Journal of Integrated Sciences
Volume 6, Issue 3, June 2026
ISSN: 2806-4801

HYBRID DEEP LEARNING FRAMEWORK FOR ACCURATE DETECTION OF BIASED NEWS TO PROMOTE TRUSTWORTHY INFORMATION CONSUMPTION

Sayma Akter Trina, Shafin Mahmood, Arpita Saha Sukanna, Md. Easin Khandokar, Sabirina Zaman Esha, Nuzhat Tabassum

Department of Computer Science and Engineering, American International University-Bangladesh (AIUB), Dhaka 1229, Bangladesh

ABSTRACT

In the digital era, news disseminates quickly on social media, online, and news collecting platform. This helps users to stay informed, but the problem of biased reporting distorts public sentiment, opinion and spreads misinformation. Existing Machine learning Deep Learning model find it challenging with the manipulative and context-dependent nature of bias in text, and detect biased data from an imbalanced dataset, achieving low accuracy. This paper proposes a hybrid approach named "CHL" that combines convolutional layers (CNN), a hierarchical attention network (HAN), and a gradient boosting approach (LightGBM). CNN picks up local language features, and HAN is used for word-level and sentence-level text to capture long-distance dependencies, and both capture local and global text. LightGBM enhances classification accuracy with computationally efficient decision boundaries and handles the minority class better. For this paper, we used a dataset collected from Kaggle with 147111 data, where biased data is labeled 1 and non-biased data is labeled as 0. Previous studies detected biased

news using a hybrid model and achieved accuracies ranging from 68\% to 82\%. Our proposed model outperforms these results and gets train accuracy 0.8688 and test accuracy 0.8607. Average interface time per sample is 0.4924ms, Precision 0.8607, Recall 0.9823, F1-Score 0.9223, MCC 0.3331, MAE 0.1393.

Keywords: *Biased News, Deep Learning, Gradient Boost, Hierarchical attention network, Machine Learning.*

Corresponding author: Shafin Mahmood can be contacted at shafin26103@gmail.com

1. INTRODUCTION

In this digital age, the use of the internet has increased people's tendency to acquire knowledge, form opinions, and engage in various social issues. As a result, the availability of accurate and unbiased information becomes important in forming a balanced perspective. However, news articles are generally considered the most reliable and quality source of information. However, media influence, funding, or political position can create a fallout (Hamborg, Donnay, & Gipp, 2018). The media can express its bias through ignoring an issue, over-emphasizing it, using selective information, or using propaganda tactics such as emotion or fear. It categorizes topics such as immigration and climate change, as well as political ideologies such as right and left. It allows journalists to analyze from either a left or right perspective, while also identifying bias in news articles and providing balanced information to news aggregator platforms to ensure media quality control and compliance (Baly et al., 2020). The main reason for the spread of low-quality and false information is the lack of proper monitoring of digital platforms.

In recent years, the spread of misleading and false information by users and public awareness has become a matter of serious concern. As the information is widely available online, machine learning technologies, especially deep learning, have emerged as a potential solution to detect fake news, playing an effective role in identifying patterns and characteristics of fake news using big data (Alghamdi et al.,2024). A fundamental task of NLP is text classification, which labels text. Recent deep learning models have advanced learning mechanisms, which incorporate knowledge of the document structure to achieve better results, while neural network-based methods are effective. Hierarchical Attention Network (HAN) is used to identify important parts of two types of attention at the word and sentence levels. To enhance the effectiveness of the HAN model, information is presented considering the hierarchical structure and relevance of the document (Yang et al., 2016). Current machine learning and deep learning systems are not useful to effectively capture the subtle, context-dependent and multi-level nature of bias in textual data, which can often lead to limited accuracy and poor generalization. Further, issues like uneven datasets and lack of the ability to model local and global linguistic features together are even more detrimental to performance. Thus, the necessity of the new model that will be able to combine various feature representations and enhance the robustness is evident. To overcome these drawbacks, a hybrid method can be used that integrates the strengths of other architectures, which can supplement each other.This study proposes a new hybrid method, named CHL, which combines CNN, HAN, and LightGBM. CNN model to capture local linguistic features and contextual cues, HAN to identify the long-range dependencies of words and dictionaries for both

local and global reception of information, and LightGBM to improve the effectiveness of the classification of disparate datasets so that presentations can be used to create fast and effective decision thresholds. These unify the components, ensuring high accuracy, good generalization, and fast inference of the model. Past research on the detection of media bias has mainly been interested in either traditional machine learning, or standalone deep learning models, which in most instances do not succeed in capturing both fine-grained linguistic features as well as long-range contextual dependence at the same time. Also, most methods fail to balance the data they deal with, and have limited extrapolation capabilities across a variety of news sources, leaving a research gap that is clear on how to achieve robust and scalable bias detection. Thus, the purpose of this study is to come up with a powerful hybrid model that combines complementary learning mechanisms to promote accuracy, capture multi-level textual characteristics, and provide reliable detection of biased news in real-life contexts.

