Detecting Political Bias in News Articles Using BERT: A Refined Approach

Jian M. Wu
jianmw123@berkeley.edu

Mehmet H. Inönü
minonu@berkeley.edu

Abstract

This study explores the application of Bidirectional Encoder Representations from Transformers (BERT) for classifying the political ideology of digital news articles. Building on the work of Baly et al. (2020), we fine-tune a BERT uncased model (Wolf et al., 2020) from Huggingface using Baly et al.'s refined, imbalanced label dataset. Our approach incorporates enhanced data preprocessing techniques and explores various BERT model architectures, hyperparameter configurations, and training/test set variations. The results showcase a substantial improvement over previous findings and the baseline model, achieving a weighted average F1 score of 91% in categorizing news articles as left, center, or right on the political spectrum. As a result, this research improves the accuracy and effectiveness of computational methods for detecting political bias in digital news media.

1 Introduction

In the past decade, the explosion of digital news media has altered how people access and engage with information. This shift has raised significant concerns about the presence of political biases in news content (Stubbs, 2020), which can distort public opinion and influence political decision-making. Detecting these biases, which are often concealed in subtle language choices and narrative strategies, can be challenging for readers. As readers increasingly turn to online sources for news, they are exposed to biased content that may distort their understanding of social events, political figures, and policy issues, potentially contributing to societal polarization.

Traditional methods for detecting bias, such as AllSides.com's manual bias rating system, have limitations. The rapid spread and overabundance of online information outstrip human evaluators' capabilities, and manual systems often fail to capture the full range of linguistic subtleties and contextual subtleties. This highlights the need for more advanced, automated approaches to analyze digital news articles and identify potential biases more effectively and efficiently.

Automated methods in Natural Language Processing (NLP) have demonstrated potential, but many still rely on predefined lexicons, basic bag-of-words models, limited datasets, or rudimentary neural networks, which often fall short in capturing the subtleties of political bias. While these approaches provided a foundation for automated text analysis, they are inadequate for detecting political bias due to their struggle with the complex, non-linear, and context-sensitive nature of bias in text.

Building on Baly et al.'s (2020) application of the BERT model for bias detection, this study seeks to advance the approach by investigating various BERT architectures, configurations, and supplementary data processing techniques. Our objective is to raise the reader's awareness of potential political bias in the news they consume and to improve prediction accuracy and create a more robust and generalizable tool for detecting political bias across a diverse array of news articles.

2 Background

In the NLP field's early work, Pang and Lee (2004) utilized machine learning techniques like Support Vector Machines and Naive Bayes for text classification, they outlay the groundwork for automated bias detection. However, these methods often struggled to capture the nuanced and context-dependent nature of political bias in textual content.

The advent of deep learning architectures marked a significant advancement in the field. Iyyer et al. (2014) employed RNNs to detect political ideology at the sentence and phrase levels has achieved promising results. Building on this, Chen et al. (2020) further refined RNN-based approaches to analyze longer text sequences and identify specific words or sentences contributing to bias.

The introduction of BERT by Devlin et al. (2019)

revolutionized various NLP tasks. Baly et al. (2020) leveraged BERT's contextual embeddings and transformer architecture to predict political ideology at the news article level, demonstrating improved performance compared to traditional methods. They also explored techniques to mitigate potential biases stemming from news sources and topics.

More recently, Hong et al. (2023) addressed the issue of domain dependencies in political bias detection by proposing a multi-head hierarchical attention model that combines BERT embeddings with BiLSTM and attention mechanisms.

This research builds upon these prior studies, particularly leveraging the BERT model as utilized by Baly et al. (2020). However, it aims to surpass previous results through innovations in data preprocessing, model architecture exploration, and diverse training/testing strategies.

3 Methods

3.1 Fine-Tuning BERT Model with Domain-Specific Knowledge for Multi-Label Classification in Political Bias

Our research involves fine-tuning the BERT model (Wolf et al., 2020) using domain-specific knowledge from the digital news article dataset provided by Baly et al. (2020), along with additional data processing and cleaning. We also explored various BERT architectures and hyperparameter settings to improve performance on the political ideology detection task. This multi-label classification problem requires assigning political ideological labels (e.g., left, right, or center) to digital news articles.

