Assessing the Impact of Feature Extraction Techniques and Machine Learning Algorithms on Fake News Detection Performance

Daniel Aditya Tumansery[a], Ida Bagus Gede Purwa Manik Adiputra[a,*], Simeon Yuda Prasetyo[a]

[a] *Computer Science Study Program, School of Computer Science, Bina Nusantara University, Jakarta, 11480, Indonesia*

Corresponding author: daniel.tumansery@binus.ac.id

***Abstract*— The prevalence of fake news on online platforms raises concerns about its potential to manipulate public opinion and disrupt societal harmony. To combat this issue, researchers and practitioners have turned to machine learning techniques for automated fake news detection. This study aims to identify the optimal combination of feature extraction techniques and machine learning algorithms for accurately detecting fake news. Using an open-source dataset of 20,800 news articles from Kaggle, the study employs three feature extraction techniques—TF-IDF vectorization, GloVe word embedding, and DistilBERT word embedding—to numerically represent the text data. Multiple machine learning algorithms, including Support Vector Machines (SVM), Random Forests, Decision Trees, and Logistic Regression, are trained and evaluated to compare their performance. The results indicate that combining TF-IDF vectorization with machine learning algorithms achieves the highest accuracy scores, surpassing the word embedding techniques. Specifically, SVM combined with TF-IDF achieves an accuracy score of over 92% and an AUC score exceeding 0.94. Among the word embedding techniques, the combination of TFDistilBERT with Logistic Regression yields the most optimal results, with 91% accuracy and a 0.97 AUC score. Future research should explore alternative algorithms like Deep Learning to further enhance the accuracy of word embedding techniques. Additionally, fine-tuning methods, hyperparameter tuning, and experimentation with different datasets are recommended to improve accuracy rates. Overall, this study provides insights into the optimal combinations of techniques and algorithms for accurate fake news detection while suggesting avenues for future research.**

***Keywords*— Fake news detection; Feature extraction; Machine learning; Natural language processing;.**

Manuscript received 15 Oct. 2020; revised 29 Jan. 2021; accepted 2 Feb. 2021. Date of publication 17 Feb. 2021.

I. Introduction

The spread of fake news cannot be underestimated. This reality shows that the circulation of fake news touches almost every aspect of life and is the most worrying in recent times. The problem of fake news is still ongoing, considering the reality that digital literacy in Indonesia is still low. Digital literacy in Indonesia [1] is ranked 56 out of 63 countries. This fact is in line with the results of the "National Hoax Outbreak" survey conducted by MASTEL [2] among 914 respondents in 2019 which revealed that 56.40 % of respondents did not always immediately know whether the information they received was categorized as fake news or not. The survey results also revealed that some people were reluctant to check the truth of the information they received [2].

Social media [3] is a place for the rapid spread of fake news. When fake news is deliberately made by individuals, then disseminated through social media, then the fake news will continue to leave digital footprints that are vulnerable to being consumed by the public [4]. Mastel's survey results show that social media and chat applications are the two most used channels in the spread of fake news. The survey results also revealed that as many as 34.60 % of respondents received fake news every day [2].

Looking back 3 years, when Indonesia was still in the status of a coronavirus disease 2019 (Covid-19) pandemic. There is a study [5] which was conducted using a quantitative method with a descriptive approach. In addition, the tool used for research is an online survey. The response results obtained through the survey method prove that fake news is at least received by more than 40% of respondents through social media. Nearly half of the respondents stated that they would think twice about receiving the COVID-19 vaccination after seeing a simulation of giving fake news on social media. This is also supported by the decrease in the number of respondents who are willing to receive the new COVID-19 vaccine compared to the number of respondents who have not received fake news about the COVID-19 vaccination.

Even the Ministry of Communication and Informatics (Kominfo) [6] explained, one of the reasons people believe they will not be exposed to Covid 19 is because they are consumed by misleading news. This statement is reinforced by data submitted by the Ministry of Communication and Informatics (Kominfo) from 23 January 2020 to 12 May 2021 which shows that there were as many as 1,587 fake issues regarding Covid-19 going around. This issue spreads in 3,377 content on various social media in Indonesian society. The data also shows that social media Facebook ranks highest in the spread of fake news about Covid-19 with a total of 2,784 content. Next, Twitter with 520 misleading content. While on Youtube found as many as 49 contents. There are also 24 content on Instagram. From that data, Kominfo has deleted 2,927 content, the remaining 450 content is still at the investigation stage.

The circulation of fake news has surged in line with the start of the national Covid-19 vaccination program since its launch on 13 January 2021[7]. There is a lot of incorrect information about vaccines in the community. Many irresponsible individuals create and disseminate misleading information to add to the commotion because many people are consumed by fake news. The police have criminalized 17 suspects who are considered to have spread fake news about Covid-19, and the other 87 have not gone to court. There have been 104 criminal cases that have been investigated by the police from January 2020 to November 2020.

The Indonesian Anti-Slander Society (Mafindo) [8] has also mapped the spread of fake news regarding Covid-19. The data shows that fake news is created with a variety of goals. As many as 48 % were found abroad, as much as 52 % were found within the country. It was also found that 40 % targeted villages, 18 % targeted markets, industrial areas and terminals, and 17 % targeted hospitals. While the distribution of types of fake news that was built based on the pattern was found in various criteria. In satirical content or there is no intention to harm but has the potential to harm as much as 1 %. There is misleading content that frames issues (Misleading) as much as 42 %. New content that is deliberately made to deceive (Fabricated Content) as much as 19 %. There is also a pattern with incorrect connections or original content disguised with erroneous information (False Connection) as much as 16 %, false content (False Context) there are 19 %.

The consequences of this rampant spread of misinformation are multifaceted and wide-ranging. Notably, it has the potential to severely undermine democratic processes, as public opinion and political discourse can be heavily influenced by false narratives. Instances of fake news have been observed to incite social unrest, exacerbate divisions within communities, and even manipulate electoral outcomes. Furthermore, studies have linked the proliferation of fake news to a decline in public trust in traditional news sources, posing a significant challenge to the foundations of reliable journalism.

Authenticating the veracity of information in the digital age presents a daunting challenge. Traditional methods of verification, such as contacting the originator of news directly, are often impractical or impossible due to the vast volume and rapid dissemination of information. As a result, the development of robust systems that can effectively verify information has become imperative. The task at hand is to design automated approaches capable of discerning between genuine and false news with high accuracy.

To address this complex problem, extensive research has been undertaken to explore innovative techniques for identifying and combating fake news. One notable area of investigation lies in the realm of Natural Language Processing (NLP) techniques, which offer promising avenues for detecting and analyzing textual data. In particular, word embedding methodologies have gained traction as powerful tools for representing words as features in fake news detection. Currently, there are two prevalent types of word embedding techniques employed: static word embedding and dynamic word embedding.

