ISSN 2085-4552

COVID-19 Fake News Detection With Pre-trained Transformer Models

Bakti Amirul Jabar[1], Seline[2], Bintang[3], Cameron Jane Victoria[4], Rio Nur Arifin[5]
[1,2,3,4]Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia
[5]Data Scientist, Telkomsel, Jakarta, Indonesia
[1]bakti.jabar@binus.ac.id, [2]seline@binus.ac.id, [3]bintang001@binus.ac.id, [4]cameron.victoria@binus.ac.id, [5]rioarifin@gmail.com

Accepted 06 September 2022
Approved 06 October 2022

***Abstract*— COVID-19 is a new virus that first appeared in the year 2020 and is still currently plaguing our world. With the emergence of this virus, much information, both fake and real, has circulated in the internet. Fake information can lead to misleading information and cause a riot in society. In this paper, we aim to build a hoax detection system using the pre-trained transformer models BERT, RoBERTa, DeBERTa and Electra. From these four models, we will find which model gives the most accurate results. BERT gives a validation accuracy of 97.15% and test accuracy of 97.01%. RoBERTa gives a validation accuracy of 97.34% and test accuracy of 97.15%. DeBERTa gives a test accuracy of 97.48% and a test accuracy of 97.25%. Lastly, Electra gives a validation accuracy of 97.95% and a test accuracy of 97.76%. Electra is one of the newer models and is proven to be the most accurate model in our experiment and the one we will choose to implement fake news detection.**

***Index Terms*— deep learning; fake news; fake news detection; hoax; hoax detection; pre-trained transformer model; transformer model**.

I. Introduction

Since the start of the pandemic, millions of people have fallen victim and lost their lives. As of 17 June 2022, according to WHO, a total of 535,863,950 confirmed cases and 6,314,972 deaths of COVID-19 were reported [1]. In these two years, cases of COVID-19 have stagnated and surged continuously, contributing to many deaths. A significant factor that has also played a part in the surge is the existence of hoaxes or misinformation that has prevented individuals from following the correct health protocols or treatment for the virus [2].

Some of the most persistent hoaxes on the internet are about the COVID-19 vaccines. People claim that a person will die within three years of receiving the vaccine. Another claim is that the vaccines contain magnetic chips to track our location. They even create videos showing proof that metals stick to them after vaccination. Although these informations are most definitely false, a lack of, changing, or conflicting information also gives birth to these misinformation [2].

In modern natural language processing, pre-training has become the standard approach. This is done by pre-training the model on large amounts of unlabelled data followed by fine-tuning using small and specific data sets [4]. A research paper on machine learning implementations for hoax detection [5] has shown that pre-trained transformer models such as BERT and RoBERTa have outperformed traditional deep learning models and neural networks such as CNN (Convolutional Neural Network).

As such, this paper will focus on designing a COVID-19 hoax detection system by experimenting with four pre-trained deep learning models: BERT, RoBERTa, DeBERTa, and Electra. These four pre-trained models will be compared, and one of the models will be chosen for implementation based on their best validation accuracy and test accuracy.

II. Literature Review

Based on the research of Oberiri Destiny Apuke and Bahiyah Omar (2020, 18 October), there are four factors of news sharing in the COVID-19 pandemic [3]. The first factor is altruism. Altruism refers to the human activity of presenting something to others without expecting something in return. By sharing the news, people feel the satisfaction of successfully contributing to their social surroundings. Social media's primary function is to create and maintain social interactions between its users. Content sharing is considered to be a form of participation in building relationships. Therefore, socialization is considered the second factor.

A practice of making a perfect image on social media is called self-promotion. By sharing news they discover with others, they want to prove they are talented, capable and/or intelligent. To maintain this image, people develop the critical thinking of choosing which news is true and can be shared. The last factor is instant news sharing. People often share news on different platforms on social media without ensuring its factuality. This can be driven by fear and anxiety due to numerous news in the tense global conditions.

