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

Ind. Journal on Computing
Vol. 5, Issue. 3, December 2020. pp. 23-36
doi:10.34818/indojc.2021.5.3.506

Implementation Information Gain Feature Selection for Hoax News Detection on Twitter using Convolutional Neural Network (CNN)

Husnul Khotimah Farid [#1], Erwin Budi Setiawan [#2], Isman Kurniawan [#3]

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

[1] husnulkhotimahh@students.telkomuniversity.ac.id
[2] erwinbudisetiawan@telkomuniversity.ac.id
[3] ismankrn@telkomuniversity.ac.id

Abstract

The development of information and communication technology is currently increased, especially related to social media. Nowadays, many people get information through social media, especially Twitter, because of its easy access and it doesn't cost much. However, it has a negative impact in the form of spreading fake news or hoaxes that are difficult to detect. In this research, the authors developed a hoax news detection model using the Convolutional Neural Network and the TF-IDF weighting method. Feature selection is performed using Information Gain with various features, such as unigram, bigram, trigram and a combination of the three. Testing is done with 3 scenarios, classification, classification by weighting, classification by weighting and feature selection. The parameter used in the information gain feature selection is the threshold 0.8. The results showed that the classification by weighting and feature selection produced the highest accuracy that is equal to 95.56% on the unigram + bigram features with a comparison of training data and test data 50:50.

Keywords: Hoax, Twitter, Convolutional Neural Network, TF-IDF, Information Gain.

Abstrak

Perkembangan teknologi informasi dan komunikasi telah mengalami kemajuan yang pesat, terutama terkait media sosial. Saat ini, banyak orang mendapatkan informasi melalui media sosial, khususnya Twitter, karena aksesnya yang mudah dan tidak mengeluarkan banyak biaya. Akan tetapi, hal tersebut memiliki dampak negatif berupa penyebaran berita palsu atau hoax yang sulit dideteksi kebenarannya. Pada penelitian ini, penulis mengembangkan model deteksi berita hoax dengan menggunakan Convolutional Neural Network dan metode pembobotan TF-IDF. Seleksi fitur dilakukan dengan menggunakan Information Gain dengan fitur yang beragam yaitu unigram, bigram, trigram dan gabungan ketiganya. Pengujian dilakukan dengan 3 skenario yaitu klasifikasi, klasifikasi dengan pembobotan, klasifikasi dengan pembobotan dan seleksi fitur. Parameter yang digunakan pada seleksi fitur information gain adalah threshold 0.8. Hasil penelitian menunjukkan bahwa klasifikasi dengan pembobotan dan seleksi fitur menghasilkan akurasi tertinggi yaitu sebesar 95.56% pada fitur unigram + bigram dengan perbandingan data latih dan data uji 50:50

Kata Kunci: Hoax, Twitter, Convolutional Neural Network, TF-IDF, Information Gain.

Received on January, 2020. Accepted on December, 2020

I. Introduction

Nowadays, the development of technology and information makes communication becomes easier so that the flow of information can be quickly spread not only in the real world but also in cyberspace, especially on social media. Access to information through social media is easier and does not cost much. However, it has a negative impact in the form of spreading fake news or hoaxes that are difficult to detect. Based on a survey from the content management service site HootSuite in 2019, the number of active social media users in Indonesia reached 150 million users out of a total population of 268.2 million where Twitter ranks in the top four of social network users on social network types and based on surveys conducted by the public Telematics (MASTEL) in 2017, social media such as Twitter and others became the most widely used to spread hoax news by 31.90% [1].

Hoax is the information or news that contains things that are uncertain or not a fact that actually occur [2]. Hoax news is very troubling for many people because it contains words that are provocative and contain discriminatory elements, which the purpose is to discredit the other party, on the other hand, glorify the other side [3]. In addition, hoax information can trigger anxiety, hatred and hostility by using unclear sources and no one is responsible or difficult to find the clarification [4].

The impact caused by hoaxes is very detrimental for many parties, therefore research related to hoax news detection is mostly done by various methods including by using Decision Tree [5], Support Vector Machine (SVM) [6], Self-Organizing Map [7], Convolutional Neural Network (CNN) [8] and various other methods. In this research, the authors used the CNN method to learn the problems above. Actually, CNN is very effective and efficient in detecting images [9], but in this time the author uses the CNN method to classify documents in text form to find out whether CNN can work well in processing text or not. Besides that, this research also uses Term Frequency Inverse Document Frequency (TF-IDF) weighting and Information Gain feature selection.

II. Literature Review

A. Related Work

Hoax news is currently very disturbing for Indonesian people. The negative impact of hoaxes causes harm to certain parties, therefore many people conduct research related to hoax detection. First, research conducted by Laode Muhammad Ikhsan et al with the Decision Tree and Analytical Hierarchy Process methods. The data used is data related to the disaster with testing data based on information from Twitter. Classification uses four different data shares in testing. From the trials that have been carried out, the highest accuracy is 98.46% with the 70% training data and 30% testing data [5].

