Penentuan *Hoax* pada Artikel Politik Berbahasa Indonesia

di Sosial Media dengan *Similarity Jaccard* dan Algoritma *Stemming*

**Indonesian Politics Articles Hoax Determination on Social Media with N-Gram Methods and Winnowing Algorithm**

# Goenawan Brotosaputro[1], Wiwin Windihastuty[2]\*, Rezza Anugrah Mutiarawan[3]

Fakultas Teknologi Informasi[1], [2], [3]

Universitas Budi Luhur Jakarta, Indonesia

wiwin.windihastuty@budiluhur.ac.id

Abstract—The rapid development of technology in the current global era has a major influence in the process of finding information. We can see a very large growth volume of online news available on the internet, as well as other networks. One of them is social media that has a lot of information about news articles. Currently the news about politics is spread mainly which can lead to misinterpretation because the news is not necessarily true or false, so it is necessary to classify political articles as belonging to the hoax or non-hoax category. Hoax is false news that is spread to gain trust so that people will feel confident that the content is true. Another impact of hoaxes can be detrimental to people's emotions and finances. The hoax classification process uses the preprossessing stage which consists of tokenization and winnowing. Followed by the process of weighting words and jaccard similarity to the classification process using N-Gram method. The evaluation results in this study using a confusion matrix, where the precision results obtained by 0.90 recall of 0.78 and accuracy obtained by 85%.

***Keywords—*** ***Stemming Algorithm, Stopword Removal, Similarity Jaccard, Machine Learning, Hoax.***

1. PENDAHULUAN

The rapid development of information and communication technology in global era had a major influence on the information. The more human needs for information increase, the more information exists, especially through online social media. Online social media is a communication forum that used as social interaction online. Online social media users can receive information, communicate with each other, interact, share, and various other activities. Online social media is a technology that can turn communication into interactive dialogue. Some social media that have many users, namely Facebook, Twitter, blogs and others.

Social media is a means for consumers to share text, image, video information with each other and acts as a medium for disseminating information from various kinds of information that is spread. The rapid development of information technology, the internet has turned into a source of information that is easy to use and can encourage the dissemination of news articles, both information of factual news and false information or hoaxes. The existence of this hoax news has indeed made online social media users anxious in particular and almost the majority of Indonesian people. The development of information technology has contributed of hoax news spread or information with the aim of creating public opinion, leading opinion, forming perceptions, for fun that tests the intelligence of internet and social media users.

In Indonesia, hoaxes have been rife since the 2014, when presidential election campaign on social media. Hoaxes are popping up and dropping the image of political opponents, as black campaigns. Based on the annual recapitulation, Kominfo received 733 hoax content complaints throughout 2018 on the instant messaging application. Meanwhile, when viewed from August 2018 to January 21 2019, Kominfo received hoax reports that spread on WhatsApp as many as 43 contents.

Based on Directorate of Informatics Application Control monitoring results, the most reports occurred in October 2018. There were 16 hoax content spread on the WhatsApp platform. In August 2018 there were reports of two hoax content, September 2018 there were five hoax content, November 2018 eight content reports and December 2018 as many as 10 hoax content reports. Meanwhile, until January 21, 2019 there were two reports of misinformation content circulating on the Facebook application. complaints submitted to the Kominfo System is shown in Figure 1 below



Figure 1. Content Complaints, 2016 – 2018

The Public Relations Bureau of the Ministry of Communications and Informatics explained, the negative content complaints by instant messaging applications in 2016 there were 14 content complaints, in the following year the number of complaints increased to 281 complaints, while in 2018 there were 1440 reports related to negative content.

The purpose of this research is to build an application that can find out an article regarding political news information and the system will systematically determine the article into a hoax or non-hoax label. The benefits of this research are to provide accuracy in the performance of determining hoax or non-hoax articles on social media and make it easier to label or categorize hoaxes and non-hoaxes on political news articles.

1. METODE PENELITIAN

In this study requires a technique or way to find, obtain, collect and record information. In developing this application refers to the waterfall model. There are five main stages in this method. It is called the waterfall model because the stages diagram is similar to a multilevel process. The waterfall model methodology is a model in software planning. The waterfall model is an example of a planning process, where all process activities must be planned and scheduled before being carried out. The stages in the waterfall model are as follows:

1. Analysis

This stage aims to find the needs of users and organizations as well as analyze the existing conditions (before implementing the new system).

