Can Deep Learning detect fake news better when adding context features?

Rachel Ladouceur
Computer Science and Mathematics
Quebec University of Chicoutimi
Saguenay, Canada
rachel.ladouceur1@uqac.ca

Abubakr Hassan
Computer Science and Software Engineering
Laval University
Quebec, Canada
abhaa1@ulaval.ca

Mohamed Mejri
Computer Science and Software Engineering
Laval University
Quebec, Canada
mohamed.mejri@ift.ulaval.ca

Fehmi Jaafar
Computer Science and Mathematics
Quebec University of Chicoutimi
Saguenay, Canada
fehmi.jaafar@uqac.ca

***Abstract*—Fake news is increasing on social media and has huge negative impacts on society. To detect fake news, different approaches and algorithms have been studied. Recently, the use of deep learning algorithms has performed very well for fake news detection. Many researchers found the importance to add the context features to have better and a rapid detection. The current paper presents two algorithms, BERT and LSTM, for the classification of fake news, by using text and context features. Moreover, we add XAI solution that determines which context features have contributed the most to the classification of fake news. We have trained our model with an extraction of 5,000 news from the COVID-19 dataset. Our result shows that BERT performs better by 2.37% than LSTM by analyzing only the content and performs better by 2.24% by analyzing both, the content and the context. Finally, SHAP is used to explain fake news by highlighting the most relevant attributes that have helped to the classification of news.**

***Index Terms*—Fake news, deep learning, network analysis, XAI**

I. Introduction

We are observing a huge increase of large-scale online fake news activities. For example, since the 2016 US presidential election, several researchers are reporting that fake news spreads faster, broader, and deeper [1]. Indeed, more users share, like, and comment on fake news than real news.

According to the European Union Agency for Cybersecurity's 2022 report, fake news is cited as the 7th top ten global cyber threats [1]. Fake news is defined as a news article that is intentionally and verifiable false and could mislead readers [2]. It is increasingly common for bad actors to seek to manipulate public debates through social media. In addition, according to NewsGuard, 20 % of videos shared on TikTok contained fake news in 2022 and 40 % of fake news topics are spread through Facebook [2]. Even though some measures are in place, this doesn't seem to counteract fake news.

Moreover, we are observing several impacts of fake news in nowadays. For example, a Canadian study indicated that the belief that COVID-19 was ”a hoax or an exaggeration” has led to that 2.35 million Canadians delayed or refused their vaccination between March and November 2021, resulting in a 22 % increase in infections (200,000) and hospitalizations by 28 % (13,000) and a 35 % number of deaths (2,300) in Canada during those 9 months alone. It also added 300 million, or 40 % more than total hospitalization [3].

Some algorithms have been studied to detect automatically and rapidly fake news. Recent deep learning algorithms, such as BERT, have a high rate of success in detecting content analysis [3]. However, based on some studies [1] [4], context features have to be added for better detection of fake news. To the best of our knowledge, we are presenting in this paper, the latest studies that employed textual and contextual analysis to detect fake news.

Moreover, it's important to explain to the users why the news is fake. The user must know when he is in front of fake news to avoid sharing it and spreading it. Since fake news is written to sow doubt in the minds of users, they have difficulty identifying the truth from the false [1]. Under several jurisdiction such as the Canadian Privacy Act and the European General Data Protection Regulation (GDPR), individuals have a ”right to explanation” for automated decisions made about them. Explainable AI, called ”XAI”, is essential to allow users to understand and trust the outcome. Indeed, only 50 % say they trust companies that use AI[4]. Some studies have examined fake news detection with an XAI solutions to explain the classification [5]. In this paper, we focus on which attribute contribute the most to this classification.

[1] https://www.enisa.europa.eu/publications/enisa-threat-landscape-2022

[2] https://www.meta-media.fr/2023/05/26/une-annee-de-desinformation-des-strategies-de-communication-tres-etablies.html. (Access June, 16th)

[3] https://www.ledevoir.com/economie/779860/coronavirus-le-cout-economique-de-la-propagation-de-fausses-informations? (access April, 23rd)

[4] https://www.mitre.org/news-insights/news-release/public-trust-ai-technology-declines-amid-release-consumer-ai-tools (access June 16 th.)

