

13_Detecting Hoaxes in Indonesian News Using TFTDM

by Erizul Eriz

Submission date: 10-Apr-2023 01:43PM (UTC+0700)

Submission ID: 2060352506

File name: 13_Detecting_Hoaxes_in_Indonesian_News_Using_TFTDM.pdf (752.6K)

Word count: 3201

Character count: 16746

PAPER • OPEN ACCESS

Detecting Hoaxes in Indonesian News Using TF/TDM and K Nearest Neighbor

To cite this article: Eri Zuliarso *et al* 2020 *IOP Conf. Ser.: Mater. Sci. Eng.* **835** 012036

View the [article online](#) for updates and enhancements.

You may also like

- [Pneumonia identification based on lung texture analysis using modified k-nearest neighbour](#)
S Kana Saputra, Insan Taufik, Mhd Hidayat et al.
- [Hoax news validation using similarity algorithms](#)
S Y Yuliani, S Sahib, M F Abdollah et al.
- [Hoax News Detection on Twitter using Term Frequency Inverse Document Frequency and Support Vector Machine Method](#)
A Fauzi, E B Setiawan and Z K A Baizal



The Electrochemical Society
Advancing solid state & electrochemical science & technology

243rd Meeting with SOFC-XVIII

Boston, MA • May 28 – June 2, 2023

Early registration discounts end **April 24!**

Accelerate scientific discovery!

Learn More & Register



Detecting Hoaxes in Indonesian News Using TF/TDM and K Nearest Neighbor

Eri Zuliarso¹, Muchamad Taufiq Anwar¹, Kristophorus Hadiono¹, Iswatun Chasanah²

¹Universitas Stikubank, Semarang, Indonesia

²Satya Wacana Christian University, Salatiga, Indonesia

eri299@edu.unisbank.ac.id, muchamadtaufiqanwar@edu.unisbank.ac.id,
kristophorus.hadiono@edu.unisbank.ac.id, uswatunnana@gmail.com

Abstract The presence of the internet and the rapid growth of social media had given rise to the blossoming of hoax creation and distribution through it. A hoax can cause anxiety and reactivity to its readers and could harm a certain party. Thereby, it is important to detect and report hoaxes to stop its spreading as soon as possible. This research aims to utilize the K Nearest Neighbor (KNN) classification algorithm to detect whether a piece of news is a hoax or not. Experiments were done by using 74 hoaxes compiled from Indonesian hoax-debunking community websites and were being compared against 74 real news from various reputable news websites in Indonesia. The result showed that the model could give detection/classification accuracy up to 83.6% and that the model is prone to false positives detections. The characteristics of the resulted model and further research directions are then discussed.

Keywords: hoax detection, fake news detection, Indonesian hoaxes, hoaxes in Bahasa, K Nearest Neighbor, KNN, Term Frequency, Term Document Matrix

I. Introduction

Hoaxes (noun) can be simply defined as fake news, or as defined in the WordNet Dictionary, “Something intended to deceive; deliberate trickery intended to gain an advantage”. Hoaxes had been present throughout the history of mankind to the nowadays internet era. The purpose of hoaxes often includes persuading or manipulating people to do or prevent pre-established actions, usually using threats or deceptions[1]. Aside from its ‘traditional’ kind, recent research had also identified other kind of hoaxes such as the humorous fakes or satire/parody [2].

The presence of the internet and the rapid growth of social media gave rise to the blossoming of hoax creation and distribution through it [3][4][5][6], as stated in [3] that “the boundaries between news production and information creation and sharing are gradually blurring in the current online news environments and social media” and [4] that social media are also increasingly used as vectors for the diffusion of misinformation and hoaxes. The spread of hoaxes has the potential to give extremely negative impacts on individuals and society[5]. The amount of disseminated information and the



Content from this work may be used under the terms of the [Creative Commons Attribution 3.0 licence](https://creativecommons.org/licenses/by/3.0/). Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

rapidity of its diffusion make it practically impossible to assess reliability in a timely manner [4]. Thereby, it is important to detect and report hoaxes to stop its spreading as soon as possible.

The K Nearest Neighbor KNN is one of the versatile and well-known clustering/classification algorithm which often used in data mining. The KNN has lower calculation time compared to other algorithms such as the Random Forest, which is useful for fast automatic detections. The KNN assigns class membership to data based on the majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small).

This research aims to utilize the KNN classification algorithm to detect whether a piece of news is a hoax or not. The model could be implemented within a social media application or even a browser or as an add-on to automatically warn the users if the news appeared in their screen is identified as a hoax.