Three of the four factors mentioned above (altruism, socialization, and instant news sharing) are positively associated with fake news sharing. These factors are those that we cannot control; however, we can also use these attributes of society to our advantage. By propagating our fake news detector system, anyone can verify the news they hear on the internet and further spread the credibility to others.

A hoax can be classified into seven categories: satire or parody, misleading content, imposter content, fabricated content, false connection, fake news, and manipulated content. The top three types of the most widely spread hoax are 1). Manipulated content (30%), 2). Misleading content (28%), and 3). Fake news (18%) [6].

Fake news is low-quality news containing intentionally false information [3]. Health has become the most widely spread fake news topic (37% of the fake news takes up health as the topic) [6]. Fake news on health is very dangerous. It will lead people to improper ways of healthcare or/and will cause discredit to the medical world.

The COVID-19 pandemic has caused thousands of deaths in various parts of the world. At the same time, this pandemic has brought another disaster in our society, called the ‘infodemic’ or an abundance of false information circulating during the pandemic. The misinformation has proliferated widely on social media. This false information ranges from fake cures, conspiracies, or dangerous health advice. The large number of deaths caused by the virus is attributed to the virus itself and to the wrong or even late medical treatment due to hoaxes spreading online [7]. This is why we need to be able to distinguish between fake and factual information. However, we humans can be influenced by emotions such as fear, panic, and sadness, making rational decisions difficult, particularly during these difficult times. As such, by creating a hoax detection system, we can rely on it to find out whether particular information is credible without further questioning the reliability of choice.

Since the deep learning concept was created, researchers have designed multiple systems for fake news detection. At first, research on fake news detection mainly used traditional machine learning models such as Support Vector Machine, Naive Bayes, and Logistic Regression [8] [9] [10]. Later on, deep learning models such as Convolutional Neural Network and Long Short Term Memory Network were used [8] [9] [11] [12] [13] [14] [15]. Recently, pre-trained transformer models were developed and found to be more accurate than the other models. Pre-trained models were created to be more efficient than other deep learning models. They were also proven to be more accurate when trained using small datasets, making them suitable for fake news detection [16].

Therefore, we have concluded that most researchers utilize pre-trained transformer models for fake news detection. Although old neural network models such as CNN and RNN are still used, pre-trained transformer models have proven to be more efficient and accurate due to their improvements [17]. Transformer models started with the invention of BERT [13] [18] [19] [20] [21] [22] during 2018 [23], followed by its variations such as BART [24], ROBERTa [5] [18] [22] [25] [26], DeBERTa [18] [26] and Electra [21] [22] [27].

In this paper, we will conduct our experiment using BERT, RoBERTa, DeBERTa and Electra. After comparing the results, we will opt for the most accurate model.

III. Material and Method

A. Data Set

The data set consists of English tweets about COVID-19 taken in 2020, which have been checked in advance for their correctness. The dataset compiled has been taken from a popular dataset website, Kaggle. The dataset has been classified into two classes that are - real and fake. The total data consists of 10700 news items, 37050 words, and 5141 common words in fake and real news data. From the data collected, 52% are classified as real news and 48% as fake news. The data has been collected from 880 unique usernames. Table I contains 2 sample data from the COVID-19 news dataset.

TABLE I. Sample News Data

ID	Tweet	Label
1	Our daily update is published. States reported 734k tests, 39k new cases, and 532 deaths. Current hospitalizations fell below 30k for the first time since June 22. https://t.co/wzSYMe0Sht	Real
2	Alfalfa is the only cure for COVID-19.	fake

B. Methodology

1) Text Pre-Processing

Text pre-processing is a process in which text data is converted to be more structured so that the data can be used for analysis or prediction. In this paper, we use nltk, pre-processor, and tweet-preprocessing library from python. Basically, those libraries did data cleaning such as removing stopwords, stemming, lemmatization, remove punctuation, lower-casing, remove numbers, and over spaces or ticks.