Static word embedding, such as the widely used GloVe model, relies on pre-trained word vectors that capture semantic and syntactic relationships between words based on their co-occurrence statistics. These embeddings provide fixed representations for words, enabling algorithms to discern underlying patterns and associations. On the other hand, dynamic word embedding, exemplified by advanced models like DistilBERT, incorporates contextualized representations that capture the meaning of words based on their surrounding context. This approach allows for a nuanced understanding of word semantics and disambiguation, thereby enhancing performance in tasks involving natural language understanding.

In this study, we propose a supervised learning-based approach that harnesses the power of NLP techniques to detect and combat fake news. Our primary focus is on achieving the highest possible accuracy in fake news detection by comparing and evaluating various feature extraction techniques and machine learning algorithms. The objective is to identify the optimal combination of these techniques that will yield the most accurate results. To this end, our research employs three distinct feature extraction techniques: vectorization, which leverages the TF-IDF method, static word embedding utilizing the GloVe model, and dynamic word embedding using the sophisticated DistilBERT model implemented with TensorFlow. The reason why we chose to utilize our research with three different feature extraction is because each method will offer unique strengths that can be beneficial for detecting fake news. For example TF-IDF is a classic technique for representing text documents as vectors, and is well-suited for identifying key terms and features that distinguish fake news from legitimate news. While GloVe is useful for understanding the relationships between words and identifying patterns that may signal fake news, such as the presence of biased language or misleading semantics. DistilBERT is a state-of-the-art method for natural language processing, and its ability to capture meaning in context can be valuable for detecting subtle nuances in language that may indicate fake news. After extracting the results of each technique, we train several machine learning algorithms, including support vector machines, random forests, decision trees, and logistic regression, to compare the accuracy of each technique and algorithm.

By delving into these feature extraction techniques and rigorously assessing their efficacy, we aim to contribute to the ever-growing body of knowledge in the realm of fake news detection. Furthermore, our study seeks to address the critical need for reliable techniques that can authenticate information in an era where the widespread dissemination of fake news poses formidable challenges to the integrity and trustworthiness of information sources.

II. Literature Review

There have been numerous approaches for detecting bogus news. This section goes over a few of them. Ahmed et al. [9] used rule-based approaches to construct a new n-gram model to automatically identify fraudulent information, with a focus on bogus comments and fake news. They used TF-IDF for feature extraction and six machine learning classification methods: SVM, Linear Support Vector Machine (LSVM), K-nearest neighbors (KNN), Decision Tree (DT), Stochastic Gradient Descent (SGD), and Logistic Regression (LR). Their experiments show highly promising and improved results when compared to standard procedures. LSVM achieves the highest accuracy of 90%.

Reis et al.[10] presented many forms of news story features, such as source and social media posts. They introduce a new set of characteristics and evaluate the predictive performance of KNN, NB, RF, SVM, and XGBoost (XGB) to detect bogus news automatically. With an accuracy of 86%, the best model is XGB. Kumari, S. et al. [11] introduced a BERT-based classification model to predict the domain and classification of fake news. They have a macro F1 score of 83.76%.

In Mansouri et al.'s [12] research, they employed a blend of semi-supervised LDA (Linear Discriminant Analysis) and convolutional neural network to identify fabricated news by utilizing an untagged dataset, which is tagged for the convolutional neural network. The precision outcome of the suggested method is 95.6%, and the recall is 96.7%, which is superior to other techniques for spotting fake news.

Numerous prior investigations [13], [14], [15], [16] have delved into the issue of counterfeit news by utilizing a genuine counterfeit news dataset, namely Fake-News. In one of these studies, Ahmed et al. [13] employed TF-IDF (Term Frequency-Inverse Document Frequency) as a technique for feature extraction, alongside various machine learning models. Through extensive experimentation with the LR (Linear-regression model), an accuracy of 89.00% was attained. Subsequently, they achieved an accuracy of 92% with their LSVM (Linear Support Vector Machine). Liu et al [14] explored the means to identify false tweets by using a corpus of over 8 million tweets amassed from supporters of presidential candidates in the US general election. In their research, they used deep CNNs for counterfeit news detection, incorporating the concept of subjectivity analysis, leading to an accuracy of 92.10%. O'Brien et al [15] implemented deep learning approaches to classify counterfeit news and were able to achieve an accuracy of 93.50% via the black-box method. Ghanem et al [17] utilized various word embeddings, including n-gram features, to identify stances in counterfeit articles and obtained an accuracy of 48.80%. Singh et al [16] used LIWC (Linguistic Analysis and Word Count) features along with conventional machine learning techniques to classify counterfeit news. They tackled the issue with SVM (support vector machine) as a classifier, resulting in an accuracy of 87.00%. Liu et al.'s [14] investigation of bogus tweet detection techniques. Authors used a corpus of more than 8 million tweets collected from supporters of the U.S. presidential candidates during the general election for their investigation. They used deep CNNs for fake news detection in their investigation. They used the idea of subjective analysis in their methodology, and their accuracy was 92.10%. Deep learning techniques have been used by O'Brien et al [15] to categorize false information. Using the black-box method, they were able to attain an accuracy of 93.50% in their study. To identify the positions in false articles, Ghanem et al. [17] have utilized various word embeddings, including n-gram characteristics. They measured accuracy at 48.80%. Singh et al [16] evaluated the use of LIWC (Linguistic Analysis and Word Count) features in classical machine learning approaches for recognizing fake news. They investigated the topic of fake news using SVM (support vector machine) as a classifier and found an accuracy of 87.00%.

The identification and categorization of false news has been automated through the application of machine learning ensemble techniques, as demonstrated in [18]. Different machine learning approaches have utilized textual features for this purpose. The present study has utilized ISOT and two open-source datasets to develop the proposed system. During data preprocessing, documents with less than 20 words are excluded. LIWC, a dialectal mechanism, is then utilized to convert textual features into numerical values. The system employs a range of machine learning algorithms, including logistic regression, SVM, KNN, random forest, and boosting classifications. Ultimately, the decision tree approach with 10-fold cross-validation yielded the highest accuracy of 94%.

Goldani et al.[19] proposed a convolutional neural network (CNN) with margin loss and several embedding models for detecting false news. Static word embeddings are compared to non-static word embeddings, which allow for gradual up training and updating of word embeddings during the training process. Two recent well-known datasets in the field, ISOT and LIAR, are used to test their suggested designs. The results on the optimal architecture indicate promising results, exceeding state-of-the-art approaches by 7.9% on ISOT and 2.1% on the LIAR dataset test set. In Vogel and Meghana's [20] investigation, it was observed that SVM attained the greatest precision of 92% in identifying counterfeit news. The researcher employed manually created characteristics derived from a news dataset such as overall word count (tokens), exclusive words, exclusive word types, type/token proportion, quantity of sentences, average sentence length, quantity of characters, average word length, nouns, prepositions, and adjectives. The counterfeit news classification models comprised XG Boost, Random Forest, Naïve Bayesian, KNN, Decision Tree, and SVM.