3.2 Illustrative Example of the Political Bias Problem and Intuition for Solution

Consider two hypothetical news headlines and their contents covering the same topic (generated by ChatGPT 3.5):

1. Headline: "Progressive Tax Plan Unveiled to Address Wealth Inequality"
   Content: "The new tax plan proposes higher rates for top earners and corporations, aiming to fund social programs and reduce wealth inequality. Proponents argue it's a fair approach, while critics worry about its impact on economic growth."

2. Headline: "Tax Hike Proposal Threatens Economic Growth, Critics Warn"
   Content: "A proposed tax increase on high-income earners and corporations is facing opposition from business leaders, who warn it could harm innovation and job creation. Critics fear it may lead to capital flight and economic downturns, while supporters dismiss these concerns."

The above examples illustrate how news articles on the same topic can employ different linguistic and discourse patterns to reveal political biases. Moreover, it's important to note that news articles often have considerable length, with political opinions subtly embedded throughout the content. Without a nuanced understanding of different political standpoints, ordinary readers may interpret the same legislation differently, and this can skew public perception and understanding.

3.3 Experimental Design

To advance the performance of the BERT uncased model (Wolf et al., 2020) for multi-label classification of political bias, we undertook a comprehensive data cleaning and preprocessing process on Baly et.al (2020)'s dataset. This involved:

1. **Excluding Incomplete Data:** articles missing essential information, such as publication dates, topics, sources, titles, content, or bias labels, were removed.

2. **Text Normalization:** applied text normalization techniques including stop word removal, duplicate elimination, expand the abbreviations, lowercase conversion, and lemmatization to the dataset.

3. **Source Anonymization:** efforts were made to obscure specific publication sources to minimize the risk of the BERT model identifying and learning patterns unique to particular media outlets.

The preprocessed dataset, with 12,000 unqualified rows excluded, is then split into training, validation, and test sets for the fine-tuning process. Our experiment includes:

Baseline Model: we established a baseline model for quick insights using a Naive Bayes classification on our preprocessed dataset. This model facilitated initial assessments of dataset quality and allowed for the identification of significant n-grams associated with different bias classes, despite its limited contextual capabilities.

Fine-Tuning the BERT Uncased Model: we employed early stopping to prevent overfitting and conducted comprehensive hyperparameter optimizations. Key parameters included dropout rates, maximum token length, number of trainable layers, hidden layer dimensions, learning rates, and training epochs. These adjustments were aimed at maximizing model performance and generalizability (Appendix Table 4).

BERT Model Architectures: we explored potential improvements in prediction performance from the perspective of BERT's architecture by integrating CNN and LSTM layer separately on top of BERT's output layer. Additionally, we evaluated the performance of using [CLS] versus [Pooler] tokens for sequence representation, evaluating their impact on capturing nuanced biases.

T**raining and Test Set Variations**: we also implemented diverse training and test set configurations to: i) continue seeking improvements in prediction performance, and ii) evaluate the generalizability of the fine-tuned BERT model.

- **Temporal Variations**: fine-tuned the BERT model on articles spanning a 4-year period of non-election years and tested it with news from a different period in the test set.
- **Event-Specific Training**: fine-tuned the BERT model exclusively on articles from the 2016 election year and tested it with news articles from the same test set, excluding those from 2016.
- **Segmented Training**: fine-tuned the BERT mode with the first six sentences of articles (as the first paragraph) and tested it with news articles from the same test set.
- **Exclusion of High-Bias Sources Training**: fine-tuned the BERT model excluding news articles published by the top 5 news organizations that consistently led to incorrect predictions. We also tested the model with those excluded news articles.

Label Rebalancing: due to a significant imbalance in label distribution among different news organizations, we applied dataset rebalancing to improve prediction performance. This involved constraining the number of articles per organization to a range of 10 to 400.

3.4 Rationale for Experiments

Our experiments are designed to enhance the BERT uncased model's ability to accurately classify political bias by addressing both internal and external factors. At the model level, figuring out the optimal mode configuration and integrating a CNN layer or LSTM layer aims to improve the model's capacity to capture both local and long-range dependencies, which are critical for detecting subtle biases. For the dataset, comprehensive data preprocessing, including cleaning, standardizing, and rebalancing, is intended to provide high-quality, balanced input, reducing potential bias from uneven data representation. Together, these approaches are expected to refine the model's precision and generalization across diverse contexts.

3.5 Measuring Success

Given the imbalanced nature of our dataset, we use the weighted average F1 score as our primary metric. This metric offers a balanced evaluation of performance across all bias classes, reflecting both precision and recall. Complementing these quantitative measures, we also conduct qualitative analysis on a small sample of predictions to evaluate the model's interpretability and ensure that its decisions are coherent and insightful.

