Classification of Hyperpartisan News using BERT-Based Techniques

Abinesh Sivakumar
CEN
Amrita Vishwa Vidyapeetham
Coimbatore, Tamil Nadu
CB.EN.U4AIE20002

Jyothis V Santhosh
CEN
Amrita Vishwa Vidyapeetham
Coimbatore, Tamil Nadu
CB.EN.U4AIE20025

Logesh KSR
CEN
Amrita Vishwa Vidyapeetham
Coimbatore, Tamil Nadu
CB.EN.U4AIE20032

Manesh Karun R
CEN
Amrita Vishwa Vidyapeetham
Coimbatore, Tamil Nadu
CB.EN.U4AIE20035

Pranav Unnikrishnan
CEN
Amrita Vishwa Vidyapeetham
Coimbatore, Tamil Nadu
CB.EN.U4AIE20053

***Abstract*—Hyperpartisan news, characterized by an extreme, one-sided perspective, poses a significant challenge to media objectivity and fairness. This research aims to investigate the efficacy of BERT-based transformer models and Longformer in detecting and classifying hyperpartisan news, in an attempt to quantify the level of bias. Leveraging the publicly available 'Hyperpartisan News Detection' dataset created for PAN @ SemEval 2019 Task 4, we apply three specific approaches - Chunk Analysis, Summarization, and Stride Chunking - to evaluate each model's performance. Additionally, Longformer is used to juxtapose the results obtained from the aforementioned techniques applied to BERT, thereby evaluating its merits and potential. This comparative study can be instrumental for future improvement of these models to enhance their accuracy and reduce their bias in news classification tasks.**

I. Objective

This study's main goal is to categorise hyperpartisan news and determine the degree of bias within those categorizations by a careful analysis of BERT-based models in conjunction with Longformer. The objective of the study is to carefully assess the strengths and weaknesses of each model with regard to the identification and classification of hyperpartisan news. Distilling knowledge that could play a crucial role in fostering the evolution of these models towards greater accuracy and objectivity is an essential part of this goal.

II. Introduction

In the current era of digital information, discerning objective reporting from hyperpartisan news, which denotes extreme bias in news content, has become crucial. The capability to detect and measure bias in news plays a vital role in preserving a balanced media environment and fostering informed discourse in democratic societies. This research delves into the utilization of BERT-based transformer models and the Longformer model for the identification and categorization of hyperpartisan news, contributing to the efforts in managing media bias.

The implications of this study extend beyond academia. The developed methodologies can provide tools to various entities, such as media organizations, policymakers, news aggregators, and educators, thereby enabling them to maintain or assess the balance of news content. Moreover, by informing the readership about potential biases, we can aid in improving media literacy and nurturing a discerning audience.

Furthermore, this research has a broad scope in the real world, paving the way for future advancements in media bias detection technology. By comparing different techniques and models, we establish a basis for the continuous evolution of these tools to match the growing demand for transparency in media. As such, this study serves as a significant stride in the application of Natural Language Processing within the realm of media analysis and beyond.

III. Literature Review

There has been an increase in recent years in NLP research pertaining to legal documents. [1] describes how to extract machine-readable rules from legal documents by combining linguistic data from WordNet with a syntax-based retrieval of rules from legal literature combined with a logic-based retrieval of dependencies from portions of such texts. [2] provides examples of several embedding-based and symbol-based approaches for case matching, judgement prediction, and answering legal questions. [3] covers five legal activities where NLP has a significant impact. These include contract review to ensure that a contract is complete, document automation to create standard legal documents, and question-and-answer dialogues for legal advice. Legal research seeks out

information pertinent to a legal decision. Electronic discovery determines the relevance of documents in an information request. [4] examines the performances of various ways for summarising legal texts and discusses the many options that are currently accessible. [5] presents a scalable and adaptable information extraction method that takes context into account to extract information from legal documents independent of format, layout, or structure. [6] develops a BERT model for German law and assesses how well it performs on downstream NLP tasks like classification, regression, and similarity.

In several NLP tasks, BERT-based models have demonstrated ground-breaking performance. They are now frequently employed in the legal sector as well. In order to capture the semantic links at the paragraph level, [7] suggests using BERT-PLI. Then, by aggregating paragraph-level interactions, the author infers the relevance between two legal cases. To support legal NLP research, [8] introduces Legal-BERT, a collection of BERT models for the legal area. [9] analyses a situation in the context of hiring a legal expert and shows how a BERT-based technique performs better than other conventional methods. BERT on their corpus of Vietnamese legal questions and answers. They also pre-train BERT on a corpus specialised to the Vietnamese legal domain, demonstrating that this new BERT outperforms the fine-tuned BERT. Legal document classification is a crucial NLP activity that can automate the alignment of legal documents with categories that have been established by humans. [10] uses BERT to categorise a proprietary corpus made up of tens of thousands of legal contracts. [11] analyses binary and multi-label categorization of legal documents utilising pre-trained BERT-based model variations as well as additional methods for dealing with lengthy documents. Although we employ a different collection of legal papers with vastly different features, our study follows a somewhat similar path. Additionally, many of our BERT-based methods differ greatly from [11].