1. Design

The design stage aims to determine the detailed specifications of the program components (people, hardware, software, network and data) and information products in accordance with the results of the analysis phase.

1. Implementation

This stage for obtaining or developing hardware and software (program coding), conducting testing, training and the move to the new system.

1. Maintenance

This stage is carried out when the system is operational. The process of monitoring, evaluating and making changes if necessary, is carried out.

In this study the authors used the N-Gram method with Winnowing algorithms and the similarity function in Jaccard. Research methodology includes research methods, namely a technique or method used in connection with research conducted has systematic steps.

The research method concerns to understand work technique process. The method includes research procedures and research techniques. The method also trying to approach experimental research, by testing the weights word in the document classification, then search effectiveness on the basis of precision.

Sample is part of the number and characteristics population possessed or a small part of the population members taken according to certain procedures so that they can represent the population. There are various kinds of sampling techniques to determine which samples can be used in research. To test results effectiveness level of determining articles from a model, the sampling method in this study used the Cluster Sampling method. The sample in this study were 200 articles obtained from online social media sources, namely turnbackhoax.com as a media that validates hoax and non-hoax articles and some from media registered at councilpers.or.id. The article used is form text without images. The reason for taking it from online social media is because a lot of information is spread through social media with the structure of the Indonesian language used. The research plan for the sample will be divided into two categories or labels, namely hoaxes and non-hoaxes.

In a research activity it is necessary to collect data information and references for materials that can support description and discussion, this makes one of steps or stages very important in carrying out research activities. The correct data collection technique will produce high credibility data, but if data collection wrong does not have high credibility. At this stage, news articles on digital media, namely turnbackhoax.com and Dewanpers.co.id, are hoax and news sites with Indonesian language content, by taking a sample of 200 articles from one of the categories on online social media, namely the political category. The 200 articles that have been compiled from each article with the aim of facilitating testing of truth and accuracy.

Documents used to classify vector space model implement with the Nazief and Andriani stemming algorithm and jaccard similarity function.

The analytical technique used for this study uses the Nazief and Andriani Stemming Algorithm approach. The design of the model in this study was made to have the following process:

1. The text preprocessing stage, namely carrying out the tokenizing process, stopword removal and the stemming process using the Nazief and Andriani Algorithms.
2. Carry out the weighting process for each word of the article using Jaccard similarity function.
3. Calculating similarity value process, where to look for differences in the articles owned by a query which includes Nazief and Andrini Algorithms and Jaccard similarity function to calculate the similarity value of the article.
4. Articles Classification model for compiling articles with agreed terms. This prototype model was chosen because it has an appropriate structure in developing a classification model simulation. The article classification prototype that will be made in this research is using a web based application.

In this study, analyzing effectiveness level of classification article results, requires a classification model, namely a web-based application that can apply vector space models, Nazief and Andraini stemming algorithm processes and weighting processes based on similarity values. The processes used in determining articles include:

1. Uploading documents
2. Tokenization
3. Removing stopwords
4. Stemming processes
5. Word weighting processes
6. Processes for calculating similarity values

The system testing technique used by the author with following provisions:

1. Precision testing is comparison the number of relevant articles obtained by the system with the total number of articles retrieved by the system, both relevant and irrelevant.

Precision: The number of articles that are relevant to the query and retrieved

The total number of articles retrieved

1. Completeness test (Recall) is a comparison the number of articles determined by the system with the total number of relevant articles in article collection (collected or not collected by the system).

Recall : Number of articles that are relevant to the query and retrieved

 Number of all relevant articles and collections of articles

2.6 Research Steps

Structured research steps can be seen in Figure 2 below;

Study of literature

Identification of problems

Prototyping

Data collection

Results and Conclusions

Figure 2. Research steps

* + 1. Problem Identification

This stage is research initialization, namely looking for problems encountered in classifying through various political news articles, to be able to understand the conditions of existing problems and conduct literature studies.

1. Literature Study

This process is carried out by reviewing various previous studies discussing the classification of Indonesian-language articles that have been carried out by several previous researchers, therefore to compare article text of classification using Nazief and Andriani stemming algorithm, with Jaccard similarity weighting function, it is expected can measure differences and provide efficient and valid results in classifying political news articles.

