

# HOAX DETECTION AT SOCIAL MEDIA WITH TEXT MINING CLARIFICATION SYSTEM- BASED

*by* Sucipto Sucipto

---

**Submission date:** 11-Dec-2020 07:03PM (UTC+0700)

**Submission ID:** 1472004535

**File name:** Extracted\_pages\_from\_9.\_Cover\_Jurnal\_Vol\_3\_No\_2.pdf (311.16K)

**Word count:** 2101

**Character count:** 11028

## HOAX DETECTION AT SOCIAL MEDIA WITH TEXT MINING CLARIFICATION SYSTEM-BASED

Aditya Gusti <sup>2</sup>mmam<sup>1)</sup>, Sucipto<sup>2)</sup>, Rini Indriati<sup>3)</sup>

<sup>1,2,3)</sup>Information System department, University of Nusantara PGRI Kediri

Jl K.H Achmad Dahlan 76 Kediri 64111, East Java, Indonesia

e-mail: gustitammam@gmail.com<sup>1)</sup>, sucipto@unpkediri.ac.id<sup>2)</sup>, rini.indriati@unpkediri.ac.id<sup>3)</sup>

### ABSTRAK

Hoax adalah isu terkini yang meresahkan masyarakat dan menyebabkan kerusuhan di berbagai bidang, mulai dari politik, budaya, keamanan dan ketertiban, hingga ekonomi. Masalah ini tidak dapat dipisahkan dari dampak penggunaan media sosial yang cepat. Akibatnya, setiap hari ada ribuan informasi yang tersebar di media sosial, yang belum tentu valid, sehingga orang-orang berpotensi terkena tipuan di media sosial. Sistem pendeteksian tipuan dalam penelitian ini dirancang dengan pendekatan Pembelajaran Tanpa Pengawasan sehingga tidak memerlukan pelatihan data. Sistem ini dibangun menggunakan algoritma Peringkat Teks untuk ekstraksi kata kunci dan algoritma Cosine Similarity untuk menghitung tingkat kemiripan dokumen. Hasil ekstraksi kata kunci akan digunakan untuk mencari konten yang terkait dengan masukan dari pengguna menggunakan mesin pencari, lalu menghitung nilai kemiripan. Jika konten yang terkait cenderung berasal dari media tepercaya, maka konten tersebut berpotensi faktual. Demikian juga, jika konten yang terkait cenderung dipublikasikan oleh media yang tidak dapat diandalkan, maka ada potensi untuk tipuan. Sistem pendeteksian tipuan telah diuji menggunakan matriks kebingungan, dari 20 data konten berita yang terdiri dari 10 masalah yang benar dan 10 masalah yang salah. Kemudian sistem menghasilkan klasifikasi dengan rincian 13 masalah termasuk salah dan 7 masalah termasuk benar, maka jumlah klasifikasi yang sesuai dengan label asli adalah 15 masalah. Berdasarkan hasil klasifikasi, nilai akurasi 75% diperoleh.

**Kata Kunci:** Hoax , text mining, textrank, keyword extraction, cosine similarity.

### ABSTRACT

Hoax is a current issue that is troubling the public and causes riot in various fields, ranging from politics, culture, security and order, to economics. This problem cannot be separated from the impact of rapid use of social media. As a result, every day there are thousands of information spread on social media, which is not necessarily valid, so that people are potentially exposed to hoax on social media. The hoax detection system in this study was designed with an Unsupervised Learning approach so that it did not require data training. The system is built using the Text Rank algorithm for keyword extraction and the Cosine Similarity algorithm to calculate the level of document similarity. The keyword extraction results will be used to search for content related to input from users using the search engine, then calculate the similarity value. If the related content tends to come from trusted media, then the content is potentially factual. Likewise, if the related content tends to be published by unreliable media, then there is the potential for hoax. The hoax detection system has been tested using confusion matrix, from 20 news content data consisting of 10 correct issues and 10 wrong issues. Then the system produces a classification with details of 13 issues including wrong and 7 issues including true, then the number of classifications that match the original label are 15 issues. Based on the results of the classification, an accuracy value of 75% was obtained.

