JOIV : Int. J. Inform. Visualization, 7(3) - September 2023 726-732

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Indonesian Fake News Detection Using Various Machine Learning Technique

Liliek Triyono [a,b,*], Rahmat Gernowo [a], Prayitno [b], Mosiur Rahaman [c], Tri Raharjo Yudantoro [a,b]

[a] *Doctoral Program of Information System School of Postgraduate Studies, Diponegoro University, Semarang, Indonesia*
[b] *Department of Electrical Engineering, Politeknik Negeri Semarang, Semarang, Indonesia*
[c] *Department of Computer Science and Information Engineering, Asia University, Taichung City, Taiwan*
*Corresponding author: *liliektriyono@students.undip.ac.id*

***Abstract*— It has become a necessity for people to communicate with each other to complete their needs. The exchange of information conveyed in communication often cannot be directly assessed, especially online news. They just get news and are unable to filter out inappropriate stuff. The media website conveys a great deal of information. Popular news websites are one source for keeping up with the newest news. It requires a significant amount of work to deliver news on prominent websites and to choose content that is not incorrect. To crawl the web and analyse enormous data, massive computer power is required, and solutions to lower the process's space and temporal complexity must be created.Data mining is seen to be a solution to the aforementioned difficulties since it extracts particular information based on defined attributes. This research investigated a model to determine the content of false news information in Indonesian popular news. Firstly, preprocessing process from dataset that collected from keaggle. Secondly, we try use classification methods to determined which the optimal method to classify fake news. Thirdly, we use another public dataset for testing method. Furthermore, five machine learning classifiers are compared: Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree Classifier (DTC), Gradient Boosting Classifier (GBC), and Random Forest (RF). These classifications are utilized independently before being compared based on receiver operating characteristic curves and accuracy. The experimental result shows that DTC has the lowest accuracy of 75.33% and SVM has the highest accuracy of 83.55%.**

***Keywords*— Data mining; hoax; false news; machine learning.**

Manuscript received 26 Sep. 2022; revised 4 Oct. 2022; accepted 1 Nov. 2022. Date of publication 10 Sep. 2023.

I. INTRODUCTION

People have always made an effort to communicate with one another. Initially, it was basic information that was passed on to each other. Unfortunately, not all of the information presented was accurate. False information was frequently transmitted as a type of gossip to mislead others orally, but as technology advanced, newer techniques of transmitting information evolved [1]–[3]. Unfortunately, not all of the information supplied was accurate [4]. False information was frequently circulated as a type of gossip in order to mislead or hurt the opponent [5]. A lot of fake material was manufactured specifically for propaganda objectives for the adversary or to hurt the opponent [6]. Digital information is absorbed on a regular basis in an apparently wide and linked globe [7], [8]. People now consume information in a variety of ways, thanks to the growing usage of mobile devices and better access to the internet [9]. Modern social media platforms such as Facebook, Instagram, and Twitter serve as a platform for digital information to grow. In social media, where enormous volumes of User Generated Content interact with one another, the danger of encountering disinformation is not insignificant. Both the credibility of information and the source of information are critical in avoiding the risk of eating fake news [10].

John McCarthy created the term "artificial intelligence" in 1955 as one of the most powerful technologies. Machine learning, deep learning, neural networks, predictive analytics and natural language processing were later discoveries. Each field has seen significant advancement due to developing technology [11]. Artificial intelligence is one of the technological innovations that has altered the way business concerns are seen [12]–[15]. In order to address challenges, a growing number of enterprises are adopting advanced analytics and machine learning techniques. Natural language processing (NLP) brings up a wide variety of opportunities for enterprises interested in deciphering human feelings

utilizing current data with this innovation in the age of artificial intelligence [11]. NLP should apply with any type of natural in social communication, including audio, video, and text. Text mining has aided in recognizing many numerous and beneficial patterns and trends in textual datasets including news documents.

Fake news changes a person's behavior [16]. It may be tough to spot fake news, not only because it can be hard to tell the difference between a legitimate depiction of a contentious viewpoint and one with malicious intent [17].

This article's contribution is to show how features like article content that use classifiers like Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree Classifier (DTC), Gradient Boosting Classifier (GBC), and Random Forest (RF) can reduce falsehoods in news received. In specifically, we did a comparison study of numerous machine learning algorithms.

The rest of this article is organized as follows: Section II explains the resources and procedure for employing the five classifiers stated before. Section III contains the results and comments, which show the outcomes of a series of experiments. Section IV summarizes the article's main points.

