Hoax Web Detection For News in Bahasa Using Support Vector Machine

Muhammad Abdillah Rahmat[1], Indrabayu[2], Intan Sari Areni[3]
[1]*Postgraduate Student of Electrical Engineering Department Hasanuddin University*
[2]*Informatics Engineering Department Hasanuddin University*
[3]*Electrical Engineering Department Hasanuddin University*
Makassar, Indonesia
abdirahmat22@gmail.com, {indrabayu, intan}@unhas.ac.id

***Abstract*— This research creates a web-based user-friendly system that aims to detect hoax and non-hoax news of Indonesian language news links. The input data is in the form of links and archive sites from the *Forum Anti Fitnah Hasut dan Hoax* (FAFHH), using 100 news for training data and 20 news for test data that is processed by crawling and then processed in the pre-processing phase, namely tokenizing, stop word and stemming. Next is the Term Frequency-Inverse Document Frequency (TF-IDF) stage to provide weighting data which will be input data at the classification stage using the Support Vector Machine (SVM) Algorithm with a linear kernel to detect hoax and non-hoax news. The experimental results show that the system can classify well with an accuracy of 85%.**

Keywords—*Text Mining, Hoax, Crawling, Support Vector Machine, TF-IDF*.

I. INTRODUCTION

The internet has been a crucial factor in human daily activities. In 2017, "eMarketer" estimated that internet users in Indonesia would reach 112 million people, defeating Japan in the 5th place with slower growth in the number of internet users [1]. With such rapid development, the internet has turned into a source of information that is remarkably fast and convenience.

Information Technology in Indonesia is also growing rapidly where internet users in Indonesia currently have reached 132.7 million or 51.7% of Indonesia's population [2]. The development of communication technology has been accelerating the dissemination of information, which can lead to the spread of disinformation called hoax. Hoax is misguided and harmful information to influence readers opinion and perception. the vast number of uses and high traffic in social media has been used as a major medium in has developed.

The results of the survey on *Wabah Hoax Nasional* conducted by Mastel showed that the three highest hoax news distribution channels are social media, which Facebook is on the highest position 92.40%, chat applications 62.80%, and websites 34.90%. The Indonesian Ministry of Communication and Informatics stated that there are nearly 800 thousand sites that carry out Hoax news dissemination activities on the internet [3].

The existence of Text Mining technology supported by increasingly advanced technological devices can help solve complex problems. The problems faced by internet users of hoax news can be solved through pattern recognition and deep learning.

The development and innovation of research on hoax detection using article data and news links have been carried out by several researchers. Yanuar et al. used data from 250 news articles from websites that were given manual tags, which consisted of 155 valid news and 95 hoax news. The system was built using Naïve Bayes Algorithm as the classification method, and it gives accuracy of 78.6% [4]. Prasetijo et al. used 200 data from 3 websites consisting of 100 hoax news and 100 real news. The data was labeled manually and processed using Support Vector Machine and Stochastic Gradient Descent methods. The accuracy of SGD modified Huber is 86% [4].

Research by Vukovic et al. used Artificial Neural Network method to classify email data into 3 classes, namely definite hoaxes, suspected hoaxes, and not hoaxes [6]. Granik et al. detected fake news using Naive Bayes and data from Facebook posts and show up with accuracy of 77% [5].

Gilda et al. evaluated several Algorithms for detecting false news including Support Vector Machines (SVM), Stochastic Gradient Descent (SGD), Gradient Boosting, Bounded Decision Trees and Random Forests using data obtained from OpenSources.co by applying TF-IDF from bi-gram detection and PCFG to the corpus about 11,000 articles. The best accuracy of 77.2% was obtained using the SGD Algorithm [6].

Meanwhile, Santoso et al. only proposed a system that used the Social Media Application Program Interface (API) such as news, Facebook, twitter to find hoax posts on other social media or news websites and also the authenticity of posts from users. The system also used websites, newspapers and popular social media to determine whether the posting is hoax or valid comment/news [7].

In this research, a web-based user-friendly system is built that can work on various platforms to detect hoax news links in Indonesian. Crawling technique is used to obtain data input from target links. The collected data are then further classified as hoax or non-hoax. Then it passes through the pre-processing process with the stages of tokenization, stop words, and stemming. Post weighting process is conducted using TF-IDF for weighting classification. The result will be used as data input for Support Vector Machine (SVM) with a linear kernel.