2. REVIEW OF LITERATURE

Media bias occurs when someone, such as a reporter or editor, presents news unfairly by using the wrong word, language, or approach, or because of any kind of political or non-political influences. While there are multiple forms of bias, e.g., bias by personal perception or by the omission of information (Spinde et al., 2022), our focus is to classify word-level and sentence-level news bias. Here, we summarize some literature based on classifying media and news bias. To predict word-level and sentence-level bias proposed a robust training expert called BABE was proposed, using a dataset of annotated media bias (MBIC) of Macro F1 score 0.804. Here, the MBIC total

annotated sentences are 1700, but this paper improves the MBIC dataset. They divided the total 3700 datasets into 2 parts and trained the BABE with distant supervision. But this paper basically focused on sentence-level bias, differentiating the crowd vs expert label quality. To predict the leading political ideology or bias of news articles, this paper (Baly et al., 2019) uses state-of-the-art pre-trained Trans-formers in this challenging setup with 34,737 articles with annotated political ideology left, center, or right, which is well-balanced across both topics and media with BERT & LSTM, adversarial adaptation (AA), triplet-loss pretraining, media-level reps (Twitter bios, Wikipedia), and results on media-based split, best model Macro F1-score 64.29, Accuracy 72.00; baseline BERT Macro F1-score 35.53, Accuracy 36.75. But this paper models overfitting to the source without debiasing. To predict lexical bias, like word, syntax, and informational bias like sentences or clauses this study (Fan et al., 2019) proposed BASIL(Bias Annotation Spans on the Informational Level) of 300 news articles with lexical and informational bias spans and benchmark it using rule-based classifiers and the BERT model with results Macro F1 47.27; lexical bias F1 31%, informational bias F1 43%, but in this case dataset size is modest and subtle bias hard to capture. Using NLP this study (Lei, Huang, Wang, & Beauchamp, 2022) proposed discourse-aware modeling via knowledge distillation based on Longformer + BiL-STM backbone to predict sentence bias, even though it seems neural with precision +2.70- 3.32 pts, recall up to +5.07 pts over strong baselines. Besides this strong performance, this study depends on accurate discourse parsing, potential error propagation, article-level constraints. Predicting media bias at the sentence level, this study (Lei & Huang, 2024) designed an event relation

graph that consists of events as nodes and four common types of event relations: coreference, temporal, causal, and subevent relations.Using two steps, an event-aware language model is built to inject the events and event relations knowledge into the basic language model via soft labels; further, a relation-aware graph attention network is designed to update sentence embedding with events and event relations information based on hard labels with result F1 +5.78 pts on BASIL; +12.86 pts on Biased Sents vs. baseline. But this study struggles with implicit (unstated) event relations and needs better extraction of implicit relations. In (Krieger et al., 2022), DA-RoBERTa represents a new state-of-the-art transformer-based model adapted to the media bias domain, which identifies sentence-level bias with an F1 score of 0.814. In addition to more transformer models like DA-BERT and DA-BART to detect prior bias. Additionally, Liu et al. (2022) works on characterizing and predicting ideology across different genres of text using Pretrained Language Models by collecting 3.6M political news articles, for pretraining, and proposed a solution named POLITICS. These results, SemEval-2019 Hyperpartisan: Accuracy 85.2, F1-score 84.9 (best across baselines). Recently, detecting hyperpartisan and propagandistic content in news articles and social media is a more focused area in research such as in Lyu et al. (2024) proposed computational method for extremely biased news titles achieving 0.84 percent accuracy. A hierarchical graph-based integration network proposed in (Ahmad et al., 2025) for propaganda detection in textual news articles on social media additionally emphasizes the relation between textual features and context-based relationships. Similarly, Ahmad et al. (2025) developed a fine-tuned deep learning model to enhance propaganda detection. Also, natural

