On the Transference and Identification of Political Bias of Transformer-Generated Summaries of News Articles

Ash Tan
MIDS W266 / Berkeley School of Information
asht@berkeley.edu

Abstract

We train several transformer-based classification models to detect bias in political news articles, which is a typically difficult task at the document-level due to the different forms bias can take. After evaluating the efficacy of these architectures, we then apply transformer-based summarization models to news articles in our test data to generate summaries. We then apply the trained classification model to the generated summaries to label them as biased or neutral. Next, we compare the generated labels of the summaries to the generated and original labels of the original news articles to study whether bias is transferred from an original text to generated summaries. Transformer-based summarization was found to generally produce debiased summaries of biased texts, with abstractive summarization performing better than extractive summarization for the purpose of debiasing text in most cases.

1 Introduction

Political polarization may not be a novel phenomenon, but it has been emboldened in recent years by the growing number of nonobjective news sources, resulting in increasingly partisan political messaging and elections (Prior, 2013). Identification of media bias through natural language processing could help combat this issue by providing a metric to quantify bias in news media.

Furthermore, as transformer-based summarization becomes more advanced, there is an increasing trend of automatic text generation being utilized for journalistic purposes (Jung et al., 2017). Understanding how bias in article text affects transformer-generated summaries will help inform decisions regarding automated news as well as our understanding of how transformer summarization processes textual bias.

2 Background

2.1 Identifying Textual Bias

Identifying bias is challenging in part because of the different forms it can take (Fan et al., 2019). Lexical bias refers to bias in word and phrasing choice, and can be locally identified without context. Informational bias refers to bias in what information the author chooses to be included in the text, and is often more difficult to detect than lexical bias. While recognizing lexical bias may only require local context, recognizing informational bias may require context than may not be included in the article (Fan et al., 2019). An article may be biased because it leaves out critical information, which would be difficult to ascertain from the article itself.

The difficulty of bias identification has led to a variety of approaches (Chen et al., 2020b). At the sentence-level, SVM and BERT have seen mixed success (Chen et al., 2020a; Fan et al., 2019), while multi-view models using hierarchical attention networks for document classification have performed very well, achieving an F1 score of 79.67 % (Kulkarni et al., 2018; Yang et al., 2016). However, there has been less research comparing BERT to its successors; RoBERTa has been shown to outperform BERT in bias classification by a significant margin (Wang, 2019), but newer models such as XLNet and ALBERT may yield better performance.

2.2 Obtaining Objectively Labeled Data

Obtaining objectively labeled data can be a difficult challenge because of implicit biases of the individual annotator. Efforts to achieve unbiased bias data have previously included using third-party news aggregators /evaluators such as Ad Fontes Media and AllSides (Chen et al., 2020b), as well as having multiple annotators both individually and collaboratively label data (Fan et al., 2019). We

argue that, while these approaches may reduce the likelihood of subjective labeling of bias, they are not optimal for creating large, accurately labeled datasets. Third party news evaluators may not accurately judge news sources for a variety of reasons (especially when their metrics are unclear) and often evaluate news sources as a whole instead of individual articles. Having multiple evaluators discuss annotations decreases the impact of personal implicit bias, but is inefficient for labeling large datasets.

2.3 Bias in Transformer Summarization

While predictive biases and bias mitigation techniques are currently subjects of much discussion (Shah et al., 2020), we found that most papers are concerned with internal model bias as the result of transformer architecture and training, and that there has been relatively little research regarding the effect of bias in text on transfer learning tasks such as generated summaries. While transformer-based models have the ability to generate fake biased news articles (Gupta et al., 2020), we know of no research that specifically documents the effect of political bias on text summarization. Biases learned from training data, such as gender stereotypes, have been shown to affect predictions in downstream tasks of BERT models if not actively mitigated (Bhardwaj et al., 2020), but how this affects news article summarization is unclear. As transformers like T5 are generally trained on politically agnostic texts (Raffel et al., 2020), we should expect models to not contain any intrinsic political bias, which means the output should depend on the bias of the input text. We predict that abstractive summarization will help reduce bias by reducing lexical bias, which may still appear in extractive summaries. While abstractive summarization models like Pegasus may be able to reduce informational bias by utilizing external learned contexts, we predict informational bias will remain challenging to identify and remove.