4 Result and Discussion

4.1 From Naive Bayes to Fine-Tuned BERT

In our analysis, the fine-tuned BERT uncased model, optimized with domain-specific data and rigorous preprocessing, achieved an impressive **weighted average F1 score of 91%** (Table 1). The result represents a significant 22 to 23 percentage point improvement over the baseline Naive Bayes model. Such an enhancement highlights BERT's capability to leverage contextual embeddings for nuanced word interpretations and its transformer architecture to model complex patterns and long-range dependencies. The domain-specific fine-tuning further refined these abilities, maximizing the model's overall classification performance.

Moreover, further cleaning of the dataset significantly enhanced the BERT model's performance. Removing incomplete articles and duplicates ensured that the model was trained on high-quality, reliable examples, thereby reducing noise and improving generalization. Text normalization helped the model focus on meaningful content and semantic relationships, while source anonymization

reduced bias by obscuring specific publication sources, which in turn facilitated more objective classifications. Collectively, these refinements led to a 10% increase in both accuracy and F1 score compared to the results reported by Baly et al. (2020).

4.2 Exploring Different BERT Model Architectures (Appendix-Table 5 for Full Experiment Results)

Maintaining consistent BERT model configurations and datasets across various experiments, we observed that incorporating an LSTM layer with an additional affine layer on top of the BERT final output layer led to improved results. Both the fine-tuned BERT model and the fine-tuned BERT model with LSTM [CLS] achieved an **overall weighted average F1 score of 91.00%** (Table 2). However, they differed in terms of the total number of incorrect predictions. The BERT model with the LSTM layer had 266 incorrect predictions out of 5,029, while the standalone fine-tuned BERT model had 287 incorrect predictions.

As illustrated in Figures 3 and 4 of the Appendix, the fine-tuned BERT model with a LSTM layer also showed slightly improved precision across each label class. This enhancement can be attributed to the LSTM's capacity to capture sequential dependencies within the text, which complements BERT's contextualized representations and potentially boosts the model's ability to learn task-specific patterns. The use of the [CLS] token is particularly advantageous here as it typically aggregates information across the entire sequence, making it more effective for classification tasks compared to the Pooler token, which focuses on fixed token representations.

We also discovered that processing the entire news article text up to the BERT model's maximum 512 tokens may not be necessary. Instead, providing just the initial portion of the article—in our case, the first six sentences—proved to be sufficient. This approach leverages the conventional newspaper structure, where key information is typically presented in the opening paragraphs. Consequently, the model achieved a promising **weighted average F1 score of 87%**.

4.3 BERT with LSTM Layer's Learning Curve

Despite implementing L2 regularization and experimenting with various dropout rates—30% overall dropout, 20% attention head dropout, and 25% hidden layer dropout—along with employing early stopping, the learning curve revealed that the BERT model began to overfit around the second epoch of fine-tuning (Figure 1). This suggests that the model had likely extracted all the learnable information from the training set within those two epochs. Given the large size of the training set, the BERT uncased model's 512-token limit may have constrained its ability to fully capture the diverse context of longer articles.

In addition, we investigated the sources of the news articles in our dataset. We found that the top 30 publishers were responsible for 91.51% of the total publications in the training set (Appendix - Figure 5), which included 18,406 news articles. This left only 8.49% of the articles distributed among the remaining 263 news organizations. Notably, 193 of these 263 organizations published fewer than 10 articles over the eight-year period covered by Baly et al.'s (2020) dataset.

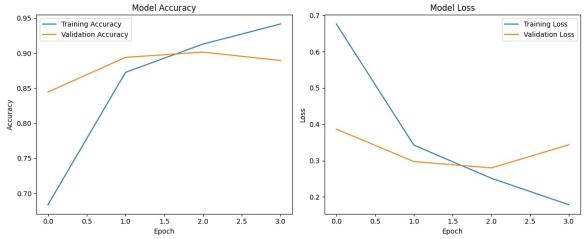

Figure 1: BERT with LSTM Layer's Learning Curve

4.4 Impact of Training BERT With Various Data Set Sizes

We also discovered that using only the digital news articles from the election year 2016, which we assumed covered a broader range of controversies and diverse voices, yielded a respectable accuracy of 79% on the test set. This result is comparable to the 79% accuracy achieved by Baly (2020) using a training set of 30,000 samples, while we utilized only 2,893 samples (one-tenth of Baly's dataset) to fine tune the model.