**Keywords:** Hoax, text mining, textrank, keyword extraction, cosine similarity.

## I. INTRODUCTION

Every day there are hundreds or even thousands of information distributed through social media by its users [1]. Information can affect emotions, feelings, thoughts, or even actions of an individual or group. It is unfortunate if the information is inaccurate or even false information (hoax) with provocative titles that lead readers and recipients to negative opinions [2].

Hoax (read: / hōks /) is a message or news that tries to convince the reader about the truth and then tries to convince the reader to take certain actions. Hoax distribution depends on the reader who intentionally sends the message or news to other potential victims who might also do the same thing [3]. In Indonesian the term hoax is absorbed into hoaks, the equivalent of the word for hoax is listed in the Indonesian Big Dictionary (KBBI) which is defined as untrue news [4]. Muhammad Alwi Dahlan thought hoax was intentionally manipulated news with the aim of giving false recognition or understanding. The communication expert from the University of Indonesia (UI) also explained that hoax tends to be planned in advance when compared to ordinary hoaxes. Hoax contains fraudulent facts that attract public attention [5].

A lecturer of communication studies of Atmajaya University in Yogyakarta, Danarka Sasongko argued that people still could not distinguish what was right and what was not true. According to him this happened due to the low public literacy of messages on social media [6]. Budi Sutedjo explained that the ability of readers to trace and criticize and rewrite the information they receive is called media literacy. The Information Technology (IT) expert from Duta Wacana Christian University in Yogyakarta also considered that media literacy could counteract the hoax news distribution [7].

The latest technology should also be able to play a role in overcoming this, one of which is a technology known as Text Mining. Text Mining is a variation of data mining that can extract useful information by identifying and exploring interesting patterns from a collection of unstructured textual data sources [8]. With the text mining capability, the author considers that there are opportunities to make machines that can help humans to do media literacy automatically.

This related research has been carried out by Dyson and Golab who explore the Natural Language Processing method to detect misleading news sources. The findings of this study indicate that the calculation of TF-IDF for bi-gram can work quite well in terms of identifying unreliable sources, while the calculation using PCFG does not give significant effects [9].

Rasywir and Purvarianti have also experimented on the hoax news classification system with machine learning-based. The experiment was conducted to select the best technique in each sub-process using 220 Indonesian-language articles in 22 topics (89 hoaks articles and 131 articles not hoaks). The result is that the Naive Bayes algorithm shows the best accuracy compared to SVM and C4.5 with an accuracy of 91.36% [10].

Previous research-studies use English language news sources and also use a Supervised Learning approach which requires training data. Although the research conducted by Rasywir and Purvarianti has used Indonesian language news sources, it is still constrained due to the lack of training data available in Indonesian. Because of these limitations, in this study the authors propose an approach without training data or called Unsupervised Learning. The author uses the TextRank algorithm for keyword extraction and the Cosine Similarity algorithm to measure the level of document similarity. With the combination of these two algorithms, a system is then built that can measure the potential of a news hoax.

## II. METHOD

## A. TextRank

TextRank is a method that includes an unsupervised learning approach and uses graph-based modeling. This method was developed based on the PageRank method [11]. The basis of the graph ranking model proposed by Mihalea & Tarau is by implementing the "voting" stage in each word (vertex) in the graph. A vertex will be considered important if the vertex is voted more than other vertices. The score on each vertex in the graph is determined from the following equation:

$$S(v_i) = (1 - d) + d \times \sum_{j \in In(v_i)} \frac{1}{|Out(v_j)|} S(v_j)$$

Where the value of  $S(V_i)$  is the value of the  $V_i$  vertex score, with the value of  $d$  as the damping factor that is set to the value of 0.85.

#### B. Cosine Similarity

Cosine similarity is a measure of similarity that is more commonly used in information retrieval and in this study will be used to calculate the similarity of documents. The formula used by cosine similarity is [12]:

$$\cos a = \frac{A \cdot B}{|A||B|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}$$

Informations :  $A$  = Vector A, which will be compared its resemblance,  $B$  = Vector B, which will be compared to the similarity,  $A \cdot B$  = dot product between vector A and vector B,  $|A|$  = length of vector A,  $|B|$  = length of vector B, and  $|A||B|$  = cross product between  $|A|$  and  $|B|$

#### C. Confusion Matrix

Confusion Matrix is a matrix that states the number of comparisons of data test that are classified [13] [14]. Calculation of confusion matrix is stated in the following equation:

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%$$

Informations : TP is True Positive, which is the number of positive data correctly classified by the system, TN is True Negative, which is the amount of negative data correctly classified by the system, FN is a False Negative, which is the amount of negative data but is incorrectly classified by the system, and FP is False Positive, namely the number of positive data but incorrectly classified by the system.