3 Methods

3.1 Data

We used a dataset[1] constructed by Budak et al. (2014), which contains bias ratings for political news articles on a broad range of topics by thirteen top US news outlets, as well as two large political blogs. Article labels were created by averaging crowdsourced ratings from Amazon Mechanical Turk to mitigate individual bias. The articles were initially graded on two Likert scales as a response to the questions: "Is the article generally positive, neutral, or negative toward members of the Democratic/Republican Party?" Responses for both parties were encoded on a five-point scale from very positive to very negative, encoded as –1, –0.5, 0, 0.5, 1 (Budak et al., 2016).

By calculating the distance between the Democrat bias rating D and Republican bias ratings R, we relabeled the data into "Biased and "Neutral" binary labels such that an article is considered "Biased" if article distance $a = |D - R| \geq 1$. We discarded all articles without strong bias directionality ($a = |D - R| < 1$), fewer than two hundred words, or that no longer exist online; the original dataset only contained URLs, so all articles were retrieved through webscraping. Article text was transformed to lower case, truncated to the first 600 words, and cleaned of HTML tags and other web elements. All mentions of news sources inside article text were replaced with "SOURCE." Headline information was excluded from the article text. This process resulted in a dataset size of 12350 labeled articles with only 2306 biased articles. We randomly selected 2306 neutral articles and concatenated them with the biased articles to form our final balanced dataset, randomly partitioned into a training set of 4000 articles, a validation set of 500 articles, and a test set of 110 articles. Each subset contains a roughly equal proportion of labels.

3.2 Classification[2]

We chose to utilize BERT, RoBERTa, XLNet, and ALBERT binary sequence classification models, all constructed and pretrained by Huggingface for standardization and ease of use. (We initally constructed custom classifiers using the Huggingface API but were unable to exceed the performance of the pretrained classifiers.) All models were initialized using the 'large' preconfiguration to optimize robustness and efficiency (Li et al., 2020). Each was compiled with sparse categorical cross-entropy loss, Adam optimizer with learning rate of 2e-7, batch size of 5, and trained for 25 epochs. (Versions of RoBERTa and XLNet with frozen layers were

[1]Data: https://deepblue.lib.umich.edu/data/concern/data_sets/8w32r569d?locale=en

[2]Code repository: https://github.com/AshQTan/W266_Project

also trained but failed to exceed normal pretrained model performance.) We also used TF-IDF to similarly predict labels as a non-transformer evaluation baseline. Models were evaluated using standard metrics (accuracy, precision, recall, F1 scores) with "Neutral" as the baseline class.

We initially attempted to classify along the original (and variations of) the 5-point Likert label schema, but models generally failed to perform multilabel classification. Despite experimenting with varying losses, learning rates, dropout rates, layer numbers, and dataset size, they often simply predicted the majority class label (56% accuracy). The singular exception was XLNet, which was able to reach validation accuracies of 62% and test accuracy of 65%. (The validation dataset contains an equal proportion of neutral and biased articles.) However, reducing the classification task to a binary choice enabled all transformer models to make actual predictions aside from the majority class.

3.3 Summarization

Summaries were generated using the transformer-based T5 and Pegasus models (Raffel et al., 2020; Zhang et al., 2020). We utilized pretrained conditional generation models from Huggingface using the 'large' preconfiguration and opted to utilize both models for zero-shot summarization instead of pretraining them on our news article dataset, as there is little distinguishable difference in quality between zero-shot and few-shot learning for both T5 and Pegasus summarization (Goodwin et al., 2020). Also, as we are interested in the effects of input text bias on transfer learning tasks, finetuning the model could affect potential phenomena. Summaries were generated with a minimum length of 35 words and a maximum length of 80 words, with beam search and n-gram repetition prevention.

3.4 Summarization Evaluation

Due to its superior performance, XLNet was chosen to predict the bias of generated summaries. In addition, we also used TF-IDF to predict bias in order to compare XLNet's predictions with that of a non-transformer-based algorithm. Both XLNet and TF-IDF were initialized in post-training states with unchanged hyperparameters.