Building on these findings, we extended our experiment by incorporating a four-year span of articles from non-election years (2017, 2018, 2021, and 2022). We then tested the model using the same test dataset, but excluded articles from these four years. This adjustment resulted in an even higher weighted average F1 score of 80% (Table 3). This improvement highlights that the model's accuracy is not solely dependent on the quality of the train-

Table 1: Comparing our chosen model to the baseline model and the reference paper's model

Experiments	Train/Test Size	Accuracy	Weighted Average F1 Score
Naive Bayes Classification (Baseline Model)	20,114/5,029	69.00%	68.00%
BERT Uncased Fine-Tuning (CLS Token)	**20,114/5,029**	**91.00%**	**91.00%**
Baly et. al (2020)'s Fine-Tuned BERT (Random Split Dataset)	33,537/1,200	79.83%	80.19%

Table 2: Exploring different BERT model architectures

Experiments	Train/Test Size	Accuracy	Weighted Average F1 Score
BERT CLS Token (12 NN Layers)			
Fine-Tuned BERT	20,114/5,029	91.00%	91.00%
Fine-Tuned BERT with a LSTM Layer	**20,114/5,029**	**91.00%**	**91.00%**
Fine-Tuned BERT with a CNN Layer	20,114/5,029	83.00%	83.00%
BERT with Fine-Tune with the First Six Sentences			
Fine-Tuned BERT with First Six Sentences	20,114/5,029	87.00%	87.00%

ing data but also on its size and diversity.

We also fine-tuned a BERT model excluding from the top 5 news sources and tested with those excluded news articles as unseen data in three different batches (Table 3) The model's performance was only 40.00%, indicating that it still heavily relies on absorbing and adapting to different text styles and contextual information. The result suggests that the model may not have generalized well across varied contexts or sufficiently learned to handle the diverse styles present in the unseen data.

Regarding efficiency and resource costs, Baly's (2020) study reported that one epoch required 22 minutes to complete using four Titan X Pascal GPUs. In contrast, our model, trained exclusively on the 2016 data, completed each epoch in approximately 4 minutes on a Google Colab with an A100 GPU. Additionally, the learning curve for the 2016 data demonstrated healthy model training, as illustrated in Figure 6 of the Appendix.

4.5 Error Analysis

4.5.1 Publication Sources, Topics, and Authors

Linking publication sources, topics, and authors with their corresponding digital news articles, we found that the fine-tuned BERT model consistently made incorrect predictions in three main areas:

1. News articles from certain organizations, including the Washington Times, CNN (web news), NPR Online, Politico, and Vox across different model architectures (Appendix Figure 7).
2. News article topics related to elections, politics, healthcare, immigration, and the US House of Representatives (Appendix Figure 9).
3. Articles by certain authors, though none had more than 5 incorrectly predicted articles (Appendix Figure 10).

Upon examining a sample batch of errors, we could not identify clear patterns explaining the model's incorrect predictions. However, we observed a notable correlation between the likelihood of errors and the volume of articles from specific sources, despite not revealing the publication source during fine-tuning. The error rate generally increased with the number of articles from a given source, roughly proportional to their total count. However when we look into more details, we found that CNN (web news) contributed 511 articles to the test set (Appendix Figure 8), with only 28 classified incorrectly (Appendix Figure 5), resulting in a robust accuracy rate of approximately 95% for this source. This strong performance indicates that the model's effectiveness varies by publication source, suggesting the presence of source-specific patterns and highlighting generalization issues. Specifically, the model might have learned implicit biases related to the frequency and characteristics of articles from different sources, affecting its overall accuracy.

Table 3: *Note, the test datasets excluded the year(s) selected for the model training

Experiments*	**Train/Test Size***	**Weighted Average F1 Score**
Fine-Tuned BERT with 2016 Data (Pooler)	2,893/4,407	79.00%
Fine-Tuned BERT **with 4-Year Data (CLS)**	**5,977/622**	**80.00%**
Fine-Tuned BERT without Top-5 Sources and Test on those Unseen Top-5 Sources' Articles	11,332/2,000 (3 batches)	40.00%

4.5.2 Smaller Size BERT Model with 512 Tokens Limitation

From the error analysis, the results revealed that the BERT uncased model (Wolf et al., 2020) has limitations in accurately classifying political ideology, primarily due to its 512-token constraint. Additionally, the smaller size of the BERT model with only 110 million parameters lacks the depth of knowledge and capacity found in larger BERT models.

