Plagiarism Checker X Originality Report

Similarity Found: 7%

Date: Sabtu, September 01, 2018
Statistics: 159 words Plagiarized / 2347 Total words
Remarks: Low Plagiarism Detected - Your Document needs Optional Improvement.

HOAX DETECTION AT SOCIAL MEDIA WITH TEXT MINING CLARIFICATION SYSTEM-BASED Aditya Gusti Tammam1), Sucipto2), Rini Indriati3) 1, 2,3)Information System department, University of Nusantara PGRI Kediri Jl K.H Achmad Dahlan 76 Kediri 64111, East Java, Indonesia e-mail: gustitammam@gmail.com1), sucipto@unpkediri.ac.id2), rini.indriati@unpkediri.ac.id3) 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. Introduction E very 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: / hoks /) 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 Purwarianti 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 Purwarianti 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. 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 Mihalcea & 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: Where the value of S (Vi) is the value of the Vi 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]: Informations :A = Vector A, which will be compared its resemblance B = Vector B, which will be compared to the similarity A • B = dot product between vector A and vector B | A | = length of vector A | B | = length of vector B | 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: 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. • 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. 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. 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. Fig. 4 The view of Keyword table in Form Pipe 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. Fig.

5 the view of Keyphhrase resulted 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. Fig. 6 The View of Cosine Similarity Result 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: Accuracy = _ = 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. Conclusion Based on the application and testing carried out, it can be concluded that the TextRank Algorithm and Similarity Algorithm can be combined to be used in helping the classification of news content whether hoax or facts with accuracy

rate of 75%. References V.

Juliswara, 2017, “Mengembangkan Model Literasi Media yang Berkebhinnekaan dalam Menganalisis Informasi Berita Palsu (Hoax) di Media Sosial,” J. Pemikir. Sosiol., vol. 4, no. 2, p. 142. Abner, Khaidir, M. R. Abdillah, R. Bimantoro, and W. Reinaldy, 2013, “Penyalahgunaan Informasi/Berita Hoax di Media Sosial,” in International Conference on Advances Science and Contemporary Engineering (ICASCE) 2013, accessed 23 November 2017 J. Hintzbergen, K. Hintzbergen, A. Smulders, and H.

Baars, 2010, Foundations of Information Security Based on ISO27001 and ISO27002, 2nd ed. Zaltbommel: Van Haren Publishing. Badan Pengembangan dan Pembinaan Bahasa, 2016, Hoaks - KBBI Daring. Jakarta: Kemdikbud. Ilham, 2017, Ahli: Hoax Merupakan Kabar yang Direncanakan [Online]. Available: http://nasional.republika.co.id/berita/nasional/hukum/17/01/11/ojm2pv361-ahli-hoax-merupakan-kabar-yang-direncanakan, accessed 26 November 2017. A. A.

Sawitri, 2017, 4 Penyebab Hoax Mudah Viral di Media Sosial [Online]. Available: https://nasional.tempo.co/read/838621/4-penyebab-hoax-mudah-viral-di-media-sosia, accessed 29 November 2017. Novaldi, 2017, Pakar IT: Tangkal Hoax dengan Literasi Media [Online]. Available: https://kominfo.go.id/content/detail/9725/pakar-it-tangkal-hoax-dengan-literasi-media/0/sorotan_media, accessed 30 November 2017. R. Feldman and J. Sanger, 2006, The Text Mining Handbook. New York: Cambridge University Press. L. Dyson and A.

Golab, 2017, “Fake News Detection Exploring the Application of NLP Methods to Machine Identification of Misleading News Sources,” CAPP 30255 Adv. Mach. Learn. Public Policy. E. Rasywir and A. Purwarianti, 2015, “Eksperimen pada Sistem Klasifikasi Berita Hoax Berbahasa Indonesia Berbasis Pembelajaran Mesin,” J. Cybermatika, vol. 3, no. 2, pp. 1–8. R. Mihalcea and P.

