Effect of Political Bias on Misinformation Identification

Shankhin Brahmavar
Student
McGill University, Department of Electrical And Computer Engineering
shankhin.brahmavar@mail.mcgill.ca

Zhe Fan
Student
McGill University, Department of Neuroscience and School of Computer Science
zhe.fan@mail.mcgill.ca

Abstract

Recent advancements in natural language processing have resulted in state-of-art models that could effectively perform fake news classification at high performance (Kumar et al., 2019). However, text classification could be prone to underlying biases of sensitive words, especially if they have political implications. For example, popular NLP word embeddings such as GloVe and BERT were revealed to contain gender and religious biases (Bolukbasi et al., 2016; Liang et al., 2020). Past studies on political bias have not been done extensively in the domain of misinformation classification; however, the presence of political bias on social media continues to influence human opinions in everyday life. In this study, we will explore the effect of political debiasing on misinformation classification from a series of news excerpts. We will first identify and summarize the presence of political bias in GloVe word embeddings and perform hard debiasing on the dataset vocabulary. Both the GloVe and debiased GloVe embeddings will be used to train a BiLSTM, whose output cell states are passed into a linear artificial neural network (ANN) for classification. We hypothesize that political bias is present in GloVe embeddings and that the classification model trained on debiased GloVe embeddings would achieve comparable or better results on misinformation classification than prior to debiasing. Furthermore, we hypothesize that the debiased network is less likely to classify misinformation based on political bias. Hopefully, this work will lead to future advancements in the establishment of equitable, impartial AI classifiers for real-world applications.

1 Introduction

Through recent decades, scientists have developed an interest in understanding the ways information could spread and influence the general population. In particular, various kinds of information and misinformation have proliferated widely on social media (van der Linden et al., 2020). Recently, misinformation about the COVID-19 pandemic has been characterized as an "infodemic" (van der Linden et al., 2020), which includes but is not limited to conspiracy theories and political accusations (Angelis et al., 2022). Disinformation like this has the potential to influence public opinion and lead to grossly distorted public perceptions, which may hinder good decision-making at an institutional level, such as in public sectors in public health, governance, and politics. As such, fair, equitable ways of identifying and mitigating misinformation may be crucial to preserving the well-being of society at large.

1.1 Motivation for Proposed Study

Recent work done on fake news classification has achieved impressive performance and fast training time. Current state-of-art technology for misinformation classification, such as various LSTM classifiers and transformers, has achieved an F1 score of 0.681 and 0.702, respectively, when trained on the ISOT data set and tested on the Combined Corpus dataset (Blackledge and Atapour-Abarghouei, 2021). However, one of the aspects that are commonly overlooked in studies of misinformation classification is the presence of bias in word embeddings. Historically, certain media and users on social networks have spread misinformation targeted toward certain political entities/concepts to influence public opinions (Angelis et al., 2022). This phenomenon has the potential to encourage an individual's belief in misinformation based on the described political concepts, as suggested by the theory of confirmation bias (Modgil et al., 2021). For example, an avid opponent of Donald Trump may subconsciously believe every negative portrayal of the political figure to be true, even when they are false. Furthermore, massive negative public opinions about a certain political campaign may lead to a disproportionate amount of fake news related to it in order to steer public opinion. In

an effort to make an objective, impartial classification of misinformation, how do we ensure that the language model is not relying on any bias to make decisions? Although it is known that word embedding algorithms have exhibited stereotypical biases such as gender and racial bias (Bolukbasi et al., 2016), the role political bias plays in natural language classification has remained unclear. Thus, our goal is to determine whether the predictions of a misinformation classification model are dependent on inherent political biases of the input word embeddings.

1.2 Investigation of Bias in GloVe Embeddings

To examine the bias problem, we hypothesize that debiasing the word embeddings of a language classification model's vocabulary could achieve bias-free classification. In particular, our goal is to determine a political subspace from GloVe embeddings based on political concepts of the U.S. governance system. We will be using a list of political attribute words determined by (Gordon et al., 2020) and additional words of our choice to generate the political bias subspace. The extent of political bias in our dataset vocabulary will be evaluated with the direct bias test proposed by (Bolukbasi et al., 2016).

1.3 Investigation of the Dataset

We will be training and testing our model on the ISOT fake news dataset (Ahmed et al., 2017). In determining our training dataset for the language model, our goal is to show that political debiasing would allow the language model to predict without relying on political bias in word embeddings. For our training and test sets, we ensure that the labels for each political category are balanced. More details about the dataset can be found in the **dataset and evaluation section**.

Our selected language processing model is a BiLSTM coupled with a linear ANN to perform the classification task. The same model will be trained using the original GloVe embeddings and the debiased GloVe embeddings under the same hyperparameter settings. Although more efficient transformer-based models of text classification exist, the study focuses on the presence of political bias; thus, as our primary goal is to examine the effect of political bias on a language classification task, the usage of a sequential model is compatible with the embedding of interest, GloVE, which we anticipate to contain political bias. We anticipate that the application of political debiasing should have a net positive effect on the performance of classification on the validation/test set, or at the very least, perform similarly to a pre-debiased model. We will be evaluating our model through standard machine learning metrics: accuracy, precision/recall, and F1 score.

