HT-MAGPIE: A Hierarchical Transformer for Article-Level Media Bias Classification

Anonymous ACL submission

Abstract

Automatic media bias classification studies typically focus on isolated sentences, presenting challenges when applied to news articles. Article-level media bias classification offers a more practical and holistic approach. However, research in this area remains under-explored, partly due to the lack of datasets. Therefore, in this paper, we first release a reconstructed version of an existing dataset, consisting of full article texts and metadata. Second, we propose HT-MAGPIE, a hierarchical transformer for article-level media bias classification, leveraging MAGPIE—a large-scale model pre-trained on bias-related tasks—to produce bias-aware representations. We demonstrate that HT-MAGPIE outperforms all baselines by at least 0.13% and surpasses fine-tuned BERT by 5.02% in F1 score. We also explore the correlation between outlet-level and article-level bias by comparing model performance with and without outlet metadata. Our findings indicate that including outlet metadata as an additional feature improves F1 scores on fine-tuned BERT by 4.32% and BigBird by 2.62%.

1 Introduction

Media bias refers to the tendency of news outlets reporting in a manner that favours certain opinions, preferences, or agendas, rather than maintaining objectivity (Mastrine, 2022). It has shaped and influenced public opinion (Castillo-Campos et al., 2025), accelerated with social media, allowing false information to circulate while remaining unchecked (Froehlich, 2024; Calvillo et al., 2021).

Organisations such as AdFontes; Allsides; MBFC manually review and assess bias in articles, but this approach is both tedious and costly (Wang et al., 2025). Automatic media bias detection presents a promising solution (Rodrigo-Ginés et al., 2024) with large language models (LLMs) enabling efficient processing and analysis of vast quantities of articles.

Nevertheless, previous studies in media bias classification largely function at the sentence level (Spinde et al., 2021; Maab et al., 2023a,b; Guo and Zhu, 2022), which is difficult to apply to news articles. These approaches often rely heavily on lexical cues (Chen et al., 2020a) and overlook the broader context (van den Berg and Markert, 2020). Moreover, individual sentences in an article may exhibit varying or contradictory biases (Lei et al., 2022).

Since media outlets present content as articles, article-level systems are more valuable (Fields et al., 2024). Yet, research in this area is still limited, complemented by the lack of adequate datasets with article-level annotations (Rodrigo-Ginés et al., 2024).

The BAT dataset (Spinde et al., 2023) offers a solution in addressing this shortage, providing expert-validated article annotations based on AdFontes. However, as it lacks article content, we reconstruct this dataset by crawling article texts and introducing a four-class label system to support classification.

In contrast, document classification is a well-established field (Ranjan and Prasad, 2023), with many promising techniques yet to be explored for article-level media bias classification. However, long text processing generally suffers from the input token limitation problem, specifically in encoder-based transformers.

Hierarchical transformers overcome the input token limitation by leveraging language models (LMs) to encode text segments and aggregate the resulting information into higher-level representations. This method has been strongly implemented in document classification tasks (Zhang et al., 2019; Pappagari et al., 2019; Su et al., 2021; Wu et al., 2021; Khandve et al., 2022), allowing full processing of long texts without truncation.

Following recent progress in developing LLMs for media bias tasks, MAGPIE (Horych et al., 2024)

was published as the first large-scale model specifically pre-trained on bias-related tasks. Specifically designed for media bias detection, MAGPIE can generate bias-aware representations, potentially enhancing performance on downstream classification tasks.

Building on these two lines of work, we propose HT-MAGPIE, a hierarchical transformer with MAGPIE encoder (Horych et al., 2024). We show that HT-MAGPIE outperforms all baselines, highlighting the effectiveness of bias-pretrained encoders for article-level media bias classification.

Lastly, as article bias tends to reflect its source outlet bias (Ganguly et al., 2020), we examine the significance of outlet information in classifying article bias by evaluating performance with the inclusion and exclusion of outlet metadata in the input features. Our results demonstrate that including outlet metadata improves performance across most fine-tuned models.

Therefore, our main contributions are summarised as follows:

1. We propose HT-MAGPIE, a novel method combining MAGPIE as a bias-pretrained encoder with a hierarchical transformer architecture, outperforming BoW baseline by 6.25% and fine-tuned BERT by 5.02% F1 score.

