ideoBERT: A Fine-Tuned BERT-Based Approach to News Ideology Classification

Jonah Roberts **David Byrne**

1 Introduction

Text classification is a pivotal area within the realm of machine learning research, with fine-tuning state-of-the-art models, such as BERT, emerging as a prevalent technique for analyzing and predicting text across various academic disciplines. Transformer models, particularly BERT, have eclipsed traditional linear regression and n-gram models in demonstrating the efficacy of text classification through machine learning methodologies.

In our study, we leverage a fine-tuned variant of BERT, labeled as ideoBERT, to ascertain the political ideology of different news sources using an extensive corpus of politically biased text. Specifically, we undertake the task of fine-tuning the distilBERT model for both binary and multiclass classification objectives. Our investigation encompasses two primary objectives: firstly, evaluating distilBERT's proficiency in predicting the general partisan ideology of news sources (e.g. conservative versus liberal); and secondly, assessing its capability to identify the source of political text among eleven distinct news outlets. Notably, in the latter task, we train ideoBERT to solely determine the source of the text without predicting its political ideology.

To benchmark the performance of our approach, we aim to compare the accuracy predictions at various levels (top-1, top-2, top-3, top-4, and top-5) with the n-gram model proposed by Wei (Wei, 2020). By juxtaposing the ideological predictions with the top-k source predictions, we intend to gain insights into the biases of different news sources within our dataset.

Given the burgeoning importance of political speech and text classification, our research holds significant implications. With the proliferation of media, ethical concerns surrounding fake news sources and biased news portrayal have become increasingly pertinent. Through our exploration, we hope to offer insightful research into this predicament by harnessing BERT models to ascertain news ideology and analyze news content. Our objective is to equip media sources and their users with the tools necessary for discerning the quality and reliability of information, thereby fostering a more informed public discourse.

2 Literature Review

Political text classification stands as a crucial focus within contemporary machine text classification research. The persistent polarization characterizing Western politics has spurred the development of various classification and prediction techniques aimed at enhancing our comprehension of news biases and detection of fake news. Among these techniques, Support Vector Machines (SVM) and neural-network n-gram models are frequently employed for large-scale text classification tasks. The NewB dataset (Wei, 2020) is as a substantial resource, comprising over 200,000 Trump-related sentences from various news sources across the political spectrum. This dataset was used in training an n-gram classification model to predict the source of the text given eleven distinct news sources. The outcomes of recurrent neural network analysis revealed top-1, top-3, and top-5 accuracy of 33.3%, 61.4%, and 77.6%, respectively.

Despite the apparent efficacy of n-gram models, transformer-based approaches have garnered prominence in the field of text classification. State-of-the-art models like BERT are now standard in text classification and prediction research. Researchers increasingly employ various pre-trained BERT models and sub-models, fine-tuning them on specific datasets to achieve superior performance in terms of both accuracy and efficiency. Often, streamlined models like distilBERT are favored for fine-tuning, offering faster training times while preserving approximately 97% of the classifica-

tion capabilities of the BERT base model (Qasim et al., 2022). Fine-tuning transfer learning models such as BERT has demonstrated significant superiority over baseline n-gram and transformer models, including discerning the ideological stance of political articles when utilizing state-of-the-art models (Liu et al., 2022). Even amongst comparisons to classical machine learning methods or deep contextualized word representations, transformer based architectures such as BERT continue to shine. In a study focused on protest news classification and sentiment analysis of product reviews, the large scale, pre-trained model out performed its modern peer, ELMo, as well as predecessors Multi Normal Bayes, and Linear Support Vector Machine. DistilBERT better transferred generic semantic knowledge to other domains while being 30% smaller and 83% faster than ELMo (Büyüköz et al., 2020).

Furthermore, BERT has found application in political text classification using social media data (Gupta et al., 2020), (Rahmati et al., 2023), (Üveges and Ring, 2023). Adopting a similar classification methodology inspired by poliBERT, we employ the Wei NewB dataset to train our ideoBERT model. While maintaining similar architectural layers and data splits, our objective is to advance the classification of news source polarity using a BERT transformer model. Our aim is to compare the performance of a distilBERT model on generic news data against existing methodologies.