II. Materials and Method

A. Fake News Detection

Fake news research has gain attention among researchers in recent years. Fake news detection has increasingly piqued the interest of the general public and academics, since the spread of disinformation on the internet has increased, notably in media channels such as social media feeds, news blogs, and online newspapers [18][19]. Fake news research is difficult and complex, from dataset preparation [20] to mitigating strategy [21]–[23]. In this paper, we use Kaggle Indonesia False News (Hoax) and Indonesian Hoax News Detection dataset. Fake news is not new phenomenon in journalism or computer science. In general, fake news checking process has been studied in two ways: (1) fake news detection by writing style journalism and (2) extracting fact in news article and factual data and work towards an automatic fact-checking to tackle misinformation using deep learning algorithm [17][24]. Given a piece of text from a news document, the method to predict it using a Convolutional Neural Network, reaching 92% f1 precision and recall [25]. The accuracy value for Google advance search utilizing the Naive Bayes Classifier for hoax classification is 78.6% [26]. There have been attempts with fake news categorization in Indonesian by comparing tree methods such as the C4.5 algorithm, naïve bayes, and SVM [27]. ClaimBuster, on the other hand, was constructed using natural language phrases to establish a classification model that distinguishes a sentence score from 0 to 1 indicating its suitability for fact checking. The system creates its own private dataset of around 8,000 argument lines annotated by students, academics, and journalists [28]. Another study employed a supervised text classification task that made use of the Keras library and two standard classifiers: Support Vector Machine (SVM) and Naive Bayes. In supervised text classification tasks, DNN models outperform classifiers, with 1D-CNN coming out on top [29].

B. Data

We used the Kaggle Indonesia False News (Hoax) Dataset, which is collected from several news article websites and social media. This False News Dataset contains thousands of articles ranging from political, climate, health and economic events. The dataset provides us with metadata such as news dates, news title, news content and a label that marks the article that indicate fake news. There are two datasets provided, for instance training set contains 4.231 news articles and testing set consists of 1.058 news articles. Table 1 present the dataset statistic that used in the Indonesia False News (Hoax) Dataset where extracted from *https://www.kaggle.com/datasets/muhammadghazimuharam/indonesiafalsenews*.

TABLE I
THE KAGGLE : INDONESIA FALSE NEWS (HOAX) DATASET

Dataset Statistic	
Indonesia False News(Hoax) Dataset	4.231
Valid	3.465
False	766

We also collected articles from Mendeley Data called Indonesian Hoax News Detection Dataset to build dynamic testing classifier. This set consists of 600 data were obtained from *https://data.mendeley.com/datasets/p3hfgr5j3m/1* in Indonesian, containing 372 valid news and 228 false news. From this dataset it will be count of accuracy in each news.

TABLE II
THE MENDELEY DATA : INDONESIAN HOAX NEWS DETECTION DATASET

Dataset Statistic	
Indonesian Hoax News Detection Dataset	600
Valid	372
False	228

C. Proposed Method

This section describes the suggested strategy, as well as the preprocessing stages and classification approaches (Fig. 1).

D. Preprocessing

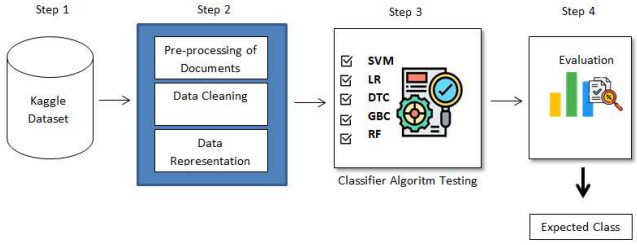

Fig. 1 Summary of the proposed approach

In Step 1 using the Kaggle Dataset, the raw data will be in the form. Data cleaning will be carried out in Step 2 until data is formed that is ready to be classified with several classification algorithms. Step 3 with the same data will run each algorithm to classify the data. Setp 4 will evaluate the results of the classification in the previous step by using a comparison of each resulting accuracy. This process will produce a classification algorithm that has high accuracy. With a high level of accuracy, it is expected to be able to distinguish hoax news and true news.

The pre-processing stage is critical in preparing the dataset for classification. The following stages are used to process documents. The HTML and XML elements are first removed from the publications. Then, terms that are deemed noise, such as "di", "yang" and "hanya" are deleted. Afterword, punctuation, and special symbols such as ".", "%" and "@" are omitted. The papers are then translated to lower case letters, with terms such as NEWS or NEWS changed to news. After that, the articles are tokenized, yielding arrays of meaningful words from which the frequency of each phrase may be calculated [30], which eliminates common morphological and inflexional ends from English words, leaving just the stem. For example, the terms "berhitung" and "hitungan" become "hitung." Some documents will become empty as a result of the preceding procedure, thus they are deleted from the array. The preprocessing processes are listed in chronological sequence in Fig. 2.

