BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **20**, No 1
Sofia • 2020 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2020-0006

An Indonesian Hoax News Detection System Using Reader Feedback and Naïve Bayes Algorithm

Badrus Zaman, Army Justitia, Kretawiweka Nuraga Sani, Endah Purwanti

Information System Study Program, Faculty of Science and Technology, Universitas Airlangga, Surabaya, 60115, Indonesia
E-mails: badruszaman@fst.unair.ac.id army-j@fst.unair.ac.id ekanuraga@gmail.com endahpurwanti@fst.unair.ac.id

Abstract**:** *Hoax news in Indonesia spread at an alarming rate. To reduce this, hoax news detection system needs to be created and put into practice. Such a system may use readers' feedback and Naïve Bayes algorithm, which is used to verify news. Overtime, by using readers' feedback, database corpus will continue to grow and could improve system performance. The current research aims to reach this. System performance evaluation is carried out under two conditions – with and without sources (URL). The system is able to detect hoax news very well under both conditions. The highest precision, recall and f-measure values when including URL are 0.91, 1, and 0.95 respectively. Meanwhile, the highest value of precision, recall and f-measure without URL are 0.88, 1 and 0.94, respectively.*

Keywords**:** *Hoax news, Indonesian news, Naïve Bayes, reader feedback.*

1. Introduction

Hoax is a malicious deception or a deceit made intentionally for malicious purposes [1]. It does not damage computer programs and operating systems, but it can destroy personal or brand image and reputation, shape public opinion, and even cause financial loss [2]. Considering the possible danger, hoaxes should be identified and classified [3]. In 2002, Yahoo email scam asked for their customers' credit card number through Yahoo PayDirect!, and requested payment for fake services [4]. Beside email, hoaxes spread through various media, such as instant messaging [5], internet chats, mobile messaging [5], social media [6], and even travel as news on a website [7].

Initial hoax detection systems were developed in the email domain [2, 3, 8, 9]. These systems receive email messages that are suspected of scamming. They compare these emails with scam emails stored in the database. There are several methods to detect e-mail hoaxes, among others, using the fuzzy logic expert approach, text similarity approach, and learning machine approach. To improve system reliability, databases need to be updated periodically.

Ishak, Chen and Yong [8] and Chen, Yong and Ishak [9] were able to identify patterns of hoax messages. However, unlike hoax messages, hoaxes on news sites do not have recognizable characteristics or patterns and do not have rigid language writing patterns [10], so readers will not be able to detect whether it is hoax news or not. A piece of news is considered as truth if there is someone who clarifies it.

Research on Indonesian hoax news has been extensive. Rasywir and Purwarianti [10] compared many types of hoax news classification methods using a machine learning approach. The results showed that Naïve Bayes yielded the best performance compared to either Support Vector Machine (SVM) algorithm or C4.5 algorithm. Prasetijo et al. [11] classified hoax news in Indonesia based on sentence features. Pratiwi, Asmara and Rahutomo [12] helped building Indonesian news datasets that are freely accessible for research purposes. Sirajudeen, Azmi and Abubakar [13] considered IP address validity and verified news source to determine whether news is hoax or not. Therefore, the database contains a list of valid IP addresses and verified news sources. Previous research in this area found it difficult to develop news hoax database collection.

In this paper, we propose a hoax detection system by using readers' feedback and text matching approach. Naïve Bayes was chosen for text matching because this method has advantages over other learning algorithms. Naïve Bayes is a simple [14] but accurate [12] learning technique in spite of false independent assumptions. Naïve Bayes learning does not require complex generalization processes [15]. It only needs to calculate the feature statistics per class by one time passing through the training document so it saves computing time [16]. Feedback from readers classifies documents that do not match any classes in the corpus, and at the same time it adds the collection of news in the corpus. The proposed system can also match the news' URL address with the trusted list of URLs and add the news' URLs to the database.

The rest of the paper is organized as follows. Section 2 reviews literature on hoax detection system. Section 3 describes the research methodology that includes system architecture and the type of document dataset. Section 4 reveals the results of the experiment. Section 5 compares our results with other studies. Chapter 6 concludes the paper.

2. Related works

Research in hoax was first developed in email domain, conducted by Hernandez et al. [14], Petkovic, Kostanjcar and Pale [3], Vukovic, Pripuzic and Belani [2], Ishak, Chen and Yong [8], and Chen, Yong and Ishak [9]. While hoax news research in Indonesian is still limited, there have been a few attempts to study the phenomenon. Table 1 shows the previous research and the results.

Indonesian hoax news research began in 2015. Rasywir and Purwarianti's experiment on several text classification methods, i.e., Naïve Bayes, SVM, and C4.5, have shown that probability-based feature selection (information gain, mutual information, and chi-square) is better than frequency-based feature selection (TF or TF-IDF).