Research in hoax detection was again carried out by A Fauzi et al using the Support Vector Machine (SVM) method and TF-IDF weighting feature to predict the possibility of Twitter users spreading hoax news. Data is taken based on the information from Twitter in the form of tweet content such as number of retweets, URLs, number of hashtags, provocations, animosity, anxiety, and inappropriate news. The test results using all content get an accuracy rate of 78.33% [6].

The other hoax research was conducted by Agung Prasetijo et al by developing hoax filters based on text vector representations on TF-IDF. The classification techniques used are Support Vector Machine (SVM) and Stochastic Gradient Descent (SGD). The highest accuracy of the two methods is 86% and obtained from the SGD classification using modified-huber. The results obtained were more than 100 hoaxes and 100 non-hoax websites randomly selected outside the dataset used in the training process [10].

Similar research was conducted by Faisal Rahutomo et al in detecting Indonesian hoax news using the Naïve Bayes classification method with term frequency features and using the library component PHP-ML or PHP-

Machine Learning. The dataset used was 600 news. After testing, the system produces an accuracy rate of 82.6% and a dynamic trial of 68.33%. [11].

Furthermore, research conducted by Lhaksmana Muslim Institute et al uses the backpropagation method to predict and classify the possibility of Twitter users spreading hoax based on user's behavior. The dataset is used for training in backpropagation using the backpropagation gradient descent algorithm and the lavenberg-marquad backpropagation algorithm. From the testing process carried out by the lavenberg-marquad backpropagation method an average accuracy of 72.19% was obtained with the lowest MSE (0.1996) compared to the backpropagation gradient descent backpropagation [12].

B. Preprocessing

The data preprocessing stage is the process of preparing raw data before another process is carried out or in other words preprocessing is the process of converting text from unstructured natural language into an input format in an algorithm [13]. Preprocessing has 4 stages of the process. Case folding is the process of changing all letters in the text into lowercase and also eliminating all characters and components such as numbers, symbols, html, links, URL. Normalization is the process of changing all the abbreviated words in the text into the appropriate words in the language dictionary. Filtering is the process of selecting important words. Selection of important words can use the stoplist / stopword algorithm (discard words that are considered not important) or the wordlist algorithm (save important words). Last, the process of stemming is to change every word into root word.

C. Term Frequency-Inverse Document Frequency (TF-IDF)

Term frequency (TF) is to give weight for terms that often appear in a document with the equation:

$$W_{dt} = tf_{dt} \tag{1}$$

where tf is the occurrence frequency of the term t in document d.

Inverse Document Frequency (IDF) is functionate to reduce the weight of a term if it widely spread throughout the document.

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

where $|D|$ is the total number of documents in the corpus and df is the number of documents that are contains term t.

Term Frequency-Inverse Document Frequency (TF-IDF) is a method used to give weights or values to terms in a document. The greater the frequency of words appearing in a document, the greater the weight. The equation is written as:

$$W_{dt} = tf_{dt}\, x\, idf_t \tag{3}$$

where W_{dt} is the total weight of the term t to document d, tf_{dt} is the occurrence frequency of the term t in document d, and idf_t is the value of Inverse Document Frequency.

D. N-Gram

N-gram is an n-word chunk obtained from a string. The n-gram method is applied for word or character generation [14]. In this research, the generation of n-gram is divided into unigram (one word), bigram (two words), trigram (three words) and a combination of all three. Examples of the n-gram application of the sentence "pemerintah berupaya menjaga masyarakat aman" can be seen in Table 1.

TABLE I
N-GRAM EXAMPLE

N-gram	N-gram formed
Unigram	“pemerintah”, “berupaya”, “menjaga”, “masyarakat”, “aman”
Bigram	“pemerintah berupaya”, “berupaya menjaga”, “menjaga masyarakat” “masyarakat aman”
Trigram	“pemerintah berupaya menjaga” “berupaya menjaga masyarakat” “menjaga masyarakat aman”

E. Information Gain

Feature selection is a dimension reduction technique that is used to reduce data matrices by paying attention to important word that needs to be processed [15]. Information gain is used to select the best features. Information gain is used to sort important words from the results of feature reduction. The results of the information gain process are important words that are informative. Attributes that meet the weighting criteria will be used in the classification process of an algorithm [16]. The following formula is to calculate information gain:

$$Gain(S, A) = Entrophy(S) - \sum_{i=1}^{n} \frac{S_i}{S} * Entrophy(S) \quad (4)$$

where the value of A is a feature. The value of n is the number of feature values in A. The value of S_i is the number of samples for feature A. And Entropy (S) is the entropy of each sample.

F. Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) is a regularization version of Multi-Layer Perceptron and is classified into a deep feed-forward Artificial Neural Network. CNN is a convolution operation that combines several layers of processing, using several elements that operate in parallel and inspired by the biological nervous system. CNN architecture consists of input layer, hidden layer and output layer. The CNN hidden layer contains Convolutional layer, Pooling layer, Activation layer (generally RelU), Fully connected layer and Loss layer [17].