1. Data Collection

Data collection was carried out by retrieving articles on digital media, namely turnbackhoax.com and several media registered at Dewanpers.co.id that is one of several digital media sites that provide articles about news or information needed by information seekers. The data that the researcher will use consists of two categories or labels from political news articles. In the category of political articles, as many as 200 will be taken

1. Pembuatan Prototype

The determination of hoax articles is a system that aims to classify Indonesian text articles. The stages of text processing process are carried out by weighting the Jaccard similarity function in each word to each text article, so that it can produce similarity in the article. Based on the Jaccard similarity function weighting and the Vector Space Model method with Nazief and Andriani's Stemming Algorithm for article determination.

1. Prototype Testing

It can be seen that the determination of political articles in this study has been running effectively and has provided the expected output results. Based on the test results for each label, calculations are carried out to find accuracy using Jaccard similarity function. The amount of data to be tested in this study is shown in Table 1 below;

Table 1. Number of Article Data

|  |  |
| --- | --- |
| Kategori Polittik (Hoax) | Kategori Politik (Non-Hoax) |
| 100 Artikel Politik | 100 Artikel Politik |

 Political Category (Hoax) Political Category (Non-Hoax)

6. Results and Conclusions

The results of manual testing to get things that can affect the truth accuracy, understand the conditions of existing problems and conduct literature studies. The data tested from online media turnbackhoax.com, so it is not certain that all the data tested will be included in the appropriate label. The data that has been tested will affect the classification. The conclusion is, all the data tested has been entered into a label or category, but there are articles that do not match that label or category.

1. **HASIL DAN PEMBAHASAN**

Pengelompokan Data, Analisis

RESULTS AND DISCUSSION

Data Grouping, Analysis

In conducting this research, the document articles that will be entered into system are pre-processed to make the text ready to be processed, this form text or terms enters the process using stemming algorithm and weighting words through the Vecor Space Model (VSM) in order to get good accuracy and model performance. The data set that will be used consists of one category of political hoax articles with two classes of hoaxes and non-hoaxes, details can be seen in table 2 below:

Table 2. Kategori dataset pengujian

|  |  |  |
| --- | --- | --- |
| **Catagory** | **Class Hoax** | **Class Non-Hoax** |
| Politics | 100 Articles | 100 Articles |

In Table 2 it is explained that political articles have 100 articles labeled hoaxes and 100 non hoax articles. Table 2 explains that political articles have 100 articles labeled hoaxes and 100 articles labeled non-hoaxes.

* + 1. Reasearch

The N-Gram is used in this study, the process of the winnowing algorithm and the term weighting process of the words resulting from the winnowing process can perform calculations. The results of the distance calculation are used to determine the level of similarity of the document.

* + 1. Pengelompokan Data

In this study using a dataset as a test obtained from online social media, namely several social media verified by the councilpers and turnbackhoax.id which had previously been taken with a total of 100 hoax articles (false news) and 100 non-hoax articles (fact news).

* + 1. Analisa Data Bahan Penelitian

Data from literature studies or research on determining articles using Jacard similarity and the winnowing algorithm are as follows:

* + 1. Stopword

Stopwords are words that are not pre-processed. The characteristic of stopwords is the high frequency of occurrence, usually in the form of conjunctions or pronouns.

1. Article files to be checked for similarity level
2. Big Indonesian Dictionary data.
3. Term for the process of winnowing
	* 1. System Design

The system design process in general for application development consists of several stages, including design.

* + - 1. Program Specifiation Design

In Figure 3 is a Class Diagram that used as the main model in modeling class diagrams with the context of its components.



Figure 3. Class Diagram

* + - 1. System Process Design

 The intended process design is how the system will work, the processes used start from the user doing input until the system produces output in the form of a similarity value using Jaccard calculations. Figures 3.1 – 3.6 explain the full process of the article classification system:

Gambar 3.1 Articles Classification Process



Figure 3.2 Tokenization Process



Figure 3.3 Stemming Process



Figure 3.4 Stopword Process



Figure 3.5 Pembobotan Process



Figure 3.6 Similarity Process

* + 1. Testing System

Testing is a important thing, testing is used to ensure the results of research objectives achieved or not. The measurement method used a confusion matrix testing method, namely recall testing precision and accuracy by doing 10 (ten) iterations. The following formula used in the measurement is found in the explanation of Figure 5