III. Methodology

In this section, we present the methodology of our study in detail. Our aim is to detect fake news using NLP techniques through a supervised learning-based approach. To achieve the highest accuracy, we compare several feature extraction techniques and machine learning algorithms. We use three different feature extraction techniques, namely vectorization and word embedding, to represent numeric text data. For vectorization, we use the TF-IDF method, and for word embedding, we use GloVe and DistilBERT. We then train several machine learning algorithms, including support vector machines, random forests, decision trees, and logistic regression, to compare the accuracy of each technique and algorithm. We will provide detailed explanations of each technique and algorithm in the following sections. Below is a visualization of the workflow methodology "Fig. 1" which this research uses.

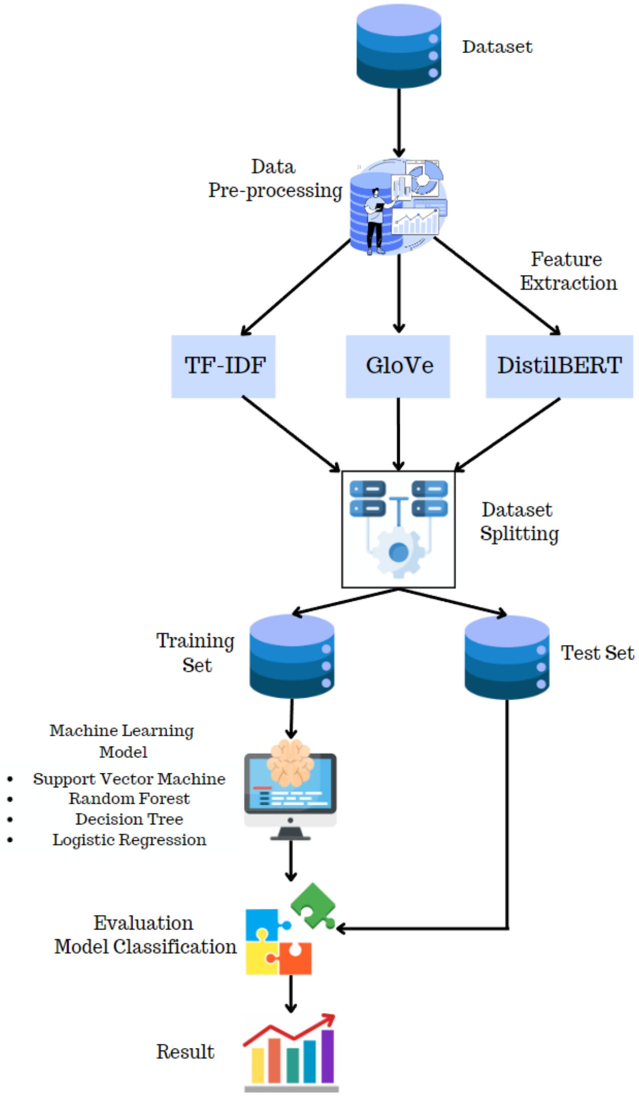

Fig. 1 Methodology Workflow

A. Dataset Description

In this work, [21] an open-source fake news dataset from Kaggle has been used. We utilized the dataset, which consists of 20800 news articles among which 10,387 are real and 10,413 are fake news. The original dataset has five columns, id, title, author, text and label. The id column represents a particular numerical label for a news article. The title holds the heading of a news article. The author column contains the information about the writer of the news item. The text column, the text of the report has been described, and finally the label column, which marks the news item as potentially unreliable or reliable that is labeled as binary 0 and 1. The sample dataset is shown in Table 1.

TABLE I
Dataset Sample

No	Variable	Data Visualization
1	Id	0
	Title	House Dem Aide: We Didn't Even See Comey's Letter Until Jason Chaffetz Tweeted It
	Author	Darrell Lucus
	Text	House Dem Aide: We Didn't Even See Comey's Letter Until Jason Chaffetz Tweeted It By Darrell Lucus on October 30, 2016 Subscribe Jason Chaffetz on the stump in American Fork, Utah (image courtesy Michael Jolley, available under a Creative Commons-BY license) \nWith apologies to Keith Olbermann, there is no doubt who the Worst Person in The World is this week–FBI Director James Comey. But according to a House Democratic aide, it looks like we also know who the second-worst person is as well. It turns out that when Comey sent his now-infamous letter announcing that the FBI was looking into emails that may be related to Hillary Clinton's email server, the ranking Democrats on the relevant committees didn't hear about it from Comey. They found out via a tweet from one of the Republican committee chairmen. \nAs we now know, Comey…
	Label	1 (reliable)
2	ID	1
	Title	FLYNN: Hillary Clinton, Big Woman on Campus - Breitbart
	Author	Daniel J. Flynn
	Text	Ever get the feeling your life circles the roundabout rather than heads in a straight line toward the intended destination? [Hillary Clinton remains the big woman on campus in leafy, liberal Wellesley, Massachusetts. Everywhere else votes her most likely to don her inauguration dress for the remainder of her days the way Miss Havisham forever wore that wedding dress. Speaking of Great Expectations, Hillary Rodham overflowed with them 48 years ago when she first addressed a Wellesley graduating class. The president of the college informed those gathered in 1969 that the students needed "no debate so far as I could ascertain as to who their spokesman was to be" (kind of the like the Democratic primaries in 2016 minus the

	terms unknown then even at a Seven Sisters school). “I am very glad that Miss Adams made it clear that what I am…
Label	0 (unreliable)

B. Pre-processing

This section applies preprocessing text data before feeding it into a feature extraction model for text analysis tasks such as fake news detection. Preprocessing techniques used include removing NaN values, concatenating multiple features into one column, removing non-alphabetic characters, lowercase letters, tokenizing, removing stopwords, and lemmatization. These techniques help remove inconsistencies and noise from data, standardize text data, and reduce the size of data sets. It also helps improve the performance and efficiency of machine learning models. Then the preprocessed data set is divided into 8:2 training and testing samples.

A notable point is the use of stop word removal. It removes common meaningless words in text documents. This helps improve system performance by reducing the weight of common but unimportant words. In addition, lemmatization is used to transform words into stemmed form to resolve data ambiguity and inflection problems. However, the decision whether to lowercase the text or keep the original case may depend on the specific task and context. Overall, these preprocessing techniques are important for text analysis tasks and can significantly improve the accuracy and efficiency of machine learning models.

C. Vectorization

Vectorization is a method of transforming text data into a numerical format by creating a vector of term frequencies. We use the Term Frequency-Inverse Document Frequency (TF-IDF) vectorization method, which computes the importance of a term in a document by scaling the term frequency by the inverse document frequency. This method allows us to represent the text data as a matrix where each row corresponds to a document and each column corresponds to a term, with the cell values indicating the TF-IDF score for the term in the document.