Convolutional layer is a layer where the input layer will be abstracted into a feature map that consists of kernel/filter (could be more than one). Like images, filters have a certain height, width, and depth and the value of this filter that becomes the parameters (height, width, depth) that will be updated in the learning process. For example when using 4 filters, the resulting map features is width x height x 4. For example, the input layer is 7x7 pixels. The part in the input layer is called the receptive field that has the same size as the kernel. Every part of the input layer will perform dot product operations on the kernel. All receptive fields in the input layer will be shifted (convolving) from top-left to bottom-right, so a feature map is generated. When convolving using stride width and zero padding, what is used generally is stride = 1 and to calculate the zero padding value, the formula P = (F-1) / 2 is used where P is the padding size and F is the receptive measure field. If the kernel is 3x3, then the receptive field used is 3x3 and zero padding = 1. The illustration can be seen in Figure 1.

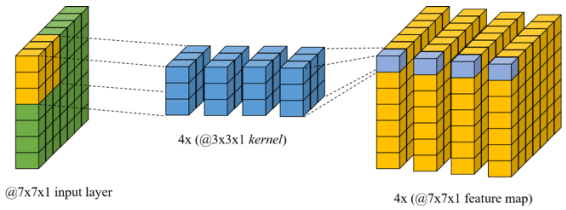

Fig. 1. Convolution layer with 4 filters 3x3x1

So, if w1 x h1 x d1 is the size of the input layer, then it can be set w2 x h2 x d2 as a feature map on the convolution layer, with the equation:

$$w_2 = \frac{(w_1 - F + 2P)}{S} + 1 \quad (5)$$

$$h_2 = \frac{(h_1 - F + 2P)}{S} + 1 \quad (6)$$

$$d_2 = K \quad (7)$$

where F is the size of the receptive field, P is the size of zero padding, S is the stride and K is the number of filters / kernels [18].

The pooling layer is used for the sample reduction (down-sampling) process. The advantage of using a pooling layer is that we can represent data to be smaller, easier to manage and easily control overfitting. There are several techniques that can be used, including maxpooling, L2-norm pooling and average pooling. The most commonly used is maxpooling, which works by selecting a maximum value in a particular area [18]. The illustration can be seen in Figure 2.

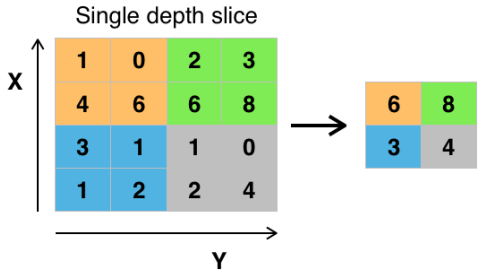

Fig. 2. Max-Pooling

In the first receptive part (top-left), a maximum value is 6, and obtained 8, 3 and 4 for the other receptive parts.

Activation RelU is an activation layer that uses the function f (x) = max (0, x). RelU Activation removes the negative value on the feature / activation map and makes 0. The output of the RelU layer will have the same dimensions as the feature map dimension before being applied to RelU. Illustration can be seen in Figure 3.

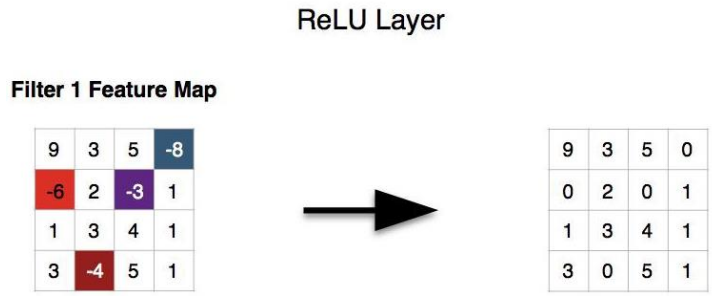

Fig. 3. Activation RelU

Fully Connected Layer is a layer where each neuron is interconnected with neurons in the previous layer. Each activation from the previous layer needs to be converted into one-dimensional data before it can be connected to all neurons in a fully connected layer [19].

Loss layer is the last layer of CNN whose task is to measure the deviation of the predicted value with the actual value (target). There are several types of loss functions that can be used, namely softmax loss, sigmoid cross-entropy, and euclidean loss.

G. Measuring Performance

Measuring performance is the stage of collecting, analyzing and evaluating system performance to be designed. Performance is measured using accuracy values. Performance measurements can be calculated using the Confusion Matrix table shown in Table II.

TABLE II
CONFUSION MATRIX

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

TP is a class that is predicted yes, and the actual is yes. FP is a class that is predicted yes, and the actual is no. TN is a class that is predicted no, and the actual is no. FN is a class that is predicted no, and the actual is yes. Accuracy can be calculated from the confusion matrix table.