The proposed approach addresses the following two main research questions:

(Q1) Can deep learning algorithms detect fake news better when adding context features ?

(Q2) Which contextual features are the most relevant to detect fake news ?

In this paper, we are reporting the results of an empirical study that we conducted using the COVID-19 dataset. We found that BERT performs better by 2.37% than LSTM by analyzing only the content and performs better by 2.24% by analyzing both, the content and the context. Finally, SHAP is used to explain fake news by highlighting the most relevant attributes that have helped to the classification of news.

The remainder of this paper is organized as follows. Section II provides background information. Section III presents the design of our empirical study. Section IV presents the results. Section V introduces threats to the validity of results. Section VI reports an overview of the related works. Finally, Section VII concludes the paper and sets our perspectives.

II. Background

This section gives background information related to the different detection approaches, the algorithms methods used for that purpose and then, the XAI solutions.

A. Classification of the detection approaches

We can classify the way to detect fake news as follows:

1) Content-based approach: This approach aims to detect fake news by analyzing the content of the initial message, the title of this message or the comments of subsequent messages.

A general challenge for the content-based approach is that the linguistic style and keywords related to the topic are constantly changing. Models that are trained on a single dataset may perform poorly on a new set of data with different content. Furthermore, the linguistic features used are mostly language-specific, so their generality is limited. Another challenge added to this approach is that it requires a large amount of training data to detect fake news. By the time these methods collect enough data, fake news has spread widely on social media [6].

2) Context-based approach: The context-based method analyzes the user profile and the network.

The analysis of the user profile includes several characteristics such as the number of messages sent per day and the number of contacts or “followers” connected to the user. So, if the user is active and has a lot of followers, he can be a propaganda agent.

The analysis of the network focus on the speed of propagation and distribution of the message between the users involved as well as the interaction between them. Indeed, the publication of a message can generate a multitude of other messages or comments associated with the first message, this is what we call the cascade of messages.

The major challenge for the context-based approach is to have rapid detection while the network features are available only after a period of time.

B. Algorithms methods

To treat the text, natural language processing (NLP) is a branch of artificial intelligence (AI) that allows computers to understand, generate and manipulate human language. Recently, an important breakthrough in the field of NLP is the BERT algorithm ”Bidirectional Encoder Representations from Transformers”, developed by Google in 2018. BERT is based on the Transformers architecture, which uses a stack of encoder layers, which increases its ability to capture long-distance dependencies between words. In addition, the model takes into account relationships between words, while calculating attention weights for each pair of words in a sequence. This allows the model to determine the importance of words with each other. Therefore, BERT offers the best results for processing text [5].

Other deep learning algorithms such as CNN (convolutional neural network), RNN (recurrent neural network), and LSTM (Long Short-Term Memory) are the most used in the field of fake news. The learning capabilities of deep learning algorithms are generally more powerful [7].

LSTM is derived from the Recurrent Neural Networks (RNN) that can process data sequences. LSTM is well-suited for processing sequences of data with long-range dependencies due to the multitude of gates which allows some relevant information to pass through the sequence. They can capture information from earlier time steps and remember it for a more extended period.

GCN stands for ”Graph Convolutional Network”. In the context of fake news detection, the nodes in the graph could represent various entities such as news articles, users, etc.

C. Explainable Artificial Intelligence

Explainable Artificial Intelligence (XAI) is a research field that studies how AI decisions and data driving those decisions can be explained to people to provide transparency, enable assessment of accountability, demonstrate fairness, or facilitate understanding [6].

Explanation can be done for a single prediction (local prediction) or for the entire model’s prediction, regardless of the input (global explanation). Various open-source solutions are available like LIME, SHAP, Deep LIFT, LRP, etc.

Local Interpretable Model-Agnostic Explanations (LIME), is an algorithm that can explain the predictions of any classifier. LIME transforms any complex model like a neural network into a simple linear model. It’s based on individual predictions and it works faster than SHAP. LIME can highlight specific words and their weights that are relevant in the choice of the classification. It sums up the probability of the outcome.

SHapley Additive exPlanations (SHAP) is very similar to LIME. It’s based on individual predictions. It can be used to explain more complex like deep learning and various NLP such as sentiment analysis.

[5] https://datascientest.com/bert-un-outil-de-traitement-du-langage-innovant (access June 16 th.)