1) *Term Frequency-Inverse Document Frequency (TF-IDF)*: A technique known as Term Frequency - Inverse Document Frequency (TF-IDF) is used to minimize the weights of less significant terms in a set of words. The TF-IDF measure provides information on the frequency of occurrence of terms, taking into account the right balance between the meaning of the local term and its meaning in the context of the entire collection of documents [22]. In the text, the Inverse Document Frequency (IDF) is the ratio of the number of processed documents n_d to the number of documents containing at least one occurrence of the word $\{d: t_i \in d\}$ and is stated as (1):

$$idf_i = log \frac{n_d}{\{d : t_i \in d\}} \qquad (1)$$

The value of TF-IDF is calculated as the product of the frequency of terms $tf_{i,j}$with the inverse frequency in the text idf_i ,which is expressed by the formula in (2):

$$(tf - idf)_{i,j} = tf_{i,j} \times idf_i \qquad (2)$$

By means of the TF-IDF algorithm, the weights of frequently occurring words in most documents are significantly lowered and become lower than words that appear several times but in one document[23].

D. Word Embedding

Word embedding is a word representation technique in the form of a multidimensional vector that captures the semantic relations between words and maps words into a high-dimensional vector space, so semantically similar words are closer together in that vector space. We use two word embedding methods in this paper: GloVe and DistilBERT. GloVe (Global Vectors for Word Representation) is a pre-trained word embedding model that uses co-occurrence statistics to learn word vectors, while DistilBERT is a pre-trained transformer-based model that learns contextualized word embeddings by processing the entire text sequence. We implement DistilBERT using the TensorFlow library (TFDistilBERT).

1) *Global Vector (GloVe):* GloVe is a model that applies weighted least squares to train the model by utilizing co-occurrence counts of words present in the input vectors [24]. It efficiently exploits the advantages of statistical information by training on the non-zero components in a word-to-word co-occurrence matrix. GloVe is an unsupervised training model that is beneficial in determining the correlation between two words based on their proximity in a vector space. These vectors are referred to as word embedding vectors. We have incorporated word embedding as semantic features in addition to n-grams since they represent the semantic distances between words in context. The smallest package of embedding is 822Mb, and it is known as “glove.6B.zip". GloVe model is trained on a dataset comprising one billion words, with a dictionary of 400 thousand words. There are different embedding vector sizes available, which include 50, 100, 200, and 300 dimensions for processing. In this study, we have utilized the 300-dimensional version. All the lines indicating the authors and their affiliations remain unchanged.

2) *TFDistilBERT:* Implementation of DistilBERT [25] in TensorFlow, and it allows for seamless integration with other TensorFlow modules. Like DistilBERT, TFDistilBERT is created by applying knowledge distillation to BERT, specifically the bert-base-uncased model, similar to how DistilBERT is created. To create a smaller version of BERT, TFDistilBERT’s creators removed the token-type embeddings and the pooler from the architecture and reduced the number of layers by a factor of 2. Hence, TFDistilBERT optimized for speed and can be easily integrated into TensorFlow models. In this study, we utilized

TFDistilBERT's last four hidden layers, which were averaged and then fed into the classifier layers for our text classification task. We used the distilbert-base-uncased model, which has a total of 66 million parameters.

E. Machine Learning Classifier Algorithms

To classify real and fake news, we have evaluated the performance of 4 machine learning classifiers for the detection of Fake News, including Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), Logistic Regression (LR). From These 4 machine learning classifiers we have used the scikit-learn python library [26] implementation of each classifier. The brief discussion of each classifier is given as below.

1) *Support Vector Machine (SVM):* The Support Vector Machine (SVM) is a supervised machine learning technique that can be utilized to solve classification and regression problems. However, its primary usage is in classification problems. In the SVM algorithm, each data point is presented as a point in n-dimensional space, with each feature value being the value of a specific coordinate. The classification is achieved by selecting the hyperplane that best separates the two classes. Support vectors are calculated using individual observation coordinates. The SVM classifier is a boundary that optimally distinguishes between the two classes (hyperplane/line). This boundary line is determined as the dividing line between different groups labeled in the given dataset. Additionally, the line is chosen in such a way that it has the maximum distance from the data points of each group and aids in classifying new input [26].

2) *Random Forest (RF):* As the name implies, a random forest is made up of a large number of individual decision trees that work together as an ensemble. The random forest creates a class prediction for each tree, and the class with the most votes becomes our model's forecast. The core premise of the Random Forest is communal knowledge, which is both simple and effective. The random forest model is very effective because it is made up of a large number of largely uncorrelated models (trees) that work together to outperform each of the individual constituent models [26].

3) *Decision Tree (DT):* Decision tree learning is a prominent categorization method. It is highly efficient and has classification accuracy comparable to other learning approaches. Every task in our daily lives consists of a series of decisions based on the availability of resources. This algorithm incorporates this idea. A tree is constructed with each internal node indicating a decision made and the leaf nodes reflecting the consequence of the decision, and the output is predicted by traversing the tree for the most likely outcome [26].

4) *Logistic Regression (LR):* Logistic regression is a widely used Machine Learning method in the Supervised Learning approach. It predicts the category dependent variable based on a set of independent variables. Using logistic regression, the outcome of a categorical dependent variable is predicted. As a result, the output must be either categorical or discrete. It can be Yes or No, 0 or 1, true or False, and so on, but instead of displaying precise values like 0 and 1, it provides probability values ranging from 0 to 1 [26].

IV. Result and Discussion

This section discusses the highest numerical results of the proposed fake news detection system with the applied machine learning approaches. After completion of the necessary processing of the dataset, all the models are assessed. In “Table II”, “Table III” and “Table IV”, we provide all the model evaluations in various ways by checking their accuracy, recall, precision, F1-score, confusion matrix, classification report and ROC AUC Curve visualization. Then the final step is to compare the evaluation results from one model to another and choose the algorithm that produces the best model.

TABLE II
TF-IDF Vectorization Result

		TF-IDF		
Model	**SVM**	**RF**	**DT**	**LR**
Accuracy	0,96268	0,93330	0,95545	0,95329
Precision	0,96198	0,96008	0,95389	0,95151
Recall	0,96337	0,90410	0,95711	0,95518
F1-Score	0,96268	0,93125	0,95550	0,95334

Based on “Table. II”, we conclude. In combination with SVM, TF-IDF can achieve accuracy above 96%. This shows that SVM using TF-IDF vectorization has a very good performance in classifying fake news. Then the combination with Logistic Regression also produces 95% accuracy, which shows that this model is also effective in the classification of fake news. Random Forest achieves 93% accuracy, which is still good, but slightly below SVM and Logistic Regression. Decision Tree, with 95% accuracy, and this shows that the TF-IDF method provides a good representation of distinguishing between fake and genuine news in the dataset we use.

SVM and Logistic Regression seem to match the TF-IDF representation and perform well in predicting fake news classification. We conclude that the reason why TF-IDF produces a high accuracy score is because TF-IDF has an advantage in highlighting keywords that may have a significant impact on the classification of fake news. In the context of fake news classification, words that appear frequently in fake news but rarely appear in genuine news are important indicators.