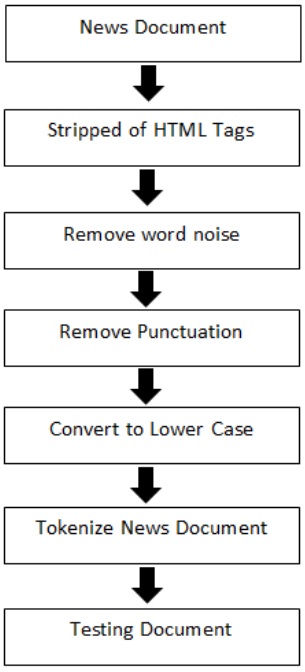

Fig. 2 The Preprocessing Step

D. Classifier

1) Support Vector Machine (SVM)

SVMs function by representing the dataset in a manner where the number of features corresponds to the dimensions of the graph. The dataset is partitioned by hyperplanes, which are composed of three parallel lines in the context of two-dimensional space. These hyperplanes separate the data into two classes. Among these hyperplanes, two are referred to as support vectors since they are next to the nearest data point(s) from different classes. The point of intersection between the two support vectors is observed in the final line. The optimal fit is determined by the hyperplane that maximises the distance between the two support vectors. If one does not exist, a kernel function must be used to scale the dataset to the next dimension. Commonly used kernel functions in machine learning include Radial Basis Function (RBF), Polynomial, Laplace RBF, Gaussian, sigmoid, and several more. It can be defined as follows Eq. (1) and Eq. (2) [31].

$$k(x_i, x_j) = (x_i x_j + 1)^d \quad (1)$$

$$k(x, y) = exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \quad (2)$$

2) Logistic Regression (LR)

Logistic regression is a classification method that is used to categorize linear logarithms. It can be defined as follows Eq. (3) and Eq. (4) [32].

$$P(x) = \frac{\exp(w.x+b)}{1+\exp(w.x+b)} \quad (3)$$

$$P(x) = \frac{1}{1+\exp(wx+b)} \quad (4)$$

Where $x \in R^n$ is the input feature, $Y \in \{0,1\}$ represents the label vector, variable "w" is used to denote the weight, "b" is used to indicate the offset value, and "w.x" denotes the dot product of the matrices. Subsequently, the logistic regression contrasts the two probabilities and allocates 'x' to the group with the greater probability.

3) Decision Tree Classifier (DTC)

A DTC generates a flow chart graphic that looks like a tree with branches for each option or variable. The root node is the tree's uppermost node; the tree's construction begins at the root node and progresses top-down. Entropy, as indicated in Eq. (5), is used to assess the unpredictability of a choice inside the tree. 1 represents an extremely uncertain 50% possibility, and 0 represents a guarantee, either a 0% or 100% chance, where p_i represents the probability of a class I and c represents the total number of classes. Eq. (6) uses information gain to assess the reduction in uncertainty when more nodes are utilized before the provided node. The collection of nodes with the highest information gain is utilized. Because the preceding method is repeated on each branch of the tree, decision trees are prone to overfitting [33].

$$E(s) = \sum_{i=1}^{c} -p_i \log_2 p_i \quad (5)$$

$$IG(X,Y) = \text{E(Y)-E(Y|X)} \quad (6)$$

4) Gradient Boosting Classifier (GBC)

Gradient tree boosting has been widely employed in industry and data mining contests, along with other tree ensemble learning techniques. It is unaffected by input scaling and is capable of learning higher level relationships between features. Gradient tree boosting, unlike earlier tree ensemble techniques, is learnt additively. It creates a new tree at each time step t in order to minimize the residual of the current model. It can be defined as follows Eq. (7) :

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{t-1} + f_t(x_i)) + \Omega(f_t) \quad (7)$$

The loss function, denoted as i, is responsible for assessing the disparity between the label of the *i*-th instance, y_i, and the prediction derived from the previous step combined with the current tree output. Additionally, the regularisation term, $\Omega(f_t)$, serves to penalise the intricacy of the newly introduced tree. Because of its excellent efficiency and success in different data mining contests, XGBoost is one of the most well-known

gradient boosting implementations. By default, it also handles missing values. From the training data, the approach precisely learns the proper default direction in each tree node. If a feature's value is absent, the instance will be categorized in the feature's default direction [24].

5) Random Forest (RF)

Breiman [34] proposed RF, a decision tree (DT) approach that works by building numerous DT. RF may be seen as a DT forest, with each tree voting on the most popular input vector class. RF requires less parameters for its definition as compared to alternative methodologies like support vector machines (SVM) and artificial neural networks (ANN). Eq. (8) denotes a collection of individual tree-arranged classifiers.

$$\{RF(y, \alpha n), n = 1, 2, \dots i, \dots\} \quad (8)$$

where RF is the classifier, n denotes the number of identical independently spread random values, and each DT chooses for the most well-known class at input variable y. The role of in DT building determines the nature and proportions of. The creation of each of the DT that make up the forest is important to achieving the RF. RF diversity can be generated by randomly changing some decision tree parameters or by sampling from the feature set or the data collection. Because of its quick execution speed, RF is an appealing classifier, and it frequently outperforms a single decision tree in classification effectiveness. In general, the more trees there are in the forest, the more prominently they appear [34].

III. Results and Discussion

For evaluating the performance of fake news detection, we using testing set provided by the Kaggle False News Dataset which contains of 4.231 news article with features such as date, title and news content. Furthermore, we extracted title and content from news article testing set as an input vector for the classifier. Submission to the competition were evaluated using accuracy score, as describe in official evaluation metrics. Our submission achieved an accuracy score of 83.55%.