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 [20]. The greater the accuracy value, the better the classification performance produced. The equation of accuracy is:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}\ x\ 100\% \quad (8)$$

Precision is the level of accuracy between the information requested by the user and the answers given by the system [20]. The formula of precision can be seen in the equation:

$$Precision(P) = \frac{TP}{TP+FP} \quad (9)$$

Recall is the number of users correctly classified in a class divided by the total number of users in that class. Recall is also often referred to as the success rate of the system in finding back an information [20]. The formula of recall can be seen in the equation:

$$Recall = \frac{TP}{TP+FN} \quad (10)$$

F-1 Score is the average comparison of precision and recall [20]. The formula of f1-score can be seen in the equation:

$$F1-Score = \frac{2\ x\ precision\ x\ recall}{precision+recall} \quad (11)$$

III. RESEARCH METHOD

A. System Flowchart

System flowchart in this research can be seen in Figure 4. The process begins by crawling the data, then labeling it. After that, preprocessing is done to convert raw data into data that can be processed by the system. Next is the weighting process using TF IDF by weighting each word. The next process is feature selection using

information that can increase the value of accuracy. The last process is the collection using convolutional neural networks. Before that, the data is divided into train data that is used by the system to conduct learning and testing data to support the system that has been created.

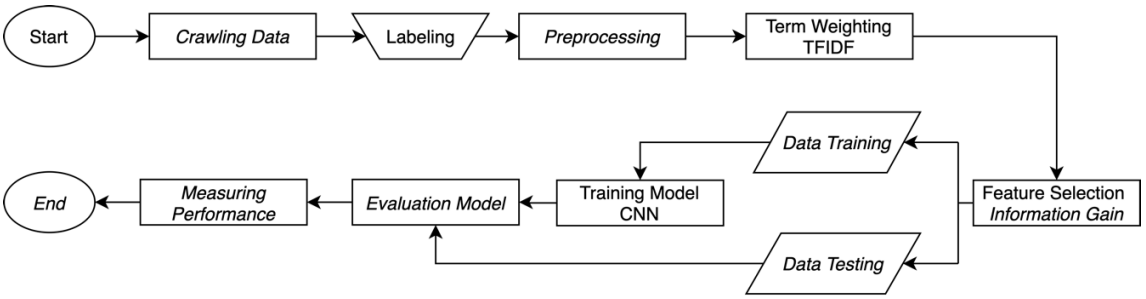

Fig. 4. Flowchart

B. Crawling Data

Data Crawling is a process of obtaining data that will be used as reference data by the system. In this research, data collection uses the Twitter API. Every one time of crawling, the tweet data obtained is 100 of the latest tweet data. The data retrieved is a tweet based on keywords by looking at hashtags that are considered to contain hoax elements. Total data taken is as many as 50,646 data.

C. Labeling

Labeling is the process of determining the class of a tweet that is done manually by labeling 1 which means hoax and label 0 which means non-hoax. There are several things that become a reference in labeling a tweet, including the use of a username checked whether using numbers or symbols and also using real names or pseudonyms, the second is whether the tweet contains elements of hatred, panic, anxiety, provocation or cornering other parties, the third is the use of location with URL in each tweet, next is the comparison of the number of followers and following and the last is whether the account is a verified account or not.

D. Preprocessing

Preprocessing data is a step of processing raw data which will be included in the classification system. There are 4 stages of the pre-processing process, Case Folding to change uppercase letters into lowercase letters and eliminate characters such as symbols, numbers, etc. Normalization to change all the abbreviated words in the text into the appropriate words in the language dictionary, filtering is the process of selecting important words and eliminating words that are not important, and Stemming to change the terms in the form of basic words. Examples of pre-processing can be seen in Table III.

TABLE III
EXAMPLE OF PREPROCESSING

Preprocessing	**Sentence**
Initial Sentence	@jokowi Lockdown nkri pak. Jgn tunggu banyak korban pak #LockdownIndonesia
Case Folding	lockdown nkri pak jgn tunggu banyak korban pak
Normalization	lockdown nkri pak jangan tunggu banyak korban pak
Filtering	lockdown nkri banyak korban
Stemming	lockdown nkri banyak korban

E. Term Weighting

In this process, data that has been preprocessed will be given a value or weight using TF-IDF weighting. This weighting is useful to measure how influential the words of a document. In this stage, the weighted features are also developed using n-grams which are divided into unigram (one word), bigram (two words), trigram (three words) and a combination of all three.

F. Feature Selection

Data obtained from the TF-IDF weighting results in the form of vectors will be reduced the dimensions which are used to reduce the data matrix by taking into the important words that needs to be processed. The process of dimension reduction uses information gain by ranking the important words from the result of dimensional reduction.

G. Classification

After all the necessary processes is finished, the next process is the classification process using the Convolutional Neural Network (CNN) method. CNN architecture used is Embedding layer - Convolution 1D - Activation Function RelU - Global Max Pooling - RelU - fully connected - Sigmoid. Other parameters used are learning rate 0.001, dropout 0.2 with epoch 30 and batch size 128.

H. Evaluation Model

This process produced a model that is used for the testing of test data in order to find out how well the model has been built.

IV. Results And Discussion

A. Dataset

The dataset is obtained by crawling data on Twitter. The total data of tweets obtained is 50,647 tweets. The data taken is a tweet based on keywords by looking at hashtags that are estimated to contain hoax elements in the span of December 2019 - March 2020. A list of keywords used in the crawling process can be seen in Table IV.

