Deep Learning for Bengali Fake News Detection: Innovative Approaches for Accurate Classification

[1]Sheikh Sadi Bandan., [2]Md Sharuf Hossain., [1]MD. Samiul Islam Sabbir., [3]Khadiza Tul Kobra

[1]Dept. of Computer Science & Engineering Daffodil International University Dhaka, Bangladesh

[2]Dept. of Data Science Loyola University Chicago, USA

[3]Dept. of Information Technology and Management, Illinois Institute of Technology Chicago, USA

DOI: https://doi.org/10.51584/IJRIAS.2024.908034

Received: 25 July 2024; Revised: 03 August 2024; Accepted: 07 August 2024; Published: 09 September 2024

ABSTRACT

A vast quantity of data and information are available on the internet. Because the internet is so widely available and has resulted in a tremendous growth in the number of online news, people are interested in reading news from online news portals. Online news portals include things like Facebook, Twitter, WhatsApp, Telegram, Instagram, blogs, and more. Both the quantity of news-on-news websites and the number of readers are increasing. But how real is online news today is a matter of thought. A huge amount of fake news is being spread in newspapers and online due to various yellow journalists. Which is having an adverse effect on society. As a result, there are many kinds of instability, bad politics, etc. problems are being created in the country. If this situation continues, our country and society will go to hell. The only solution is to ensure that yellow journalists do not spread fake news. But despite all the vigilance, fake news will spread. We can solve this by using artificial intelligence, for example, by employing various machine learning and deep learning algorithms, we can identify bogus news and take precautions against it. In this paper, fake news is detected using 4 deep learning algorithms like RNN, LSTM, BiLSTM, GRU model and 1 machine learning algorithm BERT model. RNN has an accuracy of 94.58%, LSTM has an accuracy of 92.84%, BiLSTM has an accuracy of 94.29%, GRU has an accuracy of 93.22% and BERT has an accuracy of 95%. The BERT model has the highest accuracy of 95% among all models.

Keyword*s*: NLP, Machine learning, Bangla language, Document category.

INTRODUCTION

Because social media and internet news sources are so prevalent in today's world, false news spreads quickly. This may sway public opinion, affect the results of elections, and bring in money for spammers via clickbait adverts [1]. When it comes to news, smartphone users prefer social media, but it may be difficult to verify information on sites like Facebook, WhatsApp, and Twitter. Since rumours presented as news may be detrimental to society, particularly in emerging nations like India, reliable reporting is necessary [2]. Fake news and rumours are becoming more and more common as a result of the growing dependence on online news sources. These problems have been made worse by the internet's widespread use and ease of access, which has had a detrimental effect on society [3]. The primary information source and essential component of peoples' life is the Internet. Globally, the number of online news sources is rapidly expanding, and individuals are increasingly interested in reading daily news portals as a result of internet accessibility. Bengali headlines and hourly news updates are available on thousands of portals. Nowadays, in the era of digitalization, the majority of people consume news online rather than in newspapers. The use of online portals, Twitter, Facebook, blogs, and other apps is increasing these days, which is why there is a lot of information available on websites and a growing quantity of usage of the internet. In other words, a large portion of the global population now exclusively relies on Internet news. Along with producing and accessing a lot of information every day, people are also using handheld multimedia devices and high-speed Internet. As of the end of

January 2018, there were around 80.83 million internet users in Bangladesh, of which approximately 30 million used social media. Since the paper boom has diminished, there has been a daily growth in the creation and consumption of online news, with an emphasis on the Internet. Rather than disseminating their news, many news organizations seem to be producing and uploading it to the internet. E-news is the term used to describe news that is published and made available via the Internet. The viewership of this news is growing daily because to internet user scholarship. As a result, a wide variety of news are being entered into the website's database. In emerging nations like India, Bangladesh, and Pakistan, news plays a critical role in disseminating knowledge and raising public awareness of events in neighboring countries.

One of the foundational ideas of information technology is effective information retrieval. News headlines are a more generalized kind of text content. Online sources of news include those regarding computers, social sciences, music, politics, Hollywood, Bollywood, sports, and entertainment. On the internet, users may find and see any kind of news. Users may quickly search for and see news based on their needs by using news headlines. And in addition to watching, the news can be understood whether it is real or fake thanks to the current artificial intelligence.