2 Related Work

2.1 Word Representations

Past work has been carried out for the task of misinformation identification using various natural language processing models. For example, Probierz et al. (2021) created Term Frequency Inverted Document Frequency (TF-IDF) vectors of news N-grams to train various machine learning models such as gradient boosting, random forests, classification and regression trees, and support vector machines (SVMs) to perform classification. They achieved good performance accuracy, especially with SVM. However, the limitation of the TF-IDF embedding method is clear: the method is reliant on the frequency of word occurrences and does not capture high-level characteristics of semantic meaning, meaning that similarity measurements between words cannot be performed (Qaiser and Ali, 2018). Furthermore, the TF-IDF embeddings are sparse, which augments the computational expensiveness of the processing problem (Chakraborty et al., 2019). Alternative word embedding approaches, such as word2vec and GloVe, resolved these problems elegantly by taking advantage of word co-occurrences in the trained corpus (Dessí et al., 2021). The semantic meaning of two individual words could be examined by taking their cosine similarity, and the debiasing of word embeddings could also be performed by subtracting a specific bias component from a target embedding (Bolukbasi et al., 2016). As we are interested in the identification of political bias, we choose to study GloVe embeddings that could better represent the semantic meanings of words.

2.2 Bias in Word Embeddings

The presence of bias has been a long-debated and difficult problem to combat in natural language processing. Bolukbasi et al. (2016) showed that gender bias in word embeddings can be eliminated by introducing the concept of a gender subspace, which can be computed from the collection of gender-

specific word embeddings. Bolukbasi et al. were able to show that effective debiasing of gender components from gender-neutral words did not adversely affect language coherence and analogy-solving abilities, suggesting that language learning models could still make inferences without relying on biased representations of texts.

Recently, the work of Gordon et al. (2020) provided a way of quantifying political bias in language models. Gordon et al. (2020) determined two binary political extremes along a single axis in order to generate a political bias subspace which could be used to generate political components in a list of target words. Political debiasing could be performed by removing the aforementioned components from the embeddings of the target words. Gordon et al.'s approach inspired us to adopt a similar approach for the analysis of political bias in GloVe embeddings. We will be selecting a list of politically sensitive words from our dataset vocabulary in order to produce a political bias subspace specific for GloVe embeddings.

2.3 Deep Learning for Misinformation Classification

Recent studies in misinformation classification took the deep learning approach and made use of state-of-the-art deep learning models like CNNs, LSTMs, ensemble methods, and attention mechanisms (Kumar et al., 2019). Kumar et al. (2019) used a combined CNN + BiLSTM ensembled network with attention to achieve an accuracy of 88.78% on fake news identification from a dataset of tweets as well as media sources such as PolitiFact. Liu and Guo (2019) demonstrated BiLSTM-based approaches performed well on text classification tasks, achieving an accuracy of up to 94.0% on the Subj dataset (Pang and Lee, 2004). Our study will take a similar approach, leveraging an ensemble network of BiLSTM and ANN to perform news classification over the ISOT dataset. Our model will leverage certain text processing features we believe to be important for text classification. First, the usage of a BiLSTM, as opposed to UniLSTM or RNN, ensures that our model takes advantage of parallel processing and memory to perform context-based analysis of data (Siami-Namini et al., 2019). This is important as the determination of fake news is dependent on the context of each parsed word, in a bi-directional manner. Consider the following phrase

```
The apple is blue ...
```

This is likely a false statement, however, with additional context that follows the sentence:

```
The apple is blue
because I painted it.
```

The sentence is now plausible. A sequential classifier such as an UniLSTM may not be able to leverage the additional context that follows the word "blue." As such, for the purpose of fake news classification, a bidirectional model may be more suitable for the task.

Current state-of-art language technologies such as BERT can achieve impressive performance on text classification tasks by leveraging transformers and context-based embedding creation (Devlin et al., 2018). However, as our primary goal is to examine the effect of bias in GloVE embeddings on classification, an LSTM-based approach that utilizes these embeddings is suitable for our studies.

3 Modelling

3.1 The Baseline Model

We used the following design for the baseline model's forward pass. Given an example x, which is a sentence, we embedded each word token using the 300-dimension GloVE embeddings, which include the embeddings of '<padding>' and '<unknown>' tokens for padded words and unknown words, initialized as zero vectors. The sentence embeddings were input into a BiLSTM model initialized using the following parameters: input size = 300 (to match the embedding dimension), hidden size = 50, and the number of LSTM layers = 2. Before the forward pass, the sentences were zero-padded based on the length of the longest sentence in each batch. The zero-padded tokens were masked during the training process, ensuring that backpropagation would not be performed over the parameters associated with the padded values.