3 Data

The data for ideoBERT originates from the NewB dataset utilized by the Wei n-gram classification model (Wei, 2020). The NewB dataset comprises over 200,000 sentences discussing Donald Trump, each labeled by news source and political bias. We employ two different training models, one for binary classification of partisan labeling and one multiclass classification model of text source. To construct our binary classification dataset, we merged the conservative and liberal text files available on the GitHub page `https://github.com/JerryWeiAI/NewB`. We then removed the number labels for each line of text and relabeled them as conservative or liberal based on the ideological inclination of its news source. For the multiclass classification model we used the NewB text origin file, converted to csv, with the given label and texts. Both datasets follow a 80-10-10 training, development, test split, respectively. To maintain compatibility with binary classification models, we eliminated sentences labeled as neutral along with their corresponding texts. Further preprocessing techniques are discussed in the methodology section. In the future, we hope to train and test ideoBERT's capabilities on other, similar partisan data.

Liberal (Newsday)	**Conservative (Daily Herald)**
illustration newsday photo by Jon Naso. Donald Trump, whom a casino analyst is suing for 2 million over Trump's response to the analyst's dire predictions for the Taj Mahal in Atlantic City.	Donald Trump lost a billion dollars, but he made it back, and Hillary lost 6 billion while she was in charge of the State Department, and the media is silent.

Table 1: Example sentence data by news source ideology.

4 Methodology

4.1 Preprocessing

In general, we follow similar preproccessing techniques of the data as Wei (Wei, 2020). We also follow their 80-10-10 train, dev, test split. However, to provide more context for the BERT model, we excluded any sentences that were less than 10 words, leaving us with a total of 217,151 unique sentences for the multiclass task and 230,710 samples for the binary classification task. This removed small, seemingly subset samples of text such as, "trump said this". The text was already lowercase from the NewB dataset so we continued with the distilbert-base-uncased model and tokenizer techniques. To provide visual data context for binary classification we changed the labels of the combined dataset to liberal and conservation (0, 1), although these were later converted back to binary in the tokenization. For analysis purposes of the text, the string labels were helpful.

4.2 Model Selection and Tokenization

We employ the BERT (Bidirectional Encoder Representations from Transformers) model using the distilBERT base variation. The distilBERT model is a faster and lighter weight version of the BERT base model. Text data is tokenized using the BERT tokenizer, with padding and truncation applied to ensure uniform input length. We used the "only first" truncation method in the BERT tokenizer, so

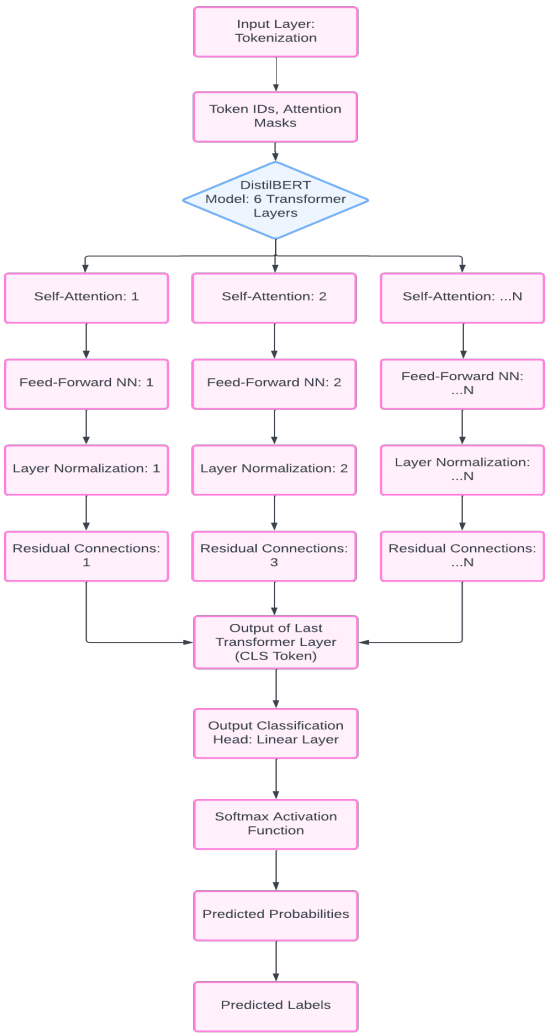

Figure 1: The model architecture of the ideoBERT multiclass model. The model is based on the distilBERT model using the HuggingFace python library. ideoBERT consists of six transformer layers that consists of self attention, feed-forward neural network, layer normalization, and finally a residual connection.