Table 1. The previous research and the results [10-13, 17]

Domain	Research	Approaches / *Methods*	*Result*
Email hoax	[14], 2002	Heuristics Approaches and Traffic Analysis	Early automated hoax detection system
Email hoax	[3], 2005	Similarity Calculation : *Levenstein Distance* Classification Method : *Modified Nearest Neighbor* and *Fuzzy Logic*	Automated hoax email detector with simple intelligence (using database)
Textual Hoaxes (email, short message service (SMS), message on internet chats and forums	[2], 2009	Similarity Calculation : *N-Gram* Classification Method : *Self Organizing Map* (SOM)	Automated hoax email detector with an integrated intelligent system
Email Hoax	[8], 2012	Similarity Calculation: *Levenstein Distance* Classification Method: *Levenstein Distance*	Email hoax detection system
Email Hoax	[9], 2014	Similarity Calculation: *Levenstein Distance* Classification Method: *Levenstein Distance*	Email hoax detection system
Indonesian Hoax News	[10], 2015	Comparison of Text Classification Performance between Naïve Bayes, SVM, and C4.5	Naïve Bayes text classification yields the best results compared to the others
Indonesian Hoax News	[11], 2017	Classification Methods : *Support Vector Machine* (SVM) and *Stochastic Gradient Descent* (SGD)	Indonesian Hoax News based on sentence feature
Indonesian Hoax News	[12], 2017	Classification Method: *Naïve Bayes*	Build an open-access Indonesian Language news dataset
Fake News	[13], 2017	Classification Method : *Multi-layered technique* (IP Address, Source of News)	Produces a generic framework to detect fake news from any news source platform
Fake News	[17], 2017	Classification Method : *Naïve Bayes*	Automated fake news detection system for Facebook News posts

Naïve Bayes provides the best accuracy by using 10-fold cross-validation. However, classification performance declines when topics are varied, so it cannot detect hoax news. Prasetijo et al. [11] has eliminated the shortcomings in Rasywir and Purwarianti's research by using full news content and by looking at sentence features more closely. Prasetijo examined news from politics, economy, sport, entertainment, and technology. The results showed that SGD with modified-huber kernels increases SVM algorithms. Pratiwi, Asmara and Rahutomo [12] used Naïve Bayes algorithm to classify Indonesian hoax news. The research outcome was an openly accessed database consisting of hoax news and a tested proportion of training and testing sets. The experimental results showed that the best accuracy was achieved when ratio between training and testing was 70:30

Sirajudeen, Azmi and Abubakar [13] argued that the barrier attribute on websites or microblogging sites can be utilized to detect fake news. Sirajudeen adopted a multi-layered technique: a combination of IP address verification and news source verification. The validity of an IP address is determined by its consistency. The validity of the news source is determined by authorship, headlines, and content. The results of IP address verification and news source verification are simultaneously evaluated using IF-THEN rule principle. If the IP address and the news source are verified, then the news is legitimate (fake or genuine). If this condition is not met, the news is a claim. Granik and Mesyura [17] used Naïve Bayes algorithm to develop a system to detect fake news on Facebook. The datasets were relatively small (around 2000 posts), compared to millions of news circulating on Facebook. Nevertheless, Naïve Bayes was able to detect fake news well, without pre-processing, i.e., removing stop words and stemming.

Previous research still faces a bottleneck issue when adding news corpus into the database. System administrator must input new news to the database regularly because hoax news is produced and circulated every day. The Indonesian government has established agencies to confirm news on official websites, but this is not sufficient. What is needed is a system that can connect all of these official websites and engage news readers to verify hoax news.

3. Research methodology

3.1. System architecture

Our proposed system consists of four stages: pre-processing, similarity calculation, classification and class determination. The architectural design of the proposed system can be seen in Fig. 1. Readers must put in the news' attributes that consist of title (mandatory), URL (optional), and news content (mandatory). URL is taken into consideration based on research of D. R. Patil and J. B. Patil [18] emphasising on the number of malicious URLs in cyberspace.

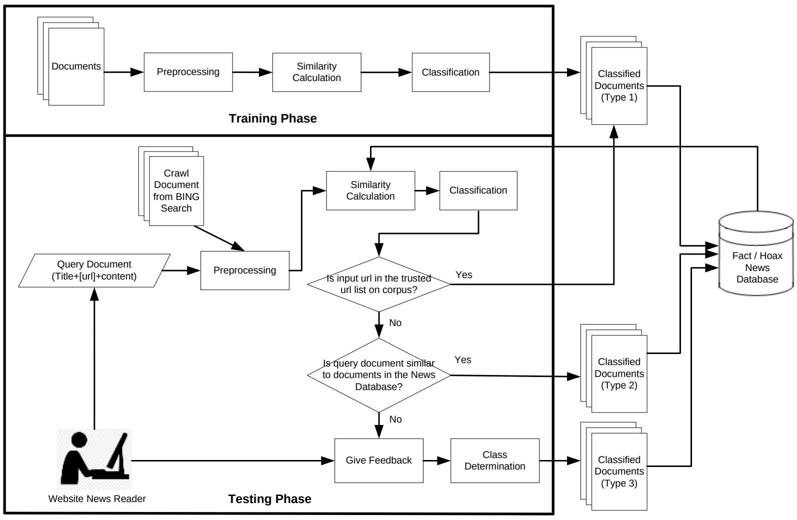

Documents
Preprocessing
Similarity Calculation
Classification
Training Phase
Classified Documents (Type 1)
Crawl Document from BING Search
Similarity Calculation
Classification
Query Document (Title+[url]+content)
Preprocessing
Is input url in the trusted url list on corpus?
Yes
No
Fact / Hoax News Database
Is query document similar to documents in the News Database?
Yes
Classified Documents (Type 2)
No
Give Feedback
Class Determination
Classified Documents (Type 3)
Website News Reader
Testing Phase