[6] https://www.uxai.design/explainable-ai (access July 18 th.

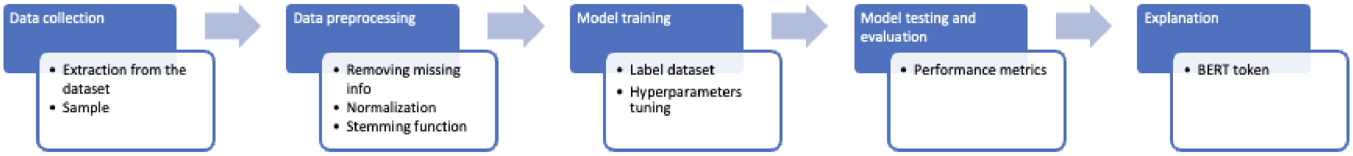

Fig. 1. Overview of the proposed approach

TABLE I
FEATURES FROM COVID-19 DATASET

Name	Name of the publisher
User location	User address or city
User description	User profile
User created	Date that account has been created
Numbers of followers	Number of persons you'd like to be informed
Numbers of friends	Number of user's contacts.
Numbers of favorites	Tweets with "like" marked
User verified?	Account verified by Twitter to attest credibility
Date of publication	Date of the news has been published
Text with URL link	Message with a link to an external website
Hashtags	Pound sign to identify posts on specific topic
Source	COVID-19 dataset from the Twitter platform
Is retweet false?	Yes or no

LRP (Layer-wise Relevance Propagation) is also an interpretability technique used in the field of machine learning. This model could use various architectures, such as CNN, RNN or BERT. A post-LRP analysis could show that the model focuses on sensationalist words or unreliable sources.

HAN (Hierarchical Attention Network) is a neural network architecture that was originally developed for the classification of hierarchical textual documents. Once the model is trained, it can be used to explain false news. Using the attention mechanisms built into the model, one can identify the parts of the text (sentences and words) that contributed the most to the prediction of the model.

III. DESIGN OF OUR APPROACH

Figure 1 shows the process of our approach which includes the BERT performance for fake news detection and SHAP solution for the explanation.

Datacollection. We used the dataset of COVID-19 which contains text and context features from the Twitter database, as follows. From that list, we took the first 11,000 news.

Data preprocessing. We removed hyperlinks, by the module "re" of Python. And also, we eliminate the stop words, like "the", usually irrelevant for the comprehension of the text, and keep the only meaning words. With the stemming function, words have been reduces to their roots. After cleaning empty cells, 5,000 news were left.

Model Training. The model is trained on a labeled dataset containing both true and fake news. The model learns to assign weights to different parts of the data (content and context features) based on their importance for the correct classification of news. We used the original BERT base model with 12 layers, 12 attention heads and a hidden size of 768. The maximum length of input sequences that BERT can process is typically set to 512 tokens. We used Google Colab to implement the Keras Python package from TensorFlow.

Model Testing and Evaluation. The dataset was split between training and testing of 80-20. To evaluate the performance of each algorithm, we utilized the following metrics presented in Table I.

TABLE II
MEASURE AND DEFINITION

Measure	**Definition**
Accuracy	Percentage of test set tuples correctly classified.
Recall	Percentage of positive tuples are labeled as such.
Precision	Percentage of tuples labeled as positive.
F-score	Combination of the precision and the recall.

Explanation. The proposed approach uses SHAP to explain the BERT result. To adapt SHAP, different steps have been followed. for example, the input text has used the same tokenizer that was used to preprocess the data for BERT. After, the model encoded the tokenized text into numerical format that BERT understands. The explanation process is done by extracting the most important attributes that contribute to the output result.

IV. RESULTS

As mentioned, the proposed approach addresses the following two research questions:

(Q1) Can deep learning algorithms detect fake news better when adding context features?

Motivation: The main purpose of this research question is the use of an recent advanced algorithm to analyze the content and the context of fake news and see if we obtain a higher performance. BERT has excellent results for detecting fake news using text messages. When we have reviewed the studies on fake news that have used deep learning for the detection [8] [9], we realized that BERT was used for the analysis of the content, and when context features were added, the authors used another algorithm.