The BiLSTM model would output a set of values for each input sentence: the output tensor, composed of the predicted value for each input token (in the case of a BiLSTM, the value is a concatenation of the forward and reverse states at each time step in the sequence); the output hidden states, which contained the final hidden state of each LSTM layer after the final time step; lastly, the final cell states, which contained the final cell state of each LSTM layer after the final time step. We chose to further

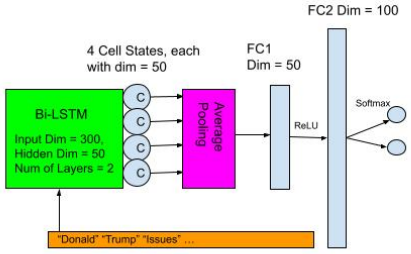

Figure 1: The structure of the BiLSTM + ANN model. Here, FC refers to a fully connected layer, one that is found in a linear artificial neural network.

process the cell states by performing an average-pooling operation over them for subsequent classification to utilize the model's long-term memory for context-dependent classification of the input text. The output of the pooling operation is then used as input to a 3-layer ANN, initialized with 100 hidden units. The input layer is activated via the ReLU activation, while the output layer is activated with the softmax activation to produce a binary vector of probabilities. The prediction of the ANN is a two-dimensional one-hot encoding of the softmax output. See Figure 1 for more details. The model was optimized to minimize the binary cross-entropy loss (BCE) with the target labels. Here, $\hat{Y}_i$ represents the predicted label for the i^{th} example, and Y_i represents the true label for the i^{th} sentence example.

$$L_{BCE} = \frac{-1}{n}\sum_{i=1}^{n}(Y_i \cdot \log \hat{Y}_i + (1 - Y_i) \cdot \log (1 - \hat{Y_i}), \tag{1}$$

This model is a suitable choice for the baseline as it employs the use of GloVE embeddings, and it is a sequential, bidirectional model suitable for parsing sentences with long-range dependencies. We hypothesize that this model would be able to perform fake news classification by optimizing the loss function and tuning all gradients in the model while preserving the embedding layer's initial embeddings.

3.2 Our Proposed Model

The proposed model used the same aforementioned parameters as the baseline, but the GloVE embeddings were debiased using (Bolukbasi et al., 2016) HardDebias strategy: we identify a collection of word-pairs $D_1, D_2, ..., D_n$ which can be used to determine the political subspace, which is given by the top principle component of the pair difference vectors calculated using each word-pair D_i.

$$\mu_i := \sum_{w \in D_i} \frac{\overrightarrow{w}}{|D_i|} \tag{2}$$

$$C := \sum_{i=1}^{n} \sum_{w \in D_i} \frac{(\overrightarrow{w} - \mu_i)^T (\overrightarrow{w} - \mu_i)}{|D_i|} \tag{3}$$

In the equations above, μ_i is the average of each word vector w in D_i. Then the political bias subspace is given by the first k rows of the singular value decomposition of C. In this case, we choose $k = 1$ to obtain the first direction of the subspace and use that as our principle bias direction, this is the axis of political biasedness for our GloVE embeddings. To perform hard debiasing over our vocabulary, we project every word embedding w in the GloVE embedding onto the bias subspace B to obtain w_B. We obtain $w_{debiased}$ by subtracting w_B from w.

$$w_{debiased} = w - proj_B(w) \tag{4}$$

The debiased embeddings would constitute the new debiased GloVE embedding for the proposed model. By using a debiased set of embeddings, we hypothesize that the proposed model needs to rely on word embeddings that lack political implications when making a classification decision.

4 Dataset and Evaluation

4.1 Dataset

The ISOT fake news dataset by Ahmed et al. (2017) compiled a total of 23481 articles classified as "fake" by the Politifact organization, a fact-checking organization based in the U.S. The remaining 21417 articles were classified to be "true," sourced from Reuters.com. Reuters news is a news editorial, as well as a signatory of the International Fact-Checking Network (FCN) (Reuters, 2022). Although it is presumptuous to say that any news source, or even news-verification service, is perfectly unbiased, the relatively high credibility of news from Reuters is supported by recent surveys conducted by multiple news credibility ranking

organizations. Factual, a news-verification organization based in the U.S., placed Reuters at the 82nd percentile for the credibility of the editorial's articles (Meylan, 2022). Furthermore, Factual classified Reuters as a "center" biased news source, meaning that news articles from Reuters are relatively politically neutral compared to other competitors, which may lean more left/right. AllSides.com, another news verification platform, gave Reuters a rating of "center" in terms of political biasedness. This conclusion is determined from three editorial reviews, 34064 community ratings, as well as a blind survey conducted in November 2020 (Gable and McDonald, 2022). Ad Fontes Media, a media bias ranking organization based in the U.S., gave Reuters a reliability rating of 47.48 and a bias of -1.45; according to Fontes Medias' rating strategy, Reuters is placed in the "Most Reliable for News" category among about 2000 news sources (Otero, 2022). Additionally, the non-profit fact-checking organization Media Bias/Fact Check (MBFC) has given Reuters a biased rating of "Least Biased" and a factual rating of "Very High" (Zandt, 2022). Overall, the ratings of the Reuters news site from various sources suggested that it has relatively high credibility.