LITERATURE REVIEW

Kingaonkar et al. [1] conducted a study categorizing 2000 news items from social media and online news sites using machine learning, including SVM, which achieved 99.90% accuracy. The dataset includes columns 5–10. Adedoyin et al. [2] is study aims to identify false and authentic news from social media and online sites, using machine learning models like RNN, Naive Bayes, Random Forest, SVM, and Logistic Regression, with SVM showing the highest accuracy. G. Senthilkumar et al. [3] employ Passive Aggressive Classifier, Naive Bayes, and Random Forest for fake news detection, with the Passive Aggressive Classifier achieving the highest accuracy of 87%. Rahman et al. [4] investigated fake news detection using LR, DT, KNN, NB, LSTM, and Bi-LSTM. Logistic Regression achieved 96% accuracy, while Bi-LSTM reached 99% in deep learning. Kulkarni et al. [5] examined various machine learning techniques for false news detection, using data from diverse sources. Logistic regression achieved the highest accuracy at 85%, outperforming Gradient Boosting, RF, DT, and KNN. Choudhury et al. [6] discuss using machine learning methods like SVM, Random Forest, Naïve Bayes, and Logistic Regression to detect false news from various sources, achieving up to 61% accuracy on the LIAR dataset. Murugesan et al. [7] found that among five machine learning algorithms—KNN, Naive Bayes, SVM, BERT, and Decision Tree—the Adaboost & Decision Tree method achieved the highest accuracy at 98.5%. Khanam et al. [8] suggest using Python libraries like Skit-Learn and NLP for fake news detection, employing machine learning methods such as XG boost, Random Forests, Naive Bayes, KNN, Decision Trees, and SVMs, with XG boost achieving 75% accuracy. Kushwaha et al. [9] employed three machine learning algorithms—Random Forest, Naïve Bayes, and Logistic Regression—to detect fake news. Logistic Regression performed the best with 65% accuracy. Ngada et al. [10] study machine learning techniques for detecting fake news, using six algorithms. The Support Vector Machine achieved the highest accuracy at 99.4%, outperforming AdaBoost, DT, KNN, RF, and XG Boost. Lakshmanarao et al. [11] analyze methods for detecting fake news using machine learning, specifically NLP. They found that the Random Forest classifier, with 90.7% accuracy, outperformed SVM, K-Nearest Neighbors, and Decision Tree algorithms. Lasotte et al. [12] used six machine learning techniques, including Naïve Bayes, SVM, Random Forest, and more, to predict false news. Soft voting achieved the highest accuracy at 93%. Sultana et al. [13] is study evaluates eight machine learning algorithms for fake news detection, with the Support Vector Classifier (SVC) achieving the highest accuracy at 96%. Md. Y. Tohabar et al. [14] present an AI model using machine learning to detect fake news. Among the methods, the support vector machine (SVM) achieved the highest accuracy at 73.20%. S. Pandey et al. [15] analyzed AI models for false news detection. Among KNN, Logistic Regression, Naïve Bayes, Decision Tree, and SVM, Logistic Regression achieved the highest accuracy at 90.46%.

RESEARCH METHODOLOGY

A. Dataset Collection

The data collection process starts with preparing a plan outlining the method and goals for gathering data. Next, identify and select appropriate news portals or websites relevant to the desired information. Once the sources are chosen, search for specific news categories within these sites. Following this, employ web scraping techniques to extract data from the targeted websites systematically. Finally, organize and store all the collected data in a spreadsheet for easy access and analysis.

B. Data Preprocessing

Information from the internet is mainly unstructured and often contains null values, redundant data, and errors. Thus, processing is essential after collection. The dataset undergoes cleaning to remove low-length, unique, and duplicate data, making it intelligible. Various features like word lists, sorted word lists, documents per class, total words per class, and dataset splitting are extracted. Preprocessing steps include removing hashtags, screen names, URLs, zero values, punctuation, symbols, emojis, integers, unnecessary symbols, retweets, and short data. These steps ensure the dataset is ready for analysis. The dataset comprises two categories: authentic and false news, with the majority in the authentic class. Cleaning involved removing emojis, symbols, and punctuation from both categories and eliminating short documents. After this process, 10321 articles remained, following the removal of 186 short ones. retweets, and short data. These steps ensure the dataset is ready for analysis. The dataset comprises two categories: authentic and false news, with the majority in the authentic class. Cleaning involved removing emojis, symbols, and punctuation from both categories and eliminating short documents. After this process, 10321 articles remained, following the removal of 186 short ones.

