ISSN 2278-3091
Volume 9, No.4, July – August 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse297942020.pdf
https://doi.org/10.30534/ijatcse/2020/297942020

Support Vector Machine-Based Hoax Detection on Indonesian Online News

Pangondian Prederikus Sihombing[1], Riyanto Jayadi[2], Edward Chandra[3], Stefanie Liu[4]
[1]Information Systems Management Department, BINUS Graduate Program – Master of Information Systems Management, Bina Nusantara University, Jakarta, Indonesia,pangondian.prederikus001@binus.ac.id[1], riyanto.jayadi@binus.edu[2], edward.chandra003@binus.ac.id[3] , stefanie.liu@binus.ac.id[4]

ABSTRACT

The rapid development of information technology and social media resulted in more information dissemination in the form of digital news and transformed analog media into online media. However, the spread of news in online media is not all true (hoax). The problems faced by internet users of hoax news can be solved through pattern recognition. Research data was conducted of 2,000 Indonesian news, 1,000 for non-hoax news, and 1,000 for hoax news. Data was collected from Indonesian online news portals by crawling text methods. The research model was built using CRISP-DM methodology. The results showed that as high as 96,01% accuracy can be achieved using Support Vector Machine. Based on these results, it can be concluded that this modeling can be used to support the detection of hoax news in Indonesia today. It is hoped that this model can also be applied to help the government filter news that will be distributed to the people of Indonesia.

Key words : CRISP-DM, Text Mining, Decision Tree, Random Forest, Support Vector Machine, Gradient Boosted Tree.

1. INTRODUCTION

Information technology is an infrastructure system such as hardware and software which functions to obtain, collect data, process, interpret, store and organize, and use the information for useful needs [1]. The development of information technology and social media is proliferating with more and more information dissemination in the form of digital news and transforming analog media into online media.

Society now quickly gets information and access to news through online media. The flow of information in digital form is experiencing rapid growth, both in terms of data volume and time of news dissemination. However, the spread of news in online media is not all true (hoaxes). Dissemination of hoaxes has adverse effects such as disunity, damage, material or non-material losses, psychological, distrust of the public, to hatred. In 2019 there were 3,801 hoax issues in Indonesia. The three most significant categories are politics, government, and health [2]. Most of the famous fake news stories were more widely shared on Facebook than the most popular mainstream news stories.

A sizable number of people who read fake news stories have reported that they believe them more than news from mainstream media. Dewey claims that fake news played a massive role in the 2016U.S. election and that it continues to affect people's opinions and decisions [3].

The attitude of the Indonesian Government in the phenomenon of hoax news has been presented and there are several articles that are ready to be applied to the disseminator of hoax written in Law No.11 of 2008 concerning Information and Electronic Transactions (ITE), Law No.40 of 2008 concerning the Elimination of Racial and Ethnic Discrimination, and actions when hatred has caused social conflict. KOMINFO also provides Trust+ services whose task is to find sites that contain negative content and then be blacklisted in Trust +.

There are 800,000 sites indicated as hoax news spreaders in Indonesia. The high number of hoax news that is supported by internet access causes the rapid spread of hoax news, both in urban and rural areas [4],[5] Its outbreak has become a national problem that causes disunity, political instability and security disturbances that potentially hamper national development[6]. This causes the need for mechanisms to detect hoax news, one of which is by using machine learning[7],[8].The application of computer based Machine Learning can be built based on certain thoughts that enable the detection of hoax news to work independently, by minimizing human intervention. The more data that is processed, the error rate decreases and inversely proportional to the higher level of accuracy.

Text Mining is a branch of data mining to analyze data in text form. Text mining has the objective to obtain useful information from a collection of documents and to find words that can have content from a document so that it can analyze the relationships between the documents. The source of data that can be used in text mining is a collection of texts that have a structured format or a semi minimal structure. Specific tasks of text mining are categorizing Text (text categorization) and grouping text (text clustering).Until now, text mining has

been widely applied in fields such as security, biomedical, software and applications, online media, marketing, education/academic, and other fields[9],[10]. Some of the practical applications of text mining techniques are spam filtering, suggestion and recommendations, monitoring public opinions, customer service, email support, fraud detection, and fighting cyberbullying or cybercrime[11].

Artificial Intelligence (AI) is a human made intelligence that was created and added to an adjustable system or machine.AI is incorporated into the machine to be able to do work like humans, especially in the field of analysis and assist decision making[12]. Examples of the application of AI to machines such as systems such as Bots or Chatbots that aredeveloping in Indonesia, Robotics Technology, Development of language detection to face detection machines.