Fig. 1. System architecture

3.1.1. Pre-processing

Pre-processing is carried out every time documents and queries are put into the system [19]. Pre-processing consists of two stages: tokenization and stemming. Tokenization aims to cut a collection of texts into several terms or words while removing certain characters, such as punctuation [20]. Stemming is a crude heuristic process that chops off words to their written word form [21]. The stemming process uses Tala algorithm because the computational speed for Indonesian language document is high [22, 23].

3.1.2. Similarity calculation

This hoax detection system uses cosine similarity algorithm to determine whether the tested document is similar to the documents in the database [24]. Cosine similarity utilizes the cosine angle value between them.

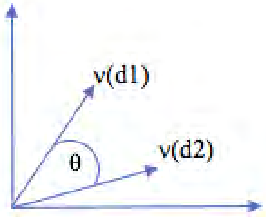


Fig. 2. Illustration of cosine similarity measurements between two vectors

Fig. 2 illustrates that each document to compare resembles a vector, called an *n*-dimensional vector. *N* is the number of words in each document (vector).

(1) $$v(d_1) = (a_1, a_2, a_3, \dots),$$

(2) $$v(d_2) = (b_1, b_2, b_3, \dots).$$

Variable a_i and b_i are the number of words in Document 1 and Document 2, respectively. The similarity of a document compared to the database is determined by using cosine similarity [25]. Theodoridis and Koutroumbas [26] defines cosine similarity measure in (3) as follows:

(3) $$\cos\theta = \frac{a\cdot b}{\|a\|\|b\|},$$

where

(4) $$\|a\| = \sqrt{a_1{}^2 + a_2{}^2 + \dots + a_n{}^2},$$

and

(5) $$\|b\| = \sqrt{b_1{}^2 + b_2{}^2 + \dots + b_n{}^2}.$$

The value of $\cos\theta$ is between 0 and 1. If the value of $\cos\theta$ is close to 1, then a document has more similarity.

In this study, the threshold cosine similarity values are compared to determine the best one. The threshold is used to calculate news similarity taken from BING API and compared to the news in the database. The threshold cosine similarity values are 0.2, 0.4, and 0.6. The threshold value of 0.2 is the most accurate. The threshold 0.6 has a sensitive result. The threshold 0.4 is between 0.2 and 0.6 results [27].

3.1.3. Classification

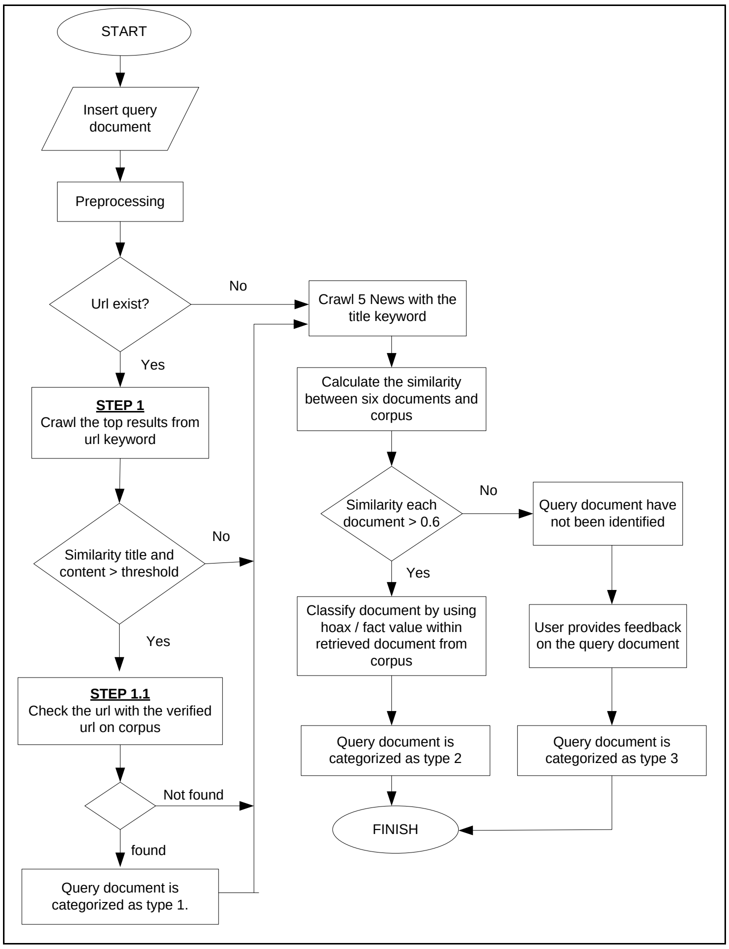

START
Insert query document
Preprocessing
Url exist?
No
Crawl 5 News with the title keyword
Yes
STEP 1
Crawl the top results from url keyword
Calculate the similarity between six documents and corpus
Similarity each document > 0.6
No
Query document have not been identified
Similarity title and content > threshold
No
Yes
Classify document by using hoax / fact value within retrieved document from corpus
User provides feedback on the query document
Yes
STEP 1.1
Check the url with the verified url on corpus
Query document is categorized as type 2
Query document is categorized as type 3
Not found
FINISH
found
Query document is categorized as type 1.