Table 1: Classifiers Description

No	Original Text	Label	Cleaned Text
1	ঢাবির ভর্তি পরীক্ষায় জালিয়াতি: আটক ২।	Real	ঢাবির ভর্তি পরীক্ষায় জালিয়াতি আটক ২
2	প্রেম করলে বাড়বে ওজন!	Fake	প্রেম করলে বাড়বে ওজন

C. Statistical Analysis

1. There are 10507 total data points in the dataset.
2. Dataset retains six columns.
3. There are 10321 original data in the collection, of which 8043 are genuine and 2278 are fraudulent.
4. There are 7430 training data points in all in the dataset.
5. There are 1033 testing data points in the dataset.
6. There are 1858 validation data in all in the dataset.
7. Real and fake labels are divided into two groups.

D. Design Approach

We employed both supervised and unsupervised learning techniques since the data presents a multivariate classification challenge. This study employs deep learning techniques such as RNN, LSTM, BiLSTM, GRU, and BERT algorithms. For every model, a confusion matrix, performance and accuracy predictions, and an analysis of the outcomes were made.

The flowchart outlines a process beginning with visiting various apps or websites to find a product of interest. Once on these platforms, the next step is to search for the specific product. After locating it, reviews and feedback from other users are gathered to assess the product's quality and performance. The relevant and important information from these reviews is then copied. Finally, this extracted data is organized and stored in a spreadsheet for easy analysis and reference. Figure displays the overall system architecture diagram:

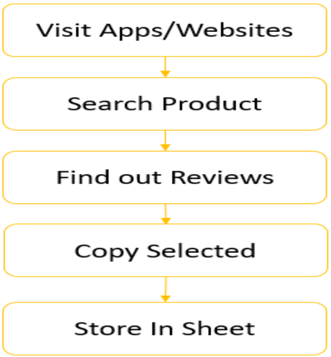


Fig. 1. Architecture of Working Process

E. Dataset Description

Six columns make up the primary dataset utilized in this work: domain, date, category, title, content, and label. Two categories comprise the label, namely- original and fake. Besides, there are 10507 rows of data in the dataset. By preprocessing or cleaning the dataset, we used a fundamental partitioning of the dataset consisting of 10321 original documents and removed 186 documents, 7430 samples for training, 1033 samples for testing and 1858 samples for validation. After data preprocessing, out of total 10321 data, 8043 real data and 2278 fake data were found.