This study included metrics such as accuracy, sensitivity, and specificity. The classification algorithm's accuracy is how close it is to producing the expected results. To evaluate the accuracy of different classification systems, use Eq. (9), where Y pre is the classifier's prediction for the test-dataset documents and Y test is the right prediction for the test-dataset documents [33]. Sensitivity is a percentage that shows the categorization algorithm's real positive rate. A genuine positive is a correctly recognized member of a positively labeled class, whereas a false positive is a correctly identified member of a positively labeled class. NP denotes false negatives, whereas FP is the total number of false positives. To calculate sensitivity, use Eq. (10), where TP is the total number of true positives and FN is the total number of false negatives [33].

Specificity is a percentage that represents the true negative rate of the classification system. A genuine negative is a correctly detected member of a negative class, whereas a fake negative is an incorrectly identified member of the same class. To calculate specificity, utilize Eq. (11) where TN represents total true negatives and FP represents total false positives [33].

$$Accuracy = \frac{\sum(Y_{pre}=Y_{test})}{Y_{test}} * 100 \quad (9)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (10)$$

$$Specificity = \frac{TN}{TN+FP} \quad (11)$$

The use of the confusion matrix is employed in this article to facilitate the comparison between the intended outcomes and the anticipated outcomes. The columns of the matrix represent samples obtained from the projected class, whereas the rows correspond to samples obtained from the actual class. The article use receiver operating characteristics (ROC) curves to assess and compare the genuine positive rate and false positive rate at different thresholds. The evaluation of a classification technique's performance often relies on two crucial measures: the area under the receiver operating characteristic (ROC) curve (AUC) and the accuracy. These metrics play a significant role in facilitating comparisons between different classification techniques.

1) Support Vector Machine (SVM)

The Support Vector Machine was implemented using the sklearn python library. The Support Vector Machine (SVM) algorithm demonstrated a notable accuracy rate of 83.55%, positioning it as the most optimal choice. The area under the curve (AUC) for the support vector machine (SVM) technique is 71.61% as seen on the receiver operating characteristic (ROC) curve in Figure 3. The Support Vector Machine (SVM) algorithm successfully classifies the dataset, achieving a sensitivity of 83.72% and a specificity of 78.12%, as seen in Figure 4.

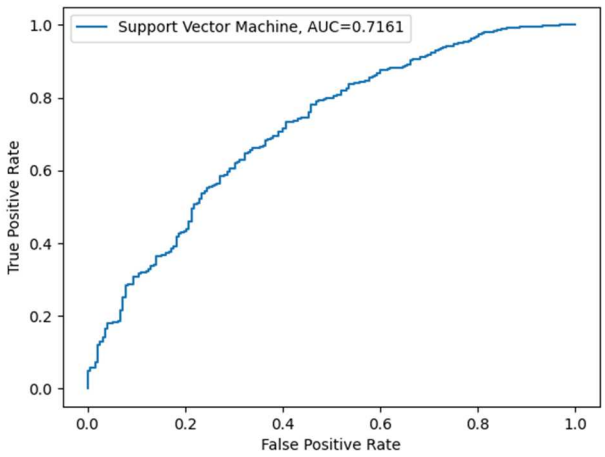

Fig. 3 ROC curve for SVM

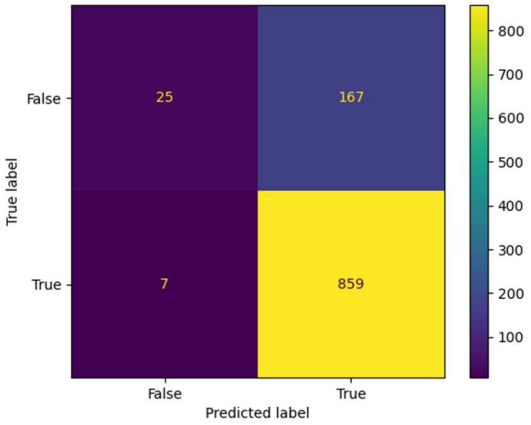

Fig. 4 Confusion matrix SVM

2) Logistic Regression Technique (LR)

The sklearn python package was used to implement logistic regression. With an accuracy of 82.61%, LR was the second-best answer. The logistic regression approach offers an average AUC of 75.26%, according to the ROC curve in Fig. 5. In Figure 6, the LR classifies the dataset with a sensitivity of 82.73% and a specificity of 75.0%.

