International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

A Pre-trained Transformer-based Ensemble Model for Automated Indonesian Fake News Classification

Pauw Danny Andersen[1], **Derwin Suhartono**[2]

Submitted: 28/04/2023 **Revised**: 26/06/2023 **Accepted**: 09/07/2023

Abstract: Fake news often aims to damage the reputation of a person or entity, or to generate personal gain. The lack of a scalable fake news classification strategy is particularly worrying. Since manually classifying fake news is a time-consuming task, automatic identification of fake news has attracted a lot of attention in the Natural Language Processing (NLP) community to help ease the activity of classifying fake news. In recent Indonesian language news dataset, existing machine learning algorithms such as KNN and Naïve Bayes are used in this task, however it suffers from the lack of the ability to capture the true (semantic) meaning of words; therefore, the context is slightly lost. To address limitations, this paper introduces a new prediction using ensemble transformer based deep learning pre-trained language model such as BERT, RoBERTa, and DistilBERT as features extraction method on social media data sources. Finally, the system takes the decision based on model averaging to make prediction. Our proposed work yields promising performance as it has outperformed similar existing works in the literature. More precisely, our results achieve a maximum accuracy of 0.887 and f1 measure score of 0.878 on the news dataset.

Keywords: *Deep Learning, Fake News, Natural Language Processing, Text Mining, Transformer Model*

1. Introduction

Newspapers and television have long been the main sources for the public to consume news. The presence of internet-based digital media has quickly shifted the role of print media and television media. Figure 1 shows the source of news in Indonesia in the year 2021 by Reuters and Oxford University. It showed that the internet was the most popular source of news in 2021, including social media which topped the list with 89%, followed by television at 58% while print media only drew 20% [1]. The presence of online media has drastically changed the way news is produced, disseminated, and consumed by the public, thus creating new, more complex challenges [2].

The main problem today is that online media is the main place for the publication of false news and information that can cause harm to others. Everyone can create their own news site and claim to be a news publisher without certain qualifications [3]. Therefore, there are many concerns about the rise of untrusted news sites and often the news can be quickly disseminated using social media. With the spread of fake news and its negative impact on society, the lack of skills and strategies to identify fake news is a critical problem.

Fake news is an article that contains false information that claims to be a news. Fake news is also often used to bring down an entity or a person, and to generate personal gain [4]. The ability to identify fake news is very important and needed at this time, but technically it is difficult to do. The difficulty lies in that humans find it difficult to distinguish fake and genuine news. For example, a study found that in viewing a fake news article, 75% of respondents stated that the news was genuine, and the study also revealed that 80% of high school students had difficulty distinguishing between genuine and fake news articles [5].

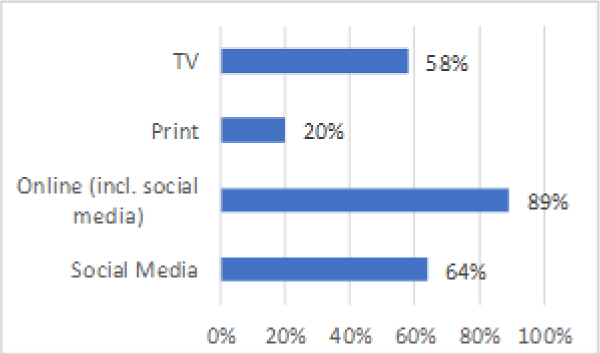

Fig. 1. Source of news in Indonesia in 2021 [1]

Since classifying fake news manually takes a lot of time and effort, the effort to automatically classify fake news and make it easier to identify fake news has attracted the attention of the Natural Language Processing (NLP) community. Even for automated systems, identifying fake news is still quite a challenge. Learning models such as the Recurrent Neural Network (RNN) and its variants as well as the Convolutional Neural Network (CNN) are widely used for this task. A study conducted by [20] proposed a hybrid Neural Network architecture, which combines the

[1] *Computer Science Department, BINUS Graduate Program, Master of Computer Science, Bina Nusantara University, Jakarta, INDONESIA*
ORCID ID: 0009-0001-1984-3628
[2] *Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, INDONESIA*
ORCID ID: 0000-0002-3271-5874
Coresponding Author Email: dsuhartono@binus.edu

capabilities of CNN and Long short-term memory (LSTM). However, the LSTM or RNN model is sequential and needs to be processed sequentially, unlike the transformer model. Due to the parallelization capability of the transformer mechanism, more data can be processed at the same time with the transformer model. In the recent Indonesian language news datasets, existing machine learning algorithms such as K Nearest Neighbor (KNN) and Naïve Bayes are used in this task [7], however, it still lacks the ability for capturing the meaning and context of words.