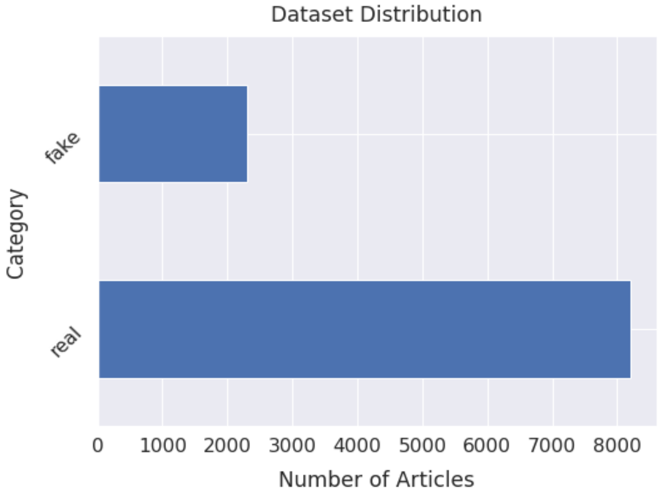


Fig. 2. Dataset Distribution

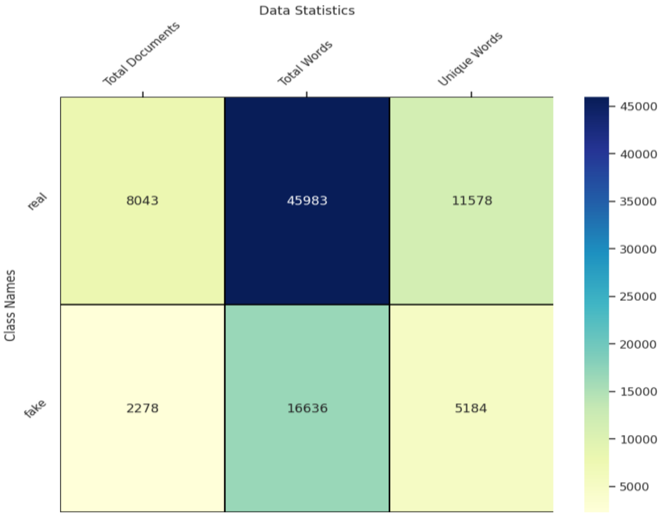


Fig. 3. Data Statistics

The dataset's document length and frequency were measured following the visualization of the dataset. There are three minimum and six average lengths for the document, with a maximum length of 135 and a minimum of 3. The length frequency distribution graph is displayed below:

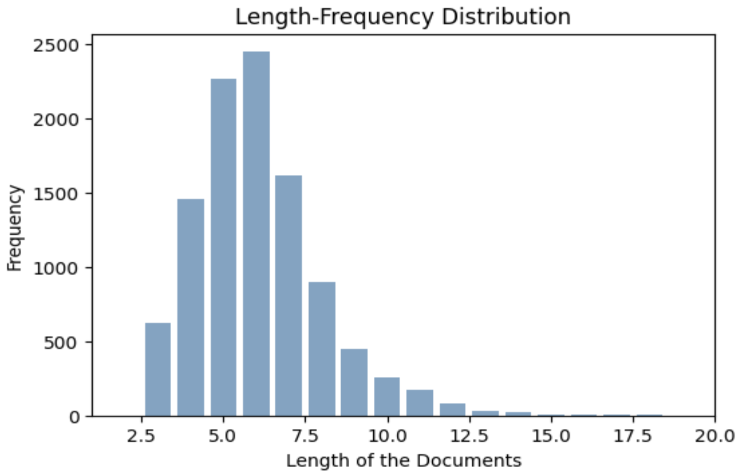


Fig. 4. Length Frequency Distribution

F. Tokenizer

Tokenization is the process of dividing a text document into smaller sections, such as a phrase, paragraph, or whole text. When dealing with textual data, tokenization is frequently utilized. We have also tokenized our dataset. Our data collection had 10321 total documents, from which 14294 distinct tokens were extracted. We tokenize our data collection in two different ways. Specifically, padded and encoded sequences.

i. Encoded Sequence

Encoded Sequences
মান্নার ১৪তম মৃত্যুবার্ষিকী
[3786, 1, 2928]

ii. Padded Sequence

Padded Sequence
মান্নার ১৪তম মৃত্যুবার্ষিকী
[3786 1 2928 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]

EXPERIMENTAL RESULT AND DISCUSSION

A. Recurrent Neural Network (RNN)

The RNN model is a type of artificial neural network designed for sequential data. It processes inputs in a loop, using its internal state to retain information from previous inputs, making it well-suited for tasks involving time series, language processing, and other sequential data.

Table 2: Classification Report of Rnn

Classes	Precision	Recall	F1_Score	Support
Fake	88.94	84.86	86.85	218
Real	96.00	97.18	96.59	815
Accuracy	94.58	94.58	94.58	1033
Macro Avg	92.47	91.02	91.72	1033
Weighted Avg	94.51	94.58	94.53	1033