The ISOT dataset is close to balanced, and the number of fake examples is comparable to the number of real examples (Figure 2). The average length of the article titles and content is 2501.139. We also gathered the frequency statistic for the top-appearing words in the dataset. We noticed several trends in both the true article set and the false article set. For the true articles, the word "Reuters" dominated the corpus, as most crawled articles from Reuters.com mentioned the name of the editorial in their content. For the false articles, the word "video" dominated the corpus, and it was the second most frequent term in the false articles even though it was not found in the most frequent terms of the true articles. To avoid training the model to classify using a heuristic based on a few dominating, yet politically neutral words that differentiated the two sets, we removed these "outliers" from the ISOT dataset. As suggested by the figures below, in both sets, the term "Trump" dominated the dataset. Among the top ten terms in each set were other political terms such as "Hillary," "Obama," and "House." Expanding the list allowed us to determine a set of top political terms that appeared in the dataset (Figure 5), which was used to determine the political subspace. See Figures 3 and 4 for more details.

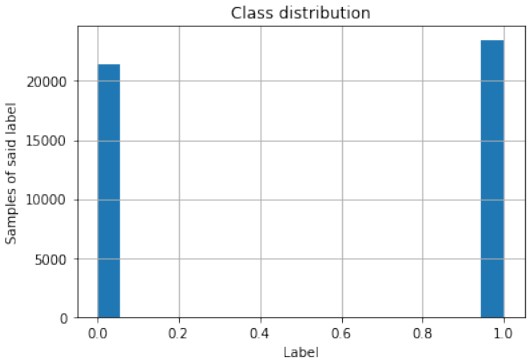

Figure 2: Class distribution for ISOT dataset

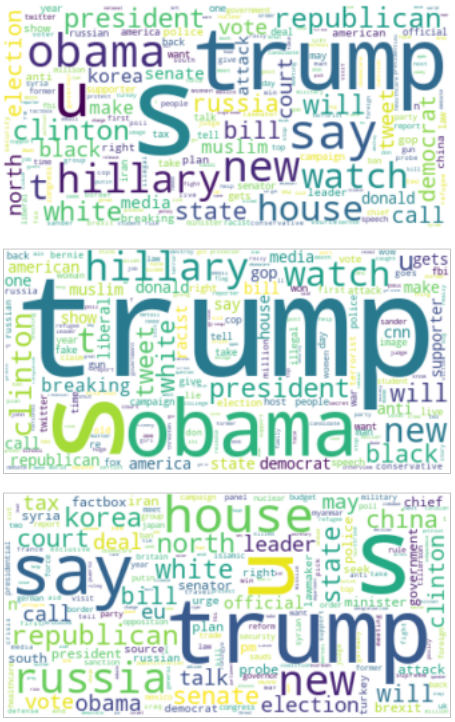

Figure 3: The word clouds generated from: (top) The entire ISOT dataset, (middle) the fake articles of the ISOT dataset, (bottom) the real articles of the ISOT dataset. Note that words with a length of 2 or less, as well as "outlier" words were ignored.

For each article entry in the ISOT dataset, the article's content was concatenated with the article's title. The dataset was shuffled and split with a train/test ratio of 70% vs. 30%. The resulting training set has 47.4% true labels, while the testing set has 48.4% true labels. Additional pre-processing of the training set was performed to remove punctuations and stop words using the Natural Language Tool Kit (NTLK) library (Steven Bird, 2009).

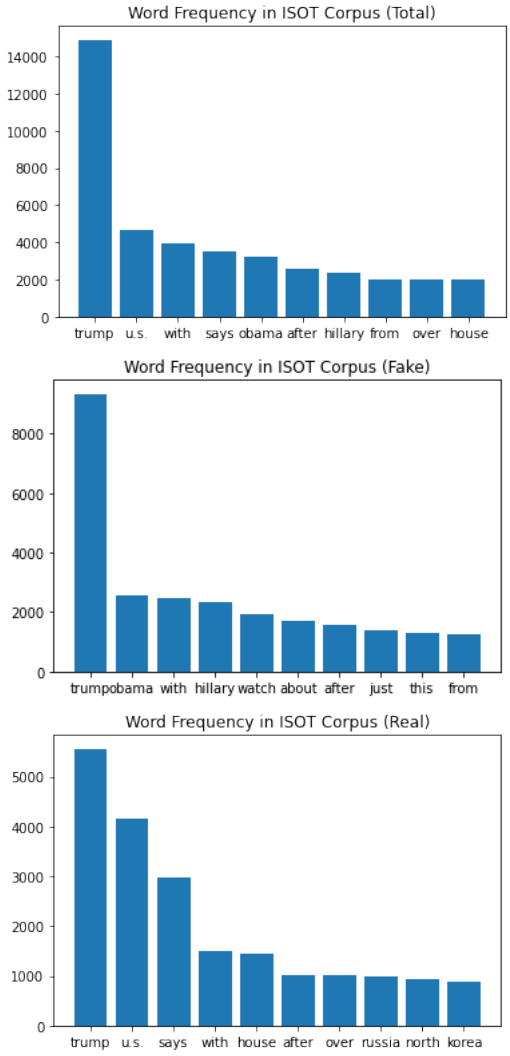

Figure 4: The top 10 word frequencies in the ISOT Corpus (ignoring words with length of 2 or less, as well as removing "outlier" words in the fake and real set), obtained from: (top) The entire ISOT dataset, (middle) the fake articles of the ISOT dataset, (bottom) the real articles of the ISOT dataset.