#### D. Desain Proses Mining

In general, the flow chart in Figure 1 illustrates how the hoax detection process occurs. Beginning with the user step to enter input in the form of news text into the system, then proceed with the Keyphrase Generation Process, which is the process by which the system will generate key phrases that will be used to search related content through the Google search engine.

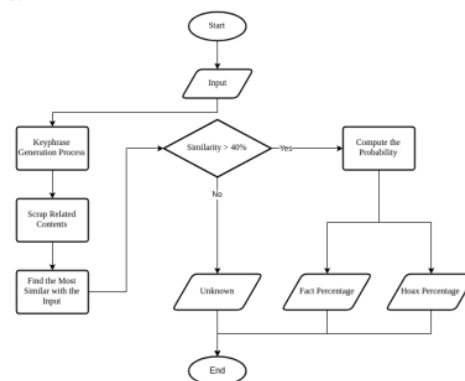


FIG. 1 DIAGRAM OF THE HOAX DETECTION FLOW GENERALLY

After the system gets a list of related content, then it will be proceed with scraping each related content. The scrap product is then calculated which is most similar to the input from the user. With a similarity tolerance limit

of 40%, it will be proceed to the next process which is calculating the percentage of probability of hoax or facts. If it turns out that all related content obtained has a tolerance limit below the predetermined then the process cannot be proceed.

### III. RESULT

#### A. Interface Implementation

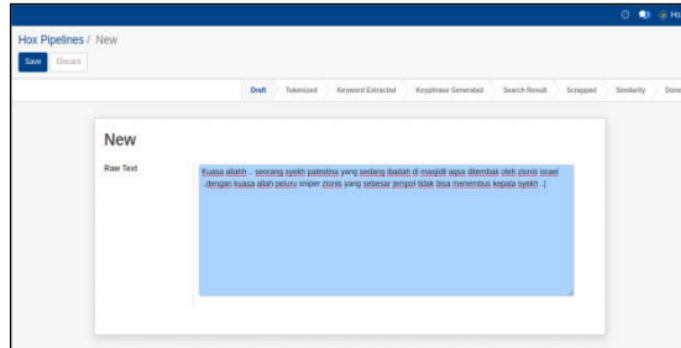
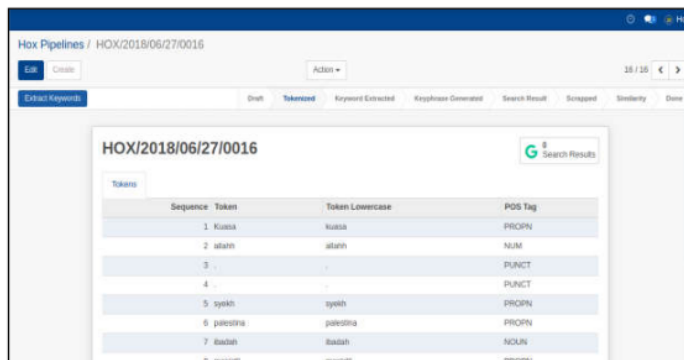


FIG. 2 THE VIEW OF FORM INPUT

Figure 2 shows the page for input content that will be calculated the probability. After the content is entered, the next step is tokenization. The results of the process are displayed in the token table containing the token equipped with POS-Tag as shown in Figure 3.



| Sequence | Token     | Token Lowercase | POS Tag |
|----------|-----------|-----------------|---------|
| 1        | Kuasa     | kuasa           | PROPN   |
| 2        | allah     | allah           | NAME    |
| 3        | .         | .               | PUNCT   |
| 4        | .         | .               | PUNCT   |
| 5        | syekh     | syekh           | PROPN   |
| 6        | palestina | palestina       | PROPN   |
| 7        | adalah    | adalah          | ADJ     |
| 8        | adalah    | adalah          | PROPN   |

FIG. 3 THE VIEW OF TOKEN TABLE IN THE FORM PIPE

The Keywords tab appears which lists keywords that come from calculations using the TextRank method. Each of these keywords has their respective scores as shown in Figure 4.