if an input sequence is longer than the maximum sequence length supported by the BERT model, only the beginning of the sequence is retained, and the remaining part is discarded. Compared to full truncation and no truncation, we found this to have the highest impact on accuracy scores.

4.3 Model Training

In the model training phase, the pre-trained distilBERT model is fine-tuned specifically for the political text classification task(s). The learning rate, epochs and batch size were derived through extensive runs of the development set. The model was fine-tuned for nine epochs for the multiclass classification task and four for binary, providing sufficient iterations for the model to learn the task-specific patterns within the data. Further epochs resulted in over fitting of the model. We fine-tuned the learning rate from a starting point of a similar, hate speech classification task we had previously developed, beginning with 1e-2. A final learning rate of 1e-5 for both tasks, were employed to control the step size during gradient descent, ensuring a balanced convergence towards optimal parameter values. The training process is facilitated using the PyTorch framework. In the multi class classification task, the batch size is 8 samples per device, optimizing memory usage and computational efficiency. The batch size is 16 for the faster, binary classification task.

4.4 Evaluation and Analysis Techniques

Model performance is evaluated using the development set from the dataset split, employing an epoch-level evaluation strategy wherein the model's performance is assessed at the conclusion of each training epoch. Following evaluation, predictions are generated for the test dataset utilizing the trained model. The primary metric utilized for test set prediction is accuracy, providing insight into the model's overall classification performance. In the future, further evaluation methods such as F-1 scores, would be helpful in further analysis of political bias in text classification.

For the multiclass task, logits are extracted from these predictions, and probabilities are computed using softmax activation, facilitating the determination of top-k predictions for each sample. These scores serve as a critical measure of the model's efficacy in correctly predicting the label within the top-k predictions, thereby offering valuable insights into its classification capabilities. We use the Wei n-gram model (Wei, 2020) as a base line for comparison to our ideoBERT model for multiclass classification.

5 Results

5.1 Binary Classification

For the binary classification task distinguishing between sentences sourced from liberal or conservative origins, the evaluation results indicate a moderate performance level achieved by the model as show in Table 4. With an accuracy of 61.72% at the end of the fourth epoch, the model demonstrates a discernible ability to differentiate between the two ideological categories. Despite not reaching

ideoBERT Training Parameters
Evaluation strategy: epoch
Learning rate: 1e-5
Per-device train batch size: 8
Per-device eval batch size: 8
Number of train epochs: 9
Padding: true
Truncation: only first

Table 2: Training and tokenization arguments for ideoBERT multiclass model. (Note: the binary model uses a batch size of 16, 1e-5 learning rate, and 4 epochs).

a high level of accuracy, the model's performance suggests a degree of effectiveness in capturing and learning the underlying patterns indicative of liberal or conservative text sources. Although we attempted to mitigate some of the task's inherent difficulty in preprocessing and model parameters, this performance level may still have been influenced by various factors such as the quality and representativeness of the training data and lack of distinct writing patterns, as well as the inherent challenges associated with classifying text based on ideological affiliations. Further analysis and refinement may be required to improve the model's accuracy and robustness in distinguishing between liberal and conservative sources accurately.

Sample 13
trumps poor standing among women millennials and hispanics will catch up to him harkin predicted as will a news media that he believes has given trump too much leeway to date
Top-5 predictions:
1. LA Times (lib): 44.33%
2. Washington Post (lib): 24.30%
3. Daily Press (con): 14.95%
4. Chicago Tribune (con): 12.92%
5. CNN (lib): 2.14%

Table 3: Sample sentence and its prediction probabilities.