language processing techniques to detect hyperpartisan news articles can effectiveness of linguistic features and semantic analysis (Naredla & Adedoyin, 2022). BERT for misinformation detection in political news is also predicted perfectly and achieves 0.79 percent accuracy (Padalko, Chomko, & Chumachenko, 2023), also demonstrating the advantages of transformer-based approaches (BERT) in text-based news verification. In Kiesel et al. (2019) achieved benchmarking performance and improving detection system in both manually labeled and distantly supervised datasets by the SemEval-2019 Task 4 on hyperpartisan news detection to predict media bias, this study (Joo & Hwang, 2019) uses a feature engineered ensemble (traditional ML) with moderate accuracy (0.745) on the test set subtask A, although this study has handcrafted features and limited generalization beyond dataset. Based on that summarization, we can say that in the topic of predicting media bias, previous studies are not enough yet. That's why we try to get more innovative and effective solutions in our study. As shown in Table I, various deep learning models have been applied to media bias detection.

3. METHODOLOGY

This study introduces hybrid framework (deep learning and machine learning approach) CHL. In this section, we discuss the dataset, data pre-processing, Model architecture and finally evaluation.

3.1 Dataset

For this research, we collect data from kaggle "Media Bias Dataset: Annotations by Experts" with 147111 data. The

dataset is labeled 1 for biased news and 0 for non-biased news with a text column.

3.2 Data pre-processing

The dataset is preprocessed using standard text-cleaning techniques to clean the data and normalize the data for suitable input into our model. First, we converted to lowercase to maintain consistency and reduce case sensitivity. Then, expanded abbreviation using predefined abbreviation map, uninformed tag was also removed, such as punctuation, URLs, hashtags, and numerical digits. After eliminating common stop words, the text was tokenized, lastly, to reduce words to their base forms, both stemming (Porter Stemmer) and lemmatization (WordNet Lemmatizer) were applied.

3.3 Model

The proposed CHL model integrates three core components - CNN, HAN, LightGBM for effectively detecting biased news. First input texts are tokenized using a vocabulary size of 20,000 and padded to a maximum length of 100 words and converted into 128-dimensional embedding vector. CNN uses a 1D convolution layer, kernel size 5, pool size 2. After that, an attention mechanism is applied on outputs. Deep features are fed into LightGBM classifier with 31 leaves, learning rate 0.05, and estimator 200. LightGBM builds an efficient boosted decision tree and handles imbalanced datasets, which improves the overall accuracy. The dataset is split into 80\% training and 20\% testing sets.

Table 1. Hyperparameters Used in the Proposed CHL Model

Component	Parameter	Value
Tokenizer	Vocabulary size	20,000
Tokenizer	Maximum sequence length	100
Embedding Layer	Embedding dimension	128
CNN Layer	Filters	128
CNN Layer	Kernel size	5
CNN Layer	Activation	ReLU
MaxPooling Layer	Pool size	2
BiLSTM Layer	Units (per direction)	64
Attention Layer	Dense units	1
Attention Layer	Activation	tanh + softmax weights
Global Pooling	Type	GlobalAveragePooling
LightGBM Classifier	num_leaves	31
LightGBM Classifier	learning_rate	0.05
LightGBM Classifier	n_estimators	200
Training Setup	Train/Test split	80% / 20% (stratified)
Training Setup	Batch size	64

Source: The author(s) own work.

3.4 Evaluation Metrics

a) Accuracy: Accuracy measures the proportion of correctly predicted samples out of the total samples.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

b) Precision: Precision indicates the proportion of positive predictions that were actually correct.

$$Precision = \frac{TP}{TP + FP}$$

c) Recall: Recall (Sensitivity) measures the proportion of actual positives correctly predicted by the model.

$$Recall = \frac{TP}{TP + FN}$$

d) F1-Score: The F1-score is the harmonic mean of precision and recall, balancing false positives and false negatives.

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

e) Mean Absolute Error (MAE): MAE quantifies the average magnitude of errors in predictions without considering their direction.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

4. ANALYSIS

Analysis of the research work being carried out here relates to assessing the efficacy of the proposed CHL (CNN–HAN–LightGBM) hybrid approach designed for biased news detection. The work here has been carried out in a systematic manner based on the nature of the data discussed within the methodology.

4.1 Data Analysis

The data set used in the research work was obtained from Kaggle's "Media Bias Dataset: Annotations by Experts", which

contains a total of 147,111 data samples, and each data set records either 1 (Biased) or 0 (Not Biased). The data set contains an imbalanced nature since biased data sources tend to be found more often in online digital sources. Furthermore, text preprocessing operations such as normalization, tokenization, stop word removal, stemming, and lemmatization were carried out on the data set, thereby generating a meaningful output from the texts that maintains uniformity. This made it fit perfectly for processing by the proposed approach based on deep learning.

