ini* *yang* *menyebutkan* *sebagai*

Classification process	News data
	pesawat
	lion
	air
	jt-610
Stopwords removing	*sosial*
	media
	beredar
	foto
	menyebutkan
	pesawat
	lion
	air
	jt-610
Stemming	*sosial*
	media
	edar
	foto
	sebut
	pesawat
	lion
	air
	jt-610

4. RESULTS AND DISCUSSION

4.1. Precision

Precision value is the level of accuracy between the information requested by the user and the answer given by the system. Calculation of precision can be written as follows (Manning et al, 2009):

$$Precision = \frac{TP}{(TP+FP)} \quad (11)$$

Equation (11) can be written as follows:

$$Precision = \frac{Relevant\ \ data\ \ found}{All\ \ data\ \ found} \quad (12)$$

Where:
TP = True Positiveis the number of relevant data that is correctly classified as matches data by the system.
FP = False Positiveis the number of irrelevant data, but classified as matches data by the system.

And based on the calculation formula, the precision obtained from each algorithm described by Table 2: Multilayer Perceptron algorithm gets a precision value of 69.12%, Naïve Bayes algorithm obtains 79.79%, and Support Vector Machine algorithm gets 70.16%, then the Random Forest algorithm gets a value precision is 79.34%, while the Decision Tree algorithm gets 74.07%. Then the algorithm that has the highest precision value in this study is the Naïve Bayes algorithm with a precision value of 79.79%.

Table 2.
Precision values on testing data.

Algorithm	*Precision*
Multilayer Perceptron	69.12%
Naïve Bayes	**79.79%**
Support Vector Machine	70.16%
Random Forest	79.34%
Decision Tree	74.07%

4.2. Recall

Recall is the system's success rate in rediscovering information. In other words, the recall shows how complete the relevant results are displayed by the system. Calculation of recall values can be written in form (Manning et al, 2009):

$$Recall = \frac{TP}{(TP+FN)} \quad (13)$$

Equation (13) can be written as follows:

$$Recall = \frac{Relevant\ \ data\ \ found}{All\ \ relevant\ \ data\ \ indatabase} \quad (14)$$

Where:
TP = True Positiveis the number of relevant data that is correctly classified as matches data by the system.
FN = False Negativeis the number of relevant data, but isn't classified as matches data by the system.

From the research that has been done, we acquired different recall values for each algorithm used described by Table 3. The Multilayer Perceptron algorithm gets a recall value of 67.16%, a Naïve Bayes algorithm of 66.86%, then the Support Vector Machine algorithm obtains the lowest recall value among the five algorithms used by us, which is 62.31%, then the Random Forest algorithm gets 73.82%, and the last is the algorithm The Decision Tree that gets the highest recall value in this study is 74.29%.

Table 3.
Recall values on testing data.

Algorithm	*Recall*
Multilayer Perceptron	67.16%
Naïve Bayes	66.86%
Support Vector Machine	62.31%
Random Forest	73.82%
Decision Tree	**74.29%**

4.3. F-1 Score

F-1 Score is a measurement value of performance performed to see the results obtained from the classification process based on the precision and recall values that have been obtained. In other words, F-1 Score is also called the harmonic mean of precision and recall. Calculations for F-1 scores can also be written as follows (Manning et al, 2008):

$$F_1 = \frac{2 \times recall \times precision}{recall + precision} \quad (15)$$

Based on the precision and recall values described by Table 4, the Multilayer Perceptron algorithm obtained F-1 Score of 67.09%, and Naïve Bayes of 67.26%, the Support Vector Machine algorithm obtained a value of 71.08%, then the Random Forest algorithm with the highest value of 74.24%, and then followed by a Decision Tree algorithm of 74.15%. Comparison of the F-1 Score the five algorithms can be seen in table (4).

Table 4.
F-1 Score on testing data.

Algorithm	*F-1 Score*
Multilayer Perceptron	67.09%
Naïve Bayes	67.26%
Support Vector Machine	71.08%
Random Forest	**74.24%**
Decision Tree	74.15%

4.4. Accuracy

Level accuracy is the level of closeness between predictive values and actual values. Calculations for value accuracy can be written as follows (Manning et al, 2008):

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (16)$$

Calculation of accuracy found in equation (16) can also be written as follows:

$$Accuracy = \frac{Data\ that\ is\ correctly\ \ classified}{Total\ \ data\ \ tested} \quad (17)$$

Where:
TP = True Positiveis the number of relevant data that is correctly classified as matches data by the system.
TN = True Negativeis the number of irrelevant data and is correctly classified as unmatched data by the system.
FP = False Positiveis the number of irrelevant data, but classified as matches data by the system.
FN = False Negativeis the number of relevant data, but was not classified as matches data by the system.

Based on the result described by Table 5, the Multilayer Perceptron algorithm receives an accuracy value of 68.63%, then the Naïve Bayes algorithm gets a value of 74.51%, followed by Support Vector Machine with the lowest accuracy value among the five algorithms compared to us, which is 60% in the study this, then the Random Forest algorithm gets the highest accuracy value with a value of 76.47%, and the last one is the Decision Tree algorithm with the accuracy value of 74.51%, the same as the Naïve Bayes algorithm.

Table 5.
Accuracy values on testing data

Algorithm	*Accuracy*
Multilayer Perceptron	68.63%
Naïve Bayes	74.51%
Support Vector Machine	60.00%
Random Forest	**76.47%**
Decision Tree	74.51%

Overall, the comparison of the five algorithms is: Multilayer Perceptron algorithm, Naïve Bayes, Support Vector Machine, Random Forest and Decision Tree algorithm based on precision, recall, F-1 Score and accuracy in this study could be seen in Table 6 below.