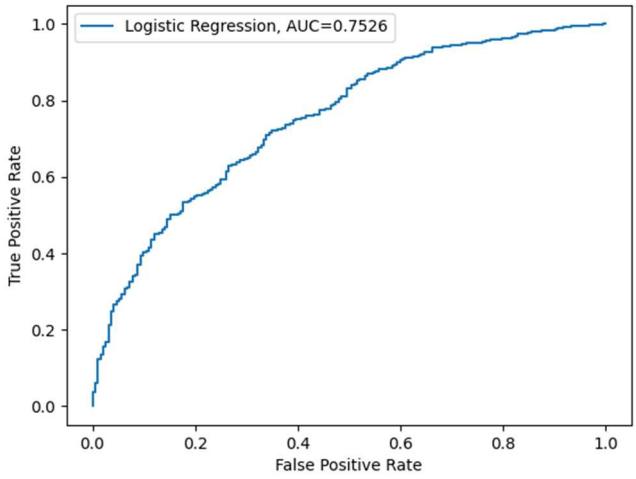

Fig. 5 ROC curve for LR

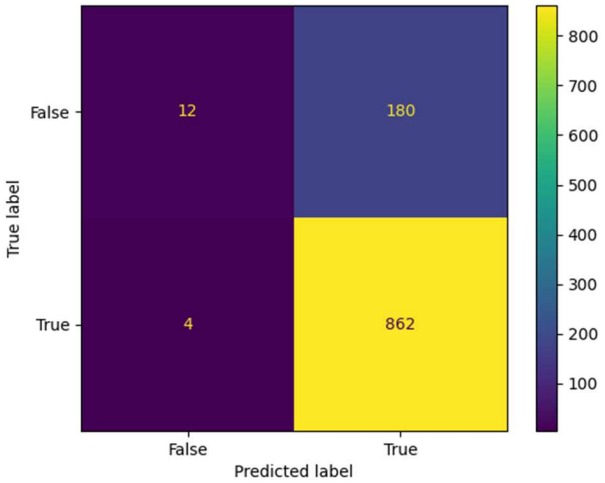

Fig. 6 Confusion matrix LR

3) Decision Tree Classifier Technique (DTC)

The sklearn python package was used to construct Decision Tree. DTC scored an accuracy of 75.33%, ranking it fifth among all solutions. The decision tree approach has an average AUC of 54.89% according to the ROC curve in Fig. 7. The DTC classifies the dataset with a sensitivity of 83.57% and a specificity of 28.03%, as shown in Fig. 8.

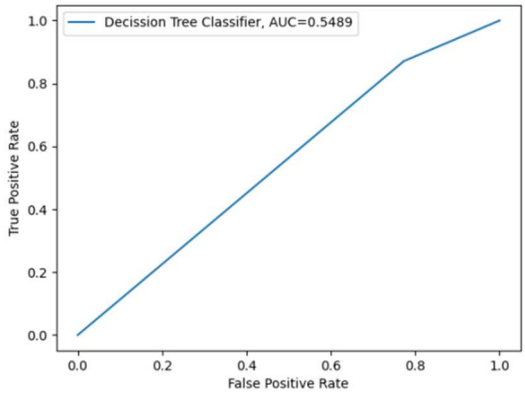

Fig. 7 ROC curve for DTC

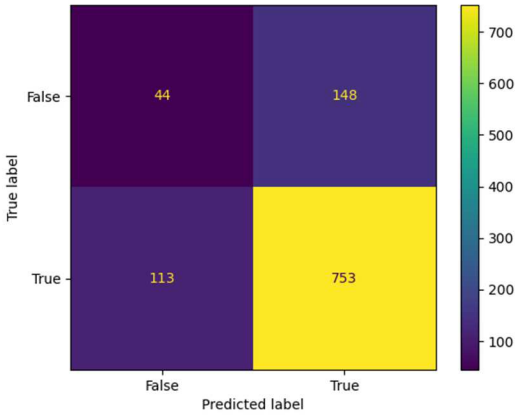

Fig. 8 Confusion matrix DTC

4) Gradient Boosting Classifier Technique (GBC)

The sklearn python package was used to construct Gradient Boosting. GBC was the third-best option with an accuracy of 82.61%. The Gradient Boosting approach has an average AUC of 69.44% according to the ROC curve in Fig. 9. The GBC classifies the dataset with a sensitivity of 82.85% and a specificity of 70.0%, as shown in Fig. 10.

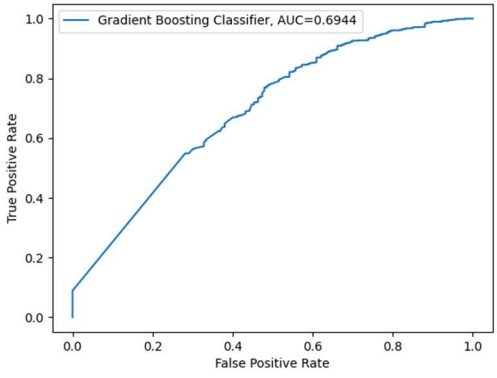

Fig. 9 ROC curve for GBC

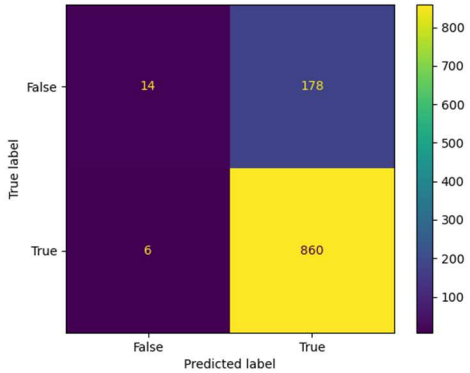

Fig. 10 Confusion matrix GBC

5) Random Forest Technique(RF)

