A Robust Hybrid Approach for Enhancing Fake News Detection and Evaluation in Digital Bengali Text

1st Argha Digar
Department of IT
BUIE, Bankura (India)
arghadigar@gmail.com

1st Soujanya Dutta
Department of IT
BUIE, Bankura (India)
soujanyadutta.8@gmail.com

1st Subrata Pan
Department of IT
BUIE, Bankura (India)
subratapan.it@buie.ac.in

2nd Pritam Roy
Department of IT
BUIE, Bankura (India)
2025it03pritam@buie.ac.in

2nd Somnath Banerjee
Department of CSE
BUIE, Bankura (India)
2025cse12somnath@buie.ac.in

2nd Priyam Banerjee
Department of CSE
BUIE, Bankura (India)
2025cse08priyam@buie.ac.in

Abstract—The purpose of this study is to enhance the discovery of fake news in Bengal, which integrates the Conventional Neural Network (CNN) and bilateral long-short-term memory (LSTM) networks by employing a DEEP Panda Education model. We used two datasets, labelathontic -7K and labelfeck -1K, which were predetermined to combine headlines and content to produce a balanced data set for training. The model architecture was designed to capture both the local and sequential features of the text, achieving an accuracy of 96% on the test set. Our findings indicate that the proposed approach effectively distinguishes between fake and authentic news, underscoring the importance of advanced Natural Language Processing (NLP) techniques in combating misinformation. This research contributes to the growing body of literature on fake news detection in low-resource languages such as bengali, highlighting the need for further exploration of deep learning techniques in this domain [1]. Future work will focus on improving model performance through data augmentation and exploring multiclass classification for various types of fake news [2].

Index Terms—Fake News Detection, Bangla Language, Deep Learning, CNN, LSTM, Text Classification.

I. Introduction

THE rapid ripening of social networks has revolutionized the dissemination of trail knowledge and stimulated widespread propagation of fake news. The influence of such information is notable because of its potential to influence public opinion, disrupt social harmony, and undermine democratic institutions. Bengali [3] is a language spoken by more than 230 million people around the world. Social networks have enabled the instantaneous reach of fake information, making it increasingly difficult for individuals to distinguish between reality and fiction

[4].Such issues increase the spread of social media platforms, where fake materials can reach millions in a few moments. Despite extensive advancements in counterfeit news detection for languages like English, Bengali lacks because of its distinctive linguistic and contextual complexities. Detecting fake news in Bengali faces significant challenges due to its semantic and contextual nuances. Ambiguous expressions like বলা হচ্ছে (it is being said) or অভিযোগ করা হয়েছে (it is alleged) are frequently used to imply unverified claims without making explicit assertions. Sarcastic remarks, such as আজ সূর্য পশ্চিমে উঠেছে (Today, the sun rose in the west!), add complexity by blurring the boundary between satire and misinformation. Additionally, contextual misrepresentation, such as presenting an old event as recent with statements like এই ঘটনা আজ সকালেই ঘটেছে (This incident just happened this morning), further complicates detection. The extensive use of idiomatic expressions and proverbs in Bengali, which are deeply context-dependent, requires sophisticated interpretation to prevent misclassification.

Although people are starting to realize how important it is to detect fake news, there has been little research on this topic for languages such as Bangla, which have fewer resources. Working with Bangla Natural Language Processing (BNLP) comes with challenges, such as not having enough labeled data, the variety of the language and its complexity [6], [7]. Standard machine learning methods often dont work well in these situations because there isnt enough training data, and Bangla has unique language features [8]. Furthermore, models created for languages with more resources, like English, might not work well for Bangla, so we need to create solutions specifically designed for it [9].

This research focuses on tackling these issues by creating a deep learning model that can accurately identify whether Bangla news articles are fake or real. By using advanced methods like Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, we aim to understand both the specific details and the overall flow of the text. This will help improve the model's ability to spot fake news [10]. The main objective of this study is to support ongoing efforts to fight misinformation in the Bangla language, helping to create a society that is better informed.

Previous research on detecting fake news in Bangla has mainly used basic machine learning methods like Support Vector Machines (SVM) and Logistic Regression, as well as more advanced deep learning models like Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. These approaches have shown different levels of success, with accuracy rates between 91% and 96% on various datasets. However, the effectiveness of these models is limited because there aren't enough large, balanced datasets, and more advanced techniques like Bidirectional Gated Recurrent Units (Bi-GRU) haven't been used much.