2) Tokenization

Tokenization is an activity of splitting an entire text into small units, also known as tokens. In this paper, we will use the tokenization concept already pre-trained from each transformer model we will use.

3) Text Pre-Processing

The deep learning model used will be the transformer models BERT, DEBERTa, ROBERTa, and Electra, which have been pre-trained. The dataset will be grouped into 32 batches and ten epochs with 0.00002 learning rate.

- *BERT (Bidirectional Encoder Representations from Transformers)*

BERT is a transformer model designed to pre-train deep bidirectional representations from the dataset by reading the entire sequence of text at once. In this way, the model will be able to learn the context of the word based on the left and right context. Correspondingly, the model can be fine-tuned based on the dataset provided to be used for many different tasks, such as language translation, question answering, or in our case, fake news detection [28].

- *RoBERTa (Robustly Optimized BERT Pre-training Approach)*

RoBERTa is a pre-trained encoder model that was built on BERT's language masking strategy. This model further optimizes BERT's architecture to shorten the pre-training time. RoBERTa is implemented in PyTorch, in which the main hyperparameter of BERT is modified, BERT's next-sentence pre-training objective is deleted and training is done with bigger batches and longer sequences. This enables RoBERTa to focus more on the language masking strategy objective compared to BERT [29].

- *DeBERTa (Decoding-enhanced BERT with disentangled attention)*

DeBERTa is a transformer model that further optimises both the BERT and RoBERTa model using two techniques. First, the disentangled attention mechanism is used, in which the content and position of every word are converted and stored into two vectors, respectively. Next, the attention weights are calculated using disentangled matrices based on the acquired contents and relative position. The following technique incorporates absolute positions in the decoding layer using an enhanced mask decoder. This is done to predict the masked tokens during the pre-training process. Furthermore, to improve the models' generalization, a newer virtual adversarial training method is implemented for fine-tuning [30].

- *Electra*

Google's engineers develop Electra. Electra uses a different pre-training approach that takes advantage of BERT, but more effectively [31].

IV. Result and Discussion

As mentioned before, the models used will be BERT, DEBERTa, ROBERTa, and Electra. The final model that will be implemented will be the model with the best validation accuracy from the ten epochs. The model will then be run with the dataset and compared with the actual result.

A. *BERT (Bidirectional Encoder Representations from Transformers)*

Figure 1 and Figure 2 illustrate the training loss graph and the validation accuracy graph respectively, of the dataset using the BERT model. The training loss graph indicates how well the model fits with the training data, shown by each epoch. The validation accuracy graph indicated how accurate the model is when tested with the validation data, also shown by each epoch.

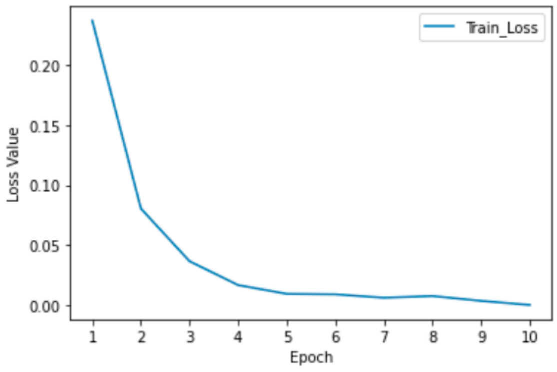

Fig. 1. Training Loss graph of BERT

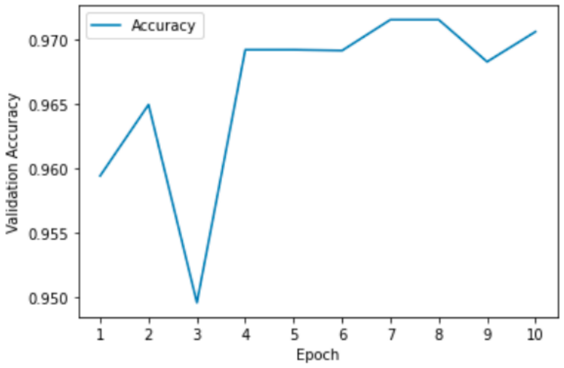

Fig. 2. Validation Accuracy graph of BERT

From Figure 2, we can infer that the best validation accuracy reaches a value of 97.15485074626866%. The test accuracy is then calculated by dividing the number of correct predictions by the total number of data, which results in 97.00826226012792%.