4.2 Evaluation Metrics

The extent of political bias in our dataset vocabulary was evaluated with the direct-bias test (Bolukbasi et al., 2016). The direct bias test was performed on a set of words that we perceived to be politically neutral. Here, N represents the set of chosen words, w is the word embedding for a word in N, g is the political subspace, and c is an exponent, chosen to be 0.25.

$$DirectBias_c = \frac{1}{|N|} \sum_{w \in N} |\cos(w, g)|^c \quad (5)$$

The direct bias metric computes bias based on the average absolute cosine similarity between the chosen word embedding and the political direction. If a word is highly biased, either positively or negatively, to the political subspace, its cosine similarity with the political subspace would be relatively high. By accumulating the cosine similarity between every word and the political subspace raised to a positive exponent, a set of words with high political bias would have a higher direct-bias score than a set of words with low political bias.We anticipate that successful debiasing of the word embeddings would result in a reduction of the direct bias between the chosen word set and the political subspace.

```
left_keywords = [ 'left','hillary',
  'obama', 'democrat', 'liberal',
      'clinton', 'socialism',
     'socialist', 'communism',
 'liberals', 'refugees', 'refugee']
right_keywords = ['right','trump',
      'republican', 'donald',
   'conservative', 'capitalism',
  'capitalist', 'bush', 'mccain',
    'conservatives', 'illegals',
          'illegal']
```

Figure 5: The list of corresponding terms used to generate the political subspace. Some of these words were chosen from the top appearing words from the ISOT corpus, while others were chosen based on their antagonistic

We evaluated and compared our original and debased model's predictions through standard machine learning metrics: accuracy, precision/recall, and F1 score over the test set. The accuracy metric captures the proportion of correct predictions over all predictions made by the model, while the F1 score captures the extent of error while taking into account the balance of labels in the dataset. Given that our training set has a balanced distribution of true and false labels, both metrics would be useful for understanding the performance of the classification model. For each of the following metrics, a well-performing model would obtain a result close to 1, while a poor-performing model would obtain a result close to 0.

$$Accuracy = \frac{(TruePositives + TrueNegatives)}{TotalPredictions} \quad (6)$$

$$Precision = \frac{(TruePositives)}{TruePositives + FalsePositives} \quad (7)$$

$$Recall = \frac{(TruePositives)}{TruePositives + FalseNegativess} \quad (8)$$

$$F1score = \frac{2 * (precision * recall)}{precision + recall} \quad (9)$$

5 Experiments

For our proposed models (see the **Modeling** section for details), we apply the following hyperparameters: input size = 300 (to match the GloVE embedding dimension), hidden size = 50, and the number of LSTM layers = 2. We chose a batch size of 32 and a learning rate of 0.01, which resulted in no problems during the optimization using stochastic gradient descent (SGD). We repeat 3 runs for each model and record the average training accuracy and training loss over 5 epochs. The model was implemented using PyTorch version 1.13.0. The GPU used was the Tesla K80 GPU provided by Google Colaboratory. Given the limited memory provided by Google Colaboratory, the number of training epochs is set to be 5 for both the baseline and the debiased model, which was sufficient for the training loss to plateau. On average, each epoch took about 2.5 minutes to train. The code repository can be found in the **Appendix** section.

6 Results & Discussions

6.1 Results

Results can be seen in Tables 1, 2 and 3.

Metric/Runs	1	2	3	Average	Standard Deviation (+/-)	p-value
Loss(training)	0.163	0.168	0.176	0.169	0.005	p>0.05
Accuracy (Training)	0.941	0.940	0.936	0.939	0.002	p>0.05
F1 (test)	0.944	0.940	0.937	0.940	0.003	p>0.05
Accuracy (test)	0.945	0.942	0.938	0.941	0.003	p>0.05
Precision (test)	0.944	0.942	0.938	0.941	0.003	p>0.05
Recall (test)	0.945	0.941	0.939	0.941	0.003	p>0.05

Table 1: The training and test performance metrics of the baseline model (no debiasing) where p-value is the two -tailed p-value taken from a t-test (compared to debiased metrics).

Metric/Runs	1	2	3	Average	Standard Deviation (+/-)
Loss(training)	0.181	0.141	0.179	0.167	0.019
Accuracy (Training)	0.937	0.951	0.939	0.942	0.006
F1 (test)	0.936	0.953	0.939	0.942	0.007
Accuracy (test)	0.937	0.954	0.940	0.943	0.007
Precision (test)	0.937	0.953	0.940	0.943	0.007
Recall (test)	0.937	0.954	0.940	0.944	0.008

Table 2: The training and test performance metrics of the debiased model.

Embeddings/Metric	Direct-Bias (Whole ISOT dataset)
Baseline	0.691
Debiased	8.04e-5

Table 3: The Direct-Bias metric of the baseline/debiased model.