In this research, we try to fill these gaps by creating a combined CNN-LSTM model to detect fake news in Bangla. We used a large collection of labeled real and fake news articles. First, we clean and prepare the data to make them consistent and ready for deep learning models. Our method combines the ability of CNNs to find important features with the strength of LSTMs in understanding sequences, which helps improve the accuracy of classifying news.

The main achievements of this study are:

1) Creating and testing a special CNN-LSTM model designed to detect fake news in Bangla.
2) Using a well-prepared and balanced dataset to make the model stronger and more reliable.
3) Comparing how well the model works using measures like accuracy, precision, recall, and F1-score.

By combining advanced deep learning techniques with the specific needs of Bangla, this research helps in the fight against fake news in languages that dont have as many resources.

II. RELATED WORK

In recent years, finding fake news has become very important, and many methods and tools have been created for languages like English and Arabic. But there hasnt been much research focused on Bangla, which shows a big missing part in this area of study.

A. Overview of Fake News Detection Techniques

Fake news detection models can be divided into two main types: traditional machine learning methods and more advanced deep learning techniques. Traditional methods, like Support Vector Machines (SVM) and Logistic Regression, are popular because they are simple and work well with structured data. For example, Hussain et al. (2020) showed that SVM and Multinomial Naive Bayes classifiers are effective in identifying fake news in Bangla, reaching an accuracy of 96.64% [11]. However, these models often have difficulty dealing with the complexities of natural language, especially in languages with limited resources, such as Bangla.

On the other hand, deep learning models, especially those using Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, have proven to be better at understanding the subtle details of language. For instance, a combined CNN-LSTM model created by Rahman et al. (2021) reached an outstanding accuracy of 99% in identifying fake news in Bangla [12]. This model takes advantage of the best features of both types of networks: CNNs are great at picking out important details, while LSTMs are good at handling the order and sequence of words in text.

B. Recent Advances in Deep Learning for Fake News Detection

New research has looked into using advanced deep learning methods, like transformer models, to better detect fake news. Models such as BERT (Bidirectional Encoder Representations from Transformers) have changed the game by offering contextual embeddings that greatly boost how accurately we can classify information. For example, a study by Abdal et al. (2023) used DistilBERT to detect fake news in Bangla, reaching an accuracy of 97.85% [13]. This shows how powerful transformer-based models can be, especially when dealing with languages that have fewer resources available.

Also, using methods that combine different models has been proven to improve detection results. A recent study by Almandouh et al. (2024) suggested a combined approach using CNN and LSTM models, which achieved high accuracy in detecting fake news in Arabic [15]. These methods could also be useful for detecting fake news in Bangla, as they use the best features of different models to make the system work better overall.

C. Challenges in Bangla Fake News Detection

Even though there have been improvements in detecting fake news, there are still some problems, especially for Bangla. One big issue is the lack of labeled data, which makes it hard to create strong detection models. The BanFakeNews dataset, which has about 50,000 news articles, has been created, but it has a problem: most of the articles are real news, and there are very few fake ones [16]. This imbalance can cause the models to be biased, so its important to find better ways to balance the data.

Also, the complexity of the Bangla language, with its rich word forms and sentence structures, makes it hard to use current models effectively. Hossain et al. (2022)

pointed out that solving these language-related issues is key to making detection more accurate [17].

In short, while we have made good progress in detecting fake news, especially with new deep learning methods, there are still specific challenges for Bangla that need more research. Using advanced models like transformers and combining different methods, along with better ways to handle data, could help improve fake news detection in Bangla. In the future, we should work on creating better datasets and try new modeling techniques to tackle the unique problems of this less-resourced language.

III. Methodology

This part explains the steps we plan to take to find fake news in Bangla. It includes getting the data ready, cleaning it, building the model, training it, and checking how well it works. The goal is to use advanced computer learning methods to make the detection of fake news more accurate and trustworthy.