5.2 Multi Class Classification

The results of the multiclass text classification task provide further insight into the model's performance both in the context of classifying text samples into liberal, neutral, and conservative categories and general source prediction. With labels 0 to 4 representing liberal sentiments, label 5 indicating neutrality, and labels 6 to 10 denoting conservative viewpoints, the model's accuracy in discerning the ideological leanings of the text samples is apparent in pattern, despite the task of source prediction which does not inherently involve ideological bias. The observed trend in accuracy across different levels of predictions suggests a nuanced understanding of the underlying sentiments within the text. Interestingly in many samples the model demonstrates a higher accuracy in predicting conservative viewpoints, as indicated by the elevated accuracy scores for labels 6 to 10. This does not align with the n-gram model of Wei (Wei, 2020) in which the accuracy scores tended to be higher for liberal news sources.

The results of the multi-class text classification task demonstrate the effectiveness of the trained BERT model in accurately predicting the labels for text samples. The top-k accuracy scores (Figure 4) reveal a progressive improvement in classification performance as the number of considered predictions increases, with the model achieving a top-1 accuracy of 44.73% , top-3 of 69.23% and 83.15% at the top-5 level. This significantly outperforms the n-gram model of Wei in Figure 5, with top-1, top-3, and top-5 accuracy of 33.3%, 61.4%, and 77.6%. Analysis of individual sample predictions further highlights the model's proficiency, as evidenced by its ability to correctly identify the most probable labels for diverse text inputs that range from political agendas to breaking headlines. These results underscore the model's robustness and efficacy in handling real-world text classification tasks across various domains.

Multiclass Classification Top-N Accuracy	
Top-1 Accuracy	44.73%
Top-2 Accuracy	59.81%
Top-3 Accuracy	69.76%
Top-4 Accuracy	77.29%
Top-5 Accuracy	83.15%
Binary Classification	
Prediction Accuracy	61.72%
End Evaluation Loss	0.697
Epochs	4
Training Time	53 min

Table 4: Results of multiclass and binary classification tasks for ideoBERT. The model highlights competent accuracy with fast run times.

N-Gram Model Top-n Accuracy (%)					
n	1	2	3	4	5
LR	18.3	32.1	42.6	52.2	60.8
CNN	34.0	50.3	61.5	70.0	77.4
RNN	33.3	50.6	61.4	70.5	77.6

Table 5: Results of the Wei multiclass classification task with n-gram model. LR: logistic regression. CNN: convolutional neural network. RNN: recurrent neural network.

5.3 Discussion

As anticipated, the distilBERT transformer model outperformed the Wei N-gram model across all three subsets: logistic regression (LR), convolutional neural network (CNN), and recurrent neural network (RNN). This highlights the effectiveness of transformer models. ideoBERT also performed impressively in terms of runtime for both multiclass and binary classification tasks, despite the extensive training and test sets. Particularly encouraging are the results of the multiclass task, given the challenge of predicting news sources from multiple ideologically similar samples. While we didn't expect high top-1 accuracy due to the number of classes, the accuracy for both top-3 and top-5 was above expectations. These results demonstrate the model's ability to predict within similar ideological groupings, mirroring the accuracy of the binary classification task.

Although the binary task yielded seemingly low accuracy, a deeper analysis of the text data (see Table 3) underscores the consistency of the ideoBERT model. While the top-5 predictions in the sample do include conservative sources, the majority align with liberal sources, particularly notable in the top-2 predictions. Interestingly, the sources 8 (Daily Press, conservative), 10 (Chicago Tribune, conservative), and 3 (LA Times, liberal) were frequently grouped together in predictions. While we only delved into this particular grouping, it suggests that political speech and text may exhibit recognizable patterns of bias, albeit not at an extremely high level, as evidenced by the competent yet moderate accuracies in the binary task. This observation implies that despite the perceived strong polarization of political ideologies in the media, the language used may not be drastically different across major news sources.