2. Related Works

Several attempts to build automatic hoax detection had been flourishing in the last decades. The object of detection varied from e-mails (spam or scam) [7][8][9][10][11], social media [5][6][12], news [2][9][13], phishing websites [9], and specific topics like Wikipedia articles [14]. The methods for detections also varied, from the well-known and established method such as the 'traditional' features [13], logistic regression [4], neural networks [11], Levenshtein Distance Method [7], Support Vector Machine (SVM) [13][15][16], Decision Tree / C4.5 [13][15], Linear Regression (LR) [15], KNN [8][9][15], Poisson probabilistic theory [8], Bayesian Classifier [9], Naïve Bayes [8][9][13]; while some also introduced novel methods like the adaptation of boolean crowdsourcing algorithms [4] or combinations of several methods [8][9]. However, the research of hoax detection in the Indonesian language (Bahasa) is still very limited. One notable research is the work by [13] which tested three algorithms, i.e the SVM, C4.5, and Naïve Bayes to detect hoax news in the Indonesian language. The position of this research is to explore the utilization of KNN method in hoax detection for news articles written in Bahasa.

3. Methods

Hoaxes were retrieved from an Indonesian hoax-debunking community website. All articles labeled as [HOAX] and containing narration published between July 31, 2015 - November 22, 2017, was copied into individual text files (or documents). Photo and video hoaxes were excluded. The hoaxes were then being compared against 74 real news from various reputable news websites in Indonesia within a similar range of publication dates. The articles for real news were randomly selected, but we also purposely added articles with similar themes/topics as found in the hoaxes. This is intended for a wildcard. Most of the collected hoaxes were in Indonesian language (Bahasa), but a small number of them were written in English. Thus, we used a similar approach for the number of English articles in the real news articles.

The next step is the creation of corpus. This includes the removal of capitalizations, punctuation marks, and stopwords. Stemming, which turns the words into their basic form, is not applied in this research since stemming could remove the lexical characteristics of the hoaxes vs non-hoaxes, and thus decrease the model accuracy[13]. Since we used two kinds of languages, we also incorporated two kinds of stopwords collections for those languages. The corpus was then being transformed into Term Document Matrix (TDM) which uses unigram (the sentences were split into individual words). The TDM consists of the word count (frequency) found in a document (article, text file) for each word listed in the corpus. The TDM is a matrix with the size of $m \times n$, where m is the number of columns which correspond to the number of words found in the corpus, while n is the number of rows resembling the number of documents (one row for each document). Not all terms found in the corpus is listed into the TDM. We used a certain limit to exclude words (terms) which have a low occurrence (high sparseness). We called this the Sparse Terms Limit (STL). The TDM was then being added an additional column to store the label for each document, whether it is 'hoax' or 'not hoax'.

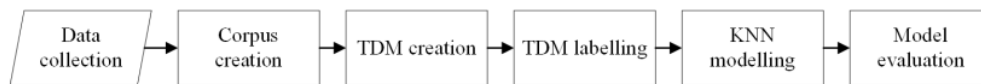


Figure 1. Research methodology

The final form of the TDM was then being randomly grouped into training data sets and the testing datasets using a certain proportion (we called this the Training/Testing Proportion, TTP). The training datasets were then being fed into the KNN algorithm with $k=1$ (Nearest Neighbor Rule). The model generated by the KNN was then being tested using the testing datasets. The accuracy of its prediction was then calculated based on its confusion matrix.

We did nine experiments with varied STL and TTP, and each experiment was repeated 1000 times to find the average accuracy for each corresponding setting of STL and TTP. The STL values used are: 0.8, 0.85, and 0.9; whereas the TTP values are: 0.7, 0.8, and 0.9. TTP value of 0.7 means that 70% of the data were used as the training sets, while the other 30% were used as the testing sets. Since the hoax articles generally have a smaller amount of word counts, in this research, the STL value below 0.8 would not give any terms, so we set the experiments with the STL value started from 0.8. STL value of 1 means that every word found in the corpus will be included in the TDM. Experiments were done using the RStudio and utilized the ‘class’ package.

4. Result and Discussion

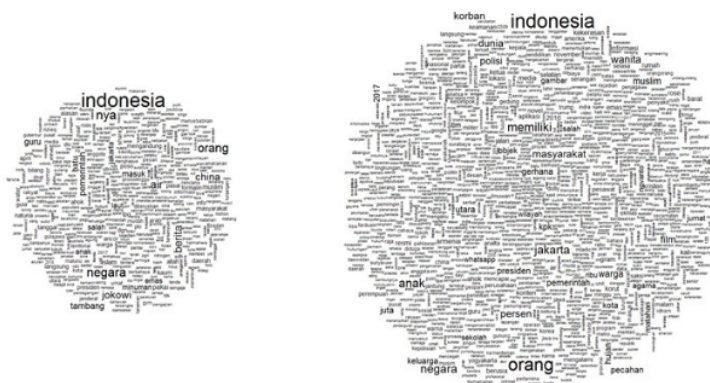


Figure 2. The comparison between word clouds for the hoaxes (left); and not hoaxes (right)