Feature Representation and Model Component Analysis: The text data was represented in a 128-dimensional embedding space, which allowed it to retain contextual semantics. The CNN part of the model used local linguistic and contextual features, which were relevant to phrases. These phrases contain indicative linguistic bias. The Hierarchical Attention Network (HAN) model allowed it to model long-range dependencies within texts by using two forms of attention mechanisms: attention at both the word and sentence levels. This made it easy for the model to recognize linguistic bias within the structure of sentence flows. Finally, LightGBM a gradient boosting decision tree algorithm, acted as a high-level classifier that received the deep-learned embeddings as input. The tree structure of LightGBM enabled it to successfully distinguish biased embeddings from non-biased ones. The combination of these components enabled the CHL model to capture both semantic depth and decision-level precision.

4.2 Result Analysis

The performance of the proposed model CHL was evaluated using performance metrics accuracy, precision, recall, F1-score, and MAE. The model achieved training accuracy of 0.8688 and test accuracy of 0.8607.

Table 2. Evaluation Metrics of the Proposed CHL Model

Metric	Value
Train Accuracy	0.8688
Test Accuracy	0.8607
Overall Accuracy	0.8607
Precision	0.8693
Recall	0.9823
F1-score	0.9223
MAE	0.1393

Source: The author(s) own work.

Table 2 presented the evaluation matrices of CHL model including train accuracy, test accuracy, overall Accuracy, precision, recall, f1-score, mae.

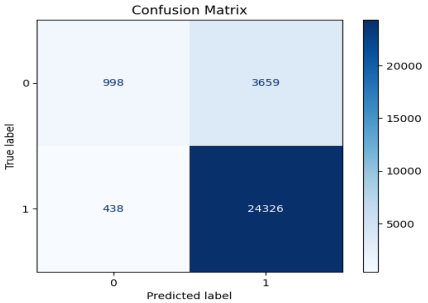

Figure 1. Confusion Matrix

Journal of Integrated Sciences
Volume 6, Issue 3, June 2026
ISSN: 2806-4801

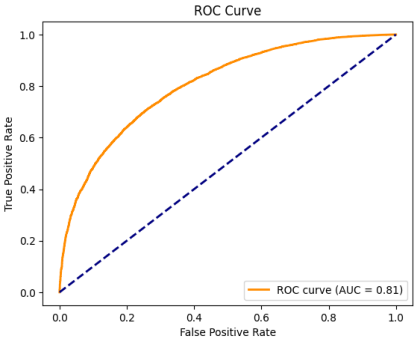

Figure 2. ROC Curve

Fig. 1 illustrates the confusion matrix of the proposed model, showing the distribution of correctly and incorrectly classified instances across all classes. Fig. 2 illustrated ROC curve for evaluating the classification performance of the proposed model by plotting the true positive rate against the false positive rate (FPR).

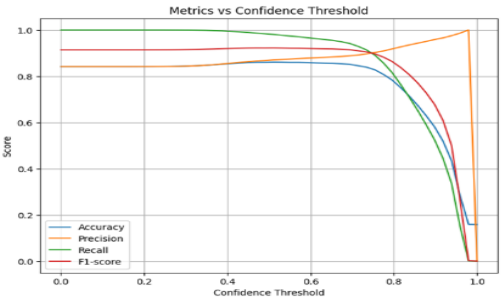

Figure 3. Metrics vs Confidence Threshold

Fig. 3 illustrates evaluation metrics such as accuracy, precision, recall, and F1-score against different confidence thresholds of the proposed model. In summary, the results show that the CHL

model significantly outperforms earlier hybrid models, and achieves a robust balance of classification accuracy, recall, and computational efficiency.

5. CONCLUSION

This research introduces CHL, an integrated deep learning model that combines CNN, HAN, and LightGBM to accurately detect biased news. By recognizing specific linguistic subtleties and larger contextual associations, and leveraging the strong features of LightGBM, the framework demonstrates exceptional performance. By enhancing predictive accuracy, the proposed framework also increases its effectiveness across various sources, presenting a practical method to encourage reliable information consumption and maintain public trust in journalism. Future studies could explore the integration of CHL with transformer models and multimodal datasets to enhance both bias detection and adaptability in real-world news contexts. When dealing with unbalanced datasets. Evaluations conducted on a comprehensive benchmark dataset indicate that CHL significantly outperforms conventional machine learning and hybrid approaches, attaining a test accuracy of 86.07\% and exhibiting remarkable precision, recall, and F1-score. The findings illustrate how hybrid models successfully tackle the complex and context-dependent aspects of news bias. By enhancing predictive accuracy, the proposed framework also improves its effectiveness across various sources, providing a practical method to encourage the consumption of reliable information and maintain public trust in journalism. Future studies could explore the integration of CHL with transformer models and multimodal datasets to advance both bias detection and adaptability in real-world news scenarios.