Fig. 3. Flowchart of the classification method

The classification method uses Naïve Bayes, with five steps as follows.

1. Determine the initial probability of the document using

(6) $$P(c|d) \propto P(c) \prod 1 \leq k \leq n_d \; P(t_k|c),$$

where $P(t_k|c)$ is a condition where probability of a word is found in a document in class c; $P(c)$ is the prior probability of documents in class c.

2. Determine Maximum A Posterior (MAP) for the best class in Naïve Bayes using

(7) $$c_{\text{map}} = \arg\max \hat{p}(c|d) = \arg\max \hat{p}(c) \prod_{1\leq k\leq nd} \hat{p}(t_k|c),$$

$\hat{P}$ is written because it does not know the truth of the parameter value $p(c)$ and $p(t_k|c)$.

3. Estimate the parameters $\hat{p}(c)$ and $\hat{p}(t_k|c)$, first it is necessary to try the *maximum likelihood estimate* using

(8) $$\hat{p}(c) = \frac{Nc}{N},$$

N_c states the number of documents in class c and N is the total number of documents available.

4. Estimate the opportunity conditions $\hat{p}(t|c)$ as the relative frequency of the t words in the document in a particular class using

(9) $$\hat{p}(t|c) = \frac{T_{c_t}}{\sum_{t' \in v} T'_{c_t}},$$

T_{c_t} states the frequency of word *t* in the training documents of class *c*, including in other documents.

5. Using *Laplace smoothing* to avoid 0 results in the dominator by using the next equation, the hoax detection systems can work accurately on scattered topics [28],

(10) $$\hat{p}(t|c) = \frac{T_{c_t}+1}{\sum_{t' \in v}(T'_{c_t}+1)} = \frac{T{c_t}+1}{(\sum_{t' \in v} T'_{c_t})+B},$$

where $B = |V|$ is the number of words in the dictionary.

Naïve Bayes algorithm results fact value and hoax value of the document. Fig. 3 describes in details the overall classification method.

3.1.4. Class determination

Documents from BING Web Search API and testing documents consist of fact-values and hoaxes obtained from Naïve Bayes algorithm. These values are then calculated to obtain the average of fact-value and hoax-value. If the fact-value is greater than the hoax-value, then the document is fact news and vice versa.

The proposed system can build a collection of news from readers' feedback and crawling BING search. Fig. 4b shows verified news from readers' feedback stored in the database. Thus, this research can improve the shortcomings of previous research [10-13, 17], where news collection was built intentionally and manually by an administrator.

3.2. Type of document dataset

The document dataset used in this study was taken from the website https://turnbackhoax.id. This is a community website that fights the circulation of hoaxes in Indonesia and has been recognized by the Ministry of Communication and Information of the Republic of Indonesia. Documents stored in the corpus database are divided into three types, namely:

1. Type-1 documents

Type-1 documents are classified news documents (hoax or fact news). This type is obtained from the training phase, and crawling news from verified websites. The attributes in this type are id, title, content, URL, category, category type, fact-value and hoax-value.

2. Type-2 documents

Type-2 documents are news documents that have been classified by the system. The attributes in this type are id, title, content, URL, category, category type, fact-value and hoax-value.

3. Type-3 documents

Type-3 documents are news documents that have been classified by readers' voting. The attributes in this type are id, title, content, URL, category and category type.

Class of Type-3 documents can changes depending on the number of votes given by the reader. Determination of Type-3 document class follows the following rules:

a. IF $\sum$ READER_FEEDBACK$_{hoax}$ > $\sum$ READER_FEEDBACK$_{fact}$ THEN DOCUMENT is classified as HOAX DOCUMENT;

b. IF $\sum$ READER_FEEDBACK$_{hoax}$ < $\sum$ READER_FEEDBACK$_{fact}$ THEN DOCUMENT is classified as FACT DOCUMENT;

c. IF $\sum$ READER_FEEDBACK$_{hoax}$ = $\sum$ READER_FEEDBACK$_{fact}$ THEN DOCUMENT is classified based on the last reader voting.

Illustration of the above rules can be seen in Table 2.