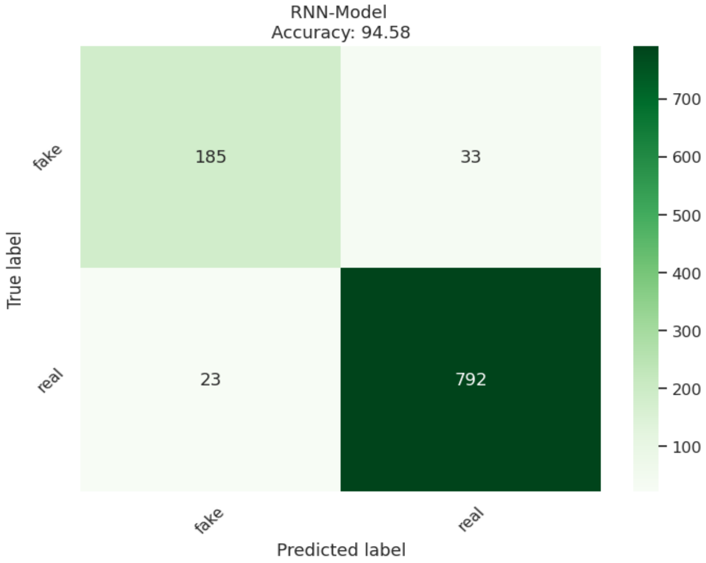


Fig. 5. Confusion Matrix for RNN Model

B. Long Short-Term Memory (LSTM)

The LSTM model is a type of recurrent neural network designed to handle long-term dependencies. It uses special units called memory cells to maintain information over extended periods, making it effective for tasks like language modeling, time series prediction, and sequence generation.

Table 3: Classification Report of Lstm

Classes	Precision	Recall	F1_Score	Support
Fake	80.25	87.61	83.77	218
Real	96.60	94.23	95.40	815
Accuracy	92.84	92.84	92.84	1033
Macro Avg	88.43	90.92	89.59	1033
Weighted Avg	93.15	92.84	92.95	1033

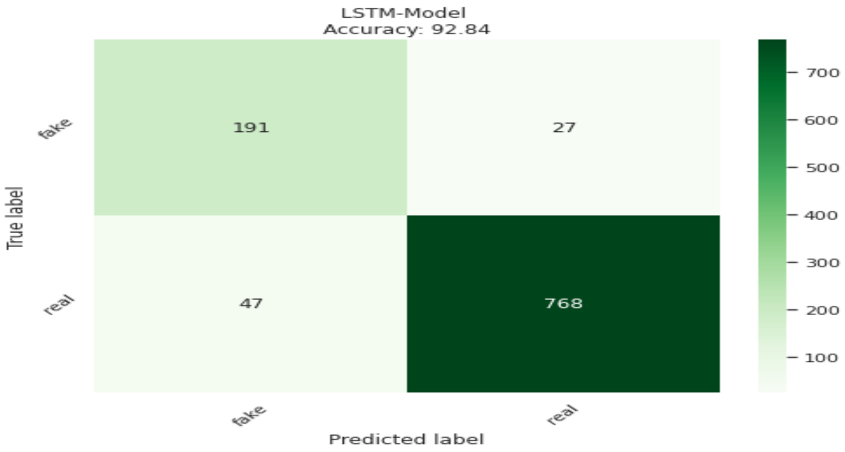


Fig. 6. Confusion Matrix for LSTM Model

C. Bidirectional Long Short-Term Memory (Bi-LSTM)

The BiLSTM model is an advanced variant of LSTM that processes data in both forward and backward directions. This bidirectional approach allows the model to capture context from both past and future sequences, enhancing its performance in tasks like language modeling and sequence prediction.

Table 4: Classification Report of Bi-Lstm

Classes	Precision	Recall	F1_Score	Support
Fake	86.98	85.78	86.37	218
Real	96.21	96.56	96.39	815
Accuracy	94.29	94.29	94.29	1033
Macro Avg	91.59	91.17	91.38	1033
Weighted Avg	94.26	94.29	94.27	1033