Addressing the issues, this paper proposes a method utilizing a pre-trained transformer-based language model such as BERT, RoBERTa, and DistilBERT on the Indonesian language news dataset. The system takes the decision based on model averaging to make the prediction. The proposed work uses an Indonesian language news dataset consisting of 228 fake news and 372 genuine news from various online news sources collected by a previous study [7].

2. Related Works

Various method of fake news classification has been proposed by many researchers. [8] developed an automatic classification of fake content in online news. This study uses 2 different datasets to classify fake news, covering 7 different news domains. Datasets are obtained by manually collecting, as well as retrieving directly from the web. After that, identification of the linguistic characteristics of fake news and real news is carried out. Then make a classification of fake news based on these linguistic features. The accuracy of this study reached 76% by comparing the results of automatic detection with subjective analysis of fake news by humans. [9] developed a prevention method for online reviews that were conducted dishonestly. Previously to detect fake reviews using syntactic and lexical pattern detection methods, this study used a neural approach by modifying the transformer-based architecture from Google (BERT). The accuracy of this method is 90% using datasets from OpSpam and Yelp. Further development in the future is carried out by understanding the relatively poor performance in some parts to get a more effective classification method.

The researcher also starts to consider other features of fake news to increase the model's performance. [10] conducted an analysis to distinguish fake news in the form of satire and genuine news by matching and analyzing 12 general news topics covering 4 domains. This is done to minimize the potential for readers to be deceived by satirical news. The study uses an SVM-based algorithm, which is enriched with 5 predictive features (Absurdity, Humor, Grammar, Negative Affect, and Punctuation) and a dataset of a combination of 360 news articles sourced from The Onion, The Beaverton, The Toronto Star, and The New York Times. The accuracy of this study reached 90%. [5] tried to classify fake news by combining 3 general characteristics, including the text of the news article, the response received by the news, and the source of the news. This study uses a model called CSI which consists of 3 modules (Capture, Score, and Integrate). The accuracy of this study ranges from 89%-95% depending on the dataset used. [11] made an effort to classify fake news by classifying the degree of fake news by combining it from different sources. This study uses the Multi-source Multi-class Fake news Detection Framework (MMFD), by combining several features to get an accurate automatic classification. [12] shows that posts on Facebook can be accurately classified as hoaxes or not hoaxes.

In a recent Indonesian language news dataset, existing machine learning algorithms such as Naïve Bayes are used in this task. [7] proposes using the Naïve Bayes algorithm for the classification method with the highest testing accuracy of 78,6%. On the other hand, the research by [13] aims to utilize the K Nearest Neighbor (KNN) classification algorithm to detect whether a news is a hoax or not. The experiment was carried out using 74 hoaxes collected from Indonesian hoax prevention community sites and compared with 74 real news from various leading sites in Indonesia. The results showed that the model can provide a detection/classification accuracy of up to 83.6%.

Recently transformer-based model has been utilized to solve this task. [14] proposes a fine-tuning approach to a transformer-based language model to detect fake news. The predictive features extracted by the RoBERTa model, and the CT-BERT model were combined. This method evaluated on the existing COVID-19 fake news dataset and showed better performance compared to other methods. [6] proposes deep learning based on BERT (Bidirectional Encoder Representations from Transformers) [7]. FakeBERT combines the different parallel blocks of the Convolutional Neural Network (CNN). The combination is useful for dealing with ambiguity, which is the biggest challenge in NLP. The classification results show that the proposed model (FakeBERT) outperforms the existing model with an accuracy of 98.90%.

The methods with the highest result are the transformer-based methods from [14] and [15] where each method has an accuracy rate of more than 90%. However, each study cannot be compared directly, considering the datasets used and the problems studied have differences. The approach used by [14] is for the classification of fake news regarding COVID-19, while [15] focuses more on the classification of fake news on social media. However, these studies conclude that the transformer-based model is the state-of-the-art approach to fake news classification tasks to date.