The existence of Text Mining technology is supported by increasingly advanced technological devices that can help solve complex problems. The problems faced by internet users of hoax news can be solved through pattern recognition and deep learning. The purpose of this paper is to determine the classification and predict hoax or non-hoax of an Indonesian news portal.

Several previous studies have been studied this problem. In another previous study conducted by [13] regarding the identification of hoaxes on social media with a machine learning approach. The hoax identification process with machine learning approach is grouped into 2 namely pre-processing and without pre-processing approaches. Approaches with higher pre-processing are higher, while approach-es without pre-processing are higher in the automation pro-cess. The aim of other researchers is the naïve Bayes experiment on hoax news detection in Indonesian [14]. The approach using the TextRank algorithm for the keyword extract-ion and the Cosine similarity algorithm is used to measure the similarity of the document and can measure the potential news hoax on the news [15]. In addition [7],[14],[16]-[18] have studied similar problem including Indonesia news classification and E-mail hoax [19].

In contrast to previous research [5], [7], [8], [20], [21], the dataset used in this study was collected from various Indonesian news portals with a total of 2000 data. News articles use data that has been proven to be a hoax category and are taken from various online sources. The dataset is enriched so that the application can run optimally in the training process. The modelling phase employs decision tree algorithm and random forest algorithms. The algorithm is optimized so that it can get high accuracy. The development of previous algorithm models can produce more significant results and make a comparison of each classification used.

This study follows the Cross-Industry Standard Process for Data Mining (CRISP-DM), which is a standard that was developed in 1966, used for the analysis process to solve a problem/research unit [22]. The business understanding part is described in this section. Data understanding presents at Section 2 and Section 3 involving data source description and tag cloud visualization. The data preparation phase presents at Section 2 as well as the modelling phase. Evaluation phase including showing the results is on Section 3. Meanwhile, Section 4 conclude this study with implications of this research and the deployment suggestions**.**

2. PROPOSED INDONESIAN HOAX DETECTION SYSTEM

To detect the news category, several criteria for hoax news and non-hoax news are needed. This criterion is used as training data so that it can distinguish between types of test news. Figure 1 shows the stages of the hoax news detection process. Further explanation of Figure 1:

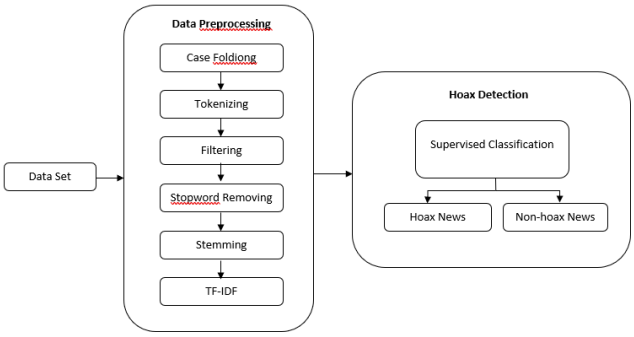

Figure1: Proposed hoax news detection stages

In pre-processing data phase, terms that do not contain content such as numbers, punctuation, conjunctions, abbreviations, uppercase and lowercase letters and erase the beginning and the end of sentences will be eliminated. The pre-processing phase will clear the news data and leave only the main word. Hoax Detection phase will train the classifier by using 80% of the dataset, and test the classifier by using 20% of the dataset. The dataset will be used as output where the results of the training data will be processed by the decision tree and random forest methods.

Decision Tree is one of the most popular classification methods and is easily interpreted by people. Decision Tree is a prediction model that uses tree structures or hierarchical structures. Decision Tree is a classification algorithm in decision making, and the data will continue to be divided with certain parameters[23]. The benefits of using decision tree are it can easily handle qualitative (categorica) features, works well with decision boundaries parallel to the feature axis and a very fast algorithm for both learning prediction [12][24].In the Decision Tree, there are parameters used such as quality measure using the Gini index which aims to measure the level of inequality, not using the pruning method which functions to reduce overfitting which can improve the quality of predictions, Reduced Error Pruning, at least number records per node 2, number records to store for view 10,000, average split point, number threads 4 and skip nominal columns without domain information.