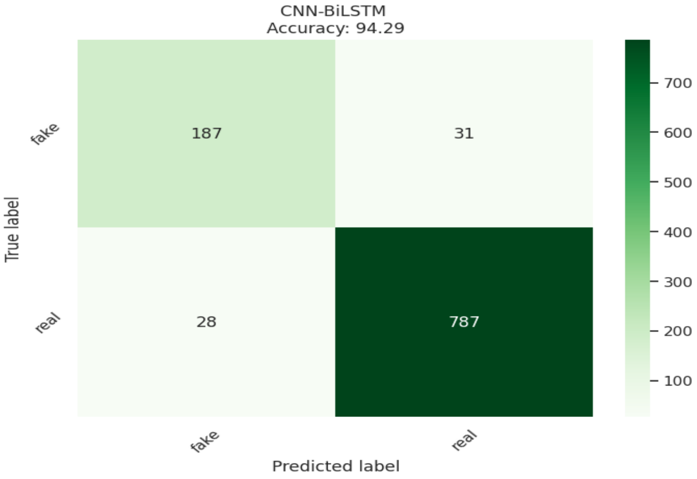


Fig. 7. Confusion Matrix for Bi-LSTM Model

D. Gated Recurrent Units (GRU)

The Bi-GRU model uses GRU cells in both forward and backward directions to capture context from both past and future inputs. This bidirectional approach enhances performance on tasks like sequence prediction by considering the full context of the input sequence.

Table 5: Classification Report of Gru

Classes	Precision	Recall	F1_Score	Support
Fake	82.17	86.70	84.38	218
Real	96.39	94.97	95.67	815
Accuracy	93.22	93.22	93.22	1033
Macro Avg	89.28	90.83	90.02	1033
Weighted Avg	93.39	93.22	93.29	1033

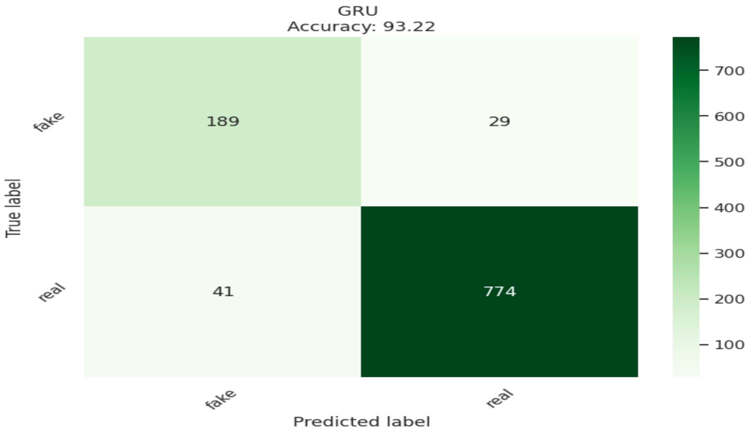


Fig. 8. Confusion Matrix for GRU

E. Prediction of All Model

The prediction accuracy of various models is: RNN at 94.58%, LSTM at 92.84%, BiLSTM at 94.29%, and GRU at 93.22%. RNN achieved the highest accuracy of 94.58%, while the lowest was 92.84% by LSTM.

Table 6: Classifiers Description

Model	Sample News	Class	Accuracy
RNN	'অবশেষে জানা গেল ‘নাবিলা জানো’ পোস্টারের রহস্য!'	Fake	94.58 %
LSTM	'অবশেষে জানা গেল ‘নাবিলা জানো’পোস্টারের রহস্য!'	Fake	92.84 %
BiLSTM	'অবশেষে জানা গেল ‘নাবিলা জানো’ পোস্টারের রহস্য!'	Fake	94.29 %
GRU	'অবশেষে জানা গেল ‘নাবিলা জানো’পোস্টারের রহস্য!'	Fake	93.22 %

The prediction accuracy of various models is: RNN at 94.58%, LSTM at 92.84%, BiLSTM at 94.29%, and GRU at 93.22%. RNN achieved the highest accuracy of 94.58%, while the lowest was 92.84% by LSTM.