6.2 Discussion & Analysis

The direct-bias metric was performed over the top 30 most appeared words (excluding words with perceived political implications and stopwords) of the ISOT corpus. In the baseline model, the 300-d GloVE embeddings accumulated a direct-bias of 0.691. After debiasing, the direct bias decreased to 8.04 e-5 (Table 3). This suggests that political bias was initially present in the words of the ISOT corpus, and the debiasing successfully removed the political bias in these words.

Interestingly, we did not observe a significant difference between the average classification metrics on the test set for the baseline and the modified model (Table 1 and 2). The performance of both models on the ISOT dataset is considered to be high. Prost et al. (2019)'s study of gender debiasing on text classification identified a similar trend between the performance metrics of the baseline GloVE embedding model and the debiased embedding models. Prost et al.'s study focused on the classification of author occupations from biographies from the BiosBias dataset (De-Arteaga et al., 2019) using a deep neural network (DNN). However, the accuracy metric of the debiased model (0.817) did not beat the baseline (0.818). The debiasing strategy we employed is analogous to what Prost et al. (2019) described as *strongly debiased embeddings*, whereby the debiasing procedure is applied to the entire vocabulary of the corpus, including the political attribute words we used to determine the political subspace. Given these results, our debiased model may not be heavily reliant on the political bias of word embeddings to make classification decisions, even though the information reduction in the political subspace component may have reduced the underlying variance of the sentence embeddings distribution that characterized true and false news. Given these results, our hypotheses are only partially confirmed: while we observed the presence of political bias in 300-dimension GloVE embeddings, the implementation of its debiased version in a BiLSTM + ANN language production model did not decrease the overall performance of the model on fake news classification.

To better understand our results, we performed direct bias computation over the set of true and false articles (named true set and fake set, respectively) in the ISOT dataset to see if the two sets had distinguishable differences. To further illustrate the potential effect of bias in classification, we computed the direct-bias scores for 30 top-appearing words (including political attribute words, excluding stop words) in the true and false subsets of the ISOT corpus.

Before debiasing, the true set had a direct-bias score of 0.699, which decreased to 7.79 e-5 after debiasing. Before debiasing, the fake set had a direct-bias score of 0.704, and a direct-bias score of 8.02e-5 after debiasing (Table 4). Thus, both sets contained a similar level of intrinsic political bias. However, since the direct bias of a word with a bias subspace is a scalar quantity, it cannot be used to determine whether the word is positively or negatively biased. Thus, we adopt a strategy to analyze the polarity of political components of the real vs. fake dataset prior to debiasing: Given the word frequency map computed from a dataset (e.g. a list containing tuples of the form (w_i, F_i), where w_i is a word and F_i is the appearance frequency of the word w_i in a dataset), we proposed a measure of the total component score, adapted from Prost et al. (2019) as follows: for each word embedding w_i, compute its political component C_i by projecting onto the political subspace, and multiply C_i by the frequency F_i, and sum the result C_iF_i for all i. The sum will then be divided by the sum of all F_i . The intuition is that for words that appear more frequently, we will accentuate them by multiplying them with their appearance frequency. The political component of the dataset would be an average of the political component of all the individual words found in a dataset:

$$PoliticalComponentOfDataset = \frac{1}{\sum_{alli} F_i} \sum_{alli} C_i F_i \tag{10}$$

The calculated component of the fake news dataset has a cosine similarity of 1 with the political subspace, and the component's magnitude is 0.793. The calculated component of the real news dataset has a cosine similarity of 1 with the political subspace, and the component's magnitude is 1.03. (Table 5)

Embeddings/Metric	Direct-Bias (True ISOT dataset)	Direct-Bias (False ISOT datset)
Baseline	0.699	0.704
Debiased	7.79e-5	8.02e-5

Table 4: The Direct-Bias over the top 30 most appeared words (including political words) in the true and false sets of the ISOT dataset.

	Data Political Component (True Label)	**Data Political Component (False Label)**
Cosine Similarity with Political Subspace	1	1
Magnitude	1.03	0.793

Table 5: The average political component of the True and False dataset from the ISOT dataset.

Thus, it can be observed that both the real and fake datasets' political components have a positive cosine similarity with the political subspace prior to debiasing, indicating that both datasets are politically biased in the same direction. With a higher component magnitude, the real dataset is slightly more biased, but since both datasets are biased in the same direction, it is less likely that the baseline model could be making classification decisions based on a difference in the polarity of the intrinsic political bias in true articles vs. false articles. Given that the difference in bias direction and magnitude is relatively small between the true and false datasets prior to debiasing, debiasing may be unlikely to affect the crucial underlying patterns that the model learned to differentiate these two datasets prior to debiasing.

In another light, strong debiasing of the dataset corpus may have removed all differences in political bias altogether, eliminating variances in the dataset that is attributed to political bias within the text. This information reduction may discourage the model from making inferences based on political information contained in an article and force it to make inferences based on some other underlying information. The fact that the debiased model was able to match the baseline in performance may suggest that it is an effective, unbiased approach to text classification as well, indicating that the model is able to make correct predictions using sentence embeddings that did not vary in the political subspace.