Reviewing incorrectly classified news articles (Sample, Appendix Figure 12), we found that key political ideological signals often require reasoning through the entire article and frequently appear in the latter portions, beyond the model's token limit. As a result, BERT typically processes only the introductory statements or background information, which are insufficient for accurately determining political bias. This constraint likely contributes to BERT's misclassifications, as it struggles to capture nuanced writing styles and critical information necessary for precise ideological assessment.

Figure 2 illustrates that among the incorrectly predicted labels of news articles with fewer than 10,000 words, none were shorter than 500 words.

To investigate this issue further, we compared

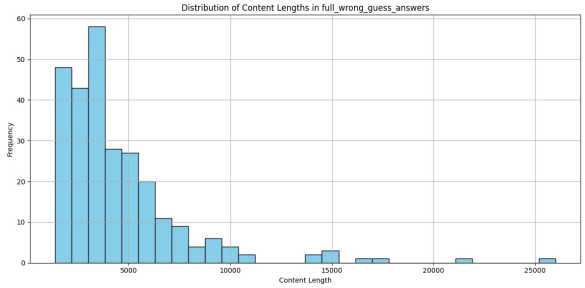

Figure 2: Distirbution of news articles' content length

BERT's performance with other larger language models that feature expanded context windows, such as ChatGPT 3.5 (200k tokens) and Claude 3.5 Sonnet (8k tokens). Both models (Appendix Figure 13), using the same input text, not only accurately predicted political bias but also provided comprehensive summaries and reasoning for their assessments of entire articles. This comparison supports our hypothesis that BERT's token limit restricts its ability to capture key ideological indicators in longer or more complex articles, resulting in misclassification.

4.5.3 News Articles' Publication Date

The likelihood of incorrect predictions showed no significant variation by publication year. The model maintained consistent performance across different years, including election and non-election periods (Appendix Figure 11). This stability indicates that the model is broadly applicable, regardless of the publication date.

4.5.4 BERT Model with Filtered Datasets

Refining our BERT model by removing articles from the top five organizations, thus reducing the dataset by 50%, achieved an accuracy of around 89% and significantly cut training time from 15 to 3.5 minutes per epoch. However, this reduction raises concerns about generalizability as the results shown in 4.4. Excluding a large portion of data could introduce biases and limit the model's exposure to diverse perspectives, potentially impacting its ability to handle a wide range of sources and nuanced ideological signals.

In the second experiment, we rebalanced the dataset by capping articles at 400 per organization and excluding those with fewer than 10 articles, resulting in a total of approximately 10,000 articles. This adjustment led to a decrease in the weighted average F1 score to 86%. This outcome indicates that achieving a 90% accuracy may require more domain-specific data, highlighting the trade-off between dataset diversity and model performance.

5 Conclusion and Future Work

Our research significantly advances the field of political bias detection by fine-tuning the BERT

model, achieving a remarkable 91% weighted average F1 score. The study's success showcases BERT's capacity to discern subtle linguistic cues and contextual nuances crucial for identifying political bias in digital news. The improved accuracy compared to previous methods highlights the effectiveness of the refined dataset and optimized model configurations.

However, challenges remain. The model's performance on non-traditional news formats and its sensitivity to dataset imbalances suggest areas for future improvement. Addressing these limitations will further enhance the model's robustness and generalizability, ensuring its applicability across a broader spectrum of news sources and styles.

6 References

AllSides. (n.d.). Media bias rating methods. Retrieved July 28, 2024, from https://www.allsides.com/media-bias/media-bias-rating-methods.

Bo Pang and Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), pages 271–278, Barcelona, Spain.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, Philip Resnik. 2014. Political Ideology Detection Using Recursive Neural Network. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 1113–1122.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (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 (pp. 4171-4186).

Baly, Ramy and Da San Martino, Giovanni and Glass, James and Nakov, Preslav. 2020. We Can Detect Your Bias: Predicting the Political Ideology of News Articles. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). EMNLP '20, pages 4982–4991.

Hong, J., Cho, Y., Jung, J., Han, J., & Thorne, J. (2023). Disentangling structure and style: Political bias detection in news by inducing document hierarchy. In Findings of the Association for Computational Linguistics: EMNLP 2023 (pp. 5664–5686). Association for Computational Linguistics.