This paper is organized into several sections: section II discusses the stages of text pre-processing and algorithms used in hoax detection systems. Section III discusses system testing and the results of the hoax detection system. This paper ends with a conclusion in section IV.

II. PROPOSED METHOD

A hoax web-based detection is proposed in this paper that focussing in bahasa Indonesia originating from the site https://turnbackhoax.id/. The data (news) which has reform from crawling results is then processed for denoising. The weighting process is carried out using TF-IDF which then computes through SVM Algorithm for classification. The System Diagram can be seen in Fig. 1.

Fig. 1. Diagram System

The input data used in this study are Indonesian news links originating from the archive *Forum Anti Fitnah, Hasut dan Hoax* (FAFHH) on the site https://turnbackhoax.id/. The interface of FAFHH website is shown in Fig. 2. Data used in this study consist of 100 news for the training data and 20 for the test data.

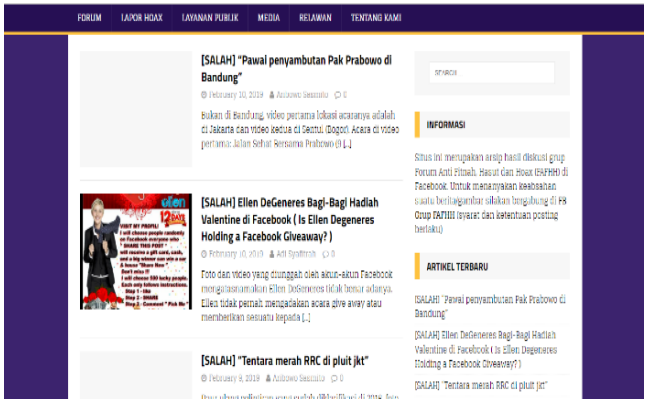

Fig. 2. Interface of site turnbackhoax.id

The news data are taken from the crawling process, where the URL to be visited is listed. The process is called as seed. The web crawler will visit the URL in the list and identify all hyperlinks on the page and add it to the list of URLs to be visited, called as the crawl frontier. Existing URLs are processed to produce information as needed. The HTTP GET protocol retrieves the needed information from the link that has to be crawled [8].

As for the preprocessing stage, the steps to be carried out consist of 3 parts:

a. Tokenizing

The tokenizing / parsing stage is the stage of cutting the input string based on each word that composes it. Non-letter characters are omitted and are considered as delimiters. Tokenizing usually used in the preprocessing stage so that the words in a document is divided into a few words according to the word delimiters predetermined divisor. Tokenizing is very useful when a text processing program requires data in a structured word and divided into arrays [9].

b. Stop word

Stop word is common words that appear frequently. Stop word removal is the process of removing words that are included in the stop word, usually done so that stemming becomes effective and efficient. Examples of Indonesian language stop word include "*yang*", "*di*", "*ke*", etc. [9].

c. Stemming

Stemming is a process to find a basic form of each word in a text document, and to reduce the number of different document index. Stemming also classifies other words that have the words and meanings of a similar base which has a similar shape or a different shape because they get a different affix to apply the rules of morphology Indonesia is good and right.

The stemming process is done by eliminating all good affixes consisting of prefixes, infixes, and suffixes. Stemming is based on the assumption that words that have the same term have the same basic meaning.

Stemming techniques can be categorized into 3, which are based on rules in certain languages, based on dictionaries, and based on shared appearance. One of the main objectives of the stemming process is to increase efficiency by sorting the contents of documents into small units that will become identifiers, for example in the form of words, phrases or sentences [10].

After the pre-processing stage, the process of calculating the weight of each word is the most common process used in information retrieval, which is called the TF-IDF method. This method will calculate the Term Frequency (TF) and Inverse Document Frequency (IDF) values for each token (word) in each document in the corpus. This method will calculate the weight of each token in the document using formula [11]:

$$idf_j = log(N/_{dfj}) \qquad (1)$$

$$w_j = tf_j \, x \, idf_j \qquad (2)$$

idf_j is the availability of a term in j document. N is total number of documents, df_j is a number of documents containing term, w is the weight of the term in a document, tf_j the number of occurrences of terms in documents obtained from the number of words j. So that the final weight of a term is to multiply both the tf_i x idf_i.