The implementation of Random Forest was carried out using the sklearn python package. The Random Forest (RF) algorithm demonstrated a classification accuracy of 82.61%, positioning it as the fourth most effective approach. The receiver operating characteristic (ROC) curve seen in Figure 11 exhibits an area under the curve (AUC) of 69.44% when utilising the Random Forest approach. Figure 12 depicts the performance of the Random Forest (RF) algorithm in categorising the dataset, achieving a sensitivity rate of 82.85% and a specificity rate of 70.0%.

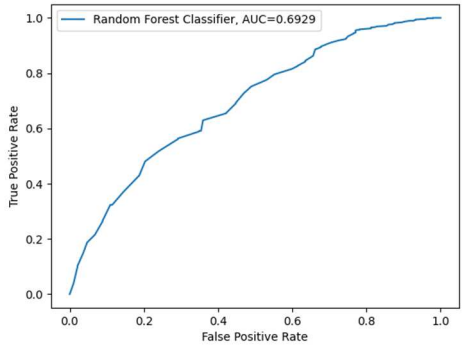

Fig. 11 ROC curve for RF

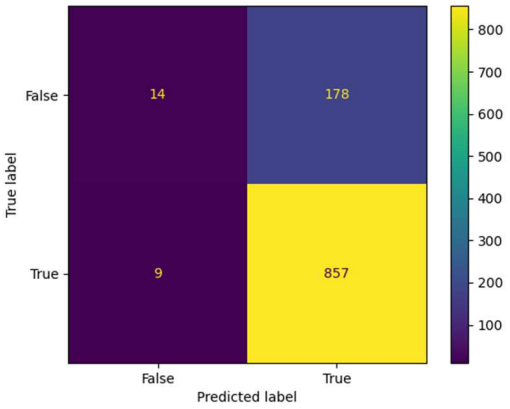

Fig. 12 Confusion matrix RF

After doing a comparison of each classification technique SVM, LR, DTC, GBC and RF, the following comparison graph on the ROC curve shown in Fig. 13. In the graph there is also an AUC value for each technique used. Logistic Regression has the largest AUC value of 0.7526 among the existing techniques. From the ROC curve, it can also be seen that the Decission Tree has the smallest value, namely AUC 0.5489.

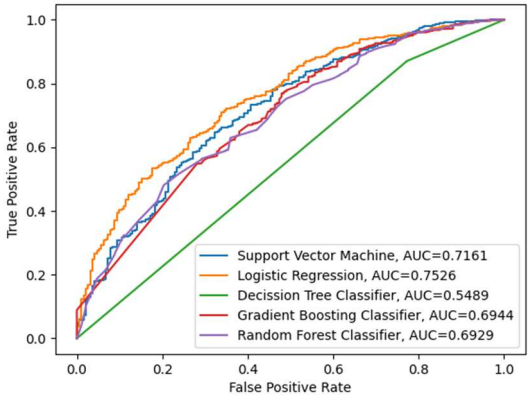

Fig. 13 Confusion matrix Classifiers AUC values

E. Implementation and testing of the proposed

Based on the results of the accuracy test, it is possible to infer that SVM is the algorithm with the best accuracy, with a value of 83.55%. Furthermore, it will be utilized for Dynamic Testing on the Mendeley Dataset.

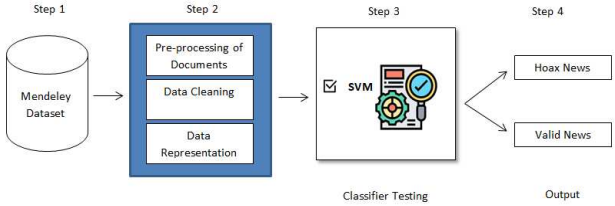

Fig. 14 Dynamic Testing with the choosen classifier

By using the Mendeley Dataset, it will be tested whether the resulting model can define according to the labels owned by each document. The calculation of the accuracy percentage of 600 documents in the dataset reaches valid condition.