Table 2. Determination of class on Type-3 news documents

Document	Feedback 1	Feedback 2	Feedback 3	Feedback 4	Result
Document 1	Fact	Fact	Fact	Hoax	Fact
Document 2	Fact	Hoax	Hoax	Hoax	Hoax
Document 3	Fact	Hoax	Hoax	Fact	Fact
Document 4	Fact	Hoax	Fact	Hoax	Hoax

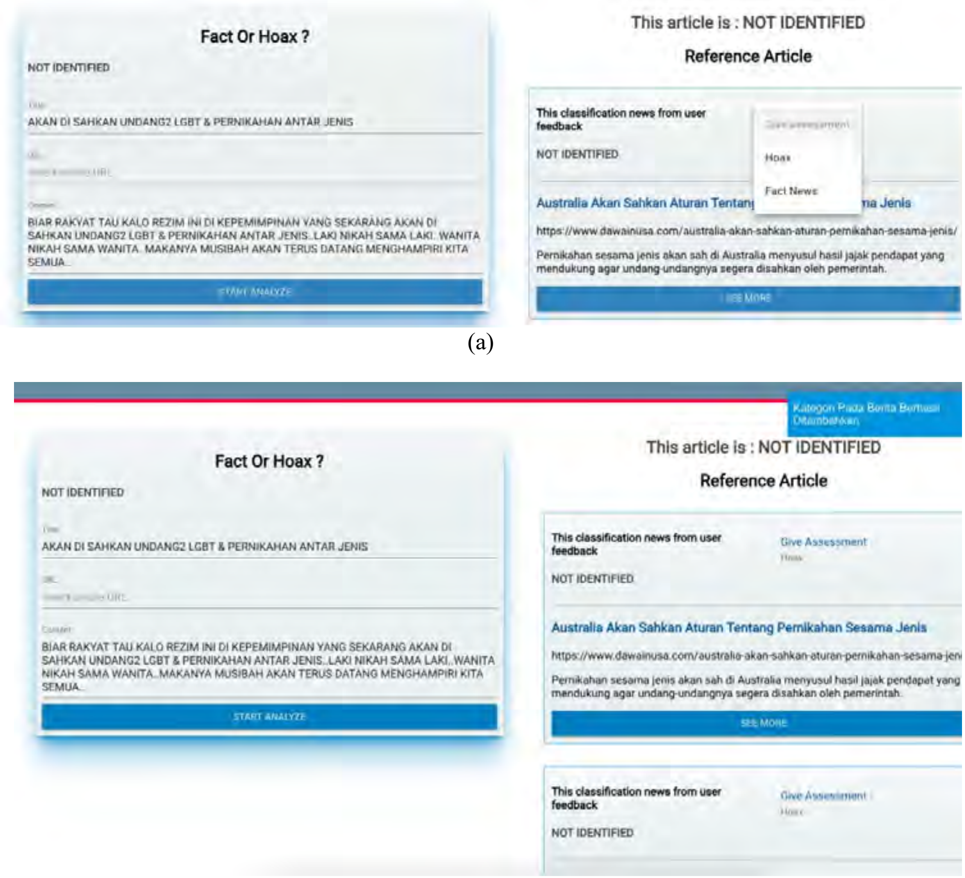
4. Results

To reduce computational complexity, this system was divided into two parts: the back-end and front-end. The back-end side system was built using nodejs programming language and expressjs framework. Nodejs was chosen because it can run code quickly so system performance will not drop due to the speed of its execution [29, 30]. The front-end side used reactjs framework. This system used MongoDB as a database system. MongoDB was chosen because this system required nosql database type, and MongoDB is one of the best nosql databases [31].

4.1. System testing

The system was tested in two phases: training and testing. The training phase was done to train the system, while the testing phase was for system evaluation. For the experiments' purpose, this research used 250 news data taken from a website **https://turnbackhoax.id/**. Training data used 108 news (32 facts news, and 76 hoax news). Testing data used 142 news (26 facts news, and 116 hoax news). Inputting initial corpus data at the training stage can be done in 2 ways: by entering documents directly into the system or by importing CSV files. Then, the system will pre-process the inputted documents, and convert them to an array of data stored in the database as a corpus.

Testing phase is done by inputting query documents into the system. The system accepted three parameters: title, content, and URL (optional). The system crawled news from BING search, pre-process, calculate the similarity, and classify the news. The system retrieves five news documents that were similar to the query document. The system prioritizes taking Type-1 news documents first. If it does not exist, it retrieves Type-2 news documents. If there are no Type-1, and Type-2 news documents, or it does not exist yet in the database, the system displays a maximum of 5 crawling reference documents that are similar to the query document. If the reference document generated by the system does not have a class yet, whether the news is fact or hoax, then the system provides voting feature for readers to clarify the news. After readers give feedback, class determination of the Type-3 documents follows suit. Then, Type-3 documents are stored in the corpus database, and increase the amount of news in the database (Fig. 4).



Fact Or Hoax ?

NOT IDENTIFIED

AKAN DI SAHKAN UNDANG2 LGBT & PERNIKAHAN ANTAR JENIS

BIAR RAKYAT TAU KALO REZIM INI DI KEPEMIMPINAN YANG SEKARANG AKAN DI SAHKAN UNDANG2 LGBT & PERNIKAHAN ANTAR JENIS..LAKI NIKAH SAMA LAKI. WANITA NIKAH SAMA WANITA. MAKANYA MUSIBAH AKAN TERUS DATANG MENGHAMPIRI KITA SEMUA..