The final stage of the proposed system is to classify whether the news is a hoax or not using SVM Algorithm. SVM is a supervised classification algorithm that works well for text classification that has large input dimensions based on text as a feature. Text documents have a few irrelevant features, unique word vectors because related words in a sentence can be different, but most text categorization problems can be separated linearly. SVM Algorithm builds a hyperplane that can divide the area into several subsets. The hyperplane is like a road that separates two categories and uses consideration of the distance of the closest feature of the hyperplane. SVM vectors are trained with vectors from two different classes. Training data is indicated by labels *x1, x2, x3, ..., xn* and classes are displayed as *y1, y2, y3, ..., yn*. Two data classes are represented as {*xi, yi*}, where *xi, yi* ∈ {-1,1} [4].

If given a training vector, $x_i \in \mathbb{R}^\rho$, i=1,…., n, in two classes and vectors $y \in \{1,-1\}^n$, SVC (Support Vector Classification) solve the primal problem as follows.

$$\min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \zeta \mathrm{i}$$
$$\text{subject to } y_i(w^T \phi(x_i) + b) \geq 1 - \zeta_i, \qquad (3)$$
$$\zeta_i \geq 0, i = 1, \ldots, n$$

Its dual is

$$\min_{\alpha} \frac{1}{2} \alpha^T Q\alpha - e^T \alpha \quad \text{subject to } y^T \alpha = 0 \quad 0 \leq \alpha_i \leq C, i = 1, \ldots, n \tag{4}$$

where e is vector of all ones, $C > 0$ is the upper bound, Q is an n by n positive semidefinite matrix, $Q_{ij} \equiv y_i y_j K(x_i, x_j)$, where $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ is kernel. Training vectors are implicitly mapped into a higher dimensional space by the function ϕ [12].

The decision function is:

$$\text{sgn}\left(\sum_{i=1}^{n} y_i \alpha_i K(x_i, x) + \rho\right) \tag{5}$$

III. Experimental and Result

This study aims to detect hoax and non-hoax news using 100 news for training and 20 news for testing. The test data used in this study and the number of words in the tokenizing, stop word, and stemming stages, are shown in Table 1.

Tabel 1. Result Of Pre-processing Data

Document	Tokenizing	Stop word	Stemming
Data 1	3818	3130	2626
Data 2	2270	1464	1187
Data 3	2264	1703	1400
Data 4	697	556	449
Data 5	2754	2051	1588
Data 6	2384	2178	1828
Data 7	2986	1862	1501
Data 8	2043	1774	1486
Data 9	1313	935	713
Data 10	21865	15392	11919
Data 11	1879	1433	1139
Data 12	1434	1048	819
Data 13	4601	3380	2711
Data 14	1515	1197	964
Data 15	1986	1638	1309
Data 16	3813	2765	2174
Data 17	4136	3558	3051
Data 18	2016	1478	1223
Data 19	2072	1662	1294
Data 20	4088	2956	2426

The steps performed to detect deception in this study are described as follows.

1. Crawling

Information text is taken from the news link. An example of a crawling process can be seen in Fig. 3.

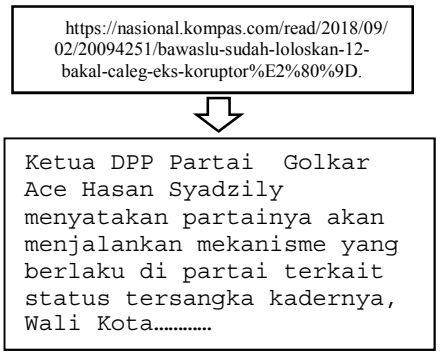

Fig. 3. Crawling Result

2. Tokenizing

The result of the process of crawling then become the input to the tokenizing stage. Example of tokenizing process can be seen in Fig. 4.

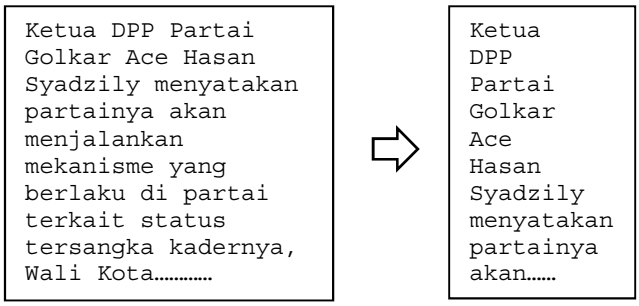

Fig. 4. Tokenizing Result

3. Stop word

The results of tokenizing process, then be input at this stage of the stop word. Removal of the words are based on the corpus NLTK (Natural Language Tool Kit). Examples of stop word processes can be seen in Fig. 5.