TABLE IV
Keyword Crawling Data

Keyword	**Amount**
#KebakaranHutan	7.002
#AniesMundurJakarTateratur	3.551
#gubernurterbodoh	4.863
#GantiGubernurDKI	4.578
#PecatAniesBaswedan	3.776
#TanpaJKWJugaBisa	6.487
#KartuSaktiJKWHanyaJanji	6.791
#KasihanJokowi	2.629
#BanjirJakart2020	3.878
#VirusCorona	7.092
Total	**50.647**

After the data from the crawling results, then it labeled 1 for hoax and labeled 0 for nonhoax. The results of the comparison between hoax and nonhoax labels are quite balanced because the comparison of the numbers is not much different, 25.021 for hoax labels and 25.625 for nonhoax labels. Labeling is done based on the references that have been described in section 3.

The data then developed using n-grams into unigram, bigram, trigram and a combination of the three. The number of features used can be seen in Table V.

TABLE V
AMOUNT OF N-GRAM FEATURES

N-gram	**Amount**
Unigram	25.538
Bigram	157.946
Trigram	169.410
Unigram + Bigram	183.484
Bigram + Trigram	327.356
Unigram + Bigram + Trigram	352.894

B. Result and Discussion

In this research, the test is carried out with three scenarios that aim to determine the performance and the success level of the system, and also to determine the effect of the feature selection in detecting hoax news. The following is the scenarios:

1. Testing of system performance using the Convolutional Neural Network (CNN) classification method. The test is carried out by comparison of training data and test data of 90:10, 80:20, 70:30, 60:40, and 50:50.
2. Testing of system performance using the Convolutional Neural Network (CNN) classification method with the TF-IDF weighting feature. The test is carried out by comparison of training data and test data of 90:10, 80:20, 70:30, 60:40, and 50:50.
3. Testing of system performance using the Convolutional Neural Network (CNN) classification method with the TF-IDF weighting feature and Information Gain feature selection with the threshold testing parameter is 0.8 Testing is done by comparing training data and test data 90:10, 80:20, 70: 30, 60:40, a 50:50.

All test scenarios are performed using n-gram features, that are unigram, bigram, trigram, unigram + bigram, bigram + trigram, and unigram + bigram + trigram. Tests are also carried out with a variation of the ratio to find out at what ratio is produced the highest accuracy.

C. CNN Classification Test Result

The results of system performance testing using the Convolutional Neural Network classification method with different test data ratio can be seen in Table VI.

TABLE VI
CNN CLASSIFICATION TEST RESULT

Feature N-gram	**Data train : data test**	**Loss**	**Acc**	**Prec**	**Recall**	**F-1 Score**
UNIGRAM	90:10	62,21%	64,27%	64,66%	63,42%	64,03%
	80:20	61,73%	64,54%	62,99%	70,92%	66,72%
	70:30	62,74%	63,48%	63,03%	64,74%	63,87%
	60:40	62,73%	63,30%	62,74%	65,15%	63,92%
	50:50	63,09%	63,48%	62,62%	66,42%	64,46%
BIGRAM	90:10	60,00%	65,85%	63,94%	67,32%	65,59%
	80:20	60,31%	65,43%	64,62%	60,80%	62,65%
	70:30	62,58%	63,25%	62,41%	57,99%	60,12%

	60:40	61,88%	64,78%	68,51%	47,99%	56,44%
	50:50	62,10%	64,53%	62,63%	60,95%	61,78%
TRIGRAM	90:10	58,71%	**68,60%**	68,37%	53,32%	59,91%
	80:20	60,56%	64,05%	57,10%	57,46%	57,28%
	70:30	60,56%	66,34%	60,16%	53,93%	56,87%
	60:40	60,39%	66,28%	63,05%	48,17%	54,61%
	50:50	61,92%	65,52%	59,79%	51,75%	55,48%
UNI + BI	90:10	61,13%	65,90%	72,14%	64,39%	68,05%
	80:20	61,71%	64,91%	64,00%	68,24%	66,05%
	70:30	63,34%	63,31%	63,04%	63,97%	63,50%
	60:40	62,34%	63,98%	62,82%	68,65%	65,61%
	50:50	62,79%	64,02%	63,10%	66,93%	64,96%
BI + TRI	90:10	61,58%	64,33%	58,80%	55,43%	57,07%
	80:20	61,16%	64,91%	59,71%	65,75%	62,58%
	70:30	62,28%	63,29%	56,80%	70,69%	62,99%
	60:40	63,04%	63,68%	59,47%	61,67%	60,55%
	50:50	63,41%	63,94%	60,06%	61,81%	60,92%
UNI + BI + TRI	90:10	61,06%	65,69%	64,80%	68,34%	66,52%
	80:20	61,35%	64,67%	63,60%	68,57%	65,99%
	70:30	62,20%	64,38%	64,96%	61,82%	63,35%
	60:40	62,78%	63,85%	62,73%	68,34%	65,41%
	50:50	63,09%	63,22%	62,76%	64,37%	63,55%

Based on Table 6, testing of test data produces the highest accuracy also on the trigram features with a ratio of 90:10 that is equal to 68.60%.