TABLE III
Glove Embedding Result

		GloVe		
Model	**SVM**	**RF**	**DT**	**LR**
Accuracy	0,80183	0,78305	0,66482	0,80929
Precision	0,80359	0,82904	0,66522	0,79740
Recall	0,79855	0,71277	0,66265	0,82892

F1-Score	0,80106	0,76652	0,66393	0,81285

Based on “Table. III”, we can conclude that the combination with SVM and GloVe results in 80% accuracy. Although there are still improvements compared to the Decision Tree, this performance is lower compared to using TF-IDF. Logistic Regression with GloVe also achieves 80% accuracy. Then Random Forest achieves 78% accuracy, which is slightly lower than SVM and Logistic Regression. Decision Tree shows the lowest performance in this combination, with 66% accuracy.

We conclude that the GloVe embedding method may be less capable of capturing context and more complex text information, so its performance is not as high as that of TF-IDF in this study. The low performance of each model combined with GloVe embedding can be caused by several factors, of course the first is due to the incompatibility of GloVe with the dataset because if the dataset has different characteristics from the GloVe learning corpus and this will be related to GloVe's ability to capture context, then the resulting representation may not be able to map text properly, then the complexity of the pattern in the text and the limitations of the GloVe representation also affect the embedding performance. Apart from the quality of word representation, the selection of simple machine learning models such as SVM or Decision Tree can be a factor, why the model is not able to extract these patterns properly.

TABLE IV
Tfdistilbert Embedding Result

	TFDistilBERT			
Model	**SVM**	**RFt**	**DT**	**LR**
Accuracy	0,89020	0,85312	0,70985	0,91572
Precision	0,91154	0,88695	0,71117	0,92467
Recall	0,86410	0,80916	0,70602	0,90506
F1-Score	0,88718	0,84627	0,70859	0,91476

Based on “Table. III”, we can conclude that SVM in combination with TFDistilBERT achieves an accuracy of 89%. Although slightly lower than the combination with TF-IDF, the performance is still very good. Then Random Forest reaches 85% accuracy, which shows good performance. Decision Tree with TFDistilBERT achieved 70% accuracy. Logistic Regression achieves 91% accuracy, which is the highest performance in this combination. Overall, TFDistilBERT can perform better than GloVe in terms of accuracy but not as well as TF-IDF.

We conclude that there are several factors that cause any combination of TFDistilBERT to produce lower performance compared to TF-IDF, the first of which could be due to the relatively small dataset because complex models like TFDistilBERT require a sufficient amount of data because TFDistilBERT and deep learning models often requires a larger volume of data to provide optimal results for good learning, then the lack of suitability of the dataset with the characteristics of the BERT model. The complexity of the model and data preprocessing also influences the evaluation results because TFDistilBERT is a more complex model layer compared to traditional methods such as TF-IDF. To exploit the full potential of TFDistilBERT requires more data processing and careful tuning of model parameters. If the machine learning model used does not take full advantage of the more abstract text representation provided by TFDistilBERT, then the resulting performance will be low.

Based on the results of the performance table, there are 4 choices to determine the most optimal combination:

- The combination of algorithms that have high accuracy if what is important is how accurately the system classifies correctly, accuracy is the ratio of correct predictions (positive and negative) to the entire data.
- Algorithm combinations that have high recall if False Positives are preferred over False Negatives. This means that it is better to predict the news wrongly when it is actually fake than an algorithm to wrongly predict that the news is fake when it is actually real.
- The combination of an algorithm that has high precision if it prefers a True Positive and really does not want a False Positive to occur, which means that it is better for the algorithm to predict fake news when it is actually genuine than to predict genuine news when it is actually fake.
- Algorithm combination with the highest F1 Score if more concerned with high recall and precision. This means that the selected algorithm gives a small False Positive value and a small False Negative as well.

Based on the four choices of evaluation algorithms, to determine the optimal combination of models, the algorithm chosen in this study is the accuracy algorithm, which is an algorithm that measures how accurately the ratio of correct predictions (positive and negative) is to the entire data. From "Table II", "Table III" and "Table IV" it can be seen that the algorithm that has the highest accuracy is the combination of TF-IDF and SVM with 96%, then the combination of algorithms between TF-IDF and Logistic Regression and also TF-IDF and Decision Tree with 95% followed by TF-IDF and Random Forest with 93%. For the highest results the combination with TFDistilBERT also gives good results, especially with Logistic Regression which achieves 91% accuracy. The following is a confusion matrix and classification report for each combination of algorithms with the highest accuracy.

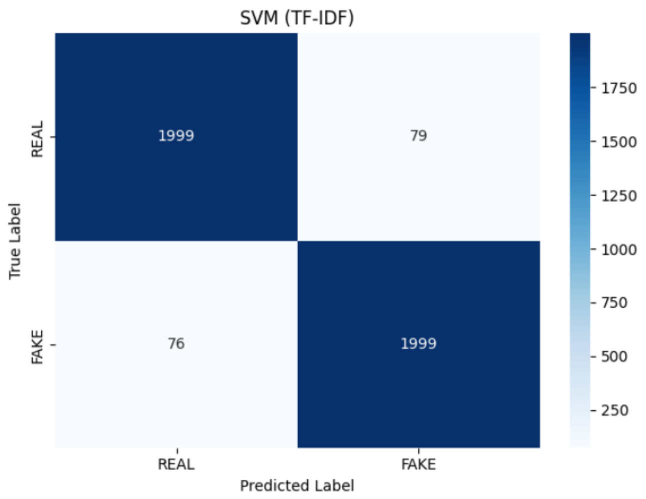

Fig. 2 Confusion matrix of TF-IDF with SVM

TABLE V
Classification Report of Tf-idf with Svm

	Precision	Recall	F1-Score	Support
0 (unreliable)	0.96337	0.96198	0.96268	2078
1 (reliable)	0.96198	0.96337	0.96268	2075
Accuracy			0.96268	4153
Macro Avg	0.96268	0.96268	0.96268	4153
Weighted Avg	0.96268	0.96268	0.96268	4153

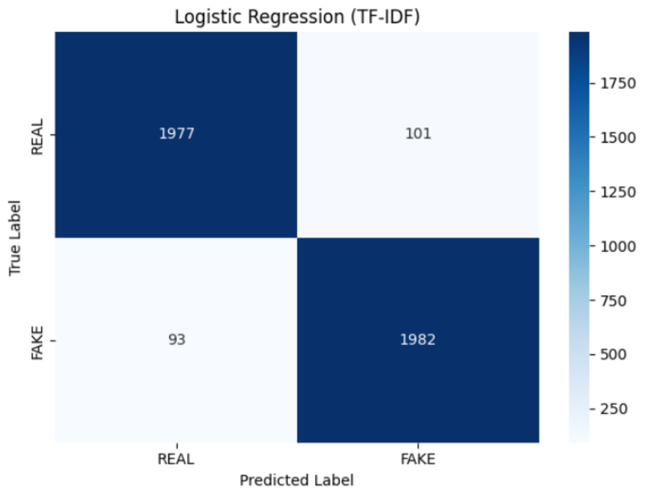

Fig. 3 Confusion matrix of TF-IDF with Logistic Regression

TABLE VI
Classification Report of Tf-idf with Logistic Regression

	Precision	Recall	F1-Score	Support
0 (unreliable)	0.95507	0.95140	0.95323	2078
1 (reliable)	0.95151	0.95518	0.95334	2075
Accuracy			0.95329	4153
Macro Avg	0.95329	0.95329	0.95329	4153
Weighted Avg	0.95329	0.95329	0.95329	4153