IV. Dataset Description

TABLE I
Data Statistics

Metric	Value
Dataset Size	10000
Min # Tokens (Unprocessed)	2589
Max # Tokens (Unprocessed)	3940
Median # Tokens (Unprocessed)	3023
Mean # Tokens (Unprocessed)	3091.6313
Min # Tokens After Pre-Processing	2022
Max # Tokens After Pre-Processing	3935
Median # Tokens After Pre-Processing	2984
Mean # Tokens After Pre-Processing	3052.495

The Dataset used is the publicly available 'Hyperpartisan News Detection' created for PAN @ SemEval 2019 Task 4. Upon evaluating a piece of news article text, it's imperative to ascertain whether it embodies hyperpartisan argumentation, that is to say, whether it displays an indiscriminate, prejudiced, or irrational devotion to a particular party, faction, cause, or individual. This assessment serves a key role in identifying and analyzing potential biases within media content.

- byarticle: Labeled through crowdsourcing on an article basis. The data contains only articles for which a consensus among the crowdsourcing workers existed.
- bypublisher: Labeled by the overall bias of the publisher as provided by BuzzFeed journalists or MediaBiasFactCheck.com.

V. Proposed Techniques

In this research study, we suggest the use of a variety of approaches to enhance the processing and comprehension of document text, principally utilising BERT-based methodologies and Longformer models.

1) Chunk Analysis: In the Hyperpartisan dataset, we refer to each 512-token segment of a given text as c_i, where 'i' denotes the segment's position in the sequence. We adhere to the maximum sequence length that BERT-based models can manage by setting the segment length to 512 tokens. We evaluate the performance of our models on these segments, allowing us to determine which portions of the document have the most influence on categorization accuracy. We assess the effectiveness of this BERT-based model on the first six segments (c_1 through c_6), where c_1, c_2, etc., represent the first, second, and so on, 512-token segments of the documents. The median document length in the Hyperpartisan dataset is roughly 3000 tokens. For documents with a length less than $i \times 512$ tokens, we choose the final 512 tokens or fewer (in the case where i equals 1) for evaluation.

2) Summarization: In this approach, we condense the documents from the Hyperpartisan dataset into summaries of 512 tokens each. We employ the summarization pipeline from Hugging Face with its default parameters. The maximum sequence length this summarization model can accommodate is 1024 tokens. Therefore, we initially divide a document into several splits based on its length. We determine the number of splits n_i for a given document d_i as l_i/1024, where l_i is the document's length. As the total length of a document's summarized version cannot exceed 512 tokens, we additionally compute the number of tokens per split nw_i for all splits of a specific document d_i. We then concatenate all nw_i's from a given document d_i to form the final summarized version. We use these summarized versions for both classification tasks.

3) Stride Chunking: In our approach, we refer to each 512-token segment of a given document from the Hyperpartisan dataset as c_i, where 'i' indicates the position of the segment in the sequence. The 'stride' in this context refers to the block of tokens that are shared between any two segments, c_i and c_{i+1}. As an example, let's take a stride value of 64. The tokens in c_1 and c_2, specifically $c_1[448:512]$ and $c_2[0:64]$, are identical if $c_i[0:512]$ signifies the 512 tokens present in c_i. In the case of a document d_i that is 1024 tokens in length, we would have three c_i's, as $c_1[448:512]$ equals $c_2[0:64]$ and $c_2[448:512]$