D. Classification Test Result with Weighting

The results of system performance testing using the Convolutional Neural Network classification method and the TF-IDF weighting feature can be seen in Table VII.

TABLE VII
CLASSIFICATION TEST RESULT WITH WEIGHTING

Feature N-gram	Data train : data test	Loss	Acc	Prec	Recall	F-1 Score
UNIGRAM	90:10	60,64%	65,65%	63,74%	69,68%	66,58%
	80:20	61,01%	65,53%	62,99%	72,36%	67,35%
	70:30	61,43%	65,18%	65,59%	62,67%	64,10%
	60:40	62,06%	64,54%	62,49%	70,52%	66,26%
	50:50	62,14%	64,74%	65,06%	62,54%	63,78%
BIGRAM	90:10	60,13%	65,96%	64,20%	66,34%	65,25%
	80:20	61,11%	64,95%	64,35%	54,88%	59,24%
	70:30	62,82%	63,12%	61,45%	60,27%	60,85%
	60:40	62,24%	64,10%	63,42%	61,82%	62,61%
	50:50	63,37%	62,87%	58,87%	69,85%	63,89%
TRIGRAM	90:10	58,81%	66,62%	62,21%	55,86%	58,86%
	80:20	58,52%	**68,94%**	63,79%	47,80%	54,65%
	70:30	59,01%	66,56%	59,64%	58,87%	59,25%
	60:40	60,05%	66,24%	62,15%	50,67%	55,83%
	50:50	59,97%	66,02%	65,51%	38,70%	48,66%
UNI + BI	90:10	61,39%	65,48%	63,90%	70,82%	67,18%
	80:20	61,45%	65,48%	64,23%	68,68%	66,38%
	70:30	62,35%	64,05%	62,91%	67,74%	65,24%
	60:40	62,89%	65,81%	63,36%	60,72%	62,01%
	50:50	63,00%	63,32%	62,41%	66,27%	64,28%
BI + TRI	90:10	60,48%	65,91%	58,37%	71,95%	64,45%
	80:20	61,23%	65,76%	60,28%	68,75%	64,24%

	70:30	62,10%	65,09%	60,41%	68,73%	64,30%
	60:40	62,97%	64,30%	60,17%	55,63%	57,81%
	50:50	63,66%	62,28%	57,67%	63,81%	60,58%
UNI + BI + TRI	90:10	61,26%	64,98%	61,66%	76,30%	68,20%
	80:20	61,91%	64,42%	60,94%	77,84%	68,36%
	70:30	62,02%	65,86%	62,47%	67,42%	64,85%
	60:40	62,15%	64,54%	63,60%	66,80%	65,16%
	50:50	62,99%	63,44%	61,55%	72,41%	66,54%

Based on Table 7, testing of test data produces the highest accuracy also on the trigram features with a ratio of 80:20 that is equal to 68.94%.

E. Classification Test Result with Weighting and Selection Feature

Next is the result of system performance testing using the Neural Network Convolutional classification method with TF-IDF weighting feature and Information Gain feature selection. Tests are carried out based on thresholds with values 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. The test results show that a threshold value of 0.8 produces the highest accuracy, so a threshold value of 0.8 will used for the next test. The result of the test can be seen in Table VIII.

TABLE VIII
CLASSIFICATION TEST RESULT WITH WEIGHTING AND FEATURE SELECTION

Feature N-gram	Data train : data test	Loss	Acc	Prec	Recall	F-1 Score
UNIGRAM	90:10	32,59%	89,22%	91,12%	75,62%	82,65%
	80:20	33,57%	88,62%	93,57%	69,09%	79,49%
	70:30	31,79%	90,40%	95,48%	77,34%	85,46%
	60:40	39,51%	83,81%	98,50%	59,65%	74,30%
	50:50	35,19%	90,16%	97,72%	74,10%	84,29%
BIGRAM	90:10	33,07%	89,25%	50,00%	45,00%	47,37%
	80:20	40,30%	85,41%	50,00%	43,00%	46,24%
	70:30	38,37%	87,41%	50,00%	44,00%	46,81%
	60:40	35,59%	88,65%	50,00%	44,00%	46,81%
	50:50	40,31%	86,15%	50,00%	43,00%	46,24%
TRIGRAM	90:10	52,18%	81,82%	50,00%	40,91%	45,00%
	80:20	61,43%	66,67%	43,75%	36,84%	40,00%
	70:30	42,77%	83,87%	57,14%	66,67%	61,54%
	60:40	54,25%	76,19%	50,00%	38,10%	43,25%
	50:50	56,72%	75,00%	50,00%	37,50%	42,86%
UNI + BI	90:10	27,36%	94,74%	96,08%	96,09%	96,08%
	80:20	41,49%	78,31%	96,54%	64,20%	77,12%
	70:30	22,45%	95,07%	94,26%	94,37%	94,31%
	60:40	19,21%	95,50%	95,51%	95,01%	95,26%
	50:50	38,33%	**95,56%**	96,65%	94,20%	95,41%
BI + TRI	90:10	2,82%	90,10%	54,55%	0,91%	1,79%
	80:20	42,96%	86,07%	50,00%	0,33%	0,66%
	70:30	41,45%	86,75%	55,56%	0,97%	1,91%
	60:40	40,36%	85,57%	49,57%	43,11%	46,11%
	50:50	42,54%	85,06%	49,77%	42,70%	45,96%
UNI + BI + TRI	90:10	29,59%	92,31%	92,31%	100,00%	96,00%
	80:20	32,18%	90,29%	90,29%	100,00%	94,90%
	70:30	34,53%	89,61%	89,15%	100,00%	94,26%
	60:40	37,35%	88,29%	87,78%	100,00%	93,49%
	50:50	33,65%	89,45%	89,45%	100,00%	94,43%