The hoax articles we had collected generally have a smaller amount of words compared to the real news articles, as shown in Figure 2, which also means that the hoaxes have a smaller amount of used vocabulary (and thus fewer words in the corpus). This agreed with the finding of [10]. We identified some most recurring themes in the hoaxes articles as shown in Table 1 (some article has overlapping themes). We also found that some of the articles themed religion, politics, and racism/ethnicity have a sentimental tone towards other groups. This finding agrees with the situation in Indonesia, which has recurring issues with religions, and the recent political ‘gossips’ regarding the religion and ethnicity of the politicians. The sentimental religion-themed hoax might not directly confront certain group, but by incorporating certain religious symbolism, to be combined with the presented bad image. The other religion-themed hoaxes did not have sentimental tone but they tried to ‘convince’ the reader to strengthen their belief towards their religion by presenting false claims. The hoax articles often use a less standardized form of written language, ‘common’ vocabularies, provocative words, capitalizations, and a lot of exclamation marks.

Table 1. The most recurring themes in the Indonesian hoax articles

Theme	Count	Percentage
Religion*	27	36%
Politics*	19	26%
Health	10	14%
Racism / Ethnicity*	9	12%
Technology	8	11%
Other	22	30%

The experimental result showed that the model could achieve up to 83% accuracy in determining whether an article is a hoax or not. The accuracy of all experiment settings is shown in Table 2. The table showed that the higher value of TTP had led to slightly better accuracy. This, of course, is not surprising, since the more data is fed into the model, the better it will learn and recognize the articles. On the other hand, the higher value of the STL led to worse prediction accuracy. This can be explained since the higher value of STL will make the TDM include more sparse words, by which that the sparse words from both 'hoaxes' and 'not hoaxes' collections would appear together in the TDM and consequently 'confused' the model and led to false detections. The accuracy of the model is also affected by the value of K, where the smaller value of K gives a better detection accuracy as shown in Table 3.

Table 2. The experiment results showing the evaluation of model accuracy (averaged from 1000 experiments) in predicting whether a piece of news is a hoax or not, for varied STL and TTP

Sparse Terms Limit	Training/Testing Proportion		
	0.7	0.8	0.9
0.80	81.4%	82.4%	83.6%
0.85	79.1%	79.7%	80.5%
0.90	67.4%	68.5%	69.4%

When we looked at the confusion matrices, we found that the models often fall for false-positive detections (the article were actually not hoaxes, but falsely identified/predicted as hoaxes). An example of the confusion matrices is shown in Table 3. The cause of false-positive detections in this research has not been investigated.

Table 3. The effect of K on model accuracy

Value of K	Model Accuracy (%)
1	83.6
2	77.9
3	75.9
4	76.4
5	74.1

Table 4. Example of Confusion Matrices for STL = 0.80 and TTP = 0.7; with the accuracy of 93% (left); and for STL = 0.90 and TTP = 0.7; with the accuracy of 64% (right) showing that the model is prone to false positive detections

Predictions	Actual		Predictions	Actual	
	Not Hoax	Hoax		Not Hoax	Hoax
Not Hoax	13	0	Not Hoax	7	2
Hoax	3	28	Hoax	14	21

5. Conclusion and further research directions

From the 74 hoaxes compiled from Indonesian hoax-debunking community websites, it was found that the most recurring themes of Indonesian hoaxes include religion, politics, health, racism/ethnicity, and technology, by which some of the religious, political, and ethnical hoaxes has some sentimental tone in them. Those hoaxes were then being compared against 74 real news from various reputable news websites in Indonesia. The KNN modeling result shows that the model is able to give detection/classification accuracy up to 83.6%. The accuracy of the detection model is affected by the value of K, Sparse Terms Limit, and the number of training data. Result also showed that the model is prone to false positives detections.

The cause of false-positive detections, which affect the model accuracy, has not yet investigated. This might be affected by the number of words included in the TDM (which is affected by article length in the article collections), the value of K, and the number of data used in the experiments. Thus, further research could experiment with varied ratios of words in TDM for both words in the hoax and non-hoax collections, experiment with different values of K, and a varied number of data. The TDM could also be fed into other classification algorithms to compare their performance. The TDM could also be fed into other machine learning algorithms such as the apriori association rules to give additional insight into the characterization of hoaxes vs real news.