A. Dataset Description

The dataset for this study was collected from the open-source Kaggle dataset [18]. Four primary files make up our suggested dataset, and each file helps us understand fake news in Bangla. In this study, we used two datasets: *LabeledAuthentic-7K* and *LabeledFake-1K*. The *LabeledAuthentic-7K* dataset has 7,000 real news articles, and the *LabeledFake-1K* dataset has 1,000 fake news articles. Both datasets include information like headlines, the main content of the articles, and labels that show if the news is real or fake. By putting these datasets together, we created a larger and more complete training set, which is important for making the model work well [19]. Below is a summarized table detailing the columns found in each of these files:

Column Name	Description of the Column
articleID	News ID
domain	Site name of news publisher
date	Publication Date
category	Category of the news
source	Source of the news
relation	Related or Unrelated
headline	Headline of the news
content	Article of the news
label	1 as authentic, 0 as fake
F-type	Fake news type (only in LabeledFake-1K.csv)

Figure 1. Dataset Columns Summary

B. Data Preprocessing

Data preprocessing is an important step to get the datasets ready for training the model. The following steps were done:

- **Loading the Datasets:** The data was loaded using the pandas library.
- **Combining the Datasets:** The two datasets were combined into one table to make it easier to work with.
- **Preparing the Text:** A new column called 'text' was made by combining the 'headline' and 'content' columns. This helps to include all the details from the news articles.
- **Tokenization and Padding:** The text was split into smaller parts (tokenized) and adjusted (padded) so that all pieces of text are the same length for the model. The limit was set to 10,000 words and 150 steps in length.
- **Splitting the Data:** The data was divided into two parts: 80% for training and 20% for testing, to check how well the model works.

C. Model Architecture

The proposed model architecture is a hybrid model that combines Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. This setup is made to pick up on both the small details and the order of words in text data, thereby enhancing the model's ability to detect fake news.

- **Model Definition:** The model was built using Keras and includes these layers:
  - **Embedding Layer:** Turns words into compact numerical vectors.
  - **Convolutional Layer:** Finds patterns or features in small parts of the text.
  - **MaxPooling Layer:** Shrinks the size of the feature maps to simplify the data.
  - **Bidirectional LSTM Layer:** Analyzes text in both forward and backward directions to understand context better.
  - **Global Average Pooling Layer:** Averages the feature maps to make the data smaller and easier to handle.
  - **Dense Layers:** Produces the final result for classification.

D. Training Process

The model was set up and trained with the following details:

- **Setup:** The model used the Adam optimizer, a binary cross-entropy loss function, and accuracy to measure performance.
- **Monitoring:** Tools like early stopping and saving the best model were used to keep track of training progress.
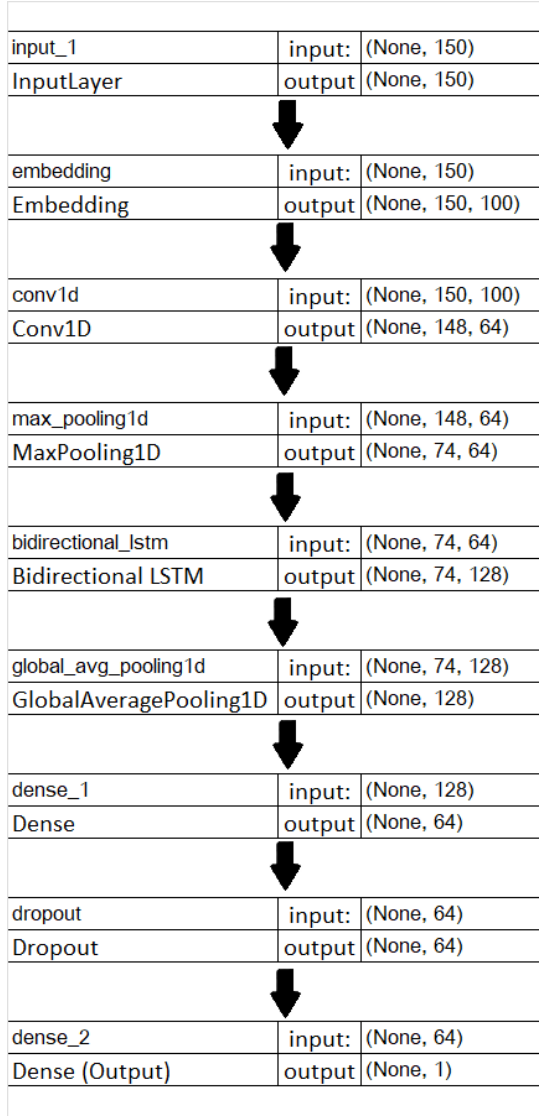
Figure 2. Model Architecture

- **Training:** This model was trained at 100 rounds (epochs) using samples of 32 (batch size) each, with 20% of the data being used for validation.

E. Evaluation of Model

After the model was trained, it was tested using another data record to see how well it performed. The results were measured using various reviews, including the accuracy, accuracy, recall, and F1 scores. These reviews were found in tools such as classification reports and confusion matrixs.