This article is : NOT IDENTIFIED

Reference Article

This classification news from user feedback

NOT IDENTIFIED

Australia Akan Sahkan Aturan Tentang Pernikahan Sesama Jenis

https://www.dawainusa.com/australia-akan-sahkan-aturan-pernikahan-sesama-jenis/

Pernikahan sesama jenis akan sah di Australia menyusul hasil jajak pendapat yang mendukung agar undang-undangnya segera disahkan oleh pemerintah.

(a)

(b)

Fig. 4. Indonesian Hoax Detection System display for website news reader feedback (a), Indonesian Hoax Detection System GUI display after website news reader feedback is successfully entered into database (b)

4.2. System evaluation

Evaluation of the system is done by comparing the results of the system with testing documents, which are 142 news documents. System performance measurement uses precision (see (11)) and recall (see (12)), with a range of values from 0 to 1 [19, 32, 33]. In this study, the evaluation adheres to 6 criteria, with and without URL. Each criterion uses three different threshold similarities, namely 0.2, 0.4, and 0.6,

$$\text{precision} = \frac{\sum \text{relevance documents retrieved by system}}{\sum \text{documents retrieved by system}}, \tag{11}$$

$$\text{recall} = \frac{\sum \text{relevance documents retrieved by system}}{\sum \text{number of relevance documents}}. \tag{12}$$

The highest precision of 0.98 is achieved when readers include the URL by applying the similarity threshold of 0.4. The second position is 0.92 under similar conditions by applying similarity threshold of 0.6. This shows that the system is able to classify data accurately based on the retrieved documents. The highest recall of 1 occurs when readers include URL by applying similarity threshold of 0.2. If readers do not include URL, then recall value is 1 by using similarity threshold of 0.2. This indicates that the system can return all relevant documents.

To obtain consistent system performance in terms of precision and recall values, *f*-measure is used the next equation:

$$f\text{-measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision}+\text{recall}}. \tag{13}$$

The best performance based on *f*-measure is 0.95, that is reached by using similarity threshold of 0.2 when readers include URL. The second-best performance is 0.94, obtained by excluding URL. Detailed results of the precision test, recall test, and *f*-measures for the all six criteria are shown in Table 3 and Fig. 5.

Table 3. Comparison of precision test, recall test, and f-measure value for the six criteria

Criteria (Threshold)	Precision	Recall	*f*-measure
URL (0.2)	0.91	1	0.95
URL (0.4)	0.98	0.73	0.84
URL (0.6)	0.97	0.63	0.76
NOURL (0.2)	0.88	1	0.94
NOURL (0.4)	0.95	0.7	0.81
NOURL (0.6)	0.96	0.56	0.71

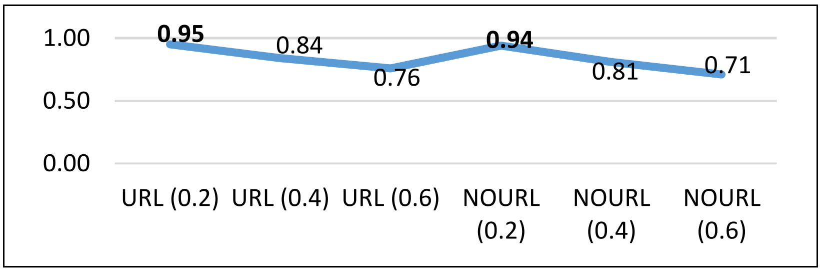


Fig. 5. Comparison of *f*-measures value against six criteria

5. Discussion

The performance of classification method in question is then compared with similar studies [10-12] by considering the precision, recall, accuracy and *f*-measure. Rasywir and Purwarianti [10] has proven that Naïve Bayes is a more reliable machine learning compared to SVM and C4.5 algorithm. The results showed that Naïve Bayes, when combined with probability-based feature selection, achieved the best accuracy value of 0.91. Rasywir did not use precision, recall and *f*-measure values, so the results in the current study cannot be compared to them. However, the classification performance went down when there was more variation in terms of topics.

Prasetijo et al. [11] substituted the feature selection method with full news content-based to improve Rasywir's text classification performance [10]. Prasetijo used news from politics, economy, sports, entertainment, and technology categories. He used SGD linear regression, SVM linear kernel and SGD modified-huber as the classification methods. The results showed that SGD modified-huber resulted in the highest precision, recall, accuracy and *f*-measure at around 0.73, 1, 0.84 and 0.82, respectively.

Pratiwi, Asmara and Rahutomo's research [12] investigated the effect of training and testing data ratio to achieve optimal performance of Naïve Bayes classification method. Pratiwi revealed that the best ratio was 70:30 which yielded values of precision, recall, accuracy and *f*-measure of around 0.67, 0.88, 0.78 and 0.76, respectively.