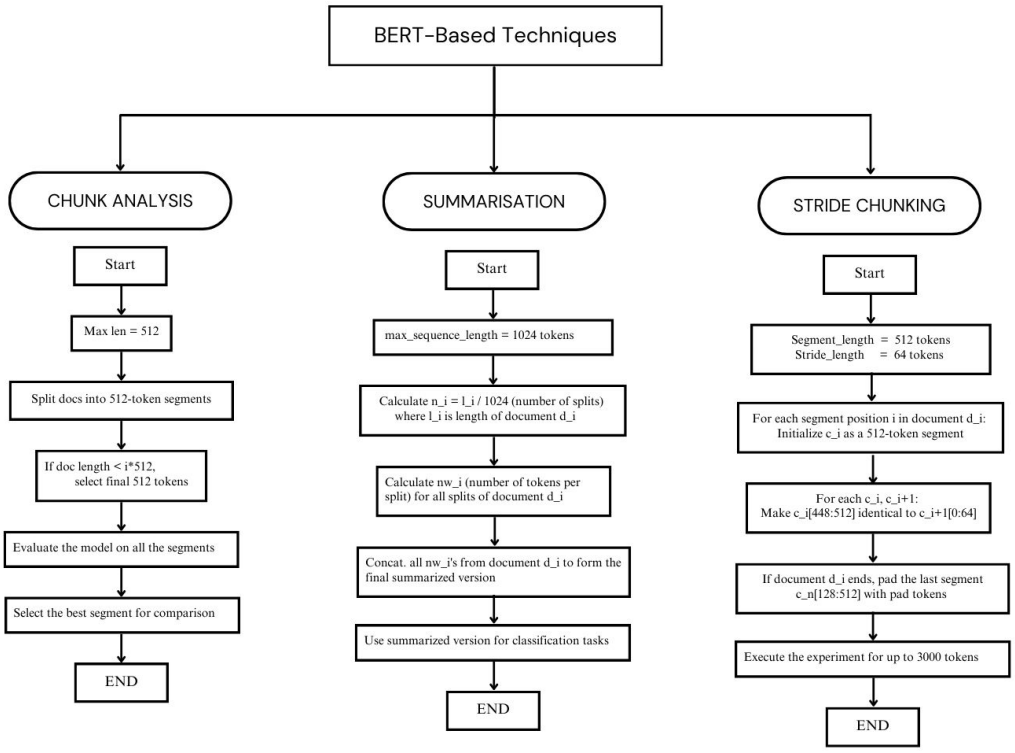

Fig. 1. Flowchart - BERT

equals $c_3[0:64]$. For instance, $d_i[0:512]$ is equivalent to $c_1[0:512]$, $d_i[448:512]$ is equivalent to $c_2[0:64]$, $d_i[512:960]$ is equivalent to $c_2[64:512]$, $d_i[896:960]$ is equivalent to $c_3[0:64]$, and $d_i[960:1024]$ equals $c_3[64:128]$. Pad tokens are found in $c_3[128:512]$. We execute our experiment up to about 3000 tokens, which mirrors the median length of documents in the Hyperpartisan dataset.

A. Longformer

In our study, we also experimented with models capable of accepting input sequences exceeding the standard 512-token limit. One such model is the Longformer. This model is specifically designed for longer text sequences, allowing for a more comprehensive analysis of larger documents in the Hyperpartisan dataset.

VI. Experiments

A. Evaluation method

We use weighted F1, accuracy and weighted precision as our evaluation metrics for both the classification tasks.

B. Experimental Details

With the exception of RoBERTa, we used the following hyperparameters for all of the aforementioned methods:

- batch size: 8
- epochs: 5
- learning rate: 3e-5
- loss function: binary cross entropy

When utilizing RoBERTa, we kept all the hyperparameters the same, with the exception of the learning rate, which we adjusted to 1e-5. Our dataset, which included 500 data points, was split into training and testing sets in a 80:20 ratio. Each experiment was run five times, and the best period scores were averaged to produce the final score. The loss function we utilised was Mean Squared Error (MSE), and Adam was the optimizer we used. By rounding the regression output to the nearest integer, the accuracy was calculated.

C. Results

TABLE II
Results for Best-512

Chunk	Results		
	Accuracy	Precision	F1
Chunk 1	**0.949**	**0.949**	**0.946**
Chunk 2	0.924	0.910	0.924
Chunk 3	0.926	0.920	0.926
Chunk 4	0.925	0.924	0.915
Chunk 5	0.929	0.922	0.929
Chunk 6	0.931	0.931	0.923

TABLE III
RESULTS FOR 5 CATEGORIES

Technique	Model	Metrics		
		Accuracy	Precision	F1
Chunk Analysis	BERT	0.949	0.949	0.946
	RoBERTa	**0.955**	**0.953**	**0.955**
Summarization	BERT	0.943	0.939	0.944
	RoBERTa	0.954	0.940	0.944
Stride Chunking (64)	BERT	0.920	0.870	0.889
	RoBERTa	0.920	0.870	0.920
Longer Sequence Model	Longformer	0.930	0.923	0.931

D. Analysis

In our research paper, we conducted a detailed analysis of the misclassified documents, examining the distribution of biases within them. Out of a total dataset of 10,000 documents, we identified 164 instances of misclassification. Our study aimed to uncover potential avenues for improving model performance and generalization in news article classification. Through a thorough analysis, we investigated the underlying patterns and factors that contribute to misclassification. Our objective was to enhance the model's ability to accurately classify a wider range of news articles, leading to improved performance and more robust generalization capabilities.