References

- [1] Hernandez, J. C., Hernandez, C. J., Sierra, J. M., & Ribagorda, A. (2002). A first step towards automatic hoax detection. In *Security Technology, 2002. Proceedings. 36th Annual 2002 International Carnahan Conference on* (pp. 102-114). IEEE.
- [2] Rubin, V. L., Chen, Y., & Conroy, N. J. (2015). Deception detection for news: three types of fakes. *Proceedings of the Association for Information Science and Technology*, 52(1), 1-4.
- [3] Chen, Y., Conroy, N. J., & Rubin, V. L. (2015). News in an online world: The need for an "automatic crap detector". *Proceedings of the Association for Information Science and Technology*, 52(1), 1-4.
- [4] Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S., & de Alfaro, L. (2017). Some Like it Hoax: Automated Fake News Detection in Social Networks. *arXiv preprint arXiv:1704.07506*.
- [5] Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter*, 19(1), 22-36.
- [6] GUNASEKARAN, K., GANESAN, G., RAMANUJAN, S. S., & SRINIVASAN, B. Fake News Detection in Social Media.
- [7] Chen, Y. Y., Yong, S. P., & Ishak, A. (2014). Email Hoax Detection System Using Levenshtein Distance Method. *JCP*, 9(2), 441-446.
- [8] Saberi, A., Vahidi, M., & Bidgoli, B. M. (2007, November). Learn to detect phishing scams using learning and ensemble? methods. In *Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Workshops* (pp. 311-314). IEEE Computer Society.
- [9] Mokhtari, M., Saraee, M., & Haghshenas, A. (2008). A novel method in scam detection and prevention using data mining approaches. *Proceedings of IDMC2008*, 1-11.
- [10] Horvitz, E., Heckerman, D. E., Dumais, S. T., Sahami, M., & Platt, J. C. (2000). U.S. Patent No. 6,161,130. Washington, DC: U.S. Patent and Trademark Office.
- [11] Vuković, M., Pripuzić, K., & Belani, H. (2009). An Intelligent Automatic Hoax Detection System. *Knowledge-Based and Intelligent Information and Engineering Systems*, 318-325.
- [12] Wang, A. H. (2010, July). Don't follow me: Spam detection in twitter. In *Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on* (pp. 1-10). IEEE.
- [13] Rasywir, E., & Purwarianti, A. (2016). Eksperimen pada Sistem Klasifikasi Berita Hoax Berbahasa Indonesia Berbasis Pembelajaran Mesin. *Jurnal Cybermatika*, 3(2).
- [14] Kumar, S., West, R., & Leskovec, J. (2016, April). Disinformation on the web: Impact, characteristics, and detection of Wikipedia hoaxes. In *Proceedings of the 25th International*

- Conference on World Wide Web (pp. 591-602). International World Wide Web Conferences Steering Committee.
- [15] Dakwala, A., & Lavingia, K. A Machine learning approach to improve the efficiency of Fake websites detection Techniques. International journal of Computer Science and Communication (IJCSC) Vol, 7, 236-243.
- [16] Ahmed, H., Traore, I., & Saad, S. (2017, October). Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In International Conference on Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments (pp. 127-138). Springer, Cham.
- [17] A. P. Windarto *et al.*, "Analysis of the K-Means Algorithm on Clean Water Customers Based on the Province," *J. Phys. Conf. Ser.*, vol. 1255, no. 1, 2019, doi: 10.1088/1742-6596/1255/1/012001.
- [18] S. Sundari, Karmila, M. N. Fadli, D. Hartama, A. P. Windarto, and A. Wanto, "Decision Support System on Selection of Lecturer Research Grant Proposals using Preferences Selection Index," *J. Phys. Conf. Ser.*, vol. 1255, no. 1, pp. 1–8, 2019, doi: 10.1088/1742-6596/1255/1/012006.
- [19] S. R. Ningsih, R. Wulansari, D. Hartama, A. P. Windarto, and A. Wanto, "Analysis of PROMETHEE II Method on Selection of Lecturer Community Service Grant Proposals," *J. Phys. Conf. Ser.*, vol. 1255, no. 1, pp. 1–7, 2019, doi: 10.1088/1742-6596/1255/1/012004.
- [20] Sudirman, A. P. Windarto, and A. Wanto, "Data mining tools | rapidminer: K-means method on clustering of rice crops by province as efforts to stabilize food crops in Indonesia," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 420, no. 1, 2018, doi: 10.1088/1757-899X/420/1/012089.
- [21] T. Imandasari, M. G. Sadewo, A. P. Windarto, A. Wanto, H. O. Lingga Wijaya, and R. Kurniawan, "Analysis of the Selection Factor of Online Transportation in the VIKOR Method in Pematangsiantar City," *J. Phys. Conf. Ser.*, vol. 1255, no. 012008, pp. 1–7, 2019, doi: 10.1088/1742-6596/1255/1/012008.

13_Detecting Hoaxes in Indonesian News Using TFTDM

ORIGINALITY REPORT

15%

SIMILARITY INDEX

16%

INTERNET SOURCES

9%

PUBLICATIONS

8%

STUDENT PAPERS

MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

14%

★ iopscience.iop.org

Internet Source

Exclude quotes On

Exclude matches < 2%

Exclude bibliography On