B. *RoBERTa (Robustly Optimized BERT Pre-training Approach)*

Figure 3 and Figure 4 define the training loss graph and the validation accuracy graph, respectively, of the dataset using the RoBERTa model. The training loss graph indicates how well the model fits with the training data, shown by each epoch. The validation accuracy graph indicated how accurate the model is when tested with the validation data, also shown by each epoch.

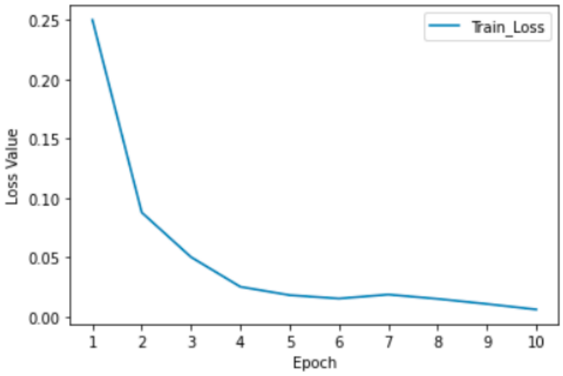

Fig. 3. Training Loss graph of RoBERTa

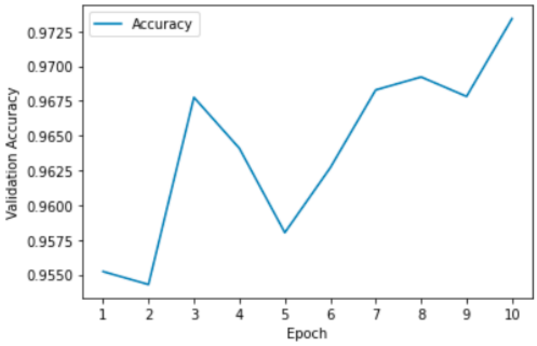

Fig. 4. Validation Accuracy graph of RoBERTa

From Figure 4, we can infer that the best validation accuracy reaches a value of 97.34141791044776%. The test accuracy is then calculated by dividing the number of correct predictions by the total number of data, which results in 97.15485074626866%.

C. *DeBERTa (Decoding-enhanced BERT with disentangled attention)*

Figure 5 and Figure 6 illustrate the training loss graph and the validation accuracy graph, respectively, of the dataset using the DeBERTa model. The training loss graph indicates how well the model fits with the training data, shown by each epoch. The validation accuracy graph indicated how accurate the model is when tested with the validation data, also shown by each epoch.