3. Methodology

The research stage consists of 3 parts, namely initiation, building the proposed model, and evaluate the proposed model. During the initiation stage, datasets from previous studies were collected. The dataset consists of 228 fake news and 372 real news. Fake news will later be represented with a value of 1 and genuine news with a value of 0. This label will later become the target prediction of the proposed model. The dataset will be augmented and preprocessed. Furthermore, the preprocessed data will go through a features extraction process, before entering the proposed deep learning model. Finally, the performance results of the proposed model will go through a model evaluation process.

3.1. Preprocessing

The dataset used in this research is called the Indonesian Hoax News Detection Dataset, which consists of 600 news articles, 228 fake news, and 372 genuine news collected by previous study [7]. This dataset is collected manually through various online media sources, the news is labeled by three referees whether it is a hoax or valid. The final label is derived from the voting procedure of the three referees. The dataset will go through an augmentation process, augmented datasets consist of 450 fake news and 450 genuine news. The dataset will be divided into 3 parts, namely training, testing, and validation sets with a distribution of 70% for the training set, 15% for the test set, and 15% for the validation set.

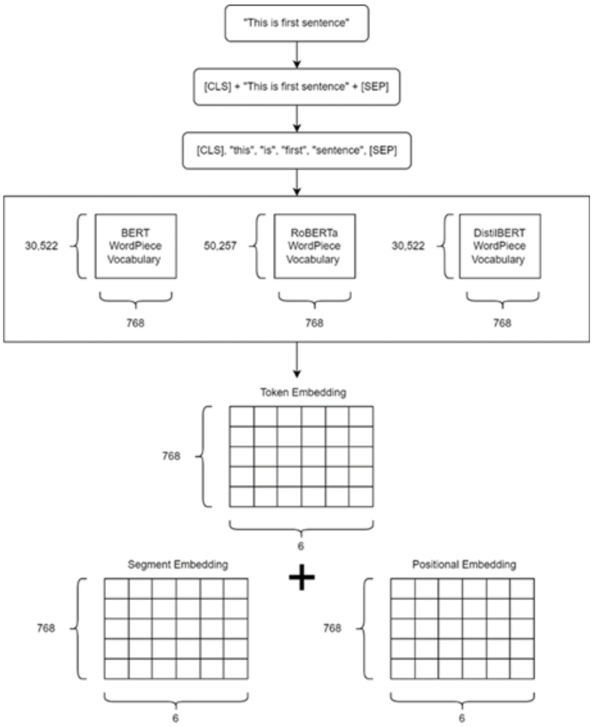
"This is first sentence"
[CLS] + "This is first sentence" + [SEP]
[CLS], "this", "is", "first", "sentence", [SEP]
30,522
BERT WordPiece Vocabulary
768
50,257
RoBERTa WordPiece Vocabulary
768
30,522
DistilBERT WordPiece Vocabulary
768
Token Embedding
768
6
Segment Embedding
768
6
+
Positional Embedding
768
6

Fig. 2. Preprocessing Stage

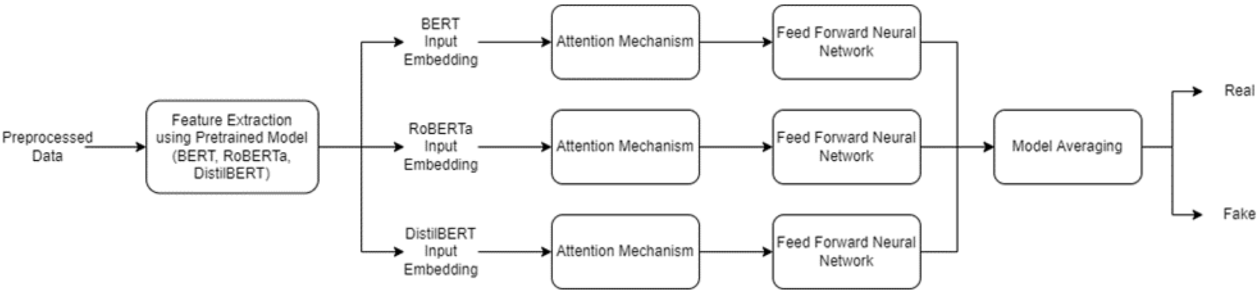
Preprocessed Data
Feature Extraction using Pretrained Model (BERT, RoBERTa, DistilBERT)
BERT Input Embedding
RoBERTa Input Embedding
DistilBERT Input Embedding
Attention Mechanism
Attention Mechanism
Attention Mechanism
Feed Forward Neural Network
Feed Forward Neural Network
Feed Forward Neural Network
Model Averaging
Real
Fake

Fig. 3. Proposed deep learning model architecture.