Figure 5. Formula Testing



Figure 4. Test Formula

Distribution of data using the holdout method, namely dividing a number of data for training and using the rest for testing. Data sharing uses a formula 70% for training and 30% for testing. In this study, testing was carried out using iteration, so the data division method used was the repeated holdout method for each iteration in the testing process. The repeated holdout method is carried out by repeating the process with different subsamples at each iteration. Before iteration testing and data sharing with repeated holdouts, the authors prepared 200 articles, which consisted of 100 hoax articles and 100 non-hoax articles. In testing hoax and non-hoax labels in each iteration which consists of 70 articles for training data and 30 articles for data testing and the articles tested are different in each iteration using repeated holdout divisions. The following is an explanation of testing on 1 iteration.

***Iteration* 1**

Table 3 are the results of hoax testing and Table 4 is non-hoax testing. The table below is a table that explains the results of testing on hoax and non-hoax articles in the first iteration.

Tabel 3. *Iteration* 1 testinghoax articles



The tests results are concluded as in Table 4 below;

Tabel 4. *Iteration* 1 accuracy hoax articles tests results



The results of testing hoax articles in the table above can be compared with the results of testing on non-hoax articles which can be seen in Table 5 below;

Table 5. *Iteration* 1 articles non hoax test



the non-hoax article testing results are concluded as in Table 6 below

Table 6. Accuracy Hasil Pengujian Artikel Hoax *Iteration* 1

So the average value of the calculation above is in iteration 1 :

Precision = 0.85 -> 85 %

Recall = 0.78 -> 78 %

Accuracy = 0.78 -> 78 %

The test was carried out for 10 iterations continuously. The results of testing as many as 10 iterations are not displayed. The average value of the 10 iterations is as shown in Table 7 below;

Table 7. Average Value of Test Calculations

|  |  |  |  |
| --- | --- | --- | --- |
|  | *Precision* | *Recall* | *Accuracy* |
| *Iteration* 1 | 0,85 -> 85 | 0,78 -> 78 | 0.78 -> 78 |
| *Iteration* 2 | 0,85 -> 85 % | 0,83 -> 83 | 0.83 -> 83 |
| *Iteration* 3 | 0,82 -> 82 | 0,76 -> 76 | 0.76 -> 76 |
| *Iteration* 4 | 0,87 -> 87 | 0,87 -> 87 | 0.87 -> 87 |
| *Iteration* 5 | 0,88 -> 88 | 0,86 -> 86 | 0.86 -> 86 |
| *Iteration* 6 | 0,82 -> 82 | 0,81 -> 81 | 0.81 -> 81 |
| *Iteration* 7 | 0,81 -> 81 | 0,80 -> 80 | 0.80 -> 80 |
| *Iteration* 8 | 0,83 -> 83 | 0,82 -> 82 | 0.82 -> 82 |
| *Iteration* 9 | 0,81 -> 81 | 0,75 -> 75 | 0.75 -> 75 |
| *Iteration* 10 | 0,85 -> 85 | 0,83 -> 83 | 0.83 -> 83 |

From the test results with 10 iterations, it can be concluded from 2 (two) sides of the aspect, namely;

System Aspect

The system aspect, expected a server that supports large data storage to store various data needed to build an article detection system

Management Aspect

The system module produced in this study is expected to be applied to various types of systems for detecting other articles in an institution.

Further Research Aspect

Researchers can use the results of this study as reference material for similar studies and become an initial model that can be developed and produce better further research.

1. KESIMPULAN

In accordance with the discussion, literature study, research review and research methods in classifying hoax articles on social media with Jacard similarity and winnowing algorithms, the conclusions that can be drawn are as follows: (1) Can facilitate the classification of articles that are not quite right, which can lead to miss information for information seekers. (2) Using the N-Gram method with winnowing and Jaccard similarity can facilitate the classification of articles in large quantities and produce a precision value of 0.92, a recall of 0.80 and an accuracy of 87%.

Suggestions for the future, described as follows; (1) In future research, it is hoped that more training data will be used, so that the similarity value is more precise, the more training data learned, the system will learn the words that are usually found in the document. (2) In related agencies, which are involved in the dissemination of article information, this classification application can also be used for other articles such as health articles, natural disaster articles and so on.