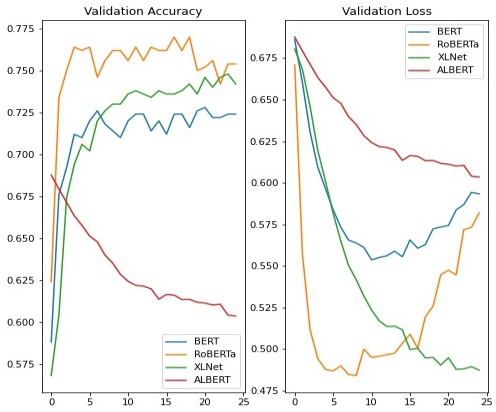

Figure 1: Model Training Comparison

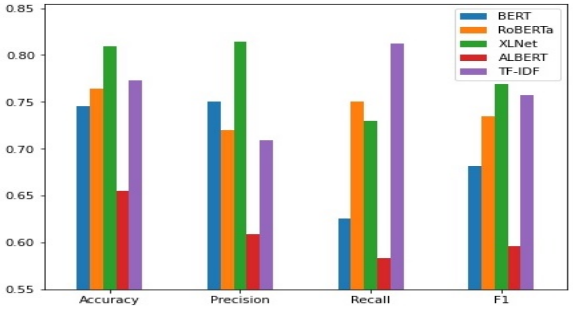

Figure 2: Model Performance Comparison

4 Results

4.1 Evaluating Classification

In this context, precision refers to the proportion of correctly labeled biased articles in relation to the total number of articles predicted to be biased; recall refers to the proportion of correctly labeled biased articles in relation to the total number of biased articles.

From Table 1, it is evident that ALBERT performed the worst across all metrics; this is surprising as it is considered an improved version of BERT that has achieved groundbreaking performance on multiple tasks (Lan et al., 2020). RoBERTa displayed improvement over the base BERT model, but TF-IDF scored better than most transformer-based models except in precision; TF-IDF is reasonably well-equipped to detect lexical bias, which is often identifiable through word choice. XLNet outperformed all other transformer-based models in all metrics, and outperforms TF-IDF in all metrics except recall. XLNet's performance is especially impressive in the context of past bias classification attempts; hierarchical attention models

Metrics	TF-IDF	BERT	RoBERTa	XLNet	ALBERT
Accuracy	0.773	0.745	0.7636	**0.809**	0.655
Precision	0.709	0.750	0.720	**0.814**	0.609
Recall	**0.813**	0.625	0.750	0.729	0.583
F1	0.757	0.682	0.735	**0.769**	0.596

Table 1: Metrics of predicted bias labels for article text compared to original article bias labels

for document-level classification have previously achieved F1 scores of 79.67 % (Kulkarni et al., 2018), with past BERT-based models scoring significantly lower (Fan et al., 2019; Wang, 2019; Chen et al., 2020a). XLNet's Transformer-XL-based architecture and permutated factorization order may allow for learning an increased number of dependencies (Yang et al., 2019), which may aid in detection of subtler informational bias (Fan et al., 2019).

Examining all instances where the real label of the article is "Neutral" but all models predicted the article as "Biased," we observe four results sourced from three distinct sources: Daily Kos, Breitbart, and CNN's 'The Political Ticker.' This is an interesting result because all three are considered to be non-objective news sources: Daily Kos and The Political Ticker are political blogs, while Breitbart is generally considered conservatively biased (Bentley et al., 2019; Kulkarni et al., 2018; Budak et al., 2016).

> i swear to you, if i didn't know any better, and i'm not a big conspiracy guy, after seeing that ad, i would think the nra is either an elaborate avant-garde joaquin phoenix-style joke, or a false flag operation run by michael moore in an attempt to discredit responsible gun owners.

Figure 3: Excerpt from example of a "Neutral" text published by *Daily Kos* (classified as "Biased" by all models)

This may suggest that transformers learn to associate bias with certain writing styles used in subjective writing. XLNet mislabeled eight "Neutral" articles as "Biased," seven of which were published by political blogs or non-objective news sources: Fox News, Huffington Post, Daily Kos, Breitbart, CNN's Political Ticker (Bentley et al., 2019; Kulkarni et al., 2018; Budak et al., 2016). This further supports that transformer-based models are able to distinguish objective and subjective writing techniques. Furthermore, by examining articles labeled as "Biased" but classified as "Neutral" by all models, we find six articles, three published by relatively objective news sources (USA Today, Yahoo News) and three published by relatively biased news sources (Fox News, New York Times, Huffington Post) (Bentley et al., 2019; Budak et al., 2016; Kulkarni et al., 2018).