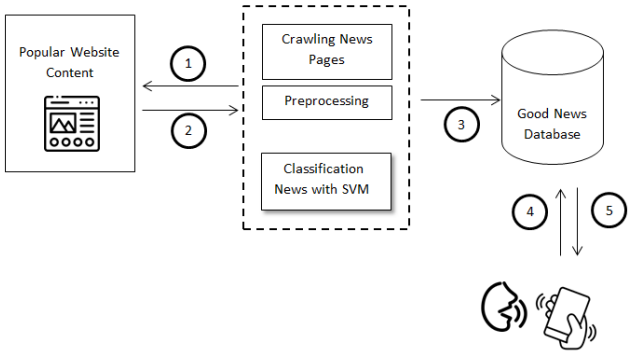

Fig. 15 Propose Architecture of Valid News Reader

In Step 1: The server will periodically fetch the latest news from the destination website. Step 2: The data will be processed in the pre-processing step, which will then be classified using the SVM classification method. Step 3: If the news is considered good then the news document will be stored on Firebase. Step 4: A person starts the app by vibrating the phone several times to bring up Voice to Text. The voice will be converted into text that will be matched with a list of keywords. These keywords are to determine which news will be played. Step 5: If the keywords match then the app will request to Firebase the appropriate data. And the app will play

the Text To Voice of the successfully fetched news from Firebase.

IV. Conclusion

Once the daily effect was acknowledged, publications were examined to identify the research gap, and several acceptable machine learning classifiers were investigated. In this study, we conducted an investigation into machine learning classifiers that exhibit enhanced accuracy while simultaneously reducing time and space complexity. The focus of our research was on their applicability to web-based big data applications. We specifically examined five classifiers: SVM, Logistic Regression (LR), Decision Tree Classifier (DTC), Gradient Boosting Classifier (GBC), and Random Forest (RF). SVM had the highest accuracy at 83.55%, while DTC had the lowest at 75.33%. It is vital to highlight that in recent years, false news in conjunction with categorization has been a popular study issue. However, to determine fake news content is very difficult, so the proposed approach works is how to classify news content using text classification methods. The results of the top classifiers demonstrated good accuracy. Deep learning and word embedding might be used in future research to extract information from news articles to enhance algorithmic decision-making.

References

[1] D. Crowe, "Flaws in Coronavirus Pandemic Theory," 2020, [Online]. Available: https://api.semanticscholar.org/CorpusID:214775362.

[2] M. N. Alenezi, H. K. Alabdulrazzaq, A. A. Alshaher, and M. M. Alkharang, "Evolution of Malware Threats and Techniques: a Review," *Int. J. Commun. Networks Inf. Secur.*, vol. 12, 2020.

[3] W. Yue, C. Li, G. Mao, N. Cheng, and D. Zhou, "Evolution of road traffic congestion control: A survey from perspective of sensing, communication, and computation," *China Commun.*, vol. 18, pp. 151–177, 2021.

[4] D. N. Rapp, S. R. Hinze, K. Kohlhepp, and R. A. Ryskin, "Reducing reliance on inaccurate information.," *Mem. Cognit.*, vol. 42, no. 1, pp. 11–26, Jan. 2014, doi: 10.3758/s13421-013-0339-0.

[5] O. Balashevych, O. Orliuk, and A. Proskurnia, "THE Development and Approbation of The Questionnaire for Detecting The Tendency to Gossip (TTGQ – Tendency to Gossip Questionnaire): A Pilot Study," *Psychol. J.*, 2023.

[6] B. Probierz, P. Stefa, J. Kozak, B. Probierz, P. Stefa, and J. Kozak, "Rapid detection of fake news based on machine learning methods," vol. 00, 2021, doi: 10.1016/j.procs.2021.09.060.

[7] M. Surve, P. Joshi, S. Jamadar, and M. M. N. Vharkate, "Automatic Attendance System using Face Recognition Technique," *Int. J. Recent Technol. Eng.*, vol. 9, no. 1, pp. 2134–2138, 2020, doi: 10.35940/ijrte.a2644.059120.

[8] P. Jha, "A Survey on various Haze and Underwater Digital Image Enhancement Techniques," 2018.

[9] J. Dunaway, K. Searles, M. Sui, and N. Paul, "News Attention in a Mobile Era," *J. Comput. Commun.*, vol. 23, no. 2, pp. 107–124, 2018, doi: 10.1093/jcmc/zmy004.

[10] M. Viviani and G. Pasi, "Credibility in social media: opinions, news, and health information—a survey," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 7, no. 5, 2017, doi: 10.1002/widm.1209.

[11] T. Chauhan and H. Palivela, "Optimization and improvement of fake news detection using deep learning approaches for societal benefit," *Int. J. Inf. Manag. Data Insights*, vol. 1, no. 2, p. 100051, 2021, doi: 10.1016/j.jjimei.2021.100051.

[12] J. L. Ruiz-Real, J. Uribe-Toril, J. A. Torres, and J. D. E. Pablo, "Artificial intelligence in business and economics research: Trends and future," *J. Bus. Econ. Manag.*, vol. 22, no. 1, pp. 98–117, 2021, doi: 10.3846/jbem.2020.13641.

[13] A. Geisel, "The Current And Future Impact Of Artificial Intelligence On Business," *Int. J. Sci. \& Technol. Res.*, vol. 7, pp. 116–122, 2018.

[14] F. Bezzazi, "The impact of artificial intelligence on business: benefits and ethical challenges on customer level," *J. Mark. Consum. Res.*, 2021.