2. We present a reconstructed BAT dataset including full article content and a four-class label system.

3. We analyse the impact of including outlet metadata in our features, showcasing its role in classifying article bias.

2 Related Work

There are only a few relevant datasets for article-level media bias classification (Rodrigo-Ginés et al., 2024). BASIL (Fan et al., 2019) is widely used but limited to only 300 articles. NLPCSS (Chen et al., 2020b) includes 6,964 articles based on AdFontes labels, classified into three broad categories: *bias*, *neutral*, or *unknown*. BAT (Spinde et al., 2023) contains 6,345 articles from 321 outlets, crawled and annotated from AdFontes with bias and reliability scores rather than class labels.

Notable works in article-level media bias classification include a Gaussian Mixture Model (Chen et al., 2020a) and decoder-based transformers (Menzner and Leidner, 2024). Alternatively, other studies instead focused on predicting political bias and ideology in articles (Kulkarni et al., 2018; Baly et al., 2019; Kim and Johnson, 2022)

Early efforts in hierarchical methods include the Hierarchical Attention Networks (HAN) (Yang et al., 2016). Pappagari et al. (2019) were among the first to implement a hierarchical transformer for long document classification, followed by several others (Su et al., 2021; Wu et al., 2021; Khandve et al., 2022).

3 BAT Dataset Reconstruction

To reconstruct the BAT dataset (Spinde et al., 2023), we crawled the missing article content from the respective websites, yielding 5,270 articles from the original 6,345 articles, with substantial data loss attributed to unavailable articles and websites. To recover the missing articles, we manually sought and retrieved an additional 226 articles, bringing the final dataset to 5,497 articles. Finally, we applied a keyword-matching system to clean the crawled texts, acknowledging that certain inconsistencies could not be resolved through automated cleaning alone—fully addressing them would likely require substantial manual effort.

The dataset contains score-based annotations from AdFontes, which rates articles through reliability scores, ranging from 0 to 64, with higher scores indicating high-accuracy articles well-supported by multiple sources (Otero, 2021). To support classification, we group the reliability scores into four simplified classes derived from the original eight-category framework (Otero, 2021): *Problematic* (scores < 24), *Questionable* (scores 24–32), *Generally Reliable* (scores 32–40), and *Reliable* (scores > 40).

4 Hierarchical Transformers

We introduce a hierarchical transformer architecture inspired by Su et al. (2021) and Wu et al. (2021), as illustrated in Figure 1. First, the input text is segmented into smaller chunks and encoded using a pre-trained LM. The last hidden states of these chunk embeddings are extracted and passed through two untrained transformer layers (Vaswani et al., 2017). A pooling operation is subsequently applied to the outputs, forming a summary representation of each chunk. Finally, the summary representations are processed through a Multi-Layer Perceptron (MLP), and a softmax function determines the final output class.

We implement two versions of hierarchical trans-

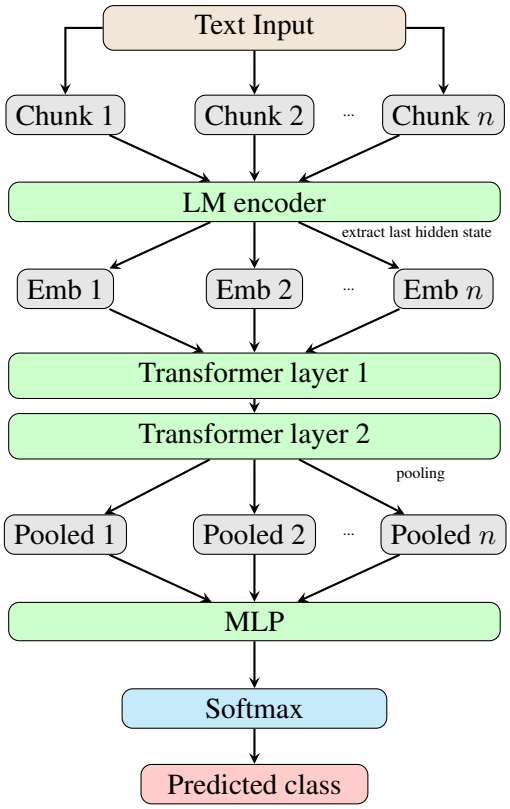

Figure 1: Hierarchical transformer architecture.

formers based on two encoders: **HT-BERT** uses BERT (Devlin et al., 2019) and **HT-MAGPIE** uses MAGPIE as the encoder. MAGPIE is pre-fine-tuned on 59 bias-related tasks (Horych et al., 2024), allowing for a comparison between a domain-specific media bias encoder and a general-purpose language model.

