Hoax Detection System on Indonesian News Sites Based on Text Classification using SVM and SGD

Agung B. Prasetijo, R. Rizal Isnanto, Dania Eridani
Department of Computer Engineering
Faculty of Engineering, Diponegoro University
Semarang, Indonesia
{agungprasetijo, rizal, dania}@ce.undip.ac.id

Yosua Alvin Adi Soetrisno, M. Arfan, Aghus Sofwan
Department of Electrical Engineering
Faculty of Engineering
Semarang, Indonesia
yosua@live.undip.ac.id, {arfan, aghus}@elektro.undip.ac.id

***Abstract*— A deliberate falsehood intentionally fabricated to appear as the truth, or often called as hoax (hocus to trick) has been increasing at an alarming rate. This situation may cause restlessness/anxiety and panic in society. Even though hoaxes have no effect on threats, however, new perceptions can be spread that they can affect both the social and political conditions. Imagery blown from hoaxes can bring negative effects and intervene state policies that may decrease the economy. An early detection on hoaxes helps the Government to reduce and even eliminate the spread. There are some system that filter hoaxes based on title and also from voting processes from searching processes in a search engine. This research develops Indonesian hoax filter based on text vector representation based on Term Frequency and Document Frequency as well as classification techniques. There are several classification techniques and for this research, Support Vector Machine and Stochastic Gradient Descent are chosen. Support Vector Machine divides a word vector using linear function and Stochastic Gradient Descent divides a word vector using non-linear function. SVM and SGD are chosen because the characteristic of text classification includes multidimensional matrixes. Each word in news articles can be modeled as feature and with Linear SVC and SGD, the feature of word vector can be reduced into two dimensions and can be separated using linear and non-linear lines. The highest accuracy obtained from SGD classifier using modified-huber is 86% over 100 hoax and 100 non-hoax websites which are randomly chosen outside dataset which are used in the training process.**

Keywords—Hoax; Support Vector Machine; Stochastic Gradient Descent

I. INTRODUCTION

Hoax is derived from the term (hocus to trick) and was created to manipulate people or invite people to perform an action, using threats or deceptions [1]. The motive of the hoax itself can be both commercial and political and can have a bad impact such as loss of reputation, material, even life-threatening. The faster the hoax news spreads the faster it will affect an existing community.

The dissemination of hoaxes, unfortunately, also supported by the social media that accelerate the spread of the news. Problems that occur at this time is the spread of news is indeed used as a means to make money so that hoaxes spread faster and uncontrolled. Every day there can be a new website spreading hoaxes, as many of the issues that occur in the world, such as elections, the spread of disease, disputes and so on. This research will help people in improving the efficiency of blocking hoax news because it will automatically detect hoax website using a machine learning approach.

Automatic hoax detection software design is done in several stages. The first stage is to automatically collect content using a crawler engine in a hoax news portal also in trusted news portal to gain a dataset that can be labeled manually. The next stage is to parse the content so that it can form the content become the features which represent the characteristic of hoax news. Using a classifier, the vectorized feature is transformed into model in training process using statistical algorithm such as SVM and SGD. Once detected as a hoax news, the address of the contents will be automatically stored into the database for then being collected and become consideration whether the whole site is a news hoax spreader.

The initial stage of text classification is pre-processing. Preprocess is the process of removing words that are commonly used (stop words) and also stemming to convert it into a basic word. The set of words will be tokenize into several n-gram in the form of bigram, then be used as a feature of text representation model such as information gain, mutual information, chi square, and TF-IDF [2]. TF-IDF is chosen as text vector representation because it performs well to increase recall and precision [3].

This system is implemented using Python programming language and for classifier model, such as Support Vector Machine and Stochastic Gradient Descent, has been developed from scikit-learn library. SVM technique using linear kernel [4] could perform better when training data has a larger feature dimensions. Linear kernel is suitable for a website which have a large number of sentences in one document.

Stochastic Gradient Descent is also chosen as a comparison classifier as this also performs well with sparse and highly dimensional data. SGD outperforms on precision and keep high recall, which mean that model running well on high priority article in addition as 'fake news' [5]. To gain the best result, then performance of two classifiers is compared. This paper is organized into the followings: Section 2 discusses related works in hoax detection system, and Section 3 reveals the methodology used for classification system. This is followed by results of the

978-1-5386-3947-4/17/$31.00 ©2017 IEEE

experiments found in Section 4. This paper ends with Conclusions (Section 5).