F. BERT Model

The Bi-BERT model leverages BERT’s bidirectional transformer architecture to understand context from both left and right of a token. It excels in capturing nuanced meaning and context in text by processing all tokens in parallel and considering their relationships comprehensively.

i. Show word cloud image in real headlines:

ii. Show word cloud image in fake headlines:

Table 7: Classification Report of Bert

Classes	precision	recall	f1-score	support
Fake	0.83	0.80	0.81	260
Real	0.97	0.98	0.97	1819
Accuracy			0.95	2079
Macro avg	0.90	0.89	0.89	2079
Weighted avg	0.95	0.95	0.95	2079

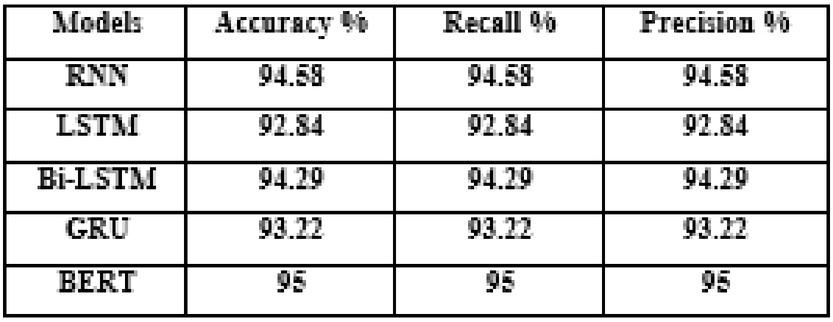
G. Discussion

In today's world, news is increasingly abundant, appearing in newspapers, websites, social media, blogs, and online portals. Unlike newspapers, there is no limit to the content published online, leading to exponential growth. If fraudsters exploit this, it could harm online news portals and impact public perception, especially for countries like Bangladesh. Identifying fake news is crucial, and modern technology offers solutions through machine learning and deep learning algorithms. This paper explores deep learning algorithms such as RNN, LSTM, BiLSTM, GRU, and BERT for detecting fake news.

4 deep learning models in the paper namely RNN, LSTM, BiLSTM and GRU; And a machine learning model used in the paper called BERT. RNN, LSTM, BiLSTM and GRU deep learning models have 94.58% accuracy

in RNN, 92.84% accuracy in LSTM, 94.29% accuracy in BiLSTM and 93.22% accuracy in GRU. Also, the machine learning model BERT has 95% accuracy. It can be seen that RNN, BiLSTM model using deep learning obtained the highest accuracy of 94.58%, 84.29% respectively and BERT model using machine learning obtained 95% accuracy. Here it can be seen that the highest accuracy is obtained using the BERT model of machine learning.

Table 8: Classifiers Description



Models	Accuracy %	Recall %	Precision %
RNN	94.58	94.58	94.58
LSTM	92.84	92.84	92.84
Bi-LSTM	94.29	94.29	94.29
GRU	93.22	93.22	93.22
BERT	95	95	95

CONCLUSION

Human activity recognition is mainly used in gaming, for news to be valuable, it must be classified into authentic categories. The need for such classification is increasing, with data science algorithms playing a key role in identifying false news through deep learning and machine learning. Our research paper categorizes news into real and fake using models such as BiLSTM, GRU, Uni-Gram, and various machine learning techniques like LR, Multinomial NB, RF, and SVM. Deep learning models RNN, LSTM, BiLSTM, and GRU show accuracies of 94.58%, 92.84%, 94.29%, and 93.22%, respectively. BERT, a machine learning model, achieved 95% accuracy. In Bangladesh, e-news is growing rapidly, with popular and reliable sources including Prothom Alo, Bangladesh Pratidin, Nayadigant, Jugantor, and Samakal, favored for their updated and trustworthy content.