5 Methodology

5.1 Dataset Split

Class	Train	Test	Val	Weight
Problematic	287	27	34	3.77
Questionable	611	54	70	1.77
G. Reliable	1033	104	128	1.05
Reliable	2394	384	371	0.45

Table 1: Number of total samples in the dataset and number of samples for each class in the train, test, and validation set, along with their class weights.

For training, we split the dataset into train, test, and validation sets, as shown in Table 1. Given the dataset size and class imbalance, we opt to maximise the use of available data by exposing models to as many patterns as possible. To achieve this, we evenly distribute articles from different outlets across the three sets, acknowledging a potential drawback that neither the test set nor the validation set contains articles from previously unseen outlets, which could introduce a limitation in performance evaluation.

5.2 Features and Baselines

Our evaluation is based on two distinct feature sets: the first includes only the article titles and contents, while the second includes article outlets, titles, and contents. We add a full stop between each article component and orderly concatenate them into a single sequence.

We employ three primary baselines: (1) a Bag-of-Words model combined with a multilayer perceptron (BoW+MLP), (2) LM fine-tuning (FT), and (3) chunked LM fine-tuning (CFT). In LM fine-tuning, we implement three models—BERT, Bigbird, and Longformer—processing only the initial tokens of an article, up to each model's maximum input length (512 for BERT, 4096 for Longformer/BigBird).

The third baseline leverages chunking techniques in fine-tuning BERT, allowing us to bypass the 512-token length limitation of BERT. Input texts are tokenised, segmented into fixed-size chunks, and padded. A 'CLS' and 'SEP' tokens are inserted at the beginning and end of each chunk, respectively. The chunks are then stacked and fed to the model as mini-batches. The logits produced by each chunk are averaged to produce the final sequence logits, which are subsequently passed through a softmax to obtain class probabilities.

For the second feature set, we introduce an additional baseline: the **outlet majority** method, which predicts the class of an article based on the most frequent class among articles from the same outlet. This baseline aims to assess outlet-level bias and examine its effect on article-level bias.

5.3 Training Details

Method	Epochs	LR	WS
BoW+MLP	10	2e-5	0
Fine-tuning	4	2e-5	500
Chunked fine-tuning	4	2e-5	216
Hierarchical models	3	1e-5	162

Table 2: Epochs, learning rate (LR), and warmup steps (WR) for evaluated methods.

Tokenisation is performed on the word level for BoW, whereas transformer-based approaches apply a sub-word tokenisation. We use the cased version

of the BERT model to retain meaningful case distinctions. For training, we use a batch size of 8 and the AdamW optimiser (Loshchilov and Hutter, 2019) along with the hyperparameters shown in Table 2. Neural networks apply the ReLU activation function (Agarap, 2019) and a dropout rate of 0.2. The BoW+MLP model consists of a single 128-unit dense layer followed by dropout. The MLP layer in our hierarchical transformer architecture includes two 768-unit dense layers, with dropout applied after each one. We experimented with various pooling strategies and chunk sizes, ultimately selecting a 156-token window for mean-pooling and a 512-token window for CLS-pooling. Additionally, we apply a weighted loss in our training, assigning higher weights to under-represented classes, as shown earlier in Table 1.

The dataset and source code used in this work are publicly available.[1]

6 Results & Analysis

Method	F1
BoW+MLP	0.7110
BERT FT	0.7193
BigBird FT	0.7131
Longformer FT	0.7544
BERT CFT, CLS-Pooling	0.7301
HT-BERT, CLS-Pooling	0.7200
HT-MAGPIE, Mean-Pooling	**0.7554**

Table 3: Results on the first feature set, **title + content**

Table 3 presents the results given articles' titles and content as features. HT-MAGPIE with mean-pooling outperforms the BoW+MLP baseline and achieves better F1 scores compared to fine-tuned BERT (BERT FT) and chunked fine-tuned BERT (BERT CFT) by 5.02% and 3.47% F1, respectively. While it only closely surpasses fine-tuned Longformer by 0.13%, HT-MAGPIE shows a notable 4.92% improvement over HT-BERT, highlighting the effectiveness of MAGPIE in capturing contextual representations for media bias classification. Interestingly, HT-BERT underperforms relative to BERT CFT, suggesting that simpler architectures may be more suitable when using BERT as the base encoder.