II. Related Works

Various research have been developed to classify hoax texts. Text classification research begins with an email classification that contains a hoax [1]. Text classification technique where used is a combination of unsupervised learning such as self-organizing map (SOM) and supervised learning such as Support Vector Machine (SVM). SVM method is suitable to classify emails into two categories: hoax and not hoax whereas SOM method is suitable to classify hoax news into several categories because it is focused on similarity level of input patterns.

Vukovic's research [1] has the disadvantage of not being able to classify new words from hoaxes email that have not been processed before. This is a problem to be solved in this research by using variation of news considered to be hoax although originated from a reputable source. SVM is also tested and compared with SGD which concrete loss function smoothed with modified-huber.

There are some advanced research using news media, social media, news commentary, and also wikipedia. Different media distributions lead to differences in characteristics, so they require different classification techniques. In Rubin's research [6] there are several linguistic features that become the characteristics of hoax news: levels of vagueness, humor, grammar, negative properties, and punctuation.

The linguistic features vector is then combined with the word vector in the form of TF-IDF (Term Frequency – Inverse Document Frequency) for later use in the classification process. The algorithm used in Rubin's research [6] is an SVM with linear kernel. The data used in the training process should be proportional so there is no information shortage about hoax news as well as real news.

The disadvantages of Rubin's research [6] is that when the tested news contains negative sentiment, the accuracy increases, but when the tested news using positive words, the accuracy is not increased. The Indonesian characteristics of the news will be different because terms of hoax vocabulary and the use of sentences are also different.

The research for hoax in Indonesian is done by Rasywir [2]. The experiment is to select the best technique by trying to combine several feature selection techniques using a combination of sections and vectors. The best feature selection technique used is a combination of mutual information and information gain because in Indonesia news there is no pattern of news that can be identified.

The vector in the information gain uses the node of the feature to find the effect of the feature on class diversity. Vectors in mutual information show how much information there, and is about the existence of a word so that it can be used to make classification decision [2]. After a word vector is combined, three classification algorithms – SVM, Naïve Bayes, and C.45 is tested and compared. Rasywir's research [2] shows that Naïve Bayes has the best accuracy among the other two algorithms.

The shortcomings of Rasywir's research [2] indicate that there is a failure of hoax classification due to error of classification when using spread up topics compared with using a particular topic. It means that when using a particular topic as a reference the similarity level is low and makes the news wrongly classified.

The contribution of this research is to improve the weaknesses that occur in Rubin's research [6] and Rasywir [2]. This research adds linguistic features by using complete content of news website in addition to the use of vector word features and will test it with SVM and SGD and see if it will produce better accuracy on a common topic as well as a specific topic.

III. Methodology

A. Tools and Materials

Tools and materials used in this research were Python version 2.7, Scrapy for web spider crawler, Cloud Flare for javascript agent, Beautiful Soup version 4.3.2 to scrape the content and HTML tag and extract the text without punctuation, and Scikit-Learn (machine-learning library) version 0.16.1 to conduct vectorization process and building classifier model with both linear SVM and SGD methods. There were 680 pages collected composed of 180 hoax news website and 500 real news website. Categories used are widespread, between politic, economy, sport, entertainment, and technology. For testing purposes, 100 hoax news website and 100 real news website are used.

B. Text Classification

Fig. 1 illustrates the main diagram of text classification used in this research. The classification process was divided into two processes, training process and testing process.