In Random Forest, there are parameters that are used such as Use column attributes or ordinary column attributes (string, double, integer) to study the model. This dialog allows us to manually move columns from the left pane to the right pane for lessons. Random Forest is the development of a decision tree by using several Decision Trees in which at each stage training has been carried out using individual samples, and each attribute is broken down in a tree chosen between the attributes of a random subset. In the classification process, individuals are based on votes of the most votes in the population tree collection [25]. In this algorithm, it yields an ensemble that can achieve both low bias and low variance [11]. The benefits of using random forest algorithm are the ensembles of decision trees are very fast to train in comparison to other techniques, reduced variance (relative to regular trees) and it does not require preparation and pre-processing of the input data [24].

In Support Vector Machine, there are parameters that are used, such as selecting the document class as the target, overlapping penalty 1.0 which is used to determine how many penalties are given for each wrongly classified point (best value 1), kernel type is polynominal with power 1.0, bias 1.0 and gamma 1.0.

In the Gradient Boosted Tree, there are parameters that are used such as the target column to be studied, attribute selection using use column attributes or ordinary column attributes (string, double, integer) to learn the model. This dialog allows us to manually move columns from the left pane to the right pane for lessons.

To evaluate the performance of algorithms used in this research, various evaluation metrics will be used. Most existing approaches consider the hoax news classification that predicts whether a news article is hoax or not [26]:

- True Positive (TP): hoax news pieces are correctly predicted as hoax news;
- True Negative (TN): non-hoax news pieces arecorrectly predicted as non-hoax news;
- False Negative (FN): non-hoax news pieces areincorrectly predicted as hoax news;
- False Positive (FP): hoax news pieces are incorrectly predicted as non-hoax news.

Accuracy measures the similarity between predicted hoax news and real hoax news, meanwhile Precision measures the fraction of all detected hoax news that are predicted as hoax news, addressing the important problem of identifying which news is hoax.The higher the value of the precision and accuracy illustrates the better the performance. The AUC-ROC will function to assess and measure performance on a classification problem by forming a probability curve, whereas AUC represents the level of class separation. The closer the ROC is to the angle at 45 degrees, the less accurate the model test accuracy is. Conversely, the higher the AUC, the higher the accuracy.

3. RESULTS AND DISCUSSION

In this research, the source of data to be processed has been collected from online news website pages. Text data comes from online news websites with non-hoax news content and news with hoax narration that is done using the Text Crawling technique. Online news sources are obtained from the pages of kompas.com merdeka.com, detik.com, cnn.com, turnbackhoax.id, republika.co.id, tribunnews.com, tempo.co, viva.co.id, metronews.co.id. The number of crawled text data is 1,000 data for non-hoax news and 1,000 data for hoax news. Non-hoax news data is collected from 2017 to 2020 and hoax data is collected from 2015 to 2020, which contains the title, content, sources and labels.

Tag Cloud is used to visualize the words most often used in online news that has been collected [27]. In this case, there are words that are most often used in the case of hoaxes and non-hoaxes.Based on the results of tag cloud visualization, there are two types of word classification that represent each category (hoax and non-hoax). Figure 2 shows the most dominant words on the hoax tag cloud are Indonesia, ahok, and president. Whereas,Figure 3 shows the most dominant words in the non-hoax tag cloud are Jakarta, Golkar, and Jokowi.