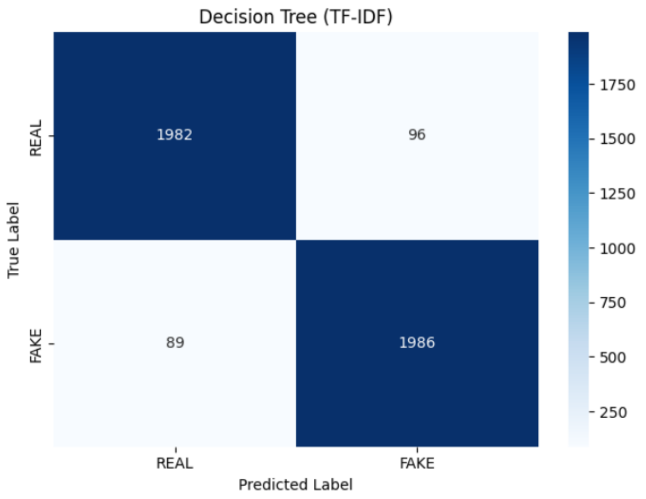

Fig. 4 Confusion matrix of TF-IDF with Decision Tree

TABLE VII
Classification Report of Tf-idf with Decision Tree

	Precision	Recall	F1-Score	Support
0 (unreliable)	0.95648	0.95188	0.95417	2078
1 (reliable)	0.95204	0.95663	0.95433	2075
Accuracy			0.95425	4153
Macro Avg	0.95425	0.95425	0.95425	4153
Weighted Avg	0.95425	0.95425	0.95425	4153

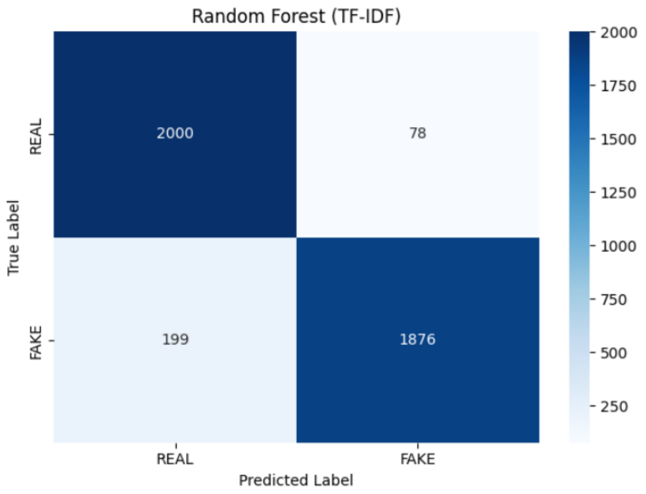

Fig. 5 Confusion matrix of TF-IDF with Random Forest

TABLE VIII
Classification Report of Tf-idf with Random Forest

	Precision	Recall	F1-Score	Support
0 (unreliable)	0.90671	0.96824	0.93647	2078

1 (reliable)	0.96587	0.90024	0.93190	2075
Accuracy			0.93426	4153
Macro Avg	0.93629	0.93424	0.93419	4153
Weighted Avg	0.93627	0.93426	0.93419	4153

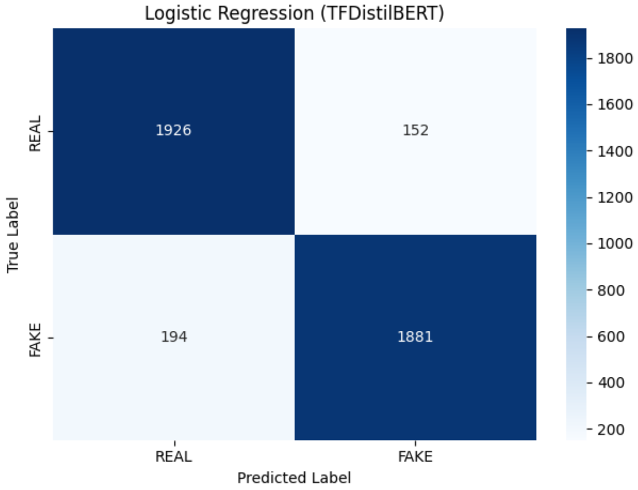

Fig. 6 Confusion matrix of DistilBERT with Logistic Regression

TABLE IX
Classification Report of DistilBERT with Logistic Regression

	Precision	**Recall**	**F1-Score**	**Support**
0 (unreliable)	0.90849	0.92685	0.91758	2078
1 (reliable)	0.92523	0.90651	0.91577	2075
Accuracy			0.91669	4153
Macro Avg	0.91686	0.91668	0.91668	4153
Weighted Avg	0.91686	0.91669	0.91668	4153

In the Confusion matrix, performance information is only presented in numbers. To display information on the performance of the classification algorithm in graphical form, you can use the Receiver Operating Characteristic (ROC) or the Precision-Recall Curve. The ROC curve is made based on the values that have been obtained from calculations with the confusion matrix, namely between False Positive Rate and True Positive Rate. To compare the performance value of each algorithm can be done by comparing the area under the curve or AUC (Area Under Curve).

The advantage of using the ROC curve to evaluate classification is that ROC is not just to find the average accuracy but ROC visualizes all possible classification thresholds, while the classifier error rate only represents the level of error, accuracy for only one threshold.

Based on the combination of algorithms that have been proven to produce high accuracy, we provide the ROC AUC Curve for the combination of algorithms that has the highest accuracy that has been selected.