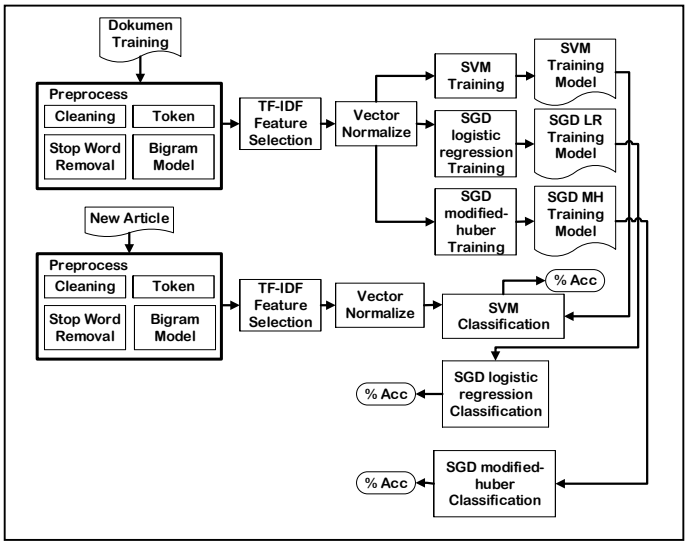

Fig. 1. The main diagram of text classification

B.1. Preprocessing

Preprocessing is the initial phase of text classification to form unstructured texts into token representation and ready to be modeled by a classifier algorithm. There are three phases done in this process: cleaning, tokenizing, and stop word removal.

Cleaning is the process to remove extra characters in HTML syntax used to design website. Tokenizing is the process of extracting serialized word in sentences to become variation of two words in many combinations, done at each sentence.

B.2. Term Weighting Method

TF-IDF (Term Frequency – Inverse Document Frequency) is the most preferred weighting method. Features are weighted locally using tfi, formula and global weighting among document collection using $idf.tfi$, term is a weight gained from the appearance frequency of word I in document j. $idfi$ term is a weight gained by deliberate number of word i (dfi) appear in all N documents.

$$idf_i = \log (N / df_i) \quad (1)$$
$$w_{i,} = tf_{i,j}\ x\ idf_i \quad (2)$$

TF-IDF has some weak points in a classification process. TF-IDF is not considered in certain category because it does weighting process globally. Information of word appearance in certain category influences the consideration to be belonging in that category. Classification algorithm will choose category which contains larger number of sample in training process when compared with keyword consideration. This research compares which classification algorithm better to be combined with TF-IDF vectorization.

B.3. Support Vector Machine

SVM is a supervised classification algorithm works well for text classification which has a large input dimension based on text as features. Text document has little features that are irrelevant, unique word vector because related words in a sentence can be different, but most of the text categorization problem can be separated linearly [7]. The main idea of SVM algorithm is to build hyper plane that can divide the area into several subsets. Hyper plane is like a road that separate two category and using consideration of the closest feature distance from the hyper plane.

SVM vector is trained with vector from two difference classes. Training data is shown with $x_1,x_2,x_3,...,x_n$ and class labels is shown as $y_1,y_2,y_3,\ldots,y_n$. Two classes of data are denoted as $\{x_i,y_i\}$, where $x_i,y_i \in \{-1,1\}$. Linear classification generates the weight vector w with sign function (w^Tx) not with sigmoid function. SVM does not use pure probability values such as Naïve Bayes but will use margin or distance vector document as the value of the truth. The farther away a testing point from the hyper plane, the higher probability of that point can be classified. Formula of the conditional probability of regression functions is defined by:

$$P(y|x) = \frac{1}{1+e^{(-yw^Tx)}} \quad (3)$$

where, y = ± 1, x is data, y is class label, and w ϵ Rn is a weight vector, ^{T} is hyper-parameter which parameter of prior distribution.

B.4. Stochastic Gradient Descent

Gradient descent is one mostly used algorithm that can offer new perspective for solving problems. Gradient descent is algorithm to minimize functions [8]. When give a function that defined by a set of parameters, gradient descent begins with an initial set of parameter values and makes iteration to move toward set of parameter values that find minimal point for the function. Minimization process uses derivation in calculus to find aligned line that approaching the minima.

Gradient descent can be slow to run on very large datasets. One iteration of gradient descent algorithm requires a prediction of each instance in the training dataset, it can take a long time when have millions of instances.

Stochastic gradient descent is a little bit different because the coefficient update occurs only when training process is running [9]. The update procedure for the coefficient is the same as Gradient Descent, except the cost is not summed over all training patterns, but only calculated for one training pattern.