Despite these results, the BoW+MLP baseline only trails transformer-based methods by 1.17% to 6.24%. This modest performance gap may be attributed to the limited sample size. We hypothesise that with larger datasets, the performance difference between transformer-based and BoW-based models would become more pronounced.

The comparable performance of HT-MAGPIE and Longformer can be attributed to the short length of most articles in the dataset, with only 138 articles exceeding 4,096 tokens. Nonetheless, HT-MAGPIE still offers better scalability for longer texts and better computational efficiency through parallelised chunk-based processing.

Method	F1
Outlet Majority	**0.7959**
BERT FT	0.7504
BigBird FT	0.7318
Longformer FT	0.7529

Table 4: Results on the second feature set, **outlet + title + content**

Table 4 presents the results given outlet metadata as an additional feature, which translates into F1 improvements of 4.32% on fine-tuned BERT, and 2.62% on fine-tuned BigBird compared to the first feature set (Table 3). However, Longformer shows no improvements.

We acknowledge that the outlet majority baseline achieves a higher F1 score than HT-MAGPIE on the first feature set. However, this baseline ignores article content and relies solely on the outlet to infer bias, which is problematic as it effectively labels outlets and overlooks the fact that outlet positions can shift.

7 Conclusion

In this paper, we propose HT-MAGPIE as a novel approach for article-level media bias classification, highlighting the potential of hierarchical methods. MAGPIE's superior performance over BERT as a hierarchical encoder suggests that bias-pretrained models can create more effective representations. Furthermore, including outlet metadata as a supplementary feature generally enhances performance across fine-tuned models. Future work should explore more robust ways to leverage outlet metadata. We also publish a reconstructed BAT dataset consisting of full article content and a four-class label system based on article reliability scores, to be used for future media bias research.

[1] https://anonymous.4open.science/r/HT-MAGPIE-F4DF

8 Limitations

The BAT dataset relies on articles that are entirely curated by AdFontes, introducing a potential selection bias that could influence model performance. Moreover, the dataset remains small and imbalanced, particularly lacking examples for the *Problematic* class. Ensuring its robustness requires substantial future work, including a more thorough cleaning of the crawled article texts. Finally, our evaluation is conducted using a test set consisting of articles from outlets that are also present in the training set, leaving the models' ability to generalise to unseen sources unknown.

References

AdFontes. Home - ad fontes media. https://adfontesmedia.com//. Accessed: 2024-05-31.

Abien Fred Agarap. 2019. Deep learning using rectified linear units (relu). *Preprint*, arXiv:1803.08375.

Allsides. Allsides | balanced news and media bias ratings. https://www.allsides.com//. Accessed: 2024-07-16.

Ramy Baly, Georgi Karadzhov, Abdelrhman Saleh, James Glass, and Preslav Nakov. 2019. Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2109–2116, Minneapolis, Minnesota. Association for Computational Linguistics.

Dustin P. Calvillo, Abraham M. Rutchick, and Reagan J. B. Garcia. 2021. Individual differences in belief in fake news about election fraud after the 2020 u.s. election. *Behavioral Sciences*, 11(12):175.

M. Castillo-Campos, D. Becerra-Alonso, and H. G. Boomgaarden. 2025. Automated detection of media bias using artificial intelligence and natural language processing: A systematic review. *Social Science Computer Review*, 0(0).

Wei-Fan Chen, Khalid Al Khatib, Benno Stein, and Henning Wachsmuth. 2020a. Detecting media bias in news articles using Gaussian bias distributions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4290–4300, Online. Association for Computational Linguistics.

Wei-Fan Chen, Khalid Al Khatib, Henning Wachsmuth, and Benno Stein. 2020b. Analyzing political bias and unfairness in news articles at different levels of granularity. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 149–154, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In plain sight: Media bias through the lens of factual reporting. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

John Fields, Kevin Chovanec, and Praveen Madiraju. 2024. A survey of text classification with transformers: How wide? how large? how long? how accurate? how expensive? how safe? *IEEE Access*, 12:6518–6531.

Thomas J. Froehlich. 2024. Misinformation in the world today. *Kent State University, College of Communication & Information*. Accessed: 2024-07-24.