XLNet's superior performance demonstrates its ability to approximate human judgement for correctly labeling biased articles, making the trained model suitable for classifying unlabeled data such as generated summaries. TF-IDF's performance is largely comparable and is generally suited for bias classification tasks; however, TF-IDF notably scored low in precision, which suggests that correctly identifying bias is more difficult for TF-IDF than transformer-based models. This could be attributed to the differences between informational and lexical bias: TF-IDF is well-equipped to identify lexical bias, but informational bias requires an understanding of context that the transformer self-attention architecture is better suited for. Both XLNet and TF-IDF perform better than the baseline of predicting the majority class (50% accuracy).

4.2 Evaluating Summarization

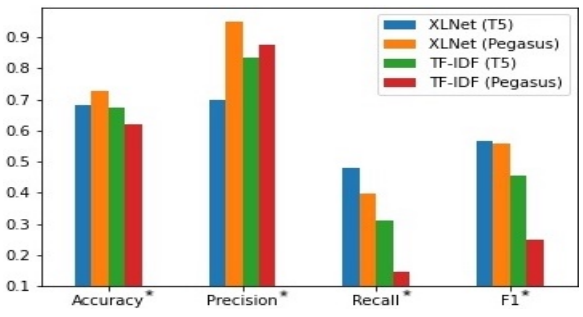

Figure 4: Metric Comparison (Classifier-generated summary labels compared to article labels)

As the generated summaries are unlabeled and we treat our classifiers as simulated human eval-

Metrics	TF-IDF (T5)	TF-IDF (Pegasus)	XLNet (T5)	XLNet (Pegasus)
Accuracy*	0.673	0.618	0.682	0.727
Precision*	0.833	0.875	0.697	0.950
Recall*	0.313	0.146	0.0.479	0.396
F1*	0.455	0.250	0.568	0.559

Table 2: Metrics of predicted bias labels for generated summary text compared to original article bias labels

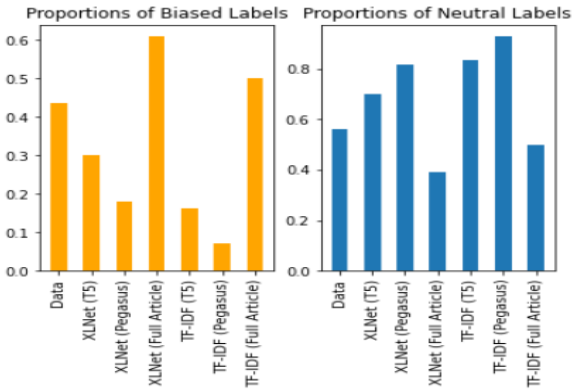

Figure 5: Comparison of Label Proportions

uators, we can use the metrics from Table 2 to compare the bias of generated summaries to the bias of the original articles. Here, precision* can be viewed as a metric of article bias salience compared to summary bias salience, since this metric refers to the proportion of summaries that are both (a) labeled as biased by the classifier and (b) generated from biased articles in relation to the total number of summaries labeled as biased by the classifier. Recall* can be viewed as a metric of the relationship between article bias and summary bias since this metric refers to the proportion of summaries that are both (a) labeled as biased by the classifier and (b) generated from biased articles in relation to the total number of biased articles.

XLNet and TF-IDF scored very differently on summary classification compared to article classification across all metrics, suggesting that the summarization process does affect bias. Notably, Figure 5 illustrates that both classifiers show reduced proportions of biased text for generated summaries in comparison to the original articles, which suggests that transformer summarization tends to result in debiased text. This is supported by high precision* and low recall* for both T5 and Pegasus, meaning (a) a high proportion of summaries that are classified as biased were generated from biased articles and (b) there is a low proportion of biased summaries from biased articles in relation to the total number of biased articles. This suggests that summarization generally succeeds in debiasing text: fewer summaries than articles were classified as biased, and the summaries that were classified as biased are very likely to have been generated from biased text. This aligns with our previous hypothesis that transformer summarization generally produces debiased summaries.