The collected dataset will go through pre-processing steps. Pre-processing will be carried out using several steps where the main purpose of pre-processing is to maximize the feature extraction stage. The steps of the pre-processing are shown in Figure 2.

Due to data scarcity and insufficient data diversity, Data Augmentation (DA) is performed on the raw data. Data Augmentation is a process that artificially increases training data size by generating different versions of real datasets without collecting the data. The data needs to be changed to preserve the class categories for better performance in the classification task. Easy Data augmentation (EDA) chooses a word randomly from the sentence and replaced it with one of these word synonyms or two words are chosen and swapped in the sentence [16]. Augmented datasets consist of 450 fake news and 450 genuine news data go through the next process.

Noise removal is one of the most important stages in the pre-processing process. Noise removal is the stage where digits, symbols, and punctuation characters are removed from the text because these characters can interfere with the analysis of the existing text. The next stage is tokenization, where the text of the sentence will be broken down into smaller parts called tokens. Next is the stop words removal process. Stop words are words that are common in a language so that their meaning is not significant in a text, so they will be omitted from the text. The next stage is lowercasing, where each letter will be converted to lowercase. The last stage is lemmatization, where the word will be changed into its root form. In this lemmatization process, part of speech tagging will be used to make the lemmatization result better. Part of speech tagging is the process of assigning part of speech to each word in a sentence.

3.2. Model Architecture

Based on several previous studies, combining several pre-trained models is very effective in various classification tasks. This study proposes the ensemble pre-trained language model, namely BERT, RoBERTa, and DistilBERT, on Indonesian language datasets. Figure 3 shows the model architecture for the proposed model.

The transformer-based models used in this research include:

BERT: Bidirectional Encoder Representations from Transformers (BERT) is a bidirectional transformer pretrained using a combination of masked language modeling objective and next sentence prediction on a large corpus comprising the Toronto Book Corpus and Wikipedia. BERT is a bidirectional model so that it considers all parts of the text to understand the meaning of each token [17]. The text is converted to lowercase and will be tokenized using WordPiece with a vocabulary size of 30,522. The input from the model is then in the form of [CLS] Sentence A [SEP] Sentence B [SEP].

RoBERTa: University of Washington researchers analyzed Google's BERT model training and identified several changes to the training procedure that improved its performance. Specifically, the usage of the new, larger data set for training, and training the model through more iterations. This optimized model is called RoBERTa (Robustly Optimized BERT Approach) [18]. Text is tokenized using a byte version of Byte-Pair Encoding (BPE) and a vocabulary size of 50,257. The input of the model takes a chunk of 512 tokens. The beginning of a new document is marked with <s> and the end with </s>.

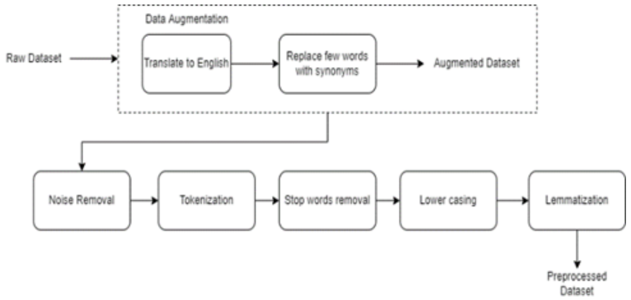

Fig. 4. Pre-trained Model Features Extraction

DistilBERT: DistilBERT is a small, fast, cheap, and light Transformer model trained by distilling BERT base. It has 40% less parameters than bert-base-uncased, and runs 60% faster while preserving over 95% of BERT’s performances as measured on the GLUE language understanding benchmark [19]. The text is converted to lowercase and will be tokenized using WordPiece with a vocabulary size of 30,000. The input from the model is then in the form of [CLS] Sentence A [SEP] Sentence B [SEP].

The feature extraction carried out by the pretrained model goes through several steps. Figure 4 illustrates the steps in the feature extraction. First, the [CLS] token will be added at the beginning of the sentence and the [SEP] token will be added at the end of the sentence. Each word in the sentence will go through a tokenization process which will later become a series of word tokens. Tokenization in the pre-trained model uses the WordPiece tokenization method. The differences in the pre-trained models used, namely BERT, RoBERTa, and DistilBERT, are found in the vocabulary wordpieces of each model. In the BERT model, which was built using a vocabulary consisting of 30,522 words, RoBERTa consisted of 50,257 words, and DistilBERT had a vocabulary of 30,522. Then the embedding token will be added with segment embedding and positional embedding, to add context to each embedding. Maximum sequence length for each model is 512, wherever sequences longer than 512 will be truncated.