Tarau, 2004, “TextRank: Bringing Order into Texts,” in Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. R. T. Wahyuni, D. Prastiyanto, and E. Supraptono, 2017, “Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen Skripsi,” J. Tek. Elektro, vol. 9, no. 1, pp. 18–23. A. Indriani, 2014, “Klasifikasi Data Forum dengan menggunakan Metode Naïve Bayes Classifier,” Semin. Nas. Apl. Teknol. Inf. Yogyakarta, vol.

21, no. 5, pp. 1907–5022. Sucipto, Kusrini, and E. L. Taufiq, “Classification method of multi-class on C4.5 algorithm for fish diseases,” in Proceeding - 2016 2nd International Conference on Science in Information Technology, ICSITech 2016: Information Science

for Green Society and Environment, 2016, pp. 5–9.

INTERNET SOURCES:

<1% -
https://www.antaranews.com/berita/690882/polri-tidak-dapat-dipisahkan-dari-globalisasi
<1% - https://www.scribd.com/document/206863375/jurn
<1% - https://www.pearson.com/corporate/efficacy-and-research/reports.html
<1% -
https://www.researchgate.net/publication/280092953_An_Overview_of_Graph-Based_Keyword_Extraction_Methods_and_Approaches
<1% - https://en.m.wikipedia.org/wiki/Transgender
1% -
https://www.researchgate.net/profile/Francisco_Oliveira/publication/272842504_Computer-aided_diagnosis_of_Parkinson%27s_disease_based_on_123IFP-CIT_SPECT_binding_potential_images_using_the_voxels-as-features_approach_and_support_vector_machines/links/55018df60cf2d60c0e5f37f8.pdf?origin=publication_detail
<1% - https://en.wikipedia.org/wiki/Big_data
<1% -
https://www.scribd.com/document/357969434/Text-Mining-Techniques-Applications-and-Issues2
<1% -
http://turnbull.dcs.st-and.ac.uk/history/Search/historysearch.cgi?SUGGESTION=Differential+equation*&CONTEXT=1
<1% - https://www.sciencedirect.com/science/article/pii/S2212420916303715
<1% - https://www.scribd.com/doc/173963311/Another-Introduction-to-Ray-Tracing
<1% - https://link.springer.com/article/10.1007/s12652-012-0143-x
<1% - https://link.springer.com/article/10.1007%2Fs11192-014-1272-0
<1% -
https://www.scribd.com/document/189460336/Making-Social-Media-Analysis-More-Efficient-Through-Taxonomy-Supported-Concept-Suggestion
<1% - http://www.irdindia.in/journal_ijraet/pdf/vol3_iss3/6.pdf
<1% -
https://www.slideshare.net/IJOERS/sentiment-analysis-methodology-of-twitter-data-with-an-application-on-hajj-season
<1% - https://www.sciencedirect.com/science/article/pii/S0920548911001097
<1% -

https://www.searchtechnologies.com/blog/natural-language-processing-techniques
<1% - https://www.scribd.com/document/307691124/Proceedings-CLICit-2014
<1% -
https://www.researchgate.net/profile/Ludovic_Kurunczi/publication/317973064_Docking_study_of_3-mercapto-124-triazole_derivatives_as_inhibitors_for_VEGFR_and_EGFR/links/596a3a100f7e9b8091945f62/Docking-study-of-3-mercapto-1-2-4-triazole-derivatives-as-inhibitors-for-VEGFR-and-EGFR.pdf
<1% - https://www.sciencedirect.com/science/article/pii/S1566253506001151
<1% - http://journals.sagepub.com/doi/10.1177/0165551511432670
<1% - https://dl.acm.org/ft_gateway.cfm?id=2010015&type=pdf
<1% -
https://www.bookdepository.com/Foundations-of-Information-Security-Hans-Baars-Jule-Hintzbergen-Kees-Hintzbergen-Andre-Smulders-Van-Haren-Publishing/9789087535681
<1% - http://www.tandfonline.com/doi/ref/10.1080/13658816.2014.978869
1% - https://www.sciencedirect.com/science/article/pii/S1877050917320975
<1% - http://eprints.undip.ac.id/view/type/article.default.html