Results : Based on Table II, we have compared the results of BERT and LSTM with different metrics. Globally, we have obtained a better result with BERT compared to LSTM, with the context and without the context.

TABLE III
BERT AND LSTM RESULTS

Classifier	**Context**	**Accuracy**	**Precision**	**F-score**	**Recall**
BERT	No Context	97.16	97.02	97.01	96.88
LSTM	No Context	94.79	94.30	94.10	93.30
BERT	Context	99.84	99.00	99.00	99.00
LSTM	Context	97.60	97.00	97.00	97.00

Effectively, without the context, BERT obtained an accuracy of 97.16% compared to LSTM with 94.79%. Our result shows that BERT performs better by 2.37% for fake news detection by analyzing the content. With the context, BERT obtained an accuracy of 99.84% compared to LSTM with 97.60%. Our result shows that BERT performs better by 2.24% for fake news detection by analyzing the context.

Indeed, we noted that the features context helps to detect fake news by 99.84% compared to 97.16% which is an increase of 2.68%.

Moreover, the precision and other metrics are also higher with BERT algorithm. Indeed, like other studies, BERT perform better than other algorithm [6].

Discussion : BERT has achieved remarkable performance due to several key factors:

- Pre-training on Large Corpus: BERT is pre-trained on a massive amount of text data which allows BERT to learn rich, contextual representations of words and phrases.
- Transformer Architecture: Highly effective in capturing long-range dependencies in sequences.
- Masked Language Model Pre-training Objective: During pre-training, BERT is trained on a masked language model (MLM) objective, where a certain percentage of words in each input sequence are randomly masked. This forces BERT to learn deep contextual representations.
- Bidirectional Context: Unlike previous models that process text in one direction (e.g., left-to-right or right-to-left), BERT utilizes a bidirectional approach which considers both left and right context.

BERT has an advantage over the classics algorithms. BERT demonstrates strong performance in text classification tasks such as sentiment analysis. Indeed, despite the used of different variations of BERT, it was the most successful algorithm in the competition regarding the detection of fake news related to COVID-19 Moreover, BERT is designed to capture contextual features from text [9].

(Q2) Which context features are the most relevant for fake news detection ?

Motivation: Our motivation is to find which attributes are the most relevant for fake news detection. This explanation can help understand why the model made a particular decision and can provide valuable information to assess the credibility of the news. As already mentioned, in several jurisdictions such as the Canadian Privacy Act and the European General Data Protection Regulation (GDPR), individuals have a ”right to explanation” for automated decisions made about them. While regulatory compliance is crucial for users, other aspects have to be considered such as error detection and debugging, trust and transparency and education.

Result: SHAP obtained a better result than LIME. Figure 2 summarizes the SHAP result which shows the most relevant context attributes that have helped to detect the news. Those two first attributes are ”user_followers” and ”hashtag”. Regarding our result, the ”user_followers” helps to detect at 25% the fake news, while the ”hashtag” helps to detect it at 20%. Both explain almost half of the fake news (45%).

Other attributes less important are user location, user friends, user verified and data text. Those attributes help to detect at 10% the fake news.

The least important attributes are user favourite, user-created and user description.

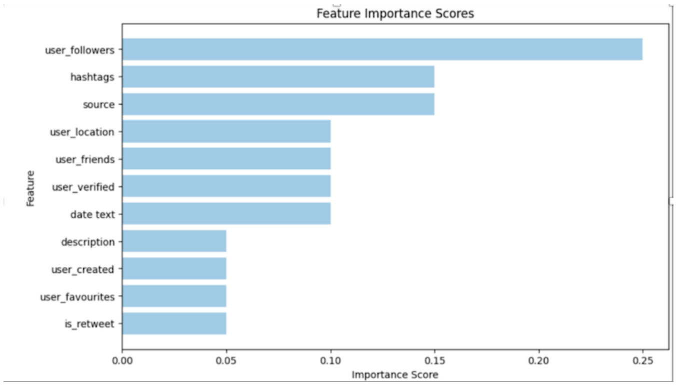

Fig. 2. SHAP result

Discussion: We explain reasons about why some attributes help to detect fake news such as users followers, hashtag, user location, user friends and user verified. Table 1 includes a short description of each attribute. User friends and user followers can be confusing terms. While they both refer to connections between users, they typically represent different types of relationships and interactions. Becoming friends usually requires mutual confirmation followers” typically refer to users who subscribe to another user’s updates without requiring mutual confirmation. It’s a one-way relationship.