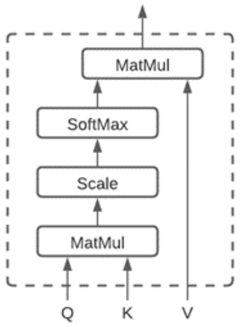

Fig. 5. Attention Mechanism [20]

Furthermore, the self-attention mechanism receives input from the resulting embedding results. Figure 5 illustrates the process that occurs in the self-attention mechanism. First, there is a matrix multiplication between the query vector (Q) and the key vector (K), where this multiplication will produce a matrix score.

This matrix score will go through a scaled function, where the score value will be divided by the square root of the query and key dimensions. The calculation of the scaled value can be seen in equation 1. This process is carried out to obtain a more stable gradient. Furthermore, the output of the scaled function will be entered into the softmax function, where the softmax function will produce an output between zero and one. The softmax function will make higher scores to one and lower scores to zero.

$$Scaled\ (x) = \frac{QK}{\sqrt{d_k}} \qquad (1)$$

$$\text{Softmax}\ (x) = \frac{\exp\ (x_i)}{\Sigma_j \exp\ (x_i)} \qquad (2)$$

A low softmax value will eliminate irrelevant words, so the model will only learn the words that are important. The softmax function can be seen in equation 2. The output of the softmax function will be multiplied by the vector value (V). The result of the matrix multiplication will be combined with the original embedding which is called the residual connection. The output of the residual connection will be entered as the input of the feed-forward neural network.

The feed-forward neural network architecture consists of two hidden layers with Rectified Linear Unit (ReLU) activation function and a dropout function that serves to reduce overfitting and generalization of training data. Dropout randomly ignores the output by changing its value to zero. Cross entropy is used as a loss function to measure how well the prediction model results at the training stage. According to previous deep learning literature [21, 22], the unweighted averaging might be a reasonable ensemble for similar base learners. The model averaging (unweighted) can be calculated by combining the softmax probabilities from three different classifications model [23]. The mean class probability is calculated as follow equation 3 and 4.

$$y^*_{i,k} = \frac{y_{i1,k}+y_{i2,k}+y_{i3,k}}{3} \forall k \in [1 \ldots K] \quad (3)$$

$$y = \arg\ \max(y^*_i, k) \quad (4)$$

The parameter tuning process is done to obtain optimal model performance. Parameter tuning is a process to get the best combination of parameters. The parameters that will be adjusted in this study are mainly batch size and learning rate. Table 1 shows the combination of parameter that will be test out for the experiment.

Table 1. Parameter tuning experiment scenarios.

Scenario	**System Baseline**	**Batch Size**	**Learning Rate**
1-12	BERT	8, 16, 32	1.00E-5, 3.00E-5, 5.00E-5, LR Scheduler (5.00E-5 – 0.0)
12-24	RoBERTa	8, 16, 32	1.00E-5, 3.00E-5, 5.00E-5, LR Scheduler (5.00E-5 – 0.0)
24-36	DistilBERT	8, 16, 32	1.00E-5, 3.00E-5, 5.00E-5, LR Scheduler (5.00E-5 – 0.0)

3.3. Evaluation Metrics

Evaluation metrics are used to evaluate and compare performance between models. The metrics used are accuracy, precision, recall, and f1 score. The accuracy value is obtained by dividing the total number of correct predictions and the total number of incorrect predictions. Precision is obtained by dividing the true positive by the total predicted positive (true positive + true negative). Precision shows how accurately the model can predict positive values. Precision is a good measure of seeing a high number of false positives. Recall is obtained from the division between true positives and total actual positives (true positive + false negative). Recall counts how many actual positives the model has successfully predicted correctly. F1 score is obtained from 2 times of precision times recall divided by precision plus recall. The F1 score is a good measurement if there is an uneven class distribution. The F1-score is used to calculate the class accuracy as an evaluation metric to demonstrate the completeness of the proposed model.