REFERENCES

Abreu, J., Fred, L., Macêdo, D., & Zanchettin, C. (2019). Hierarchical attentional hybrid neural networks for document classification. In *Lecture Notes in Computer Science* (pp. 396–402). Springer. https://doi.org/10.1007/978-3-030-30493-5_39

Ahmad, P. N., Guo, J., AboElenein, N. M., et al. (2025). Hierarchical graph-based integration network for propaganda detection in textual news articles on social media. *Scientific Reports, 15*, 1827. https://doi.org/10.1038/s41598-024-74126-9

Ahmad, P. N., Shah, A. M., & Lee, K. (2025). Enhanced propaganda detection in public social media discussions using a fine-tuned deep learning model: A diffusion of innovation perspective. *Future Internet, 17*(5), 212. https://doi.org/10.3390/fi17050212

Alghamdi, J., Alqarni, M., Alharbi, A., & Alotaibi, F. (2024). Enhancing hierarchical attention networks with CNN and stylistic features for fake news detection. Expert Systems with Applications, 257, 125024.

Baly, R., Da San Martino, G., Glass, J., & Nakov, P. (2019). We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1454–1464). Association for Computational Linguistics.

Baly, R., Da San Martino, G., Glass, J., & Nakov, P. (2020, October 11). We can detect your bias: Predicting the political ideology of news articles. arXiv*. https://arxiv.org/abs/2010.05338

Fan, L., White, M., Sharma, E., Su, R., Choubey, P. K., Huang, R., & Wang, L. (2019). In plain sight: Media bias through the lens of factual reporting. arXiv preprint.

Hamborg, F., Donnay, K., & Gipp, B. (2018). Automated identification of media bias in news articles: An interdisciplinary literature review. *International Journal on Digital Libraries, 20*(4), 391–415.

Joo, Y., & Hwang, I. (2019). Steve Martin at SemEval-2019 Task 4: Ensemble learning model for detecting hyperpartisan news. In *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 967–972). Association for Computational Linguistics.

Kiesel, J., Mestre, M., Shukla, R., Vincent, E., Adineh, P., Corney, D., Stein, B., & Potthast, M. (2019, June). SemEval-2019 Task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation* (pp. 829–839). Minneapolis, MN: Association for Computational Linguistics.

Krieger, J. D., Spinde, T., Ruas, T., Kulshrestha, J., & Gipp, B. (2022). A domain-adaptive pre-training approach for language bias detection in news. *arXiv preprint arXiv:2206.01234*.

Lei, Y., & Huang, R. (2024). Sentence-level media bias analysis with event relation graph. *arXiv preprint arXiv:2404.12345*.

Lei, Y., Huang, R., Wang, L., & Beauchamp, N. (2022). Sentence-level media bias analysis informed by discourse structures. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1234–1245.

Liu, Y., Zhang, X. F., Wegsman, D., Beauchamp, N., & Wang, L. (2022). POLITICS: Pretraining with same-story article comparison for ideology prediction and stance detection. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5678–5689.

Lyu, H., Pan, J., Wang, Z., & Luo, J. (2024). Computational assessment of hyperpartisanship in news titles. *Proceedings of the International AAAI Conference on Web and Social Media, 18*(1), 999–1012. https://doi.org/10.1609/icwsm.v18i1.31368

Naredla, N. R., & Adedoyin, F. F. (2022). Detection of hyperpartisan news articles using natural language processing technique. *International Journal of Information Management Data Insights, 2*(1), 100064. https://doi.org/10.1016/j.jjimei.2022.100064

Padalko, H., Chomko, V., & Chumachenko, D. (2023, November). Misinformation detection in political news using BERT model. In *Proceedings of the 3rd International Workshop of IT-Professionals on Artificial Intelligence (ProfIT AI)* (Vol. 3641, pp. 117–127). Waterloo, Canada: CEUR Workshop Proceedings.

Spinde, T., Plank, M., Krieger, J. D., Ruas, T., Gipp, B., & Aizawa, A. (2022). Neural media bias detection using distant supervision with BABE – Bias Annotations by Experts. *arXiv preprint*.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1480–1489)