There are many reasons why the numbers of followers are important to consider to detect fake news. We already mentioned that the more the publisher has followers, the more fake news can spread. There is tricky way that users can do to increase their number of followers. One technique is to create fake accounts that can be an automated tool behind completely created characters. The advantage of using automated tools is the fastest of spreading messages on social media. However, most users that spread fake news (around 78%) are still more likely to be humans than bots [1]. Moreover, research indicates that repeating the same fake news ideas increases the probability that humans will believe it [7].

Hashtags can also be used to detect fake news. A hashtag is a word or phrase preceded by the pound sign (#) used on social media to identify posts on a specific topic. When a user creates a post that contains a hashtag, that message becomes associated with all other posts that contain the same hashtag. Bad actors can use popular hashtags to promote their fake news, taking advantage of the audience already engaged around this topic. Moreover, hashtags sent by trolls spread fake news exponentially [10]. The study of Zhou et al. [11] observed that hashtags related to COVID-19 conspiracy theories are mainly involved in the spread of fake news.

[7] https://counterhate.com/wp-content/uploads/2023/03/Toxic-Twitter-II-Final-Report.pdf (access July, 23rd)

Another interesting attribute in the case of fake news is the user location. User can include the city they live for different purposes such as connecting with others in your area who have similar interests. Sharing your location can provide context for your tweets. Indeed, fake news often contains information about events or incidents in specific locations. Understanding where a user is located can help in assessing the relevance and potential impact of the news they are sharing.

Studying the user friend feature, where fake news can spread rapidly, is important for several reasons. Users often rely on their social networks to source information called ”echo chamber” that aligns with their existing beliefs and opinions. Researching the friend feature can help uncover patterns of information propagation. Vosoughi et al. [1] stated in 2018 that fake news spreads faster, by more people and deeper than real news.

V. Threats to validity

In this section, we discuss about some threats that limit the validity of our paper.

Construct validity: Construct validity concerns the extent to which your test or measure accurately assesses what it's supposed to. We used a classified dataset that already categorized news between true and fake news. We have assumed that this classification is accurate.

Internal Validity: Internal validity refers to the extent to which a study accurately establishes a causal relationship between the variables under investigation, without the influence of confounding variables. In the context of COVID-19 dataset, internal validity would ensure that any observed effects of fake news exposure on COVID-19 are indeed due to this factor and not to other one.

Reliability validity: Reliability validity threats concern the possibility of replicating this study. The COVID-19 dataset is a huge database which contains 1 million tuples. We extract only 11,000 tuples. It means that we only have a sample of 0.11%. Since this sample is very small, we supposed that it represents the whole dataset. Moreover, after the pre-processing of our data collection, 5,000 news left. Our sample is very small and our results depend on this data collection.

External validity: External validity threats refer to the extent to which the results of a study can be generalized or applied to other subject. Indeed, it could be possible that the result would be different with another topic.

VI. Related work

In this section, we reviewed studies regarding fake news detection with deep learning algorithms and XAI solutions. We have summarized in Tables III and IV those studies along with their metrics related to each of them.

The study of Kula et al. [9] was based on the content approach using the titles and the messages. They used the dataset of ISOT which contains 44,898 labelling news. The authors experimented various transformers architecture, including BERT and RoBERTa, to classify the information with an accuracy of 98.8%. The RoBERTa differs from the standard BERT by the improved pre-training process, which includes bigger batches over more data. It was developed by Facebook AI research.

TABLE IV
Deep learning algorithms and their metrics

Authors	**Accuracy**	**Algorithm**	**Approach**
Kula et al.[9]	98.80	BERT	Content
Shu et al.[12]	86.1	HPFN	Context
Shu et al.[13]	72.30	CNN	Context
Kaliyar et al.[8]	98.90	BERT-CNN	Content
Dou et al.[14]	97.23	BERT-GCN	Context

Shu et al. [12] analyzed the network's propaganda and they split the attributes into three categories: structural, temporal and linguistic. They analyzed the importance of the features in different levels, macro and micro, to understand how each type of features contributes to the detection performance. Among structural features extracted from both dataset, Gossipcop and Politifact, the fraction of cascades with retweets, the number of people engaged in the initial message and the number of comments (posts, likes, shares) committed to the initial message are those determinant features to detect. They have used HPFN algorithm with two different dataset: Politifact and Gossipcop with an accuracy of 84.3% and 86.1% respectively.