Also apparent are the differences between the precision* and recall* of Pegasus and T5. Pegasus scored higher precision* and lower recall* with both XLNet and TF-IDF. XLNet predicted different labels for Pegasus and T5 summarizations of 33 articles; out of these 33, 10 Pegasus summaries were classified as biased (10 T5 summaries as unbiased) and 23 Pegasus summaries were classified as biased (23 T5 summaries as unbiased).

> republicans have been pressuring regulators for years to exempt derivative products from new rules . financial reform advocates say initiative is padding wall street profits at the expense of important public protection, and democratic support has eroded

Figure 6: T5 summary of "Biased" text "GOP Wall Street Bill Would Eviscerate Dodd-Frank," published by *The Huffington Post* (summary classified as "Biased" by XLNet)

> As the US Congress prepares to return from its summer break, the fate of the Dodd-Frank financial reform law is in the hands of a small group of lawmakers and the White House.

Figure 7: Pegasus summary of "Biased" text "GOP Wall Street Bill Would Eviscerate Dodd-Frank," published by *The Huffington Post* (summary classified as "Neutral" by XLNet)

Pegasus' abstractive summarization enables it to rephrase article content using new language (Zhang et al., 2020). T5, while able to achieve abstractive summarization, is a multipurpose model that was not specifically trained for abstractive summarization and may output extractive summaries (Raffel et al., 2020). As seen in Figures 6 and 7, extraction may result in lexical and/or informational bias being transferred from articles to summaries, depending on what the transformer chooses to extract, while abstractive summarization reduces the possibility of lexical bias transference.

> president george w. bush spent more time outside of washington than his predecessor, said the post in 2005 and 2006 respectively–but nearly all were trips back to their ranch or family vacation homes as well-asked for by other members on families with kids at home from school (and secret service) white house tours have been shut down since last spring because they are too expensive

Figure 8: T5 summary of "Biased" text "Media Ignores Lavish Obama Vacations, Slammed Bush for Mountain Biking," published by *Breitbart News* (summary classified as "Neutral" by XLNet)

> As President Barack Obama prepares to leave the White House for an eight-month summer vacation at his ranch in Crawford, Texas, the president's aides and reporters are trying to figure out how to cover up his lavish summer vacations.

Figure 9: Pegasus summary of "Biased" text "Media Ignores Lavish Obama Vacations, Slammed Bush for Mountain Biking," published by *Breitbart* (classified as "Biased" by XLNet)

However, as seen in Figures 8 and 9, if the entire narrative of an article is biased, abstractive summarization is likely to summarize the narrative instead of relevant facts, while it remains possible for extractive summarization to avoid lexical or informational bias even in this context, depending on the amount of bias in the article and the transformer's choice of extracted terms and phrases. However, our data suggest that these cases are atypical, and that Pegasus will likely generate an unbiased summary of the average news article. Notably, TF-IDF, which scored much higher on recall than XLNet on article bias, scores extremely low recall* for Pegasus-generated summaries and slightly higher recall* for T5-generated summaries. As TF-IDF is theoretically better at detecting lexical bias, this difference in recall* supports our claim that T5-generated summaries are more likely to contain lexical bias than Pegasus-generated summaries: TF-IDF struggles to identify any bias in Pegasus-generated summaries, which, due to abstrative summarization, is more likely to contain informational bias. Additionally, this difference could be attributed to abstract summarization utilizing language outside of TF-IDF's vocabulary, which would make XLNet a better choice for classifying bias of abstractive models. We can further explore the differences between abstractive and extractive summarization by comparing XLNet-predicted summary labels to XLNet-predicted article bias.