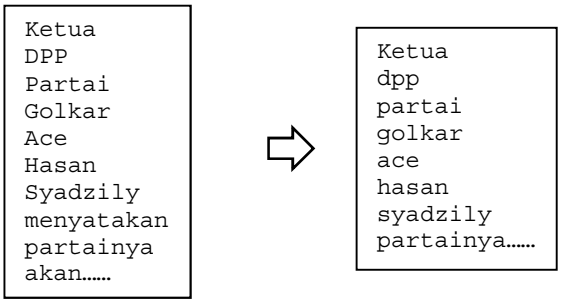

Fig. 5. Stop word Result

4. Stemming

The results of the stop word process then become the input to stemming process by eliminating affixes that consist of prefixes, inserts, and ends of words based on the corpus Sastrawi. The example of stemming process can be seen in Fig. 6.

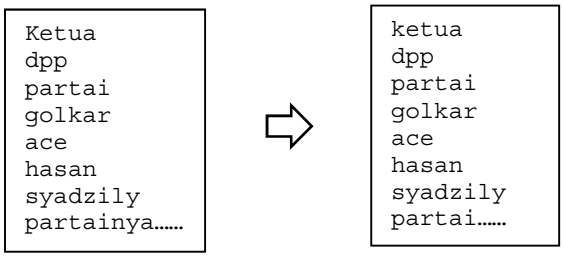

Fig. 6. Stemming Result

5. TF-IDF

The next process is to give weights to the words in a news story. The word's weight is greater for the term in a news that often appears in a document. Examples of TF-IDF process can be seen in Fig. 7.

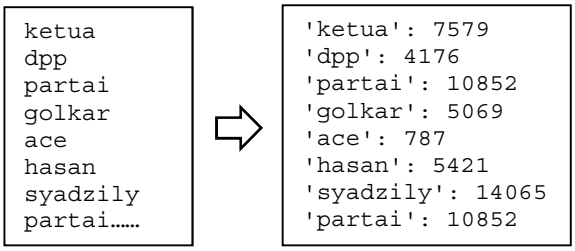

Fig. 7. TF-IDF Result

From the TF-IDF process, patterns are obtained from features and labels so that hoax news patterns and non-hoax news patterns are obtained. On the testing data, the classification process will be carried out using SVM method. Input data is the result of the TF-IDF process with linear kernel SVM.

The formula for calculating the accuracy (Ac) of the system is as follows.

$$Ac = \frac{Nc}{Nt} x\ 100 \qquad (6)$$

Where *Nc* is the amount of correctly detected data and *Nt* is the amount of overall data.

The results of the news classification process is shown in Fig. 8.

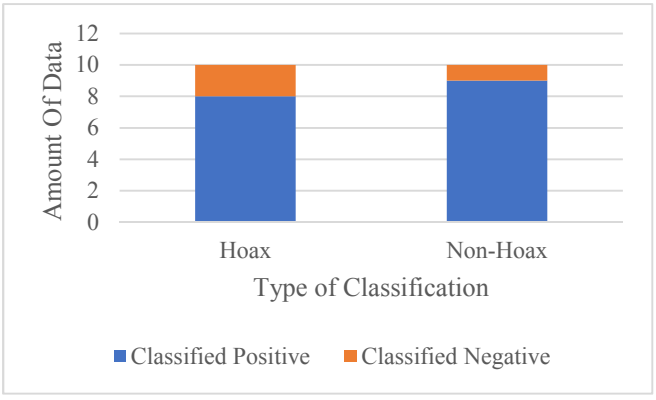

Fig. 8. System Test Results

In Fig. 8. the result of the hoax news classification that is detected correctly is as much as 80%, while the classification of non-hoax news detected is as much as 90%. In the news that is detected incorrectly on the testing data processed through crawling, there are tags that do not need to be processed or not informative. For example, the system still processed the function tag. The examples of classification error caused by unimportant words are shown in Fig. 9, marked in red.

```
saran anda konsumsi sayap ayam kaki saat
80 wanita milik fibroid rahim mudah kista
coklat ketika anda terima pesan apa yg
andalaku terus teman keluarga teman
perempuanatau end-chat skeptis ekspektasi
jika friend-list 851 kontak 10 orang yg
men-share ulang minimal 10 orang yg baca
snow ball effect tau kelipat sekian yg
tolong karena peduli indah bagi i vit vit
skybanner width 160px googletag cmd push
function googletag display div-gpt-ad-
1540809567840-0
```