6.3 Qualitative Analysis

Several examples of title + article predictions from both the biased and debiased models were studied. These examples can be found in Table 7 of the **Appendix** section. In the second sentence example containing multiple instances of the word "Trump," one of the right-representing political words included in the political attribute word list, and the top occurring word in the ISOT dataset, resulted in an incorrect prediction of "True" by the biased model but a prediction of "False" by the debiased model, which was correct. In the first sentence example, in which the word "Trump" did not appear, the biased model gave a correct prediction of "False," and the debiased model gave an incorrect prediction of "True." Although these qualitative analyses are not sufficient to conclude about the pattern of sentences classified as true or false by each model, it may be interesting to pose the future study on the word distribution of the sentences that received a true prediction or false prediction for each model to learn if there exists any bias for particular words/concepts for producing a classification decision.

7 Conclusion

In this study, we explored the process of political debiasing in co-occurrence-based word embeddings such as GloVE. We also investigated the effect of political debiasing on fake news classification using a BiLSTM + ANN model. The study's main takeaways include the following: 1. The presence of political bias in GloVE embeddings can be quantified and removed using the direct-bias metric and a computed political subspace; 2. The removal of political bias does not decrease the performance of fake news classification on the ISOT dataset using a BiLSTM-based model. In any case, the removal of political bias is a simple and fast task that can be performed on language classification models that uses word embeddings, creating fairer representations of an AI without harming its performance. However, there are several potential future directions to improve this study further. For example, one could compare the model performance by including or mixing the ISOT dataset with other fake news classification datasets, such as PolitiFact (Shu et al., 2017), and Fakeddit (Nakamura et al., 2019). This may help improve the model's ability to generalize over more article examples, and these datasets may also contain more varied political bias between the true and false subsets. Aside from GloVE embeddings, it may also be interesting to examine transformer-based approaches to fake news classification and whether intrinsic political bias could be determined from a transformer-based embedding to produce more equitable and efficient

models of language AIs. Overall, we see hope in the boundless potential of fair, bias-free models of natural language processing.

8 Contributions

The authors contributed equally to this project. Zhe designed and tested the classification model while Shankhin performed bias analysis on GloVE word embeddings.

9 Code Availability

The code used to generate the results of this study is accessible at: https://github.com/Zivvvo/COMP-599-Project

References

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2017. Detection of online fake news using n-gram analysis and machine learning techniques. *Lecture Notes in Computer Science*, pages 127–138.

Andrea De Angelis, Christina E. Farhart, Eric Merkley, and Dominik A. Stecula. 2022. Editorial: Political misinformation in the digital age during a pandemic: Partisanship, propaganda, and democratic decision-making. *Frontiers in Political Science*, 4.

Ciara Blackledge and Amir Atapour-Abarghouei. 2021. Transforming fake news: Robust generalisable news classification using transformers. *2021 IEEE International Conference on Big Data (Big Data)*.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings.

Rupak Chakraborty, Ashima Elhence, and Kapil Arora. 2019. Sparse victory – a large scale systematic comparison of count-based and prediction-based vectorizers for text classification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 188–197, Varna, Bulgaria. INCOMA Ltd.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 120–128, New York, NY, USA. Association for Computing Machinery.

Danilo Dessí, Rim Helaoui, Vivek Kumar, Diego Reforgiato Recupero, and Daniele Riboni. 2021. Tf-idf vs word embeddings for morbidity identification in clinical notes: An initial study.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Joseph Gable and Scott McDonald. 2022. Reuters media bias rating. Retrieved December 10, 2022, from https://www.allsides.com/news-source/reuters.

Joshua Gordon, Marzieh Babaeianjelodar, and Jeanna Matthews. 2020. Studying political bias via word embeddings. In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 760–764, New York, NY, USA. Association for Computing Machinery.

Sachin Kumar, Rohan Asthana, Shashwat Upadhyay, Nidhi Upreti, and Mohammad Akbar. 2019. Fake news detection using deep learning models: A novel approach. *Transactions on Emerging Telecommunications Technologies*, 31(2).

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations.

Gang Liu and Jiabao Guo. 2019. Bidirectional lstm with attention mechanism and convolutional layer for text classification. *Neurocomputing*, 337:325–338.

Phillip Meylan. 2022. Is reuters reliable? Retrieved December 10, 2022, from https://www.thefactual.com/blog/is-reuters-reliable/.

Sachin Modgil, Rohit Kumar Singh, Shivam Gupta, and Denis Dennehy. 2021. A confirmation bias view on social media induced polarisation during covid-19. *Information Systems Frontiers.*

Kai Nakamura, Sharon Levy, and William Yang Wang. 2019. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *arXiv preprint arXiv:1911.03854*.

Vanessa Otero. 2022. Reuters bias and reliability. Retrieved December 10, 2022, from https://adfontesmedia.com/reuters-bias-and-reliability/.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts.