4. Result and Discussion

Prediction results of all models will be evaluated with accuracy, precision, recall, and f1 measure metrics. The results of the performance evaluation of the proposed deep learning model can be seen in table 2, where the DistilBERT model has the highest performance results with 0.887 accuracy and 0.878 f1 score. The proposed model has the second highest performance evaluation result after Distillbert with an accuracy score of 0.831 and an f1 score of 0.815. The difference in performance results between pre-trained transformer models is quite far, namely BERT with an accuracy of 0.803 and f1 of 0.778, RoBERTa with an accuracy of 0.739 and f1 of 0.766, and DistilBERT with an accuracy of 0.887 and f1 of 0.878, causing the ensemble model to have less than ideal results, with performance below the single model. The ensemble model averaging method will be better if it is used for models that have comparable results so that the ensemble model will have an increase in performance from a single model. However, the proposed model with the averaging ensemble model and the DistilBERT model has higher accuracy and f1 score than previous studies with the same dataset, namely an accuracy of 0.831 and f1 score of 0.815 for the proposed ensemble model and an accuracy of 0.887 and f1 score of 0.878 for the DistilBERT model compared to an accuracy of 0.786 and f1 score of 0.798 from previous studies.

The hyperparameter tuning is done by using 5-fold cross-validation. In determining the batch size and learning speed, the data validation mentioned in the previous section was used. From these results, the model used requires a batch size of 16 to get optimal results. As for the learning rate, all models use a learning rate scheduler with a learning rate value of 5.00E-5 – 0.0 for optimal performance.

Table 2. Deep learning classification performance result and comparison with previous work

System Baseline	Metric			
	Accuracy	Precision	Recall	F1 Measure
BERT	0.803	0.772	0.785	0.778
RoBERTa	0.739	0.726	0.766	0.766
DistilBERT	**0.887**	**0.902**	**0.857**	**0.878**
Proposed Model	0.831	0.864	0.772	0.815
Pratiwi et al. [7]	0.786	0.793	0.804	0.798

5. Conclusion

This research shows a comparison of several approaches to developing a fake news classification system using an Indonesian language news dataset. From the experimental results, the deep learning approach using BERT, RoBERTa, DistilBERT as a pre-trained language model and the proposed ensemble model with the model averaging method outperform other approaches in terms of average accuracy and f1 scores. However, if the proposed ensemble model compared with pre-trained single model, DistilBERT has better performance result, due to the significantly different result of each pretrained model that being ensembled. In the future, experiments can be carried out using a larger Indonesian dataset. Comparisons and ensemble with other pre-trained models such as IndoBERT and Multilingual BERT may increase the performance of the prediction, where the Indonesian language text does not need to be translated into English, thus minimizing the shift in context and meaning resulting from inaccuracies in the translation process.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Steele, J. (2021, June 23). Indonesia Digital News Report 2021. Reuters Institute for the Study of Journalism. Retrieved August 13, 2022, from https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2021/indonesia

[2] Reis, J. C., Correia, A., Murai, F., Veloso, A., & Benevenuto, F. (2019). Supervised learning for fake news detection. IEEE Intelligent Systems, 34(2), 76–81. https://doi.org/10.1109/mis.2019.2899143

[3] Ribeiro, F., Henrique, L., Benevenuto, F., Chakraborty, A., Kulshrestha, J., Babaei, M., & Gummadi, K. (2018). Media Bias Monitor: Quantifying Biases of Social Media News Outlets at Large-Scale. Proceedings of the International AAAI Conference on Web and Social Media, 12(1).

[4] Gelfert, A. (2018). Fake news: A definition. Informal Logic, 38(1), 84–117. https://doi.org/10.22329/il.v38i1.5068

[5] Ruchansky, N., Seo, S., & Liu, Y. (2017). CSI: A Hybrid Deep Model for Fake News Detection. Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. https://doi.org/10.1145/3132847.3132877

[6] Umer, M., Imtiaz, Z., Ullah, S., Mehmood, A., Choi, G. S., & On, B.-W. (2020). Fake news stance detection using Deep Learning Architecture (CNN-LSTM). IEEE Access, 8, 156695–156706. https://doi.org/10.1109/access.2020.3019735

[7] Pratiwi, I. Y., Asmara, R. A., & Rahutomo, F. (2017). Study of hoax news detection using naïve Bayes classifier in Indonesian language. 2017 11th International Conference on Information & Communication Technology and System (ICTS). https://doi.org/10.1109/icts.2017.8265649

[8] Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2018). Automatic Detection of Fake News. Proceedings of the 27th International Conference on Computational Linguistics, 3391–3401.