Our proposed method did not perform feature selection as that of in Rasywir's research and Prasetijo's research, so computing time is faster yet performance is better. Although we used documents from various criteria, our classifier was able to detect hoax news. We used the best ratio of training data and testing data, which is 70:30. The comparison of our classifier's performance with other studies is summarized in Table 4, which shows that precision, recall and *f*-measure of our classifier are better than others.

Table 4. Comparison between the performance of our method and those of other studies

Research	Precision	Recall	Accuracy	*f*-measure
Our proposed method	0.91	1	0.87	0.95
Rasywir and Purwarianti [10]	–	–	0.91	–
Prasetijo et al. [11]	0.73	1	0.84	0.82
Pratiwi, Asmara and Rahutomo [12]	0.67	0.88	0.78	0.76

It is important to note that our proposed method is able to append corpus dataset in news database and collect readers' feedback (shown in Fig. 4a and b). Classified documents obtained from readers' feedback also contribute to the system to detect hoax news. This should overcome the problem of text classification that only relies on manual input by administrator to constantly improve the corpus dataset.

6. Conclusion

The rapid increase of hoax news circulation in cyberspace has become a major concern of the Indonesian government. Previous attempts to reduce the circulation of hoax news did not work effectively. There is still a lot of hoax news spread in the community, whether it is reappearing hoax news or news that is not very legitimate and reliable. Developing a hoax news detection system is one of the solutions offered by information technology to help readers classify, verify, compare and cross-check news.

This study proposes a detection system which involves users' feedback feature to verify unclassified news and uses Naïve Bayes algorithm. The best performance of the system in detecting hoax news is with a threshold value of 0.2. The values of precision, recall, and *f*-measures are 0.91, 1, and 0.95, respectively if URL is included; and 0.88, 1, and 0.94, respectively if URL is excluded. From these results it can be concluded that with or without URL, the system can yield good results. In addition, readers' feedback feature can run well in verifying news documents and build the corpus database.

Future research should improve system performance by using other classification methods, so the highest precision, recall and *f*-measure can be obtained. We also suggest combining content analysis with sentiment analysis to deal with different hoax patterns.

***Acknowledgements*:** This research was made possible by research grant from Faculty of Science and Technology, Universitas Airlangga, Grant No 1909/UN3.1.8/LT/2018 under the schema Excellent Faculty Research.

References

1. Hornby, A. S. Oxford Advanced Learner's Dictionary of Current English. Oxford University Press, UK, 1995.
2. Vukovic, M., K. Pripuzic, H. Belani. An Intelligent Automatic Hoax Detection System. – Knowledge-Based and Intelligent Information and Engineering Systems, 2009, pp. 318-325.
3. Petkovic, T., Z. Kostanjcar, P. Pale. e-Mail System for Automatic Hoax Recognition. – In: 27th MIPRO International Conference, 2005, pp. 117-121.
4. Anonymous. Hoax Email Captures Yahoo's Customer Credit Numbers, 2002, p. 3.
5. Almeida, T. A., T. P. Silva, I. Santos, J. M. G. Hidalgo. Text Normalization and Semantic Indexing to Enhance Instant Messaging and SMS Spam Filtering. – Knowledge-Based System Journal, Vol. **108**, 2016, pp. 25-32.
6. Purnomo, M. H., S. Sumpeno, E. I. Setiawan, D. Purwitasari. Biomedical Engineering Research in the Social Network Analysis Era : Stance Classification for Analysis of Hoax Medical News in Social Media. – In: 2nd International Conference on Computer Science and Computational Intelligence, Bali, 2017.
7. Cunningham, E., W. Marcason. Internet Hoaxes: How to Spot Them and How to Debunk Them. – Journal of the American Dietetic Association, Vol. **101**, 2001, No 4, p. 460.
8. Ishak, A., Y. Y. Chen, S.-P. Yong. Distance-Based Hoax Detection System. – In: International Conference on Computer & Information Science (ICCIS'12), Kuala Lumpur, Malaysia, 2012, pp. 215-220.
9. Chen, Y. Y., S.-P. Yong, A. Ishak. Email Hoax Detection System Using Levenshtein Distance Method. – Journal of Computers, Vol. **9**, 2014, No 2, pp. 441-446.
10. Rasywir, E., A. Purwarianti. Eksperimen Pada Sistem Klasifikasi Berita Hoax Berbahasa Indonesia Berbasis Pembelajaran Mesin. – Jurnal Cybermatika, Vol. **3**, 2015, No 2, pp. 1-8.
11. Prasetijo, A. B., R. R. Isnanto, D. Eridani, Y. A. A. Soetrisno, M. Arfan, A. Sofwan. Hoax Detection System on Indonesian News Sites Based on Text Classification Using SVM and SGD. – In: 4th International Conference on Information Technology, Computer, and Electrical Engineering, Semarang, Indonesia, 2017, pp. 45-49.
12. Pratiwi, I. Y. R., R. A. Asmara, F. Rahutomo. Study of Hoax News Detection Using Naive Bayes Classifier in Indonesian Language. – In: International Conference on Information & Communication Technology and System, 2017, pp. 74-78.
13. Sirajudeen, S. M., N. F. A. Azmi, A. I. Abubakar. Online Fake News Detection Algorithm. – Journal of Theoretical and Applied Information Technology (JATIT), Vol. **95**, 2017, No 17, pp. 4114-4122.
14. Hernandez, J., C., Hernandez, C. J., Sierra, J. M., Ribagorda. A First Step towards Automatic Hoax Detection. – In: Proc. of 36th Annual 2002, Atlantic City, NJ, USA, 2002, pp. 102-114.