The algorithm process of stochastic gradient is to choose θ to minimize $J(\theta)$. A search algorithm is used to make some initial guess for θ, and change repeatedly the value of θ to make output from J is minimal [10]. The update process which repeatedly done in SGD is formulated as follows:

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \quad (4)$$

The update is simultaneously performed for all values of j= 0,…,n. α is called learning rate.

Partial derivative is used on the right hand side. With example of one training example (x,y), the sum of definition of J can be ignored. By using power rule and chain rule it can be obtained:

$$\begin{aligned}\frac{\partial}{\partial \theta_j} J(\theta) &= \frac{\partial}{\partial \theta_j} \frac{1}{2} (h_\theta(x) - y)^2 \\ &= 2 \frac{1}{2} (h_\theta(x) - y) \frac{\partial}{\partial \theta_j} (h_\theta(x) - y) \\ &= (h_\theta(x) - y) \frac{\partial}{\partial \theta_j} (\sum_{i=0}^{n} \theta_i x_i - y) \\ &= (h_\theta(x) - y)\, x_j \end{aligned} \quad (5)$$

Where, $h_\theta(x)$ is guess and y is output so SGD is also known by the least mean square of the calculated error. Error is come from the shifted value of gradient from x and also constant. For single training example this gives an update rule:

$$\theta_j := \theta_j + (y^{(i)} - (h_\theta(x^{(i)}))x_j^{(i)} \quad (6)$$

IV. Experiment Results

The performance of the system using SVM classifier will be compared with SGD technique using two kernels: linear regression and modified-hurbe. We have tested each algorithm with 200 websites consist of 100 hoax news websites and 100 real news websites which are collected outside the dataset. There are some false positive different between SVM, SGD

linear regression, and modified-hurbe. SGD linear regression have 48 wrongly hoax news detected as real news, SVM have 38 wrongly hoax news detected as real news, SGD modified-hurbe has 28 wrongly hoax news detected as real news. Comparison of accuracy, precision, and recall of each classifier are shown in Fig. 2.

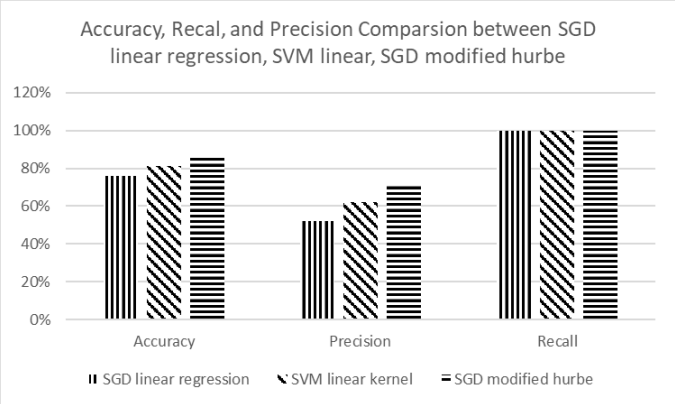

Fig. 2. Performance Measures between Algorithms

Accuracy shows that performance to detect real news is better than to detect the hoax news. Based on the result in Fig. 2, it can be seen that SGD with modified-hurbe outperformed both SVM linear and SGD with linear regression. SGD with modified-hurbe smoothed the loss caused by hyper plane or divider line so it can detect hoax more precisely.

F-measure is a harmonic mean of precision (P) and recall (R). F-measure used to compare experimental result of each classification algorithms because accuracy does not have a meaning so much if the precision and recall is not balance. Hence, ß=1 as parameter is used to gain balance between precision and recall because the aim is to gain higher precision and recall [11]. Fig. 3 shows the f-measure comparison of classifier algorithm.

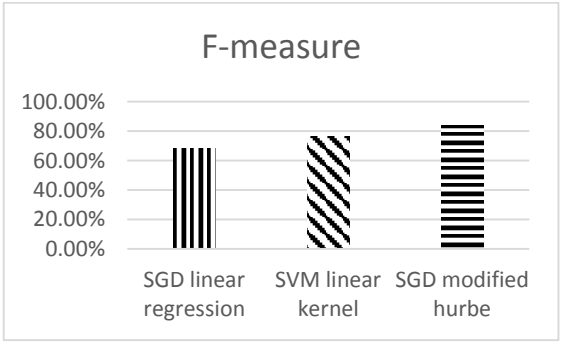

Fig. 3. F-Measure between Algorithms

Over-blocking is also an important point of defining error of the system. High value of recall in detecting non-hoax category is a thing that can be connected to over-blocking issue. Classifier is better to detect wrongly hoax than real news detected as hoax. Wrong detection can occur because the unique condition of hoax news. Sometimes, trusted sources also write hoax news based on public opinion.