Fig. 9. Classification Error caused by unimportant words

The average result of the system classification process proposed in this study is 85%.

CONCLUSIONS

The hoax news detection system has been carried out in this study with a dataset of 100 news for training data and 20 news for testing data. Input data are in the form of links that are processed by crawling. Furthermore, it is processed through pre-processing which consists of tokenizing, stop word, and stemming stages. In the final stage, word weighting is done using TF-IDF to be inputted to the classification process using SVM Algorithm with a linear kernel. The results of the classification of hoax detection system reached an accuracy rate of 85%.

The development of this research is to improve the crawling output results of a link, such as excluding unimportant words from the crawling process, as it affects the accuracy of hoax detection system. This system is intended only for Indonesian news and Indonesian websites because the corpus used is Indonesian and the characteristics from the Indonesian news website are different from the non-Indonesian website characteristics. The author hopes this research can be done in various languages.

REFERENCES

[1] P. KOMINFO, "Indonesia's Number Six World Internet User," *Official Website of the Indonesian Ministry of Communication and Information*. [Online]. Available: https://kominfo.go.id:443/content/detail/4286/pengguna-internet-indonesia-nomor-enam-dunia/0/sorotan_media. [Accessed: 01-Apr-2019].

[2] APJII, "BULETIN APJII (Association of Indonesian Internet Service Providers) ISSUE 05 November 2016." 05-Nov-2016.

[3] "There are 800 thousand Hoax Spreader Sites in Indonesia." [Online]. Available: https://www.cnnindonesia.com/teknologi/20161229170130-185-182956/ada-800-ribu-situs-penyebar-hoax-di-indonesia. [Accessed: 01-Apr-2019].

[4] A. B. Prasetijo, R. R. Isnanto, D. Eridani, Y. A. A. Soetrisno, M. Arfan, and A. Sofwan, "Hoax detection system on Indonesian news sites based on text classification using SVM and SGD," in *2017 4th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, Semarang, 2017, pp. 45–49.

[5] M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," in *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, Kiev, 2017, pp. 900–903.

[6] S. Gilda, "Evaluating machine learning algorithms for fake news detection," in *2017 IEEE 15th Student Conference on Research and Development (SCOReD)*, Putrajaya, 2017, pp. 110–115.

[7] I. Santoso, I. Yohansen, Nealson, H. L. H. S. Warnars, and K. Hashimoto, "Early investigation of proposed hoax detection for decreasing hoax in social media," in *2017 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*, Phuket, 2017, pp. 175–179.

[8] B. J. Wijaya and H. A. Santoso, "Crawling Of The Government Of Jawa Tengah Regional E-Government Website Based On Ontology," p. 8, 2015.

[9] R. J. Mooney, "Book Recommending Using Text Categorization with Extracted Information," *Proc. AAAI-98ICML-98 Workshop Learn. Text Categ. AAAI-98 Workshop Recomm. Syst.*, pp. 49–54, 2000.

[10] A. Dwiyoga Tahitoe and D. Purwitasari, "Implementation Of Modification Of Enhanced Confix Stemmer Stripping For Indonesian Using Corpus-Based Stemming Method," *Inst. Teknol. Sepuluh Nop.*, pp. 1–15.

[11] A. B. Prasetijo, R. R. Isnanto, D. Eridani, Y. A. A. Soetrisno, M. Arfan, and A. Sofwan, "Hoax detection system on Indonesian news sites based on text classification using SVM and SGD," in *2017 4th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, Semarang, 2017, pp. 45–49.

[12] S. Javadi, H. Moosaei, and D. Ciuonzo, "Learning Wireless Sensor Networks for Source Localization," *Sensors*, vol. 19, no. 3, p. 635, Feb. 2019.