Based on Table 8, testing of test data produces the highest accuracy on the uni + bi features with a ratio of 50:50 that is equal to 95.56%.

The features that influence the hoax distribution that are obtained from the highest occurrence frequency values can be seen in Table IX.

TABLE IX
LIST OF INFLUENTIAL FEATURES

Unigram	Bigram	Trigram
jokowi	bakar hutan	nkri harga mati
corona	virus corona	bakar hutan lahan
presiden	presiden Jokowi	sebar virus corona
jakarta	anies baswedan	peta sebar corona
virus	gubernur dki	akibat bakar hutan

F. Analysis of Testing Results

The results of the accuracy comparison of the three scenarios shown in Figures 5 and 6.

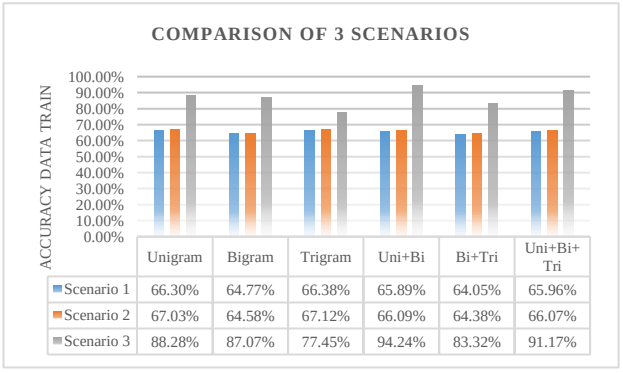

Fig. 5. Comparison of 3 Scenarios Results (Data Train)

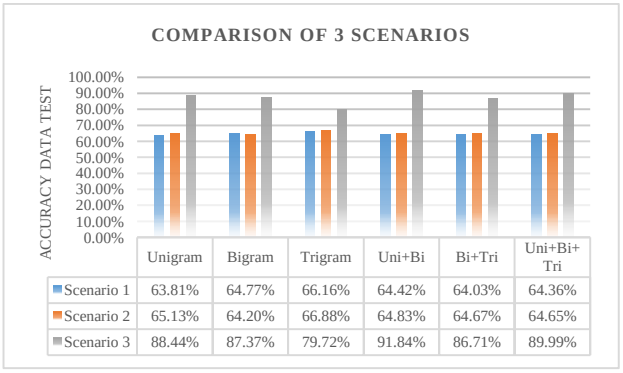

Fig. 6. Comparison of 3 Scenarios Results (Data Test)

To see the accuracy value in general, the accuracy values in each data comparisons are averaged, after that the n-gram values are also averaged to get the average accuracy values of each scenario. This is done to compare the three scenarios and measure how far the difference in value from the third accuracy scenario. The results show that scenario 3 gets the highest accuracy value which is 91.84% on the unigram + bigram feature. It can be seen that feature selection has an effect to increase the value of accuracy, because the information gain feature selection can take the best features that have important information in classification so that the performance of prediction models increases. In addition to feature selection, the use of n-gram can also improve the value of accuracy that is proven in scenario 1 and 2, the highest accuracy is obtained on the trigram feature and in scenario 3 the highest accuracy value is obtained from the unigram + bigram feature. Although the

highest accuracy value lies in the different ratio of data comparison for each scenario, the average value shows that the 90:10 ratio is the best ratio. The system is also proven to be able to produce a pretty good model, because the accuracy value generated by the training data is not much different from the accuracy value generated by the test data as shown in Figures 5 and 6.

V. CONCLUSION

In this research, it can be concluded that the Convolutional Neural Network (CNN) method can properly classify hoax news. The use of the TF-IDF weighting and Information Gain feature selection also greatly affects the results of the classification because the test obtained the highest accuracy value of 95.56% on the unigram + bigram feature with a ratio of training data and test data 50:50. The average accuracy also increased significantly by 22.75% compared to testing without using TF-IDF weighting and Information Gain feature selection. In addition, the n-gram feature also affects the classification, as evidenced by the highest accuracy value found in the unigram + bigram feature.