Supervised algorithm can predict more precisely when there is a lot of example. Example of a unique case, news from a trusted site detected as hoax news is key to give a knowledge to the classifier. Writing pattern at some news portal is unique and can become a dataset.

In an English article there are some unique keywords that can be rearranged as key terms of text classification. Text classification an English news uses only news title not the overall content [12]. In an Indonesian news sometimes, a hoax news uses fact for news title and in another time uses opinion words. For gaining insight, overall sentences in a news site must be used in dataset to gain better understanding of hoax context.

V. Conclusions

In this research, classification system on Indonesian hoax sites based on sentence feature is built with Python programming language. This research tries to testing the sentence features and SVM as linear function to separate the content outperformed the others. Algorithm chosen is SGD with two kernel variations to gain best result.

Based on the result, it can be concluded that using SGD with modified-huber kernel increases the accuracy and precision of SVM for about 4% and 20% respectively. The accuracy of TF-IDF is better combined with SGD when the sample of hoax classification does not have specific terms, instead, it has a unique pattern in the same news portal provider. For further research, the combination of text feature, sentiment analysis, and also voting based on search engine [13] can be done to overcome many patterns of hoax news especially which do not have a specific term.

Acknowledgment

This work was supported by The Ministry of Research, Technology and Higher Education, Republic of Indonesia.

References

[1] Y. Y. Chen, S.-P. Yong, and A. Ishak, “Email Hoax Detection System Using Levenshtein Distance Method,” J. Comput., vol. 9, no. 2, Feb. 2014.

[2] E. Rasywir and A. Purwarianti, “Eksperimen pada Sistem Klasifikasi Berita Hoax Berbahasa Indonesia Berbasis Pembelajaran Mesin,” J. Cybermatika, vol. 3, no. 2, 2016.

[3] Y. Zhang, L. Gong, and Y. Wang, “An improved TF-IDF approach for text classification,” J. Zhejiang Univ. Sci., vol. 6, no. 1, pp. 49–55, Jan. 2005.

[4] D. Michie, D. J. Spiegelhalter, and C. C. Taylor, Eds., Machine learning, neural and statistical classification. New York: Ellis Horwood, 1994.

[5] L. Dyson and A. Golab, “Fake News Detection Exploring the Application of NLP Methods to Machine Identification of Misleading News Sources,” CAPP 30255 Adv. Mach. Learn. Public Policy.

[6] V. L. Rubin, N. J. Conroy, Y. Chen, and S. Cornwell, “Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News.,” in Proceedings of NAACL-HLT, 2016, pp. 7–17.

[7] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” J. Mach. Learn. Res., vol. 12, no. Oct, pp. 2825–2830, 2011.

[8] M. Nedrich, “An Introduction to Gradient Descent and Linear Regression.”,<URL: https://spin.atomicobject.com/2014/06/24/gradient-descent-linear-regression/> Accessed October 1, 2017

[9] J. Brownlee, “Gradient Descent For Machine Learning.” <URL: https://machinelearningmastery.com/gradient-descent-for-machine-learning/> (Accessed October 1, 2017)

[10] A. Ng, “CS229 Lecture notes,” CS229 Lect. Notes, vol. 1, no. 1, pp. 1–3, 2000.

[11] Y. Sasaki and others, “The truth of the F-measure,” Teach Tutor Mater, vol. 1, no. 5, 2007.

[12] S. Bajaj, “‘The Pope Has a New Baby!’ Fake News Detection Using Deep Learning.”

[13] N. J. Conroy, V. L. Rubin, and Y. Chen, “Automatic deception detection: Methods for finding fake news,” Proc. Assoc. Inf. Sci. Technol., vol. 52, no. 1, pp. 1–4, 2015.