Shu et al. [13] also published in 2020 another study which had a deeper look into the user profiles and assessed the social bots' impacts. They observed that there is a sudden increase in the number of retweets for fake news and it does remain constant beyond a short time. They have used their own built dataset named FakeNewNet, which contains news content, social context and spatiotemporal information for a total of 5,778 labelling news. They also observed that fake news tends to receive fewer replies than real news. They have used different machine learning (ML) algorithms such as CNN. However, the accuracy metrics were not so good: 72.3% with the Gossipcop dataset.

The study of Kaliyar et al. [8] was based on the content approach using the titles and the messages. They used the dataset of Kaggle which contains 20,800 labelling news. The BERT algorithm was used to encode the text and was subsequently paired with the LSTM algorithm to classify the information with a rate of accuracy of 97.55%. They also merged BERT with CNN and obtained a success rate of 98.9%.

The study of Dou et al. [14] was based on content (comment analysis) and context (number of retweets, number of friends, etc.). They used their own dataset called FakeNewsNet that contained dataset from two fact-checking websites (Politifact-314 news and Gossipcop - 5,464 news). The BERT algorithm encoded the comment text and was subsequently paired with the GCN (Graph Convolution Network) algorithm with an accuracy of 97.23%. The study concluded that merging different algorithmic methods, such as those dealing with text and context, can increase the detection of fake news.

A. XAI

Shu et al. [15] proposed a DeFEND ”Explainable Fake News Detection” solution. The model examines the users’ comments and the textual message. The users’ comments include texts, hashtags, and emojis which are significant as it helps to understand the context of the comments. The model learned the sentence representation via a recurrent neural network (RNN) based word encoder. Depending on the dataset, the best accuracy was 90.4%. They choose the hierarchical attention network (HAN) for news sentence explainability, and HPA-BLSTM as the baselines for users’ comments explainability. DeFEND doesn’t indicate the probability of the news being fake. Another concern is the missing evaluation of the credibility of the comments users.

Mohseni et al. [16] have trained a BiLSTM network with Word2Vec word embedding. They obtained an accuracy of 73.65%. They also used the hierarchical attention network (HAN) to explain the sentences and added different visual elements. The major issue of this study is their dataset is quite small and it includes information related to politics. The model has only analyzed the textual content.

Szczepanski et al. [17] used BERT with the content-based approach. BERT has been trained on 4,652 data and has an accuracy of 98%. They added XAI, LIME and Anchors. The LIME explainability module was configured to present the five most relevant words with their probability. According to their result, keywords found by LIME are not relevant to explain the falsity of news. Anchors acts like LIME: it highlights the keywords without the probability of that relevance.

Chien et al. [18] proposed XFlag, which used LSTM with a layer-wise relevance propagation (LRP) to explain the fake news and situation awareness-based agent transparency (SAT) to represent and explain the LRP-retrieved factors. LRP has been extended to LSTM and confirmed to perform effectively [19]. LRP calculates which input features were most relevant between the content, the user and the sentiment. The proposed model is been trained on the Weibo dataset which contains 4,664 labeled news with an accuracy of 93.70%.

TABLE V
XAI SOLUTIONS AND THEIR METRICS

Authors	**Accuracy%**	**F-score%**
Shu et al. [15]	90.40	92.80
Mohseni et al. [16]	73.65	
Szczepanski et al. [17]	98.00	98.00
Chien et al. [18]	93.70	93.80

VII. CONCLUSION AND FUTURE WORKS

In conclusion, we have the same result as the study of Shu et al. in 2019 [4]: adding context helps to detect and classify fake news with a better result. Indeed, our BERT result is better by 2.61% (99.84% vs 97.23%) than the study of Dou et al. [14].

There is substantial empirical evidence that fake news spreads differently and faster than real news forming automatically detectable diffusion patterns [16]. Fake news tends to show multiple, periodic spikes rather than a single spike as usually occurs for real news [17]. There are the reasons why temporal features are important to consider in the early stage of fake news detection [12]. Thus, in future work, it would be interesting to design an approach that will include time series features to the other approaches.