Figure 2: Tag Cloud for Hoax

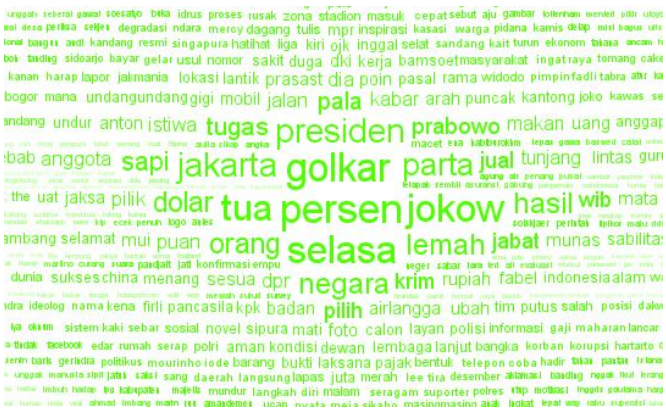

Figure 3: Tag Cloud for Non Hoax

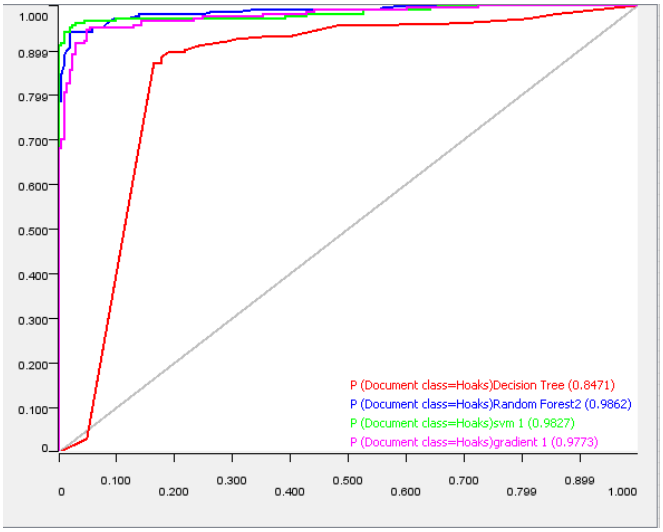

Figure 4: ROC Decision Tree, Random Forest, SVM Predictor and Gradient Boosted Trees

Tabel 1: Confusion Matrix Decision Tree

Document	Hoax (P)	Non Hoax (N)
Hoax (P)	174 (TP)	26 (FP)
Non Hoax (N)	36 (FN)	165 (TN)

Tabel 2: Confusion Matrix Random Forest

Document	Hoax (P)	Non Hoax (N)
Hoax (P)	190 (TP)	10 (FP)
Non Hoax (N)	15 (FN)	186 (TN)

Tabel 3: Confusion Matrix SVM Predictor

Document	Hoax (P)	Non Hoax (N)
Hoax (P)	193 (TP)	7 (FP)
Non Hoax (N)	9 (FN)	192 (TN)

Tabel 4: Confusion Matrix Gradient Boosted Tree

Document	Hoax (P)	Non Hoax (N)
Hoax (P)	190 (TP)	10 (FP)
Non Hoax (N)	26 (FN)	175 (TN)

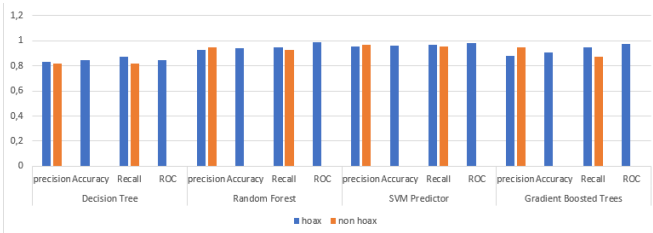

Figure 5: Modelling Result

Based on table 5, the model evaluation table using different data there are results that indicate precision, recall, accuracy and ROC.[14]used a dataset of 600 valid news and hoaxes. This research uses Naive Bayes classification and library component PHP-ML. This test produces precision according to the system of 68.33% and an accuracy of 82.6%.

Research conducted by [15] was designed with an Unsupervised Learning approach that did not use training data. In the construction of the model using the TextRank algorithm to extract keywords and Cosine Similarity algorithm to check the similarity of documents. The study used 20 news data which had been divided into 10 valid and 10 hoaxes. Based on the results of the classification, researchers get an accuracy of 75%.

The proposed method on [8] is the Naives Bayes algorithm to verify news. In this study there are two performance evaluation work, namely with and without a source (URL). This system is able to detect hoax news very well in both conditions. Based on the results obtained, Precision of 91%, 100% recall and accuracy of 87%.

In this research, table 1 shows the decision tree method resulted 87% of precision, 84% of accuracy, 82,8% recall and the ROC is 84,7%. Meanwhile, table 2 shows the random forest method resulted 95% of precision, 93,8% of accuracy, 92,6% recall and the ROC is 98,6%.Meanwhile, table3 shows the SVM Predictor method resulted 96% of precision, 96% of accuracy, 95,5% recall and the ROC is 98,2%.Table 4 shows Gradient Boost Trees method resulted 95% of precision, 91% of accuracy, 87,9% recall and the ROC is 97,7%. The SVM Predictor method provided higher accuracy than the other method and previous research.