Wei-Fan Chen, Khalid Al Khatib, Henning Wachsmuth, and Benno Stein. 2020. Analyzing political biasand 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.

AlDahoul, N., Rahwan, T., & Zaki, Y. (2024). A novel BERT-based classifier to detect political leaning of YouTube videos based on their titles. arXiv preprint arXiv:2404.04261v1.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020). Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 38-45).

Stubbs, H. (2020, September 10). Bias in others' news a greater concern than bias in own news. Gallup News. https://news.gallup.com/poll/319724/bias-others-news-greater-concern-bias-own-news.aspx.

7 Appendix

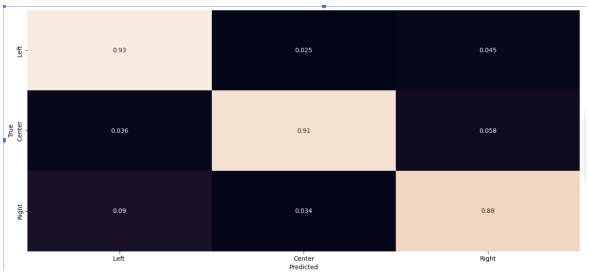

Figure 3: BERT Confusion Matrix

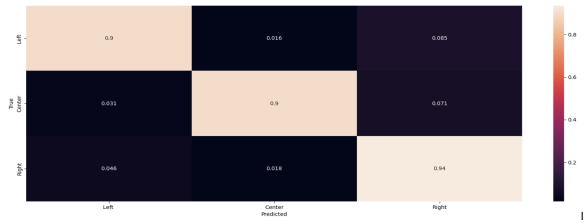

Figure 4: BERT with LSTM Layer Confusion Matrix

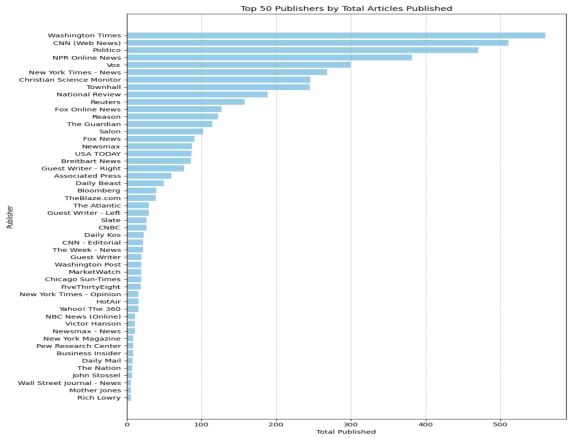

Figure 5: Top 30 Publishers of the News Articles

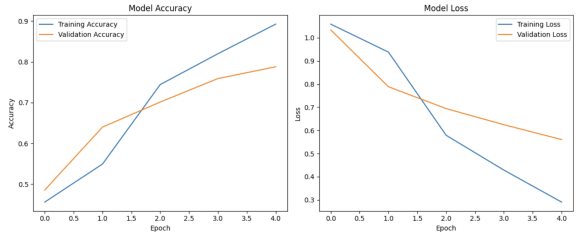

Figure 6: Learning Curve with 2016 Training Data Only

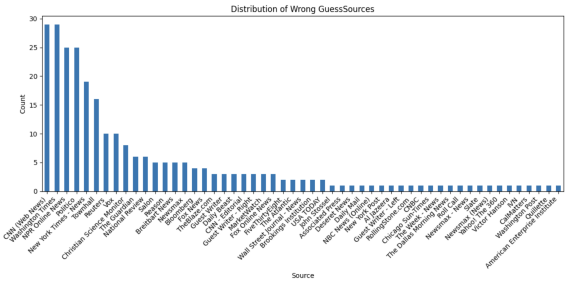

Figure 7: BERT & LSTM Wrong Prediction by the Top 30 News Organizations

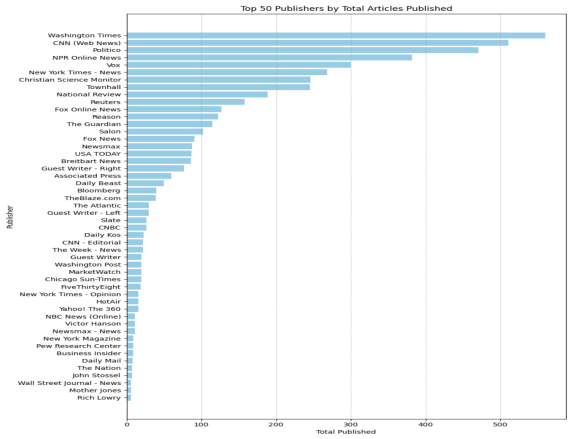

Figure 8: BERT & LSTM: Test Dataset's Top 50 News Organizations

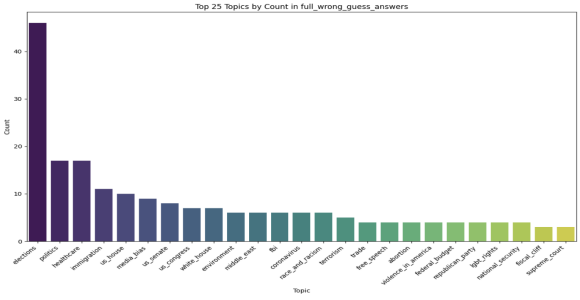

Figure 9: BERT & LSTM's Wrong Prediction by topics

Table 4: BERT Model Configuration Experiment Table

Experiments	**Train/Test Size**	**Accuracy**	**Weighted Average F1 Score**
Naive Bayes Classification (Baseline Model)	20,114/5,029	69.00%	68.00%
BERT Uncased Fine-Tuning (Pooler Token)	20,114/5029	90.00%	90.00%
BERT Uncased Fine-Tuning (CLS Token)	**20,114/5,029**	**91.00%**	**91.00%**
BERT Uncased Fine-Tuning (CLS Token & Filetered Label Counts)	8,845/2,213	86.00%	86.00%
BERT Uncased Fine-Tuning with CNN Layer (CLS Token)	-	83.00%	83.00%