REFERENCES

1. Maxmanroe. 2017. Pengertian Media Sosial Secara Umum, Fungsi, Tujuan, Jenis Sosmed. *Maxmanroe.* [Online] 2017. [Dikutip: 21 Juli 2019.] https:/[/ww](http://www.maxmanroe.com/vid/teknologi/internet/pengertian-)w[.maxmanroe.com/vid/teknologi/internet/pengertian-](http://www.maxmanroe.com/vid/teknologi/internet/pengertian-) media-sosial.html#.
2. Afriza, A. (2019) ‘Metode Klasifikasi Rocchio untuk Analisis Hoax Metode Klasifikasi Rocchio untuk Analisis Hoax Rocchio Classification Method for Hoax Analysis’, *Metode Klasifikasi Rocchio untuk Analisis Hoax Metode Klasifikasi Rocchio untuk Analisis Hoax Rocchio Classification Method for Hoax Analysis*, (February), pp. 0–10. doi: 10.29244/jika.5.1.1-10.
3. TAMMAM, A. G. (2018) ‘Deteksi Hoaks Pada Media Sosial Berbasis Text Mining Classification System’, *Deteksi Hoaks Pada Media Sosial Berbasis Text Mining Classification System*.
4. Weddiningrum, F. G. *et al.* (2018) ‘Deteksi Konten Hoax Berbahasa Indonesia Pada Media Sosial Menggunakan Metode Levenshtein Distance’, *Deteksi Konten Hoax Berbahasa Indonesia Pada Media Sosial Menggunakan Metode Levenshtein Distance*
5. Roger S. Pressman, Ph.D. 2012. *Rekayasa Perangkat Lunak.* s.l. : Andi , 2012. 9789792931044.
6. Annisa, Nurul, D. (2017) ‘Implementasi Algoritma Vector Space Model Dalam Pencarian E-Book’, (October 2014).
7. E. Garcia, P. (2006) ‘The Classic TF-IDF Vector Space Model’, *The Classic TF-IDF Vector Space Model*, (1), pp. 1–15.
8. Irmawati (2017) ‘Sistem Temu Kembali Informasi Pada Dokumen Dengan Metode Vector Space Model’, *Jurnal Ilmiah FIFO*, IX(1), pp. 74–80. doi: 10.1093/clinids/13.Supplement\_1.S109.
9. Lutfi and Musodo, K. A. (2013) ‘Trigram Pada Vector Space Model Dengan Pembobotan Dalam Dokumen Dan Fungsi Kesetaraan Cosine’, *Trigram Pada Vector Space Model Dengan Pembobotan Dalam Dokumen Dan Fungsi Kesetaraan Cosine*.
10. Maulina, D. *et al.* (2018) ‘Klasifikasi Artikel Hoax Menggunakan Support Vector Machine Linear Dengan Pembobotan Term Frequency – Inverse Document’, *Klasifikasi Artikel Hoax Menggunakan Support Vector Machine Linear Dengan Pembobotan Term Frequency – Inverse Document*, 2(1), pp. 35–40.
11. Prasetyo, A. R. and Adikara, P. P. (2018) ‘Klasifikasi Hoax Pada Berita Kesehatan Berbahasa Indonesia Dengan Menggunakan Metode Modified K-Nearest Neighbor’, *Klasifikasi Hoax Pada Berita Kesehatan Berbahasa Indonesia Dengan Menggunakan Metode Modified K-Nearest Neighbor*, 2(12), pp. 7466–7473.
12. Immersa.lab. 2018. Pengertian Artificial Intelligence. *immersa-lab.* [Online] February 15, 2018. [Cited: Juli 23, 2019.] http://www.immersa- lab.com/pengertian-artificial-intelligence.htm.
13. Nasrullah, Rulli. 2015. *Media Sosial; Persfektif Komunikasi, Budaya, dan Sosioteknologi.* Bandung : Simbiosa Rekatama Media, 2015.
14. Purbolaksono, Arfianto. 2017. Anti Hoax. *theindonesianinstitute.* [Online] Januari 01, 2017. [Cited: Juli 21, 2019.] https:/[/ww](http://www.theindonesianinstitute.com/anti-hoax/)w[.theindonesianinstitute.com/anti-hoax/.](http://www.theindonesianinstitute.com/anti-hoax/)