Table 6.
Comparison among precision, recall, F-1 score and accuracy values.

Algorithm	*P*	*R*	*F-1*	*A*
Multilayer Perceptron	69.12%	67.16 %	67.09%	68.63%
Naïve Bayes	79.79%	66.86 %	67.26%	74.51%
Support Vector Machine	70.16%	62.31 %	71.08%	60%
Random Forest	**79.34%**	**73.82 %**	**74.24%**	**76.47%**
Decision Tree	74.07%	74.29 %	74.15%	74.51%

5. CONCLUSION

In this study, a hoax news classification system in Indonesian has been formed using machine learning with 5 different types of algorithms, namely: Multilayer Perceptron, Naïve Bayes, Support Vector Machine, Random Forest, and Decision Tree algorithm, with a

total of 251 articles which consists of: 151 non-hoax articles and 100 hoax articles obtained from online news articles. This research also through tokenizing process, case folding, normalization, filtering, stopwords removing, stemming, and TF-IDF weighting using unigram and bigram features.

It can be concluded that the best classification results are achieved by the Random Forest algorithm if compared with the Multilayer Perceptron algorithm, Naïve Bayes, Support Vector Machine, and the Decision obtained the highest accuracy level of 76.47%, and the highest F-1 score is 74.24%, where F-1 Score is the average between precision and recall. The higher the precision and recall, the higher the accuracy obtained. Based on the comparison in table 6, the writer prefers to use algorithms with the highest accuracy, because on average if the accuracy level is high then the level of precision and precision also high.

6. REFERENCES

[1] Asiyah. S. N., &Fithriasari. K., (2016): Klasifikasi Berita *Online* Menggunakan Metode *Support Vector Machine* dan *K-Nearest Neighbor*, Surabaya: Jurnal Sains dan Seni ITS.

[2] Binarwati. L., Mukhlash. I., &Soetrisno. S., (2017): Implementasi Algoritma Genetika untuk Optimalisasi *Random Forest* dalam Proses Klasifikasi Penerimaan Tenaga Kerja Baru :Studi Kasus PT.XYZ, Surabaya: Jurnal Sains dan Seni ITS.

[3] Breiman. L., (2001): *Random Forests*, *Berkeley*: *University of California*.

[4] Dahlan. M. A., (2017): Ahli: "*Hoax*" MerupakanKabar yang Direncanakan, Jakarta: ANTARA News.

[5] Ghosh. S., Biswas. S., Sarkar. D., & Sarkar. P. P., (2014): *A Novel Neuro-fuzzy Classification Technique for Data Mining*, India: *Egyptian Informatics Journal*, 129-147. Harlian. M., (2006): *Machine Learning Text Kategorization*, *Austin*: *University of Texas*.

[6] Juditha. C., (2018): Interaksi Simbolik dalam Komunitas Virtual Anti Hoak suntuk Mengurangi Penyebaran Hoaks, Jakarta: Jurnal PIKOM, vol. 19, no. 1, Kementerian Komunikasi dan Informatika RI.

[7] Manning. C. D., Raghavan. P., &Schutze. H., (2009): *An Introduction to Information Retrieval*, *Cambridge*: *Cambridge University Press*.

[8] Mitchell. T. M., (1997): *Machine Learning*, *Singapore*: *McGraw-Hill*.

[9] Monohevita. L., (2017): Stop Menyebarkan *Hoax*, Depok: Universitas Indonesia.

[10] Negnevitsky. M., (2005): *Artificial Intelligence*: *A Guide to Intelligent System* (*2nd Ed*), *Harlow*: *Pearson Education*.

[11] Nugroho. Y. S., (2014): Penerapan Algoritma C4.5 untuk Klasifikasi Predikat Kelulusan Mahasiswa Fakultas Komunikasi dan Informatika Universitas Muhammadiyah Surakarta, Yogyakarta: Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST).

[12] Nugroho. Y. S., & Emiliyawati. N., (2017): Sistem Klasifikasi Variabel Tingkat Penerimaan Konsumen Terhadap Mobil Menggunakan Metode *Random Forest*, Surakarta: Jurnal Teknik Elektro, vol. 9, no. 1.

[13] Prasojo. L. D., &Riyanto, (2011): Teknologi Informasi Pendidikan, Yogyakarta: Gava Media. ISBN: 978-602-8545-28-0.

[14] Rasywir. E., & Purwarianti. A., (2015): Eksperimen pada Sistem Klasifikasi Berita *Hoax* Berbahasa Indonesia Berbasis Pembelajaran Mesin. Bandung: Jurnal Cybermatika, vol. 3, no. 2.

[15] Republik Indonesia, (1946): Undang-Undang Republik Indonesia No. 1 Tahun 1946 Tentang Peraturan Hukum Pidana, Jakarta: Sekretariat Negara.

[16] Republik Indonesia, (2008): Undang-Undang Republik Indonesia No. 11 Tahun 2008 Tentang Informasidan Transaksi Elektronik, Jakarta: Sekretariat Negara.

[17] Warsita. B., (2008): Teknologi Pembelajaran: Landasan dan Aplikasinya, Jakarta: Rineka Cipta.

[18] Weddiningrum. F. G., (2018): Deteksi Konten *Hoax* Berbahasa Indonesia pada Media Sosial menggunakan Metode *Levenshtein Distance*, Surabaya: Universitas Islam Negeri sunan Ampel.

[19] Harlian. M., (2006): *Machine Learning Text Kategorization*, *Austin*: *University of Texas*