15. Kim, H., J. Kim, J. Kim, P. Lim. Towards Perfect Text Classification with Wikipedia-Based Semantic Naïve Bayes Learning. – Neurocomputing, Vol. **315**, November 2018, pp. 128-134.
16. Rasjid, Z. E., R. Setiawan. Performance Comparison and Optimization of Text Document Classification Using k-NN and Naive Bayes Classification Techniques. – Procedia Computer Science, Vol. **116**, Bali, Indonesia, 2017, pp. 107-112.
17. Granik, M., V. Mesyura. Fake News Detection Using Naive Bayes Classifier. – In: IEEE First Ukraine Conference on Electrical and Computer Engineering, Kiev, Ukraine, 2017, pp. 900-903.
18. Patil, D. R., J. B. Patil. Malicious URLs Detection Using Decision Tree Classifiers and Majority Voting Technique. – Cybernetics and Information Technologies, Vol. **18**, 2018, No 1, pp. 11-29.
19. Fatmawati, T., B. Zaman, I. Werdiningsih. Implementation of the Common Phrase Index Method on the Phrase Query for Information Retrieval. – In: Proc. of International Conference on Mathematics: Pure, Applied and Computation, AIP Conference, 2017.
20. Christopher, M., P. Raghavan, H. Schütze. An Introduction to Information Retrieval. – Natural Language Engineering, Vol. **16**, 2010, No 1, pp. 100-103.
21. Kallimani, J. S., K. G. Srinivasa, E. B. Reddy. Summarizing News Paper Articles: Experiments with Ontology-Based Customized, Extractive Text Summary and Word Scoring. – Cybernetics and Information Technologies, Vol. **12**, 2012, No 2, pp. 35-50.
22. Tala, F. Z. A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia. Thesis, Universiteit van Amsterdam, Netherlands, 2003.
23. Agusta, L. Perbandingan Algoritma Stemming Porter Dengan Algoritma Nazief & Adriani Untuk Stemming Dokumen Teks Bahasa Indonesia. – In: Konferensi Nasional Sistem dan Informatika, Bali, Indonesia, 2009, pp. 196-201.
24. Li, B., L. Han. Distance Weighted Cosine Similarity Measure for Text Classification. – In: International Conference on Intelligent Data Engineering and Automated Learning, Berlin, Heidelberg, 2013, pp. 611-618.
25. Al-Anzi, F. S., D. AbuZeina. Toward an Enhanced Arabic Text Classification Using Cosine Similarity and Latent Semantic Indexing. – Journal of King Saud University – Computer and Information Sciences, Vol. **29**, April 2016, pp. 189-195.
26. Theodoridis, S., K. Koutroumbas. Pattern Recognition: Second Recognition. Academic Press, 2003.
27. Tung, K. T., N. D. Hung, L. T. My Hanh. A Comparison of Algorithms Used to Measure the Similarity between Two Documents. – International Journal of Advanced Research in Computing Engineering and Technology (IJARCET), Vol. **4**, 2015, No 4, pp. 1117-1121.
28. Han, J., M. Kamber, J. Pei. Data Mining: Concepts and Techniques. – In: The Morgan Kaufmann Series in Data Management Systems, 2011, pp. 83-124.
29. Bermudez-Ortega, J., E. Besada-Portas, J. A. Lopez-Orozco, J. A. Bonache-Seco, J. M. de la Cruz. Remote Web-Based Control Laboratory for Mobile Devices Based on EJsS, Raspberry Pi and Node.js. – Elsevier, Ltd., Vol. **48**, 2015, No 29, pp. 158-163.
30. Pasquali, S., K. Faaborg. Mastering Node.js: Build Robust and Scalable Real-Time Server-Side. Web Applications Efficiently, Vol. **2**. Birmingham, Mumbai, Packt Publishing, 2017.
31. Chodorow, K. MongoDB: The Definitive Guide Powerful and Scalable Data Storage. O'Reilly Media, Inc., 2013.
32. Priandini, N., B. Zaman, E. Purwanti. Categorizing Document by Fuzzy c-Means and k-Nearest Neighbors' Approach. – In: AIP Conference Proceedings, 2012.
33. Zaman, B., E. Winarko. Analisis Fitur Kalimat untuk Peringkas Teks Otomatis pada Bahasa Indonesia. – International Journal of Computing and Cybernetics System, Vol. **5**, 2011, No 2, pp. 60-68.

Received: 30.07.2019; Second Version: 17.12.2019; Accepted: 27.12.2019