F. Ethical Considerations

Ethical issues are very important when doing research with human data. This study followed ethical rules by protecting the privacy of participants, making sure they joined voluntarily, getting their permission after explaining everything, and keeping their data safe. All partici-

pants were told about the study's goals and their rights, making sure everything was clear and responsible during the research [20].

The proposed methodology outlines a complete way to find fake news in the Bangla language using advanced computer learning techniques. By using a combined CNN-LSTM model and tackling the special difficulties of Bangla language processing, this research hopes to contribute to ongoing efforts to combat misinformation in digital spaces.

IV. Discussion

The findings of this study show that the combined CNN-LSTM model works well for spotting fake news in Bangla. The model reached a test accuracy of 96%, and the precision, recall, and F1-score results showed it performed strongly for both types of news articles. The classification report highlighted these important results:

Label	Precision	Recall	F1-Score	Support
Fake News	0.95	0.75	0.84	249
Authentic News	0.96	0.99	0.98	1452
Macro Average	0.95	0.87	0.91	1701
Weighted Average	0.96	0.96	0.96	1701

Table I
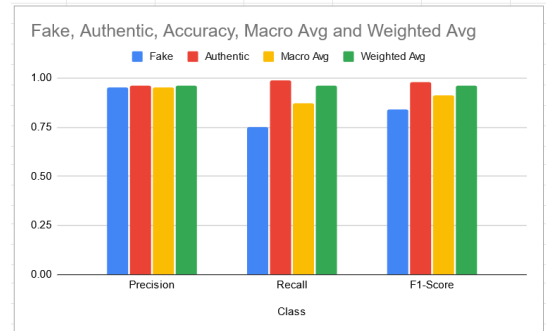CLASSIFICATION REPORT FOR FAKE AND AUTHENTIC NEWS DETECTION

Figure 3. Fake, Authentic, Accuracy, Macro Avg, and Weighted Avg

V. Performance Analysis

The model's precision for detecting fake news is 0.95, meaning it is very good at spotting fake news without mistakenly labeling too many real articles as fake. However, its recall of 0.75 shows that while it does well at finding fake news, it still misses some. This gap might be because detecting subtle misinformation is tricky and often needs a deeper understanding of context and language details [21]. On the other hand, the model performs exceptionally well with real news, achieving a recall of 0.99 and an F1-score of 0.98. This means it is very good at identifying genuine news, which is important to avoid wrongly flagging real articles as fake in practical use [22]. The high precision and recall for real news show that the model is effective at telling the difference between real and fake news, which is key for maintaining trust and accuracy.
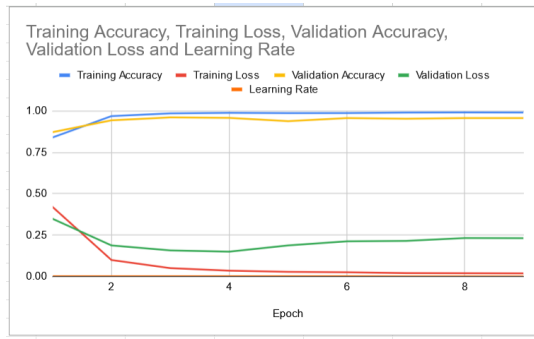
Figure 4. Training Accuracy, Training Loss, Validation Accuracy, Validation Loss, and Learning Rate

VI. Implications of Findings

The results of this study are very important for creating systems that can detect fake news, especially for languages with fewer resources, like Bangla. The hybrid CNN-LSTM model worked well, showing that deep learning methods can help solve the problem of identifying fake news in different languages. As false information keeps spreading, being able to correctly tell if a news article is real or fake is crucial for keeping the public informed and maintaining trust in information [23]. The results also show how important it is to keep improving how we train the model and how we prepare the data. Since the model isn't as good at spotting fake news, future studies should work on adding more varied examples of fake news to the dataset. This could include tricky examples designed to test how strong the model is [24]. Also, trying out methods that combine different models or using advanced architectures like transformers could help make the detection more accurate and reliable [25].

VII. Limitations and Future Work

The findings are encouraging, but this study has some limitations. The dataset had more real articles than fake ones, which might have affected how well the model performed. Future research should try to balance the dataset or use methods like oversampling or creating synthetic data to make sure the model learns equally from both types of articles [26].