Barbara Probierz, Piotr Stefański, and Jan Kozak. 2021. Rapid detection of fake news based on machine learning methods. *Procedia Computer Science*, 192:2893–2902.

Flavien Prost, Nithum Thain, and Tolga Bolukbasi. 2019. Debiasing embeddings for reduced gender bias in text classification. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 69–75, Florence, Italy. Association for Computational Linguistics.

Shahzad Qaiser and Ramsha Ali. 2018. Text mining: Use of tf-idf to examine the relevance of words to documents. *International Journal of Computer Applications*, 181.

Thomson Reuters. 2022. About reuters fact check. Retrieved December 10, 2022, from https://www.reuters.com/fact-check/about.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36.

Sima Siami-Namini, Neda Tavakoli, and Akbar Siami Namin. 2019. A comparative analysis of forecasting financial time series using arima, lstm, and bilstm.

Edward Loper Steven Bird, Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media, Inc.

Sander van der Linden, Jon Roozenbeek, and Josh Compton. 2020. Inoculating against fake news about COVID-19. *Frontiers in Psychology*, 11.

Dave M. Van Zandt. 2022. Reuters. Retrieved December 10, 2022, from https://mediabiasfactcheck.com/reuters/.

A Appendix

The real news is collected from Reuters and the fake news is collected from various sources that were flagged to be unreliable by Politifact (a fact-checking organization in the USA) and Wikipedia. Each entry consists of the label (1 for fake, 0 for real), article text, article title, the article type (World news etc.) and the publishing date of the article as seen in the example set of data in Table 6.

Label	Article Title	Article Text	Type	Date Published
1	Donald Trump Sends Out Embarrassing New Year,Äôs Eve Message; This is Disturbing	Donald Trump just couldn t wish all Americans a Happy New Year and....	News	December 31, 2017
0	As U.S. budget fight looms, Republicans flip their fiscal script	WASHINGTON (Reuters) - The head of a conservative Republican faction in the U.S....	politicsNews	December 31, 2017
1	Drunk Bragging Trump Staffer Started Russian Collusion Investigation	House Intelligence Committee Chairman Devin Nunes is going to have a bad day....	News	December 31, 2017
0	U.S. military to accept transgender recruits on Monday: Pentagon	WASHINGTON (Reuters) - Transgender people will be allowed for the first time to....	politicsNews	December 29, 2017
1	Sheriff David Clarke Becomes An Internet Joke For Threatening To Poke People ,ÄòIn The Eye,Äô	WASHINGTON (Reuters) - On Friday, it was revealed that former Milwaukee Sheriff David Clarke, who was being considered for...	News	December 30, 2017
0	Senior U.S. Republican senator: 'Let Mr. Mueller do his job'	WASHINGTON (Reuters) - The special counsel investigation of links between Russia....	politicsNews	December 31, 2017

Table 6: Sample from dataset in tabular format (Article text shortened for brevity)

Example Sentence	Baseline (Biased) Prediction	Debiased Prediction	True Label
benghazi spokesliar susan rice tells cnn we should expect iran to use funds it gets for terrorist operations tell us susan what s worse iran with a nuclear weapon and billions of dollars to help fund muslim terrorists or iran with a nuclear weapon and frozen assets president obama s national security advisor susan rice told cnn s wolf blitzer on wednesday that we should expect iran to use the money it gets under sanctions relief for military and even terrorist operations.as part of obama s nuclear agreement with iran tens of billions of dollars frozen as part of the sanctions on iran will be released over time provided iran complies with a list of deadlines outlined in the agreement. what do we think they ll spend that money on we think for the most part they re going to need to spend it on the iranian people and...	False	True	False
wow the washington post publishes realnews story about president trump under trump s leadership gains against isis have dramatically accelerated the washington post nearly a third of territory reclaimed from the islamic state in iraq and syria since 2014 has been won in the past six months due to new policies adopted by the trump administration a senior state department official said friday.brett mcgurk the state department s senior envoy to the anti islamic state coalition said that steps president trump has taken including delegating decision making authority down from the white house to commanders in the field have dramatically accelerated gains against the militants.combined islamic state losses in both countries since the group s peak control in early 2015 total about 27 000 square miles of territory...	True	False	False
all whites in back democrats prove their obsession with race in one ridiculous photo nothing says embracing diversity like dividing interns by color and kicking the white interns to the back of the photo.it s pretty fitting that the race obsessed us rep. from texas sheila jackson lee would post such a telling photo on twitter the democratic interns on capitol hill 2016 deminternselfie pic.twitter.com zzuml4hkoc sheila jackson lee jacksonleetx18 july 20 2016the photo of dem interns was supposed to be in response to a paul ryan selfie that the democrats who can never see past the color of one s skin posted to show how much more diverse they are.so diversity is kicking white interns to the back of the photo h t weasel zipperssomething about this picture is eerily similar to the picture taken of the crowd taken that exposed the segregation of women and men during london s new muslim mayor sadiq khan s speech on the benefits of britain sticking with the eu. the muslim women were noticeably segregated from the men as the photo showed them standing behind the men like second class...	False	False	False

Table 7: Sample predictions from each model (articles shortned for brevity)