Soumen Ganguly, Juhi Kulshrestha, Jisun An, and Haewoon Kwak. 2020. Empirical evaluation of three common assumptions in building political media bias datasets. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):939–943.

Shijia Guo and Kenny Q. Zhu. 2022. Modeling multi-level context for informational bias detection by contrastive learning and sentential graph network. *CoRR*, abs/2201.10376.

Tomáš Horych, Martin Paul Wessel, Jan Philip Wahle, Terry Ruas, Jerome Waßmuth, André Greiner-Petter, Akiko Aizawa, Bela Gipp, and Timo Spinde. 2024. MAGPIE: Multi-task analysis of media-bias generalization with pre-trained identification of expressions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10903–10920, Torino, Italia. ELRA and ICCL.

Snehal Ishwar Khandve, Vedangi Kishor Wagh, Apurva Dinesh Wani, Isha Mandar Joshi, and Raviraj Bhuminand Joshi. 2022. Hierarchical neural network approaches for long document classification. In *2022 14th International Conference on Machine Learning and Computing (ICMLC)*, ICMLC 2022. ACM.

Michelle YoungJin Kim and Kristen Marie Johnson. 2022. CLoSE: Contrastive learning of subframe embeddings for political bias classification of news media. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2780–2793, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Vivek Kulkarni, Junting Ye, Steve Skiena, and William Yang Wang. 2018. Multi-view models for political ideology detection of news articles. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3518–3527, Brussels, Belgium. Association for Computational Linguistics.

Yuanyuan Lei, Ruihong Huang, Lu Wang, and Nick Beauchamp. 2022. Sentence-level media bias analysis informed by discourse structures. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10040–10050, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Iffat Maab, Edison Marrese-Taylor, and Yutaka Matsuo. 2023a. An effective approach for informational and lexical bias detection. In *Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER)*, pages 66–77, Dubrovnik, Croatia. Association for Computational Linguistics.

Iffat Maab, Edison Marrese-Taylor, and Yutaka Matsuo. 2023b. Target-aware contextual political bias detection in news. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 782–792, Nusa Dua, Bali. Association for Computational Linguistics.

Julie Mastrine. 2022. What is media bias? [Online; accessed 28-April-2024].

MBFC. Media bias fact check. Accessed: 2024-07-24.

Tim Menzner and Jochen L. Leidner. 2024. Improved models for media bias detection and subcategorization. In *Natural Language Processing and Information Systems: 29th International Conference on Applications of Natural Language to Information Systems, NLDB 2024, Turin, Italy, June 25–27, 2024, Proceedings, Part I*, page 181–196, Berlin, Heidelberg. Springer-Verlag.

Vanessa Otero. 2021. Multi-analyst content analysis methodology. Technical report, Ad Fontes Media. Revised September 2021.

Raghavendra Pappagari, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical transformers for long document classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 838–844.

Nihar M. Ranjan and Rajesh S. Prasad. 2023. A brief survey of text document classification algorithms and processes. *Journal of Data Mining and Management*.

Francisco-Javier Rodrigo-Ginés, Jorge Carrillo de Albornoz, and Laura Plaza. 2024. A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it. *Expert Systems with Applications*, 237:121641.

Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2021. Neural media bias detection using distant supervision with BABE - bias annotations by experts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1166–1177, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Timo Spinde, Elisabeth Richter, Martin Wessel, Juhi Kulshrestha, and Karsten Donnay. 2023. What do twitter comments tell about news article bias? assessing the impact of news article bias on its perception on twitter. *Online Social Networks and Media*, 37-38:100264.

Xin Su, Timothy Miller, Xiyu Ding, Majid Afshar, and Dmitriy Dligach. 2021. Classifying long clinical documents with pre-trained transformers. *Preprint*, arXiv:2105.06752.

Esther van den Berg and Katja Markert. 2020. Context in informational bias detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6315–6326, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jenny S Wang, Samar Haider, Amir Tohidi, Anushkaa Gupta, Yuxuan Zhang, Chris Callison-Burch, David Rothschild, and Duncan J Watts. 2025. Media bias detector: Designing and implementing a tool for real-time selection and framing bias analysis in news coverage. *arXiv [cs.HC]*.

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2021. Hi-transformer: Hierarchical interactive transformer for efficient and effective long document modeling. *ArXiv*, abs/2106.01040.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019. HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages

5059–5069, Florence, Italy. Association for Computational Linguistics.