Tabel 5: Comparison between the performance of our method and those of other studies

Research	Method	Data Set	Precision (%)	Recall (%)	Accuracy	ROC
Our Proposed Method	Decision Tree	2000	89.9	89.9	87.7	0,89
	Random Forest		93.5	97	95.8	0,74
	SVM		96	95,5	96	0,98
	Gradient Boost		95	87,9	91	0,97
[14]	Naïve Bayes	600	68.33	90.6	82.6	-
[15]	TextRank and Cosine Similarity	20	-	-	75	-
[8]	Naïve Bayes	250	91	100	87	-
[28]	SVM	608.738	-	-	97,28	-
[29]	Decision Tree	-	61	70	58	-

4. CONCLUSION

In this research, a hoax news classification system in Indonesian news portal has been formed using machine learning with different types of algorithms, namely: Decision Tree, Random Forest, SVM Predictor and Gradient Boosted Tree with a total of 2,000 news which consists of 1,000 non-hoax news and 1,000 hoax news obtained from Indonesia online news portals. This research also has been through the tokenizing process, case folding, normalization, filtering, stopwords removing, stemming, and TF-IDF.

The results show that the process of news sentiment analysis is done by calculating the amount of weight of hoax and non-hoax sentiments contained in Indonesian language news content. After that, there is a classification process through the preprocessing stage, sentiment analysis and calculating the weight of words with TF-IDF. Based on these weights will be calculated the closeness of the word with test data and training data. With the process of analyzing the words in the news will be able to analyze the news sentiments contained in the word. Based on the results of the sentiment analysis using the SVM Predictor methodresulted 96% of precision, 96% of accuracy, 95,5% recall and ROC 0,98.

Based on these results, it can be concluded that this modeling can be used to support the detection of hoax news in Indonesia today. It is hoped that this model can also be applied so that it helps the government filter news that will be distributed to the people of Indonesia. That way the government can minimize news that is not properly spread as the consumption of Indonesian