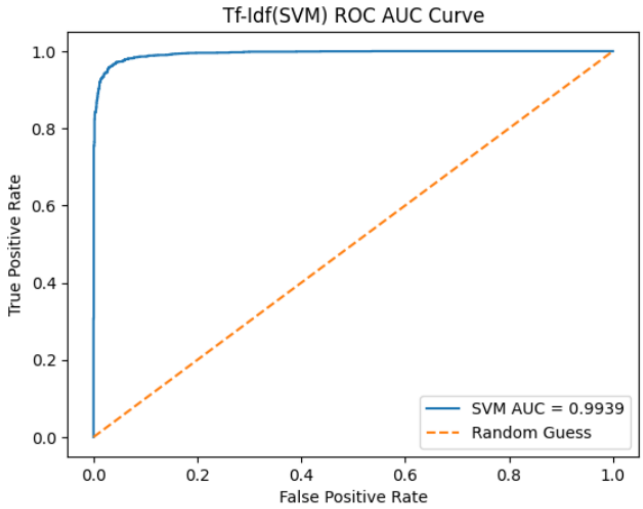

Fig. 7 ROC Curve of TF-IDF with SVM

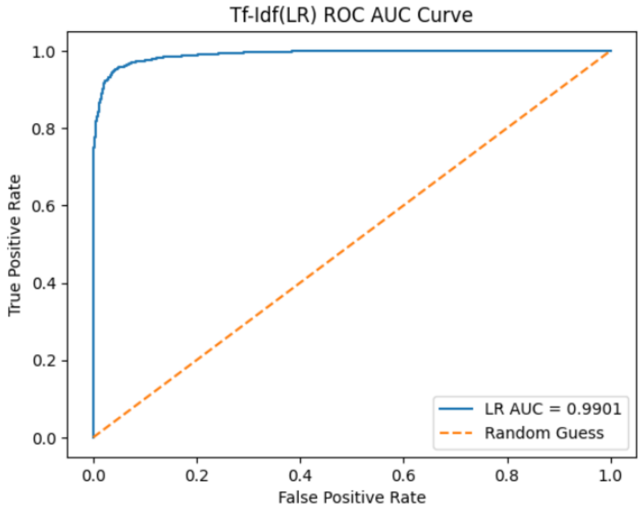

Fig. 8 ROC Curve of TF-IDF with Logistic Regression

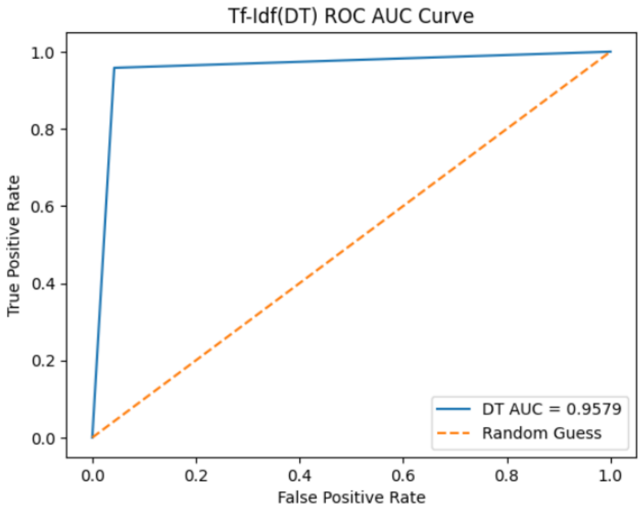

Fig. 9 ROC Curve of TF-IDF with Decision Tree

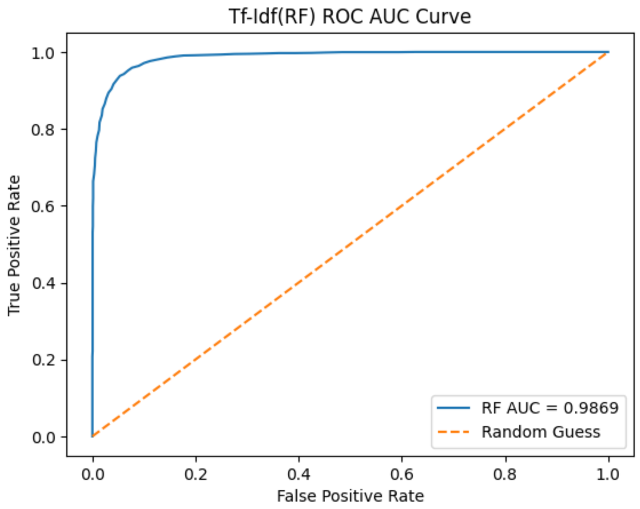

Fig. 10 ROC Curve of TF-IDF with Random Forest

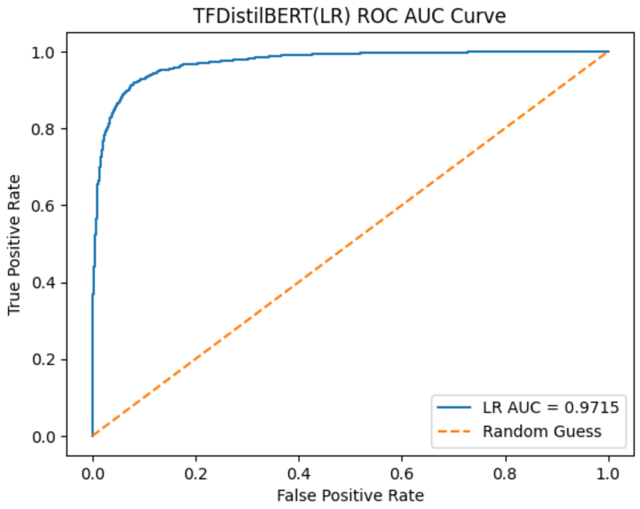

Fig. 11 ROC Curve of DistilBERT with Logistic Regression

Diagnostic test research will be better if the AUC value is close to 1. _AUC value: 0.5 ≤ AUC < 0.6 very weak, 0.6 ≤ AUC < 0.7 weak, 0.7 ≤ AUC < 0.8 moderate, 0.8 ≤ AUC < 0.9 good and 0.9 ≤ AUC ≤ 1 very Good. So if based on the evaluation results, the ROC-AUC curve, the best model is TF-IDF with SVM then TF-IDF with Logistic Regression then TF-IDF with Random Forest and finally DistilBERT with Logistic Regression. The AUC value provides an overview of the overall measurement of the suitability of the model used. The greater the Area Under Curve (AUC), the better the model is in predicting news.

V. Conclusion

In this research, we compared various feature extraction techniques and machine learning algorithms performance. To identify the optimal combination of feature extraction technique and machine learning algorithm that would produce the highest accuracy in detecting fake news. We utilize three different feature extraction techniques. With vectorization and word embedding to represent numeric text data. The vectorization method used was TF-IDF, and for word embedding, we used GloVe and DistilBERT. The results of each technique were then used to train several machine learning algorithms, including Support Vector Machines, Random Forests, Decision Tree, and Logistic Regression, to compare their score rate.

Based on the results of the evaluation by looking at the accuracy score and the AUC ROC curve, the conclusions in this study, each machine learning algorithm that is combined with TF-IDF gets the highest accuracy score compared to the word embedding technique with GloVe and TFDistilBERT with a score above 92% with an AUC score above 0.94 and SVM is the best choice to be combined with TF-IDF followed by Logistic Regression, Decision Tree, and finally Random Forest. As for the results of the combination of word embedding with machine learning models that are most optimal and have high accuracy is the combination of the TFDistilBERT algorithm with Logistic Regression with 91% for accuracy and 0.97 for the AUC score. However, for further research purposes in order to obtain a better accuracy value for the word embedding technique, it is proposed to try using other algorithms such as Deep Learning, applying fine-tuned methods for the models, consider using hyperparameter tuning for the models to increase the accuracy rate and trying with different datasets.