In Figure 5, the Keyphrase tab shows the keyphrase that results from a combination of keywords generated in the previous stage. Next, the keyphrase will be used to search related content through the Google search engine.

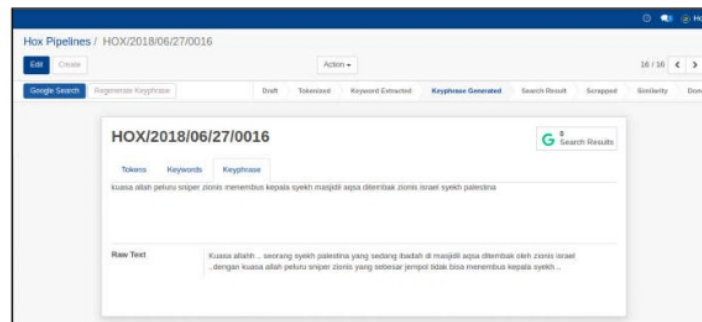


Figure 6 shows that after all the content is scraped, then the calculation of cosine similarity is done to find out which content is most similar or most relevant to the input of the user.

| Name   | External Link   | Similar Percentage (%) |
|--|---|------------------------|
| 1. @KontemporerKipapa #Syekh #Palestina... - Indonesian Hoaxes   | <a href="https://www.facebook.com/7urnidhikHias/posts/1384905001561861?comment_id=183084229056696&amp;comment_facking=67842264229438422990422674">https://www.facebook.com/7urnidhikHias/posts/1384905001561861?comment_id=183084229056696&amp;comment_facking=67842264229438422990422674</a>             | 9                      |
| 2. Peluru Senjata Terampas di Kepala Seorang Syekh Palestina ...   | <a href="https://maksudnya.bekas.peluru-senjata-terampas-di-kepala-seorang-syekh-palestina/">https://maksudnya.bekas.peluru-senjata-terampas-di-kepala-seorang-syekh-palestina/</a>   | 17                     |
| 3. Kuasa Allah - seorang syekh palestina saat ditembak oleh israel ...   | <a href="https://www.youtube.com/watch?v=3M5588Zg8i8">https://www.youtube.com/watch?v=3M5588Zg8i8</a>   | 68                     |
| 4. SUBHANALLAH.. ALLAHU AKBAR Kuasa Allah - seorang Syekh ...  |   | 54                     |
| 5. Foto Peluru Menampas di Kepala Bukan Hoax - Pseudos   | <a href="https://www.presswater.com/indonesia/jawa-prod/2008/12/01/30948267350">https://www.presswater.com/indonesia/jawa-prod/2008/12/01/30948267350</a>   | 28                     |
| 6. Kepala Syekh Palestina Tidak Terbunuh Ditembak Senjata Zionis - Tembak Dikonfirmasi Faktanya? 05/17/2013 15:08:08 WIB | <a href="https://www.uchronia.id/news/Kepala-Syekh-Palestina-Tidak-Terbunuh-Ditembak-Senjata-Zionis-Tembak-Dikonfirmasi-Faktanya?05/17/2013/150808-wib">https://www.uchronia.id/news/Kepala-Syekh-Palestina-Tidak-Terbunuh-Ditembak-Senjata-Zionis-Tembak-Dikonfirmasi-Faktanya?05/17/2013/150808-wib</a> | 28                     |
| 7. Tembak, Peluru Menampas di Kepala Pria Ini Bukan Hoax ...   | <a href="http://www.gmn.com/news/tembak-peluru-menampas-di-kepala-pria-ini-bukan-hoax">http://www.gmn.com/news/tembak-peluru-menampas-di-kepala-pria-ini-bukan-hoax</a>   | 21                     |
| 8. Tembak, Peluru Menampas di Kepala Pria Ini Bukan Hoax   KASKUS  | <a href="https://www.kaskus.co.id/thread/361124d8e2c0cfdb4556/tembak-peluru-menampas-di-kepala-pria-ini-bukan-hoax">https://www.kaskus.co.id/thread/361124d8e2c0cfdb4556/tembak-peluru-menampas-di-kepala-pria-ini-bukan-hoax</a>   | 26                     |
| 9. Foto Peluru Menampas di Kepala Bukan Hoax   Radar Pekanbaru   | <a href="http://radarpekanbaru.com.co.id/542538foto-peluru-menampas-di-kepala-bukan-hoax/">http://radarpekanbaru.com.co.id/542538foto-peluru-menampas-di-kepala-bukan-hoax/</a>   | 13                     |