We believe our findings are important in the larger realm of political text and speech classification along with fake news detection. The promising results of ideoBERT hint at further developmental research for political language classification. Large improvements in this field could lead to strong media detection methods that could allow users to understand and contextualize political polarity in news. It may show that political differences are not as large as they are highlighted to be, further aiding in progressive political action within the United States. Our findings here indicate the ability of transformer models to aid in the detection of fake news by comparing to similar ideological news sources. We hope our work will be further expanded on to analyze BERT's ability in day to day applications.

6 Ethical Considerations

As with many BERT models, considering the ethical implications of large machine learning models is import when conducting predictive research. Training machine models on human classified data can lead to significant and potentially harmful bias if not understood and handled correctly. ideoBERT is trained on a large set of data that is roughly split 50-50 between labeled "conservative" and "liberal" ideological news sources. Although the news sources are largely accepted by the general public to lean one way or the other, it certainly does not mean that everything produced or released by each source is only liberal or only conservative. The NewB dataset is largely vacant of harmful or hateful speech, though the topics in some of the sentences may be considered difficult or offensive to some populations given their political position. We acknowledge this and have taken it into consideration when rendering our predictive results and future ideas.

Political speech is largely partisan in nature, so we also recognize the value of partisan predictions. Understanding and learning about political biases in news may help avoid public echo chambers, in which people are solely exposed to information and sources of similar political leanings. News sources capitalize on public echo chambers, portraying heavily biased information to create partisan bubbles. In times of vast political differences we hope that using transformer models can aid in removing and classifying unethical and hateful political text while exposing intentional news biases towards the public.

7 Conclusion

Our study shows the effectiveness of the distilBERT model in both binary and multi-class text classification tasks, showcasing its prowess in discerning ideological leanings and predicting news sources. Despite the moderate accuracy achieved in the binary classification task, ideoBERT demonstrates a discernible ability to differentiate between liberal and conservative text sources, indicating its potential in capturing underlying patterns indicative of ideological affiliations. As with the Wei n-gram model, further refinement and analysis are needed to enhance accuracy and robustness.

In the multi-class classification task, ideoBERT's progressive improvement in classification performance, particularly evident in top-k accuracy scores, highlights its proficiency in accurately predicting labels for diverse text inputs. The model's ability to correctly identify the most probable labels for varied text samples emphasizes its robustness and efficacy in real-world text classification tasks across different domains. These findings, coupled with the superior performance of ideoBERT over the n-gram models, contribute to a deeper understanding of transformer model capabilities and their potential applications in media analysis and beyond.

References

Berfu Büyüköz, Ali Hürriyetoğlu, and Arzucan Özgür. 2020. Analyzing ELMo and DistilBERT on socio-political news classification. In *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*, pages 9–18, Marseille, France. European Language Resources Association (ELRA).

Shloak Gupta, Sarah Bolden, Jay Kachhadia, A Korsunska, and J Stromer-Galley. 2020. Polibert: Classifying political social media messages with bert. In *Social, cultural and behavioral modeling (SBP-BRIMS 2020) conference. Washington, DC*.

Yujian Liu, Xinliang Frederick Zhang, David Wegsman, Nick Beauchamp, and Lu Wang. 2022. Politics: Pretraining with same-story article comparison for ideology prediction and stance detection. *arXiv preprint arXiv:2205.00619*.

Rukhma Qasim, Waqas Haider Bangyal, Mohammed A Alqarni, and Abdulwahab Ali Almazroi. 2022. A fine-tuned bert-based transfer learning approach for text classification. *Journal of healthcare engineering*, 2022.

Ali Rahmati, Ehsan Tavan, and Mohammad Ali Keyvanrad. 2023. Predicting content-based political inclinations of iranian twitter users using bert and deep learning. *AUT Journal of Mathematics and Computing*, 4(2):145–154.

Jerry Wei. 2020. Newb: 200,000+ sentences for political bias detection. *arXiv preprint arXiv:2006.03051*.

István Üveges and Orsolya Ring. 2023. Hunembert: A fine-tuned bert-model for classifying sentiment and emotion in political communication. *IEEE Access*, 11:60267–60278.