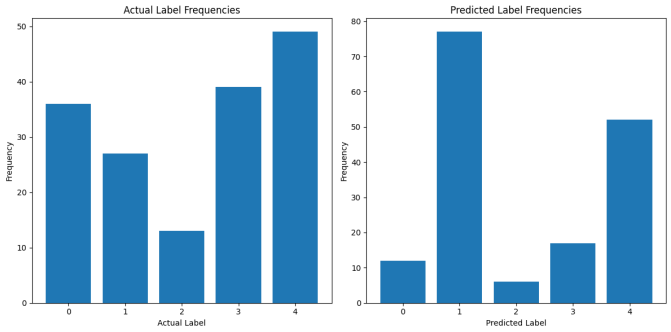

Fig. 2. Distribution of biases over Actual vs Predicted

Based on these statistics, the following inferences can be drawn:

- The predicted labels do not perfectly match the actual labels, indicating that the model makes some errors in classification.
- Labels 4 and 3 are frequently predicted and observed in the actual labels, suggesting a reasonable level of agreement between the model's predictions and the ground truth.
- Labels 0 and 2 have lower counts in both the predicted and actual labels, indicating that they may be more challenging for the model to accurately classify.
- **Increasing the training data specifically for labels with high misclassification rates has the potential to improve the model's performance. By providing more examples of these challenging labels, the model can learn more accurate representations and potentially achieve better generalization.**

VII. CONCLUSION

This research evaluated the efficacy of BERT-based transformer models and Longformer for the detection and classification of hyperpartisan news. Among the techniques used - Chunk Analysis, Summarization, and Stride Chunking - Chunk Analysis demonstrated the highest accuracy. Furthermore, this BERT-based chunking method proved to be more accurate even when compared to the Longformer approach, which processes the entire dataset.

This study underscores the significance of data segmentation in improving classification accuracy. While no new methods were introduced, the research offers invaluable insights for future studies, emphasizing the potential of optimizing Chunk Analysis and exploring other models to further enhance performance in classifying hyperpartisan news.

REFERENCES

[1] M. Dragoni, S. Villata, W. Rizzi, and G. Governatori, "Combining nlp approaches for rule extraction from legal documents," in 1st Workshop on MIning and REasoning with Legal texts (MIREL 2016), 2016.

[2] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, and M. Sun, "How does nlp benefit legal system: A summary of legal artificial intelligence," arXiv preprint arXiv:2004.12158, 2020.

[3] R. Dale, "Law and word order: Nlp in legal tech," Natural Language Engineering, vol. 25, no. 1, pp. 211–217, 2019.

[4] A. Kanapala, S. Pal, and R. Pamula, "Text summarization from legal documents: a survey," Artificial Intelligence Review, vol. 51, no. 3, pp. 371–402, 2019.

[5] M. Garcia-Constantino, K. Atkinson, D. Bollegala, K. Chapman, F. Coenen, C. Roberts, and K. Robson, "Cliel: context-based information extraction from commercial law documents," in Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law, 2017, pp. 79–87.

[6] C. M. Yeung, "Effects of inserting domain vocabulary and fine-tuning bert for german legal language," Master's thesis, University of Twente, 2019.

[7] Y. Shao, J. Mao, Y. Liu, W. Ma, K. Satoh, M. Zhang, and S. Ma, "Bert-pli: Modeling paragraph-level interactions for legal case retrieval." in IJCAI, 2020, pp. 3501–3507.

[8] I.Chalkidis,M.Fergadiotis,P.Malakasiotis,N.Aletras,andI.Androutsopoulos, "Legal-bert: The muppets straight out of law school," arXiv preprint arXiv:2010.02559, 2020.

[9] L. Sanchez, J. He, J. Manotumruksa, D. Albakour, M. Martinez, and A. Lipani, "Easing legal news monitoring with learning to rank and bert," in European Conference on Information Retrieval. Springer, 2020, pp. 336–343.

[10] C.-N. Chau, T.-S. Nguyen, and L.-M. Nguyen, "Vnlawbert: A vietnamese legal answer selection approach using bert language model," in 2020 7th NAFOSTED Conference on Information and Computer Science (NICS). IEEE, 2020, pp. 298–301.

[11] E. Elwany, D. Moore, and G. Oberoi, "Bert goes to law school: Quantifying the competitive advantage of access to large legal corpora in contract understanding," arXiv preprint arXiv:1911.00473, 2019.