Figure 7 shows that at the final stage, the percentage of the calculation results is likely to have a fact or vice versa with the chance to hoax.

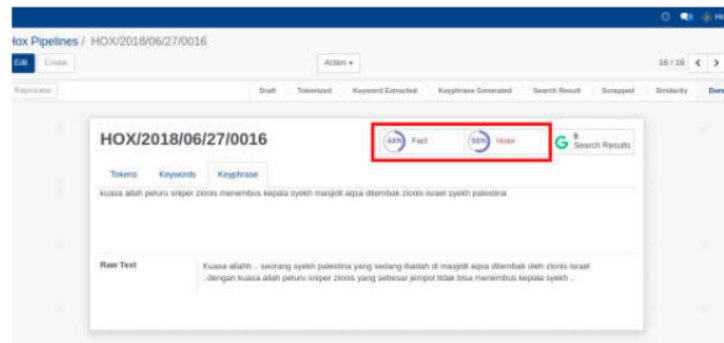


FIG. 7 THE VIEW OF CALCULATING RESULT

### B. System testing

The test uses 20 random content data / issues that have been verified by CekFakta.com and labeled TRUE or FALSE. The content data will then be used as testing data to be compared with the results of the classification carried out by the system. The following is a table of confusion matrix:

TABLE 1 CONFUSION MATRIX

|              |       | Predicted Class |       |
|--------------|-------|-----------------|-------|
|              |       | True            | False |
| Actual Class | True  | 6               | 4     |
|              | False | 1               | 9     |

After the system performs classification, then calculate the accuracy value with Confusion Matrix:

$$\text{Accuracy} = \frac{6+9}{6+9+1+4} \times 100\% = 75\%$$

Accuracy testing data from table 1 which contains 20 issues, consisting of 10 correct issues and 10 wrong issues. Then the system produces a classification with details of 13 issues including wrong and 7 issues including true, then the number of classifications that match the original label are 15 issues. Based on the results of the classification, an accuracy value of 75% was obtained.

### IV. CONCLUSION

Based on the application and testing carried out, it can be concluded that the TextRank Algorithm and Similarity

# HOAX DETECTION AT SOCIAL MEDIA WITH TEXT MINING CLARIFICATION SYSTEM-BASED

## ORIGINALITY REPORT

6%

SIMILARITY INDEX

3%

INTERNET SOURCES

5%

PUBLICATIONS

3%

STUDENT PAPERS

## PRIMARY SOURCES

1

Submitted to Binus University International

Student Paper

1%

2

[www.walsnet.org](http://www.walsnet.org)

Internet Source

1%

3

Yakup Kutlu, Damla Kuntalp. "A multi-stage automatic arrhythmia recognition and classification system", Computers in Biology and Medicine, 2011

Publication

1%

4

Mia Huljanah, Zuherman Rustam, Suarsih Utama, Titin Siswantining. "Feature Selection using Random Forest Classifier for Predicting Prostate Cancer", IOP Conference Series: Materials Science and Engineering, 2019

Publication

1%

5

[www.popekim.com](http://www.popekim.com)

Internet Source

1%

6

[www.springerprofessional.de](http://www.springerprofessional.de)

Internet Source

1%



---

7

Des Suryani, Ahmad Hidayat, Ause  
Labellapansa, Setia Wulandari. "Data Mining  
Classification of Intelligence Quotient in High  
School Students", 2018 Third International  
Conference on Informatics and Computing  
(ICIC), 2018

Publication

---

<1%

---

Exclude quotes Off

Exclude matches Off

Exclude bibliography Off