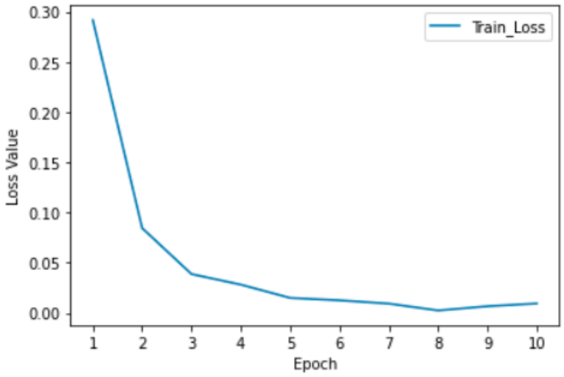

Fig. 5. Training Loss graph of DeBERTa

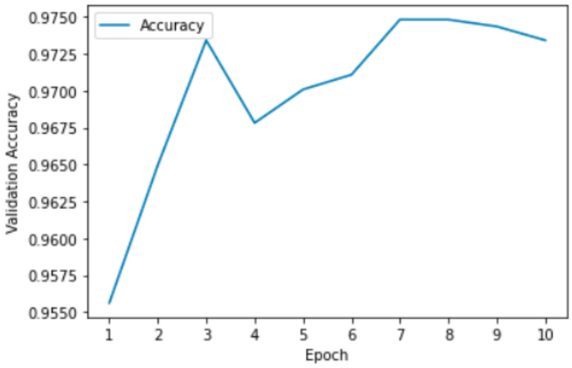

Fig. 6. Validation Accuracy graph of DeBERTa

From Figure 6, we can infer that the best validation accuracy reaches a value of 97.48134328358209%. The test accuracy is then calculated, by dividing the number of correct predictions by the total number of data, which results in 97.2481343283582%.

D. *Electra*

Figure 7 and Figure 8 illustrate the training loss graph and the validation accuracy graph respectively, of the dataset using the Electra model. The training loss graph indicates how well the model fits with the training data, shown by each epoch. The validation accuracy graph indicated how accurate the model is when tested with the validation data, also shown by each epoch.

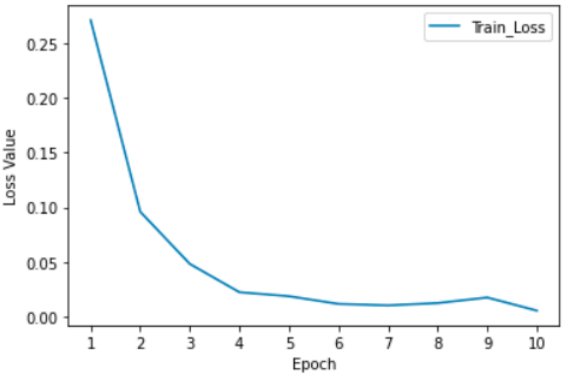

Fig. 7. Training Loss graph of Electra

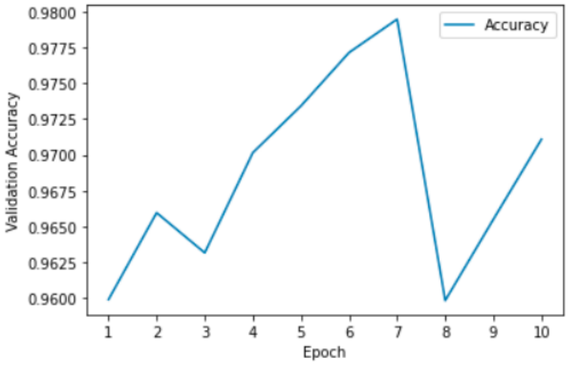

Fig. 8. Validation Accuracy graph of Electra

From Figure 8, we can infer that the best validation accuracy reaches a value of 97.94776119402985%. The test accuracy is then calculated by dividing the number of correct predictions by the total number of data, which results in 97.76119402985076%.

E. Discussion

TABLE II. VALIDATION ACCURACY AND TEST ACCURACY

Model Accuracy	Best Validation	Best Test Accuracy
BERT	97.15%	97.01%
DeBERTa	97.48%	97.25%
RoBERTa	97.34%	97.15%
Electra	97.95%	97.76%

Table II summarizes the validation accuracy and test accuracy results obtained from the experiments. From the table, we can conclude that Electra provides the best validation accuracy and test accuracy compared to BERT, DEBERTa, and ROBERTa. This is because Electra implements a pre-training task called Replaced Token Detection (RTD) which will train the model towards two outputs - real and fake. This task makes Electra a better, more suitable, and efficient model for implementing a fake news detection model. Although our experiments have shown that pre-trained transformer models are highly accurate, note that we have chosen to focus on a specific topic for hoax detection. Future researchers should aim to create a more general hoax detector system that can detect false information from any category or topic.