BERT Uncased Fine-Tuning with CNN Layer (CLS Token & Filetered Label Counts)	8,845/2,213	83.00%	83.00%
BERT Uncased Fine-Tuning with LSTM Layer (Pooler Token)	-	**91.00%**	**91.00%**
BERT Uncased Fine-Tuning with LSTM Layer (CLS & Filetered Label Counts)	8,845/2,213	85.00%	85.00%
BERT Uncased Fine-Tuning with 1st Six Sentences (Pooler Token)	-	86.00%	86.00%
BERT Uncased Fine-Tuning with 1st Six Sentences (CLS Token)	-	87.00%	87.00%
BERT Uncased Fine-Tuning with 2016 Data Only (Pooler Token)	-	79.00%	79.00%
BERT Uncased Fine-Tuning with 2016 Data Only (CLS Token)	-	78.00%	78.00%
BERT Uncased Fine-Tuning with 4-Year Data Only	-	80.00%	80.00%
Distill RoBERTA NLi Fine-Tuning (CLS Token)	-	80.00%	80.00%
Distill RoBERTA NLi Fine-Tuning (CLS Token & Filetered Label Counts)	8,845/2,213	40.00%	23.00%
BERT Uncased Fine-Tuning excluding those outlier news articles	-	89.00%	89.00%
BERT Uncased Fine-Tuning excluding outlier news orgs. and use the outliers as testdataset	11,332/2,000	38.00%	40.00%

Table 5: BERT Model Fine Tuning Results

Experiments	Train/Test Size	Accuracy	Weighted Average F1 Score
BERT CLS Token vs BERT Pooler Token (12 NN Layers)			
BERT Uncased Fine-Tuning (Pooler Token)	20,114/5,029	90.00%	90.00%
BERT Uncased Fine-Tuning (CLS Token)	20,114/5,029	91.00%	91.00%
BERT Uncased Fine-Tuning with LSTM Layer (Pooler Token)	20,114/5,029	91.00%	91.00%
BERT Uncased Fine-Tuning with LSTM Layer (CLS Token)	**20,114/5,029**	**91.00%**	**91.00%**
BERT Uncased Fine-Tuning with CNN Layer (CLS Token)	20,114/5,029	83.00%	83.00%
BERT with Fine-Tune with the First Six Sentences			
BERT Uncased Fine-Tuning with First Six Sentences (Pooler Token)	20,114/5,029	86.00%	86.00%
BERT Uncased Fine-Tuning with First Six Sentences (CLS Token)	20,114/5,029	87.00%	87.00%