REFERENCES

1. P. G. McKeown, ***Information technology and the networked economy***. Harcourt College Publishers Fort Worth, TX, 2001.
2. KOMINFO, "**Data Statistik Hoax Agustus 2018 - 31 Desember 2019**," 2019.
3. H. Kudarvalli and J. Fiaidhi, "**Detecting Fake News using Machine Learning Algorithms**," 2020.
4. I. Nadzir, S. Seftiani, and Y. S. Permana, "**Hoax and Misinformation in Indonesia: Insights from a Nationwide Survey**," *Perspective*, no. 92, pp. 1–12, 2019, [Online]. Available: https://www.iseas.edu.sg/images/pdf/ISEAS_Perspective_2019_92.pdf.
5. S. Y. Yuliani, M. F. Bin Abdollah, S. Sahib, and Y. S. Wijaya, "**A framework for hoax news detection and analyzer used rule-based methods**," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 10, pp. 402–408, 2019, doi: 10.14569/ijacsa.2019.0101055.
6. A. Nugroho, "**The Analysis of Hoax Spread in Social Media**," *J. Humanit. Soc. Sci.*, vol. 23, no. 6, pp. 50–60, 2018, doi: 10.9790/0837-2306065060.
7. A. B. Prasetijo, R. R. Isnanto, D. Eridani, Y. A. A. Soetrisno, M. Arfan, and A. Sofwan, "**Hoax detection system on Indonesian news sites based on text classification using SVM and SGD**," *Proc. - 2017 4th Int. Conf. Inf. Technol. Comput. Electr. Eng. ICITACEE 2017*, vol. 2018-Janua, pp. 45–49, 2017, doi: 10.1109/ICITACEE.2017.8257673.
8. B. Zaman, A. Justitia, K. N. Sani, and E. Purwanti, "**An Indonesian Hoax News Detection System Using Reader Feedback and Naïve Bayes Algorithm**," *Cybern. Inf. Technol.*, vol. 20, no. 1, pp. 82–94, 2020, doi: 10.2478/cait-2020-0006.
9. N. W. S. Saraswati, "**Naïve Bayes Classifier Dan Support Vector Machines Untuk Sentiment Analysis**," *SESINDO 2013*, vol. 2013, 2013.
10. S. A. Salloum, M. Al-Emran, A. A. Monem, and K. Shaalan, "**A survey of text mining in social media: facebook and twitter perspectives**," *Adv. Sci. Technol. Eng. Syst. J*, vol. 2, no. 1, pp. 127–133, 2017.
11. S. Sheela and T. Bharathi, "**Analyzing Different Approaches of Text Mining Techniques and Applications**," *Int. J. Comput. Sci. Trends Technol.*, vol. 6, no. 4, pp. 23–29, 2018.
12. S. J. Russell and P. Norvig, "**Artificial Intelligence (A Modern Approach).**" Prentice Hall, 2010.
13. P. K. L. Utama, "**Identifikasi Hoax pada Media Sosial dengan Pendekatan Machine Learning**," *Widya Duta J. Ilm. Ilmu Agama dan Ilmu Sos. Budaya*, vol. 13, no. 1, pp. 69–76, 2018.
14. F. Rahutomo, I. Yanuar, R. Pratiwi, and D. M. Ramadhani, "**Eksperimen Naïve Bayes Pada Deteksi Berita Hoax Berbahasa Indonesia Naïve Bayes's Experiment On Hoax News Detection In Indonesian Language**," *vol*, vol. 23, pp. 1–15, 2019.
15. A. . G. Tammam, "**Deteksi Hoaks Pada Media Sosial Berbasis Text Mining Classification System**," no. 3. Thesis, Fakultas Teknik, Nusantara PGRI, Kediri, pp. 1–13, 2018, doi: 10.1093/imamci/dnt037.
16. H. A. Santoso, E. H. Rachmawanto, A. Nugraha, A. A. Nugroho, D. R. I. M. Setiadi, and R. S. Basuki, "**Hoax classification and sentiment analysis of Indonesian news using Naive Bayes optimization**," *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 18, no. 2, pp. 799–806, 2020 doi: 10.12928/TELKOMNIKA.V18I2.14744.
17. S. Yuliani, S. Sahib, M. Faizal, B. Abdollah, and F. Z. Ruskanda, "**Hoax News Classification using Machine Learning Algorithms**," *Int. J. Eng. Adv. Technol.*, vol. 9, no. 2, pp. 3938–3944, 2019 doi: 10.35940/ijeat.b3753.129219.
18. P. WiraBuana, S. Jannet D.R.M., and I. Ketut Gede Darma Putra, "**Combination of K-Nearest Neighbor and K-Means based on Term Re-weighting for Classify Indonesian News**," *Int. J. Comput. Appl.*, vol. 50, no. 11, pp. 37–42, 2012, doi: 10.5120/7817-1105.
19. S. Yuliani, S. Sahib, M. Faizal Abdollah, M. Nasser Al-Mhiqani, and A. Rialdy Atmadja, "**Review Study of Hoax Email Characteristic**," *Int. J. Eng. Technol.*, vol. 7, no. 3.2, p. 778, 2018, doi: 10.14419/ijet.v7i3.2.18754.
20. I. Y. R. Pratiwi, R. A. Asmara, and F. Rahutomo, "**Study of hoax news detection using naïve bayes classifier in Indonesian language**," *Proc. 11th Int. Conf. Inf. Commun. Technol. Syst. ICTS 2017*, vol. 2018-Janua, no. October, pp. 73–78, 2018, doi: 10.1109/ICTS.2017.8265649.
21. J. Andrean and S. Suharjito, "**Detection of Hoax Spread in The Whatsapp Group with Lexicon Based and Naive Bayes Classification,**" *Int. J. Eng. Adv. Technol.*, vol. 9, no. 4, pp. 506–511, 2020 doi: 10.35940/ijeat.c6587.049420.
22. P. Chapman *et al.*, "**CRISP-DM 1.0: Step-by-step data mining guide,**" *SPSS inc*, vol. 9, p. 13, 2000.
23. T. Pranckevičius and V. Marcinkevičius, "**Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for**

text reviews classification," *Balt. J. Mod. Comput.*, vol. 5, no. 2, p. 221, 2017.
24. K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "**Text classification algorithms: A survey**," *Information*, vol. 10, no. 4, p. 150, 2019.
25. I. M. B. Adnyana, "**Prediksi Lama Studi Mahasiswa Dengan Metode Random Forest (Studi Kasus: STIKOM Bali)**," *CSRID (Computer Sci. Res. Its Dev. Journal)*, vol. 8, no. 3, pp. 201–208, 2016.
26. K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "**Fake News Detection on Social Media: A Data Mining Perspective**," vol. 19, no. 1, pp. 22–36, 2017, [Online]. Available: http://arxiv.org/abs/1708.01967.
27. J. C. Cuizon, J. Lopez, and D. R. Jones, "**Text mining customer reviews for aspect-based restaurant rating**," *Int. J. Comput. Sci. Inf. Technol. Vol*, vol. 10, 2018.
28. Tuan, Nguyen Anh. "**Detecting Botnet based on Network Traffic**." *International Journal* 9.3 (2020).
29. Alroobaea, Roobaea. "**An Empirical combination of Machine Learning models to Enhance author profiling performance**." *International Journal* 9.2 (2020).