Additionally, the model's effectiveness should be tested in real-world situations, where the way news spreads and how people react to it might be different from the controlled conditions of this study. Testing the model in a live environment could offer useful insights into how practical it is and where it can be improved [27].

VIII. Conclusion

In this study, we introduced a CNN LSTM model combining CNN LSTM to recognize fake Bangla messages. This closes the important gap in the detection of false information. This model worked very well, achieving 96% test accuracy and exhibited strong accuracy and recall

values in both fake and real news. For fake messages, the model's accuracy was 0.95 and the recall was 0.75, but the actual message achieved an accuracy of 0.96 and a recall of 0.99. These results show that the model is particularly good for identifying actual messages. This is important to reduce actual use errors..

The results of this study go further than what was found right away. Using deep learning methods for languages like Bangla, which dont have a lot of resources, shows that these same methods could work for other languages and situations too. With false information spreading more and more, creating strong systems to detect it is very important. This helps keep peoples trust in the media and makes sure information is shared honestly.

However, this study has a few limitations. The dataset had many more real articles than fake ones, which might have affected how well the model worked. Future studies should try to add more varied examples of fake news to the dataset and look into different methods to make the detection more accurate. Also, the model should be tested in real-world situations to see how useful it is in practice.

In conclusion, this research helps in the fight against false information online. By using smart computer programs, we can create better tools to spot fake news, helping people stay well-informed. Moving forward, its important to keep improving these tools and solving the problems that come with understanding language, so they work well in different situations.

References

[1] Enhancing Bangla Fake News Detection Using Bidirectional Gated Recurrent Units and Deep Learning Techniques | Proceedings of the 7th International Conference on Networking, Intelligent Systems, and Security. (n.d.). Retrieved January 16, 2025, from https://dl.acm.org/doi/10.1145/3659677.3659703

[2] Enhancing Bangla Fake News Detection Using Bidirectional Gated Recurrent Units and Deep Learning Techniques | Proceedings of the 7th International Conference on Networking, Intelligent Systems and Security. (n.d.). Retrieved January 16, 2025, from https://dl.acm.org/doi/10.1145/3659677.3659703

[3] Alam, F., Hasan, M.D., Alam, T., Khan, A., Tajrin, J., Khan, N., & Chowdhury, S. A. (2021). A Review of Bangla Natural Language Processing Tasks and the Utility of Transformer Models. arXiv.Org, abs/2107.03844.

[4] Raman, R., Nair, V. K., Nedungadi, P., Sahu, A. K., Kowalski, R., Ramanathan, S., & Achuthan, K. (2024). Fake news research trends, linkages to generative artificial intelligence, and sustainable development goals. Heliyon, 10.

[5] Wani, M. A., ELAffendi, M. A., Shakil, K. A., Abuhaimed, I., Nayyar, A., Hussain, A., & El-latif, A. (2024). Toxic Fake News Detection and Classification for Combating COVID-19 Misinformation. IEEE Transactions on Computational Social Systems, 11, 51015118.

[6] Alam, F., Hasan, Md. A., Alam, T., Khan, A., Tajrin, J., Khan, N., & Chowdhury, S. A. (2021). A Review of Bangla Natural Language Processing Tasks and the Utility of Transformer Models. arXiv.Org, abs/2107.03844.

[7] Kowsher, Md., Sami, A. A., Prottasha, N. J., Arefin, M., Dhar, P. K., & Koshiba, T. (2022). Bangla-BERT: Transformer-Based Efficient Model for Transfer Learning and Language Understanding. IEEE Access, 10, 9185591870.

[8] Kowsher, Md., Sami, A. A., Prottasha, N. J., Arefin, M., Dhar, P. K., & Koshiba, T. (2022). Bangla-BERT: Transformer-Based Efficient Model for Transfer Learning and Language Understanding. IEEE Access, 10, 9185591870.

[9] Kowsher, Md., Sami, A. A., Prottasha, N. J., Arefin, M., Dhar, P. K., & Koshiba, T. (2022). Bangla-BERT: Transformer-Based Efficient Model for Transfer Learning and Language Understanding. IEEE Access, 10, 9185591870.

[10] Abdal, Md. N., Oshie, Most. H. K., Haque, Md. A., & Rahman, S. (2023). A Transformer Based Model for Bangla Fake News Detection Using DistilBERT. 2023 5th International Conference on Sustainable Technologies for Industry 5.0 (STI), 16.