V. CONCLUSIONS

Of the four models, the best results on validation and test data were obtained using the Electra model with epoch 7. This model can still be improved by increasing the number of iterations on the epoch to get a more stable training loss, then validation is carried out at each epoch with cross validation. Future researchers should aim to create a more general hoax detector system that can detect false information from any category or topic.

REFERENCES

[1] “WHO Coronavirus (COVID-19) Dashboard”. Covid19.who.int. https://covid19.who.int (accessed Jun. 17, 2022).

[2] J. L. Ravelo. "'Hoax killed my father': Indonesia's other pandemic". www.devex.com. https://www.devex.com/news/hoax-killed-my-father-indonesia-s-other-pandemic-100488 (accessed Jun. 17, 2022).

[3] O. D. Apuke and B. Omar. “User motivation in fake news sharing during the COVID-19 pandemic: an application of the uses and gratification theory”, Online Information Review, vol. 45, no. 1, pp. 220-239, 2020, doi: 10.1108/OIR-03-2020-0116.

[4] K. S. Kalyan, A. R. and S. Sangeetha, “AMMUS : A Survey of Transformer-based Pretrained Models in Natural Language Processing”, arXiv:2108.05542v2 [cs.CL], Aug. 2021.

[5] S. Malla and P. Alphonse, "Fake or real news about COVID-19? Pretrained transformer model to detect potential misleading news", The European Physical Journal Special Topics, 2022, doi : 10.1140/epjs/s11734-022-00436-6.

[6] Fardiah and F. Darmawan, “Hoax Digital Literacy on Instagram”, Jurnal Komunikasi Ikatan Sarjana Komunikasi Indonesia, vol. 6, no. 2, pp. 171-186, 2021. doi : 10.25008/jkiski.v6i2.581

[7] S. van der Linden, J. Roozenbeek and J. Compton, "Inoculating Against Fake News About COVID-19", Frontiers in Psychology, vol. 11, 2020. doi : 10.3389/fpsyg.2020.566790

[8] S. Vijayaraghavan et al., “Fake News Detection with Different Models”, arXiv:2003.04978v1 [cs.CL], Feb. 2020.

[9] B. Nayoga, R. Adipradana, R. Suryadi, and D. Suhartono, "Hoax Analyzer for Indonesian News Using Deep Learning Models", Procedia Computer Science, vol. 179, pp. 704-712, 2021. doi : 10.1016/j.procs.2021.01.059

[10] A. Kumar, S. Singh and G. Kaur, "Analyzing And Detecting The Fake News Using Machine Learning", International Journal of Computer Sciences and Engineering, vol. 7, no. 5, pp. 1044-1050, 2019. doi : 10.26438/ijcse/v7i5.10441050

[11] J. Nasir, O. Khan and I. Varlamis, "Fake news detection: A hybrid CNN-RNN based deep learning approach", International Journal of Information Management Data Insights, vol. 1, no. 1, pp. 100007, Apr. 2021. doi : 10.1016/j.jjimei.2020.100007

[12] A. Thota, P. Tilak, S. Ahluwalia, and N. Lohia, "Fake News Detection: A Deep Learning Approach," SMU Data Science Review, vol. 1, no. 3, 2018, Art. no. 10.

[13] Á. I. Rodríguez and L. L. Iglesias, "Fake News Detection Using Deep Learning", arXiv:1910.03496v2 [cs.CL], Sep. 2019.

[14] Y. Yang, et al.,“TI-CNN: Convolutional Neural Networks for Fake News Detection”, arXiv:1806.00749v1 [cs.CL], Jun. 2018.