References

[1] A. Kamaliah, “Indonesia’s Digital Literacy Is Far Behind in the World,” Detikcom, 2020. Accessed: Jun. 10, 2023. [Online]. Available: https://inet.detik.com/cyberlife/d-4933782/literasi-digital-indonesia-ketinggalan-jauh-di-dunia

[2] Mastel, “Results of the 2019 National Hoax Outbreak Survey,” Masyarakat Telematika Indonesia, 2019. https://mastel.id/hasil-survey-wabahhoax-nasional-2019/

[3] O. D. Apuke and B. Omar, “Fake news and COVID-19: modelling the predictors of fake news sharing among social media users,” Telematics and Informatics, vol. 56, p. 101475, Jan. 2021, doi: 10.1016/j.tele.2020.101475.

[4] Arifa Rachma Febriyani and Rintulebda Anggung Kaloka, “Communication Strategy of the Batang Regency Communication and Informatics Service in Counteracting Hoaxes,” Sosiohumaniora, vol. 8, no. 1, pp. 33–45, Feb. 2022, doi: 10.30738/sosio.v8i1.11853.

[5] H. Naufal Marbella, N. Hanifah Nur’aini, S. Agung, and N. Aini Rakhmawati, “Analysis of the influence of fake news on social media on the Indonesian people’s decision to vaccinate against Covid-19,” Jurnal Indonesia Sosial Teknologi, vol. 2, no. 11, pp. 1951–1966, Nov. 2021, doi: 10.36418/jist.v2i11.267.

[6] A. Vidi, “Hoaxes Regarding Covid-19 Still Proliferating, Kominfo Drops 2,927 content on Social Media,” Liputan6, May 15, 2021. Accessed: May. 28, 2023. [Online]. Available: https://m.liputan6.com/cekfakta/read/4558123/hoaks-seputarcovid-19-masih-menjamur-kominfoturunkan-2927-konten-di-mediasosial

[7] I. E. Alamsyah, “Police Detain 17 Suspects of spreading Covid-19 Hoaxes,” Republika Online, Nov. 25, 2020. Accessed: May. 28, 2023. [Online]. Available: https://news.republika.co.id/berita/qkbdgc349/polisi-tahan-17-tersangka-penyebar-hoaks-covid19

[8] N. Hidayah et al., “2020 COVID-19 Hoax Mapping Report – MAFINDO,” Mafindo, 2020. https://www.mafindo.or.id/2021/10/18/laporan-pemetaan-hoaks-covid-19-tahun-2020/ (accessed May. 28, 2023).

[9] H. Ahmed, I. Traore, and S. Saad, “Detecting opinion spams and fake news using text classification,” Security and Privacy, vol. 1, no. 1, p. e9, Dec. 2017, doi: 10.1002/spy2.9.

[10] J. C. S. Reis, A. Correia, F. Murai, A. Veloso, and F. Benevenuto, “Supervised Learning for Fake News Detection,” IEEE Intelligent Systems, vol. 34, no. 2, pp. 76–81, Mar. 2019, doi: 10.1109/mis.2019.2899143.

[11] S. Kumari, “NoFake at CheckThat! 2021: Fake News Detection Using BERT,” arXiv.org, Aug. 11, 2021. https://arxiv.org/abs/2108.05419

[12] R. Mansouri, M. Naderan-Tahan, and M. J. Rashti, “A Semi-supervised Learning Method for Fake News Detection in Social Media,” in 2020 28th Iranian Conference on Electrical Engineering (ICEE), Aug. 2020. Accessed: May 28, 2023. [Online]. Available: http://dx.doi.org/10.1109/icee50131.2020.9261053

[13] H. Ahmed, I. Traore, and S. Saad, “Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques,” in Lecture Notes in Computer Science, Cham: Springer International Publishing, 2017, pp. 127–138. Accessed: May 28, 2023. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-69155-8_9

[14] Y. Liu and Y.-F. Wu, “Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, Apr. 2018, doi: 10.1609/aaai.v32i1.11268.

[15] N. O’Brien, S. Latessa, G. Evangelopoulos, and X. Boix, “The Language of Fake News: Opening the Black-Box of Deep Learning Based Detectors,” DSpace@MIT, Nov. 01, 2018. http://hdl.handle.net/1721.1/120056 (accessed May 28, 2023).

[16] V. K. Singh, R. Dasgupta, D. Sonagra, and I. Ghosh, “Automated Fake News Detection Using Linguistic Analy-sis and Machine Learning,” unknown, Jul. 05, 2017. https://www.researchgate.net/publication/318541620_Automated_Fake_News_Detection_Using_Linguistic_Analy-sis_and_Machine_Learning (accessed May 28, 2023).

[17] B. Ghanem, P. Rosso, and F. Rangel, “Stance Detection in Fake News A Combined Feature Representation,” in Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), 2018. Accessed: May 28, 2023. [Online]. Available: http://dx.doi.org/10.18653/v1/w18-5510

[18] I. Ahmad, M. Yousaf, S. Yousaf, and M. O. Ahmad, “Fake News Detection Using Machine Learning Ensemble Methods,” Complexity, vol. 2020, pp. 1–11, Oct. 2020, doi: 10.1155/2020/8885861.

[19] M. H. Goldani, R. Safabakhsh, and S. Momtazi, “Convolutional neural network with margin loss for fake news detection,” Information Processing &amp; Management, vol. 58, no. 1, p. 102418, Jan. 2021, doi: 10.1016/j.ipm.2020.102418.

[20] I. Vogel and M. Meghana, “Detecting Fake News Spreaders on Twitter from a Multilingual Perspective,” in 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), Oct. 2020. Accessed: May 28, 2023. [Online]. Available: http://dx.doi.org/10.1109/dsaa49011.2020.00084

[21] W. Lifferth, “Fake News,” Kaggle, 2018. https://kaggle.com/competitions/fake-news (accessed May 28, 2023).

[22] G. Sidorov, “Vector Space Model for Texts and the tf-idf Measure,” in Syntactic n-grams in Computational Linguistics, Cham: Springer International Publishing, 2019, pp. 11–15. Accessed: May 29, 2023. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-14771-6_3

[23] N. S. Mohd Nafis and S. Awang, “An Enhanced Hybrid Feature Selection Technique Using Term Frequency-Inverse Document Frequency and Support Vector Machine-Recursive Feature Elimination for Sentiment Classification,” IEEE Access, vol. 9, pp. 52177–52192, 2021, doi: 10.1109/access.2021.3069001.

[24] T. Shi and Z. Liu, “Linking GloVe with word2vec,” arXiv.org, Nov. 20, 2014. https://arxiv.org/abs/1411.5595

[25] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” arXiv.org, Oct. 02, 2019. https://arxiv.org/abs/1910.01108

[26] Pedregosa, “scikit-learn: machine learning in Python — scikit-learn 1.2.2 documentation,” Scikit-learn, 2011. https://scikit-learn.org/stable/ (accessed May 28, 2023).