[11] Hussain, M. G., Hasan, Md. R., Rahman, M., Protim, J., & Hasan, S. (2020). Detection of Bangla Fake News using MNB and SVM Classifier. 2020 International Conference on Computing, Electronics & Communications Engineering (iCCECE), 8185.

[12] Adib, Q. A. R., Mehedi, M. H. K., Sakib, Md. S., Patwary, K. K., Hossain, Md. S., & Rasel, A. A. (2021). A Deep Hybrid Learning Approach to Detect Bangla Fake News. 2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), 442447.

[13] Abdal, Md. Nazmul, Most. Humayera Kabir Oshie, Md. Azizul Haque, and Salman Rahman. A Transformer Based Model for Bangla Fake News Detection Using DistilBERT. In 2023 5th International Conference on Sustainable Technologies for Industry 5.0 (STI), 16, 2023.

[14] Abdal, Md. Nazmul, Most. Humayera Kabir Oshie, Md. Azizul Haque, and Salman Rahman. A Transformer Based Model for Bangla Fake News Detection Using DistilBERT. In 2023 5th International Conference on Sustainable Technologies for Industry 5.0 (STI), 16, 2023.

[15] Aboulola, Omar Ibrahim, and Muhammad Umer. Novel Approach for Arabic Fake News Classification Using Embedding from Large Language Features with CNN-LSTM Ensemble Model and Explainable AI. Scientific Reports 14 (2024).

[16] Hossain, Md. Zobaer, Md Ashraful Rahman, Md Saiful Islam, and Sudipta Kar. BanFakeNews: A Dataset for Detecting Fake News in Bangla. International Conference on Language Resources and Evaluation abs/2004.08789 (2020).

[17] Hossain, Md. Muzakker, Zahin Awosaf, Md. Salman Hossan Prottoy, Abu Saleh Muhammod Alvy, and Md. Kishor Morol. Approaches for Improving the Performance of Fake News Detection in Bangla: Imbalance Handling and Model Stacking. arXiv.Org abs/2203.11486 (2022).

[18] (iJET), 2020, Accessed: Aug. 15, 2023. [Online]. Available: https://www.researchgate.net/publication/347886404 BanFakeNews | Kaggle. Accessed: Aug.15,2023. [Online]. Available: https://www.kaggle.com/datasets/cryptexcode/banfakenews

[19] How to Write Research Methodology in 2025: Overview, Tips, and Techniques | Research.Com. Accessed January 16, 2025. https://research.com/research/how-to-write-research-methodology.

[20] How to Write Research Methodology in 2025: Overview, Tips, and Techniques | Research.Com. Accessed January 16, 2025. https://research.com/research/how-to-write-research-methodology.

[21] Recalling Fake News during Real News Corrections Can Impair or Enhance Memory Updating: The Role of Recollection-Based Retrieval - PMC . Accessed January 16, 2025. https://pmc.ncbi.nlm.nih.gov/articles/PMC9481799/.

[22] Recalling Fake News during Real News Corrections Can Impair or Enhance Memory Updating: The Role of Recollection-Based Retrieval - PMC . Accessed January 16, 2025. https://pmc.ncbi.nlm.nih.gov/articles/PMC9481799/.

[23] Recalling Fake News during Real News Corrections Can Impair or Enhance Memory Updating: The Role of Recollection-Based Retrieval - PMC . Accessed January 16, 2025. https://pmc.ncbi.nlm.nih.gov/articles/PMC9481799/.

[24] Recalling Fake News during Real News Corrections Can Impair or Enhance Memory Updating: The Role of Recollection-Based Retrieval - PMC . Accessed January 16, 2025. https://pmc.ncbi.nlm.nih.gov/articles/PMC9481799/.

[25] What Are the Performance Metrics Used to Compare Fake News Detection Models? - Consensus. Accessed January 16, 2025. https://consensus.app/results/?q=What%20are%20the%20performance%20metrics%20used%20to%20compare%20fake%20news%20detection%20models%3F.

[26] What Are the Performance Metrics Used to Compare Fake News Detection Models? - Consensus. Accessed January 16, 2025. https://consensus.app/results/?q=What%20are%20the%20performance%20metrics%20used%20to%20compare%20fake%20news%20detection%20models%3F.

[27] Recalling Fake News during Real News Corrections Can Impair or Enhance Memory Updating: The Role of Recollection-Based Retrieval - PMC . Accessed January 16, 2025. https://pmc.ncbi.nlm.nih.gov/articles/PMC9481799/.