REFERENCES

1. Kingaonkar, S. et al. (2023) ‘Fake News Detection using Machine Learning’, INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT, 07(03). doi:10.55041/ijsrem18181.
2. Senthilkumar and D. Ashok Kumar (2023) ‘A comparative study on various machine learning algorithms for the prediction of fake news detections using bring feed new data sets’, International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 9(1), pp. 131–142. doi:10.32628/cseit228691.
3. Rahman, M. et al. (2022) ‘Political fake news detection from different news source on social media using Machine Learning Techniques’, AIUB Journal of Science and Engineering (AJSE), 21(2), pp. 110–117. doi:10.53799/ajse. V 21i1.383.
4. Kulkarni, P. et al. (2021) ‘Fake news detection using machine learning’, ITM Web of Conferences, 40, p. 03003. doi:10.1051/itmconf/20214003003.
5. Choudhury, D. and Acharjee, T. (2022) ‘A novel approach to fake news detection in social networks using genetic algorithm applying machine learning classifiers’, Multimedia Tools and Applications, 82(6), pp. 9029–9045. doi:10.1007/s11042-022-12788-1.
6. Murugesan, S. and Pachamuthu, K. (2022) ‘Fake news detection in the medical field using machine learning techniques’, International Journal of Safety and Security Engineering, 12(6), pp. 723–727. doi:10.18280/ijsse.120608.
7. Khanam, Z. et al. (2021) ‘Fake news detection using machine learning approaches’, IOP Conference Series: Materials Science and Engineering, 1099(1), p. 012040. doi:10.1088/1757-899x/1099/1/012040.
8. Kushwaha, N.S. and Singh, P. (2022) ‘Fake news detection using machine learning: A comprehensive analysis’, Journal of Management and Service Science (JMSS), 2(1), pp. 1–15.

doi:10.54060/jmss/002.01.001.

9. Ngada, O. and Haskins, B. (2020) 'Fake news detection using content-based features and machine learning', 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE) [Preprint]. doi:10.1109/csde50874.2020.9411638.
10. Lakshmanarao, A., Swathi, Y. and Kiran, Dr. T. (2019) 'An effecient fake news detection system using machine learning', International Journal of Innovative Technology and Exploring Engineering, 8(10), pp. 3125–3129. doi:10.35940/ijitee.j9453.0881019.
11. Lasotte, Y.B. et al. (2022) 'An Ensemble Machine Learning Approach for Fake News Detection and Classification Using a Soft Voting Classifier', European Journal of Electrical Engineering and Computer Science, 6(2). doi://dx.doi.org/10.24018/ejece.2021.6.2.409.
12. Sultana, R. et al. (2022) 'An effective fake news detection on social media and online news portal by using Machine Learning', Australian Journal of Engineering and Innovative Technology, pp. 95–106. doi:10.34104/ajeit.022.0950106.
13. Md. Y. Tohabar, N. Nasrah, and A. M. Samir, "Bengali Fake News Detection Using Machine Learning and Effectiveness of Sentiment as a Feature," 2021 Joint 10th International Conference on Informatics, Electronics & Vision (ICIEV) and 2021 5th International Conference on Imaging, Vision & Pattern Recognition (icIVPR), Aug. 2021, doi: https://doi.org/10.1109/icievicivpr52578.2021.9564138.
14. S. Pandey, S. Prabhakaran, N. V. Subba Reddy, and D. Acharya, "Fake News Detection from Online media using Machine learning Classifiers," Journal of Physics: Conference Series, vol. 2161, no. 1, p. 012027, Jan. 2022, doi: https://doi.org/10.1088/1742-6596/2161/1/012027.
15. S. S. Bandan, M. Rahman Ajmain, A. R. Rejuan, M. Farhana Khatun and S. A. Khushbu, "State of Survey: Advancement of Knowledge Environmental Sustainability in Practicing Administrative Apps," 2022 13th International Conference on Computing Communication and Net- working Technologies (ICCCNT), 2022, pp. 1-8, doi: 10.1109/ICC-CNT54827.2022.9984416.
16. M. R. Ajmain, M. F. Khatun, S. S. Bandan, A. R. Rejuan, N. J. Ria and S. R. H. Noori, "Enhancing Sentiment Analysis using Machine Learning Predictive Models to Analyze Social Media Reviews on Junk Food," 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2022, pp. 1-7, doi: 10.1109/ICC- CNT54827.2022.9984355.
17. M. A. R. Rejuan, S. S. Bandan, M. A. Rakib and M. Assaduzzaman, "A Comparative Study for Measuring the Quality of Dhaka City Transportation System: Survey Based," 2023 International Conference for Advancement in Technology (ICONAT), Goa, India, 2023, pp. 1-8, doi: 10.1109/ICONAT57137.2023.10080188.
18. Sheikh Sadi Bandan, Sabid Ahmed Sunve, and Shaklian Mostak Romel, "A Deep Learning Approach for Bengali News Headline Categorization," Jul. 2023, doi: https://doi.org/10.1109/icccnt56998.2023.10307776.