[15] C. Chan and D. Petrikat, "Impact of Artificial Intelligence on Business and Society," *J. Bus. Manag. Stud.*, 2022.

[16] Z. Bastick, "Would you notice if fake news changed your behavior? An experiment on the unconscious effects of disinformation," *Comput. Human Behav.*, vol. 116, no. November 2020, p. 106633, 2021, doi: 10.1016/j.chb.2020.106633.

[17] Y. Wang, M. McKee, A. Torbica, and D. Stuckler, "Systematic Literature Review on the Spread of Health-related Misinformation on Social Media," *Soc. Sci. Med.*, vol. 240, no. August, p. 112552, 2019, doi: 10.1016/j.socscimed.2019.112552.

[18] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, "Automatic detection of fake news," *COLING 2018 - 27th Int. Conf. Comput. Linguist. Proc.*, pp. 3391–3401, 2018.

[19] M. Aldwairi and A. Alwahedi, "Detecting Fake News in Social Media Networks," *Procedia Comput. Sci.*, vol. 141, pp. 215–222, 2018, doi: https://doi.org/10.1016/j.procs.2018.10.171.

[20] T. Murayama, S. Hisada, M. Uehara, S. Wakamiya, and E. Aramaki, "Annotation-Scheme Reconstruction for 'Fake News' and Japanese Fake News Dataset." 2022.

[21] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," vol. 1151, no. March, pp. 1146–1151, 2018.

[22] K. Sharma, F. Qian, H. Jiang, N. Ruchansky, M. Zhang, and Y. Liu, "Combating Fake News: A Survey on Identification and Mitigation Techniques." 2019.

[23] W. Shahid *et al.*, "Detecting and Mitigating the Dissemination of Fake News: Challenges and Future Research Opportunities," *IEEE Trans. Comput. Soc. Syst.*, 2022.

[24] Z. Shae, C. Shyu, and M. W. Kearney, "Automatic Fake News detection in News Article with Opinion Classifier using XGBoost Algorithm.

[25] Y. Yang, L. Zheng, J. Zhang, Q. Cui, Z. Li, and P. S. Yu, "TI-CNN: Convolutional Neural Networks for Fake News Detection," 2018, [Online]. Available: http://arxiv.org/abs/1806.00749.

[26] 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. February 2018, pp. 73–78, 2018, doi: 10.1109/ICTS.2017.8265649.

[27] F. Rahutomo, I. Y. R. Pratiwi, and D. M. Ramadhani, "Eksperimen Naïve Bayes Pada Deteksi Berita Hoax Berbahasa Indonesia," *J. Penelit. Komun. Dan Opini Publik*, vol. 23, no. 1, 2019, doi: 10.33299/jpkop.23.1.1805.

[28] N. Hassan *et al.*, "Claim buster: The firstever endtoend factchecking system," *Proc. VLDB Endow.*, vol. 10, no. 12, pp. 1945–1948, 2017, doi: 10.14778/3137765.3137815.

[29] B. P. Nayoga, R. Adipradana, R. Suryadi, and D. Suhartono, "Hoax Analyzer for Indonesian News Using Deep Learning Models," *Procedia Comput. Sci.*, vol. 179, no. 2020, pp. 704–712, 2021, doi: 10.1016/j.procs.2021.01.059.

[30] M. S. S. Nur Hayatin, Suraya Alias, Lai Po Hung, "Sentiment Analysis Based On Probabilistic Classifier Techniques In Various Indonesian Review Data," *Jordanian J. Comput. Inf. Technol.*, vol. 8, no. 3, pp. 171–175, 2022.

[31] T. H. J. Hidayat, Y. Ruldeviyani, A. R. Aditama, G. R. Madya, A. W. Nugraha, and M. W. Adisaputra, "Sentiment analysis of twitter data related to Rinca Island development using Doc2Vec and SVM and logistic regression as classifier," *Procedia Comput. Sci.*, vol. 197, no. 2021, pp. 660–667, 2021, doi: 10.1016/j.procs.2021.12.187.

[32] A. G, B. Ganesh, A. Ganesh, C. Srinivas, Dhanraj, and K. Mensinkal, "Logistic regression technique for prediction of cardiovascular disease," *Glob. Transitions Proc.*, vol. 3, no. 1, pp. 127–130, 2022, doi: 10.1016/j.gltp.2022.04.008.

[33] A. Mulahuwaish, K. Gyorick, K. Z. Ghafoor, H. S. Maghdid, and D. B. Rawat, "Efficient classification model of web news documents using machine learning algorithms for accurate information," *Comput. Secur.*, vol. 98, 2020, doi: 10.1016/j.cose.2020.102006.

[34] M. M. Hasan, G. J. Young, M. R. Patel, A. S. Modestino, L. D. Sanchez, and M. Noor-E-Alam, "A machine learning framework to predict the risk of opioid use disorder," *Mach. Learn. with Appl.*, vol. 6, no. August, p. 100144, 2021, doi: 10.1016/j.mlwa.2021.100144.