[15] J. Tembhurne, M. Almin and T. Diwan, "Mc-DNN: Fake News Detection Using Multi-Channel Deep Neural Networks", International Journal on Semantic Web and Information Systems, vol. 18, no. 1, pp. 1-20, 2022. doi : 10.4018/ijswis.295553

[16] J. Khan, M. Khondaker, S. Afroz, G. Uddin and A. Iqbal, "A benchmark study of machine learning models for online fake news detection", Machine Learning with Applications, vol. 4, pp. 100032, 2021. doi : 10.1016/j.mlwa.2021.100032

[17] D. Fox, "4 Reasons Transformer Models are Optimal for NLP". https://www.eweek.com/big-data-and-analytics/reasons-transformer-models-are-optimal-for-handling-nlp-problems/#:~:text=Transformer%20models%20are%20excellent%20at,by%20way%20of%20the%20decoder (Accessed: Jun. 19, 2022).

[18] S.M.S. Shifath., M. Faiyaz Khan and S.Md. Islam. “A transformer based approach for fighting COVID-19 fake news”, arXiv:2101.12027v1 [cs.CL], Jan. 2021.

[19] H. Jwa, D. Oh, K. Park, J. Kang and H. Lim, "exBAKE: Automatic Fake News Detection Model Based on Bidirectional Encoder Representations from Transformers (BERT)", Applied Sciences, vol. 9, no. 19, pp. 4062, 2019. doi : 10.3390/app9194062

[20] Y. Wang, Y. Zhang, X. Li and X. Yu, “COVID-19 Fake News Detection Using Bidirectional Encoder Representations from Transformers Based Models”, arXiv:2109.14816v2 [cs.CL], Sep. 2021.

[21] P. Singh, R. Srivastava, K.P.S.Rana and V. Kumar, “SEMI-FND: Stacked Ensemble Based Multimodal Inference For Faster Fake News Detection”, arXiv:2205.08159v1 [cs.CL], May. 2022.

[22] E. Shushkevich, M. Alexandrov and J. Cardiff, "Covid-19 Fake News Detection: A Survey", Computación y Sistemas, vol. 25, no. 4, Dec. 2021. doi : 10.13053/cys-25-4-4089.

[23] A. Chernyavskiy, D. Ilvovsky, and P. Nakov, “Transformers:’The End of History’ for Natural Language

Processing?”, in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2021, pp. 677-693. doi: 10.1007/978-3-030-86523-8_41

[24] K. Huang, K. McKeown, P. Nakov, Y. Choi, and H. Ji, “Faking Fake News for Real Fake News Detection: Propaganda-loaded Training Data Generation”, arXiv:2203.05386v1 [cs.CL], Mar. 2022.

[25] A. Pritzkau, “Multi-class fake news detection of news articles and domain identification with RoBERTa - a baseline model”, in CEUR Workshop Proc., Sept. 21-24, 2021. [Online]. Available: http://ceur-ws.org/Vol-2936/paper-46.pdf

[26] S. D. Das, A. Basak, and S. Dutta, "A Heuristic-driven Uncertainty based Ensemble Framework for Fake News Detection in Tweets and News Articles", arXiv:2104.01791v2 [cs.CL], Dec. 2021.

[27] K. A. Das, A. Baruah , F. A. Barbhuiya, and K. Dey, “Ensemble of ELECTRA for Profiling Fake News Spreaders”, in CEUR Workshop Proc., Sept. 22-25, 2020. [Online]. Available: http://ceur-ws.org/Vol-2696/paper_193.pdf

[28] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding”, arXiv:1810.04805v2 [cs.CL], May. 2019.

[29] Y. Lie, et al., “RoBERTa: A Robustly Optimized BERT Pretraining Approach”, arXiv:1907.11692v1 [cs.CL], Jul. 2019.

[30] P. He, X. Liu, J. Gao, and W. Chen, “DeBERTa: Decoding-Enhanced BERT With Disentangled Attention”, arXiv:2006.03654v6 [cs.CL], Oct. 2021.

[31] K. Clark and T. Luong. "More Efficient NLP Model Pre-training with ELECTRA". Google AI Blog. https://ai.googleblog.com/2020/03/more-efficient-nlp-model-pre-training.html (accessed Jun. 19, 2022).