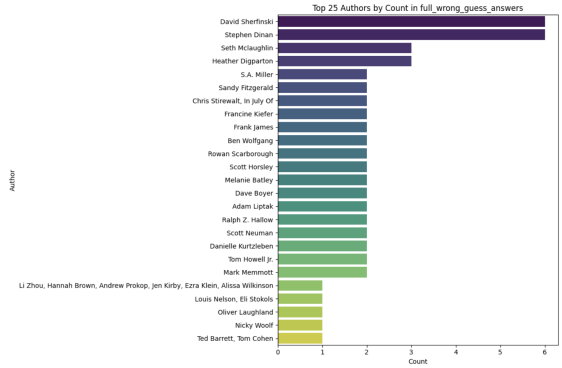

Figure 10: BERT & LSTM Model's Wrong Prediction Spread by Authors

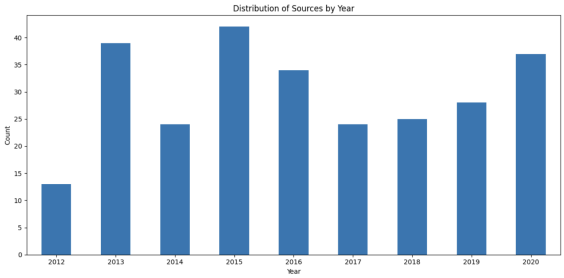

Figure 11: BERT & LSTM Model's Wrong Prediction Spread by Years

"President barack obama ' s administration reportedly misled congress — american people — crucial element iran nuclear deal. a gopled investigation senate permanent subcommittee investigation revealed wednesday obama ' s treasury department secretly worked license allow iran access american financial system. a license, according treasury department, authorization [treasury] engage transaction otherwise would prohibited. " it ' s big claim. recall iran deal u lift certain sanction tehran exchange iranian curbing nuclear program. u didn ' t remove every penalty signing agreement 2015 — meaning iran ' s access u financial system still limited. but according new report, u treasury department issued license oman ' s bank muscat february 24, 2016. license would let iran change $ 5. 7 billion worth omani rial held bank euro — first changing u dollars. iran could change money dollar without license, since would violated sanction still place. however, two u bank refused help conversion iranianowned omani rial dollar — even obama official approached — reportedly citing severe hit reputations. all, knew would perceived circumventing iran nuclear deal aiding tehran. the associated press first report senate subcommittee ' s findings. sean kane, former sanction official obama ' s treasury department lawyer dechert llp, told treasury not practice allowing iran even indirect access u financial system. " he continued, reported appears limited authorization would allowed onetime conversion specific iranian reserve held abroad, something allowed continue accessing u financial system goingforward basis. " kane also noted treasury issue license based america ' s changing foreign policy needs. that ' s critic see it. obama effort boost iran ' s economy ran counter american interests, skirted law, " jonathan schanzer, iran expert conservative foundation defense democracy think tank, tweeted wednesday. did obama administration lie helping iran? what make revelation worse, republican lawmaker say, obama administration may best misled — worst, deliberately lied — granting iran limited access u financial system. the obama administration misled american people congress desperate get deal iran, " sen. rob portman (roh), subcommittee ' s chair, said statement. here ' s portman others make claim : wednesday ' s report, note statement top obama official saying type action wasn ' t possibility. iran continue denied access [us ' s] financial commercial market, " thentreasury secretary jack lew told senate foreign relation committee july 2015. and adam szubin, top treasury official obama, said senate banking committee later month : iranian bank able clear u dollar new york, hold correspondent account relationship u financial institutions, enter financing arrangement"

Figure 12: Sample News Articles that the first portion is background information (True Label is Center; BERT model prediction is Right)

Figure 13: Chatgpt and Claude AI Model Experiments (3 Samples)

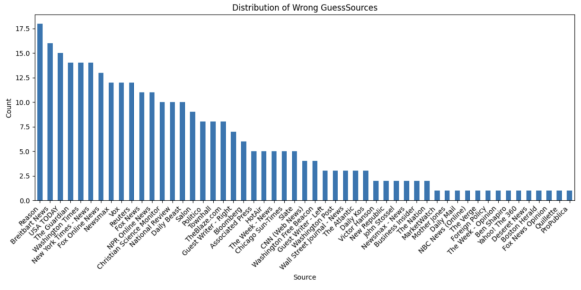

Figure 14: Spread of Incorrect Classifications After Rebalancing the Labels