[9] Kennedy, S., Walsh, N., Sloka, K., McCarren, A., & Foster, J. (2019). Fact or factitious? contextualized opinion spam detection. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. https://doi.org/10.18653/v1/p19-2048

[10] Rubin, V., Conroy, N., Chen, Y., & Cornwell, S. (2016). Fake news or truth? using satirical cues to detect potentially misleading news. Proceedings of the Second Workshop on Computational Approaches to Deception Detection. https://doi.org/10.18653/v1/w16-0802

[11] Karimi, H., Roy, P., Saba-Sadiya, S., & Tang, J. (2018). Multi-Source Multi-Class Fake News Detection. Proceedings of the 27th International Conference on Computational Linguistics, 1546–1557.

[12] Tachhini, E., Ballarin, G., Della Vedova, M. L., Moret, S., & de Alfaro, L. (2017). Some Like it Hoax: Automated Fake News Detection in Social Networks. Proceedings of the Second Workshop on Data Science for Social Good (SoGood), 1960.

[13] Zuliarso, E., Anwar, M. T., Hadiono, K., & Chasanah, I. (2020). Detecting hoaxes in Indonesian news using

TF/TDM and K nearest neighbor. IOP Conference Series: Materials Science and Engineering, 835(1), 012036. https://doi.org/10.1088/1757-899x/835/1/012036

[14] Chen, B., Chen, B., Gao, D., Chen, Q., Huo, C., Meng, X., Ren, W., & Zhou, Y. (2021). Transformer-based language model fine-tuning methods for COVID-19 fake&nbsp;news detection. Combating Online Hostile Posts in Regional Languages during Emergency Situation, 83–92. https://doi.org/10.1007/978-3-030-73696-5_9

[15] Kaliyar, R. K., Goswami, A., & Narang, P. (2021). Fakebert: Fake news detection in social media with a Bert-based deep learning approach. Multimedia Tools and Applications, 80(8), 11765–11788. https://doi.org/10.1007/s11042-020-10183-2

[16] Wei, J., & Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). https://doi.org/10.18653/v1/d19-1670

[17] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. CoRR, abs/1810.04805. https://doi.org/https://doi.org/10.48550/arXiv.1810.04805

[18] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (n.d.). RoBERTa: A Robustly Optimized BERT Pretraining Approach, abs/1907.11692.

[19] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. CoRR, abs/1910.01108. https://doi.org/10.48550/ARXIV.1910.01108

[20] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. CoRR, abs/1706.03762.

[21] Ju, C., Bibaut, A., & van der Laan, M. (2018). The relative performance of ensemble methods with deep convolutional neural networks for Image Classification. Journal of Applied Statistics, 45(15), 2800–2818. https://doi.org/10.1080/02664763.2018.1441383

[22] Lynn, V., Balasubramanian, N., & Schwartz, H. A. (2020). Hierarchical modeling for user personality prediction: The role of message-level attention. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.472

[23] Christian, H., Suhartono, D., Chowanda, A., & Zamli, K. Z. (2021). Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging. Journal of Big Data, 8(1). https://doi.org/10.1186/s40537-021-00459-1

[24] Mrs. Leena Rathi. (2014). Ancient Vedic Multiplication Based Optimized High Speed Arithmetic Logic . International Journal of New Practices in Management and Engineering, 3(03), 01 - 06. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/29

[25] Ramasamy, J. ., Doshi, R. ., & Hiran, K. K. . (2023). Three Step Authentication of Brain Tumour Segmentation Using Hybrid Active Contour Model and Discrete Wavelet Transform. International Journal on Recent and Innovation Trends in Computing and Communication, 11(3s), 56–64. https://doi.org/10.17762/ijritcc.v11i3s.6155

[26] Prema, K. ., & J, V. . (2023). A Novel Marine Predators Optimization based Deep Neural Network for Quality and Shelf-Life Prediction of Shrimp. International Journal on Recent and Innovation Trends in Computing and Communication, 11(3s), 65–72. https://doi.org/10.17762/ijritcc.v11i3s.6156

[27] Mr. Vaishali Sarangpure. (2014). CUP and DISC OPTIC Segmentation Using Optimized Superpixel Classification for Glaucoma Screening. International Journal of New Practices in Management and Engineering, 3(03), 07 - 11. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/30