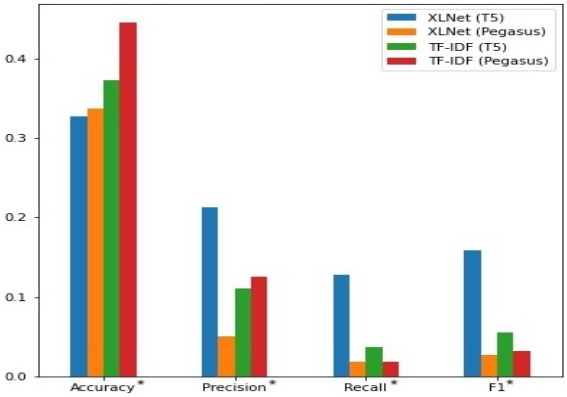

Figure 10: Metric Comparison (XLNet-generated summary labels compared to XLNet-generated article labels)

Figure 10 illustrates that, from the perspective of the XLNet classifier, most summaries generated from biased articles are unbiased text. However, precision* and recall* remain higher for T5-generated summaries than for Pegasus-generated summaries, which further supports our interpretation that extractive summarization is more prone to inheriting bias from parent text. Figure 5 also displays reduced proportions of biased summaries for Pegasus-generated summaries in comparison to T5-generated summaries.

5 Conclusion

The transformer architecture is a powerful tool that is capable of both classifying bias and also debiasing text. XLNet was able to successfully differentiate biased and neutral articles, and was able to subsequently identify a reduction in bias in transformer-generated summaries. While transformer-based summarization generally produces debiased summaries, abstractive summarization is more likely to produce unbiased summaries of articles than extractive summarization due to the reduction of lexical bias. The possibilities of more powerful, fine-tuned models include efficiently creating large, labeled datasets, improving social discourse through protecting journalistic integrity, and reducing political polarization by providing easy access to automated bias ratings of individual articles.

A follow-up study might focus on designing a transformer model specifically for debiasing text, potentially in the form of a generative adversarial network which competes with itself to produce and discriminate unbiased texts. A GAN was initially attempted for this paper, but the necessary computing power was unavailable at the time.

References

F. Bentley, Katie Quehl, Jordan Wirfs-Brock, and M. Bica. 2019. Understanding online news behaviors. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*.

Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. 2020. Investigating gender bias in bert. *ArXiv*, abs/2009.05021.

C. Budak, S. Goel, and Justin M. Rao. 2016. Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, 80:250–271.

Wei-Fan Chen, Khalid Al Khatib, Benno Stein, and Henning Wachsmuth. 2020a. Detecting media bias in news articles using gaussian bias distributions. *ArXiv*, abs/2010.10649.

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. *ArXiv*, abs/2010.10652.

Lisa Fan, M. White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and L. Wang. 2019. In plain sight: Media bias through the lens of factual reporting. In *EMNLP/IJCNLP*.

Travis Goodwin, Max E. Savery, and Dina Demner-Fushman. 2020. Flight of the pegasus? comparing transformers on few-shot and zero-shot multi-document abstractive summarization. *Proceedings of COLING. International Conference on Computational Linguistics*, 2020:5640 – 5646.

S. Gupta, Huy H. Nguyen, J. Yamagishi, and I. Echizen. 2020. Viable threat on news reading: Generating biased news using natural language models. *ArXiv*, abs/2010.02150.

Jaemin Jung, Haeyeop Song, Youngju Kim, Hyun-Ju Im, and Sewook Oh. 2017. Intrusion of software robots into journalism: The public's and journalists' perceptions of news written by algorithms and human journalists. *Computers in Human Behavior*, 71:291 – 298.

Vivek Kulkarni, Junting Ye, S. Skiena, and William Yang Wang. 2018. Multi-view models for political ideology detection of news articles. In *EMNLP*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. *ArXiv*, abs/1909.11942.

Zhuohan Li, Eric Wallace, Sheng Shen, Kevin Lin, K. Keutzer, D. Klein, and J. Gonzalez. 2020. Train large, then compress: Rethinking model size for efficient training and inference of transformers. *ArXiv*, abs/2002.11794.

M. Prior. 2013. Media and political polarization. *Annual Review of Political Science*, 16:101–127.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, W. Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Deven Shah, H. A. Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. *ArXiv*, abs/1912.11078.

W. Wang. 2019. Calculating political bias and fighting partisanship with ai. *The Bipartisan Press*.

Z. Yang, Zihang Dai, Yiming Yang, J. Carbonell, R. Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*.

Zichao Yang, Diyi Yang, Chris Dyer, X. He, Alex Smola, and E. Hovy. 2016. Hierarchical attention networks for document classification. In *HLT-NAACL*.

Jingqing Zhang, Y. Zhao, Mohammad Saleh, and Peter J. Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *ICML*.