The author's suggestion for further research is that it should work towards for making a better dictionary in the preprocessing stage so as to reduce unclear words in order to make the system built optimally. In addition, because there are so many parameters on CNN, it is recommended to optimize the parameter tuning process in order to produce a better accuracy.

ACKNOWLEDGMENT

The authors would like to thank Allah SWT, parents who always support, lectures who always guide and give direction, and all the author's friends who always accompany and help in this research.

REFERENCES

[1] 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]
[2] C. Juditha, "Hoax Communication Interactivity in Social Media and Anticipation (Interaksi Komunikasi Hoax di Media Sosial serta Antisipasinya)," J. Pekommas, vol. 3, no. 1, p. 31, 2018.
[3] J. J. S. S. M. U. B. Kariaman Sinaga, "Pelatihan Meminimalisir Efek Hoaks Media Sosial di Desa Namo Sialang Kecamatan Batang Serangan Kabupaten Langkat, Sumatera Utara." E-Dimas: Jurnal Pengabdian kepada Masyarakat, 10(2), 150-159, 2019.
[4] Okezone, (2017, Mei). 7 ciri berita hoax. [Online]. Dipetik Agustus 21, 2018. Available at: https://news.okezone.com/read/2017/05/02/337/1680830/7-ciri-berita-hoax- seperti-ini-lho. [Accessed 20 September 2019]
[5] E. B. S. Laode Mohammad Ikhsan, "Deteksi hoax pada twitter menggunakan metode Decision Tree dan Analytical Hierarchy Process.", *Open Library Telkom*, 2019.
[6] E. B. S. Z. K. A. Achmad Fauzi, "Hoax News Detection on Twitter using Term Frequency Inverse Document Frequency and Support Vector Machine Method," J. Phys. Conf. Ser., vol. 1192, no. 1, 2019.
[7] K. P. H. B. Marin Vuković, "An intelligent automatic hoax detection system," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 5711 LNAI, no. PART 1, pp. 318–325, 2009.
[8] A. D. T. M. Jeqwalin Claudya, "Klasifikasi Spam Pada Email Menggunakan Algoritma Convolutional Neural Network." *Open Library Telkom*, 2019.
[9] I. W. S. E. Putra, "Klasifikasi citra menggunakan convolutional neural network (CNN) pada caltech 101" Doctoral dissertation, Institut Teknologi Sepuluh Nopember, 2016.
[10]R. R. I. D. E. Y. A. A. S. M. A. A. S. Agung. B. Prasetijo,"Hoax detection system on Indonesian news sites based on text classification using SVM and SGD," Proc. - 2017 4th Int. Conf. Inf. Technol. Comput. Electr. Eng. ICITACEE 2017, vol. 2018-Janua, pp. 45– 49, 2018.
[11] I. Y. R. P. D. M. R. Faisal Rahutomo, "Eksperimen Naïve Bayes Pada Deteksi Berita Hoax Berbahasa Indonesia," J. Penelit. Komun. dan Opini Publik, vol. 23, no. 1, pp. 1–15, 2019.
[12]F. N. A. B. Kemas Muslim Lhaksmana, "Klasifikasi Pengguna Media Sosial Twitter Dalam Persebaran Hoax Menggunakan Metode Backpropagation Classification of Users Social Media Twitter in the Hoax Spread," vol. 4, no. 2, pp. 3082–3090, 2017.
[13]Torunoğlu, Dilara, et al. "Analysis of preprocessing methods on classification of Turkish texts." 2011 International Symposium on Innovations in Intelligent Systems and Applications. IEEE, 2011.
[14]D. N. I. G. S. I. N. Chandra, "Klasifikasi Berita Lokal Radar Malang Menggunakan Metode Naïve Bayes Dengan Fitur N-Gram". Jurnal Ilmiah Teknologi Informasi Asia, 10(1), 11-19, 2016.

[15]A. S. H. R. H. Maulida Indah, “Seleksi Fitur Pada Dokumen Abstrak Teks Bahasa Indonesia Menggunakan Metode Information Gain,” JSM (Jurnal SIFO Mikroskil), vol. 17, no. 2, pp. 249–258, 2016.
[16]M. A. A. K. Much. Rifqi Maulana, “Information Gain Untuk Mengetahui Pengaruh Atribut,” J. Litbang Kota Pekalongan, vol. 9, 2015.
[17]H. H. Erlyn Nour Arrofiqoh, “Implementasi Metode Convolutional Neural Network Untuk Klasifikasi Tanaman Pada Citra Resolusi Tinggi,” Geomatika, vol. 24, no. 2, p. 61, 2018.
[18]Dharmadi, R. 2018. Mengenal Convolutional Layer Dan Pooling Layer [online] available at: https://medium.com/nodeflux/mengenal-convolutional-layer-dan-pooling-layer-3c6f5c393ab2 [accessed 29 September 2019].
[19]D. Kefin Pudi. Implementasi Deep Learning Menggunakan Convolutional Neural Network untuk Klasifikasi Citra Candi Berbasis GPU. Diss. UAJY, 2017.
[20]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.