REFERENCES

[1] S. Vosoughi, D. Roy, and S. Aral, ”The spread of true and false news online,” *Science*, vol. 359, no. 6380, pp. 1146-1151, 2018.

[2] K. Shu and H. Liu, ”Detecting fake news on social media,” *Springer Nature*, 2022.

[3] J. Alghamdi, Y. Lin, and S. Luo. ”Modeling Fake News Detection Using BERT-CNN-BiLSTM Architecture,” In *2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 354-357. IEEE.

[4] K. Shu, S. Wang, and H. Liu. ”Beyond news contents: The role of social context for fake news detection,” In *Proceedings of the twelfth ACM international conference on web search and data mining*, 2019,pp. 312-320.

[5] G. Srivastava, R. H. Jhaveri, S. Bhattacharya, S. Pandya, P. Maddikunta, P. K. R., G. Yenduri, and T. R. Gadekallu, ”XAI for cybersecurity: state of the art, challenges, open issues and future directions,” *arXiv preprint arXiv:2206.03585*, 2022.

[6] S. Raza, and C. Ding. ”Fake news detection based on news content and social contexts: a transformer-based approach,” *International Journal of Data Science and Analytics* vol. 13, no. 4, pp. 335-362, 2022.

[7] E. Aïmeur, S. Amri, and G. Brassard. ”Fake news, disinformation and misinformation in social media: a review,” *Social Network Analysis and Mining*, vol. 13, no. 1, p. 30, 2023.

[8] R. K. Kaliyar, A. Goswami, and P. Narang. ”FakeBERT: Fake news detection in social media with a BERT-based deep learning approach,” *Multimedia tools and applications*, vol. 80, no. 8, pp. 11765-11788. 2021.

[9] S. Kula, R. Kozik, and M. Choraś. ”Implementation of the BERT-derived architectures to tackle disinformation challenges,” *Neural Computing and Applications*, pp. 1-13.2022.

[10] M. Alizadeh, J. N. Shapiro, C. Buntain, and J. A. Tucker, ”Content-based features predict social media influence operations,” *Sci. adv.*, vol. 6, no.30, eabb5824, 2020.

[11] X. Zhou, R. Zafarani, and E. Ferrara. ”From Fake News to# FakeNews: Mining Direct and Indirect Relationships among Hashtags for Fake News Detection,” *arXiv preprint arXiv:2211.11113*, 2022.

[12] K. Shu, D. Mahudeswaran, S. Wang, and H. Liu. ”Hierarchical propagation networks for fake news detection: Investigation and exploitation,” In *Proceedings of the international AAAI conference on web and social media*, Vol. 14, pp. 626-637, 2020.

[13] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu. ”Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media,” Big data, vol. 8, no. 3, pp. 171-188, 2020.

[14] Y. Dou, K. Shu, C. Xia, P. S. Yu, and L. Sun. ”User preference-aware fake news detection,” In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pp. 2051-2055, 2021.

[15] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu. ”Defend: Explainable fake news detection,” In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 395-405, 2019.

[16] S. Mohseni, F. Yang, S. Pentyala, M. Du, Y. Liu, N. Lupfer, X. Ji, and E. Ragan. ”Machine learning explanations to prevent overtrust in fake news detection,” In *Proceedings of the international AAAI conference on web and social media*, vol. 15, pp. 421-431, 2021.

[17] M. Szczepanski, M. Pawlicki, R. Kozik, and M. Choraś. ”New explainability method for BERT-based model in fake news detection,” *Scientific reports*, vol. 11, no. 1, p. 23705, 2021.

[18] S. Y. Chien, C. J. Yang, and F. Yu. ”XFlag: Explainable fake news detection model on social media,” *Int. J. Hum.-Comput. Interact.*, vol. 38, no.18-20, pp. 1808-1827, 2022.

[19] L. Arras, J. Arjona-Medina, M. Widrich, G. Montavon, M. Gillhofer, K. R. Müller, S. Hochreiter, and W. Samek. ”Explaining and interpreting LSTMs,” *Explainable ai: Interpreting, explaining and visualizing deep learning*, 211-238, 2019, *Springer*.