An Effective Approach for Informational and Lexical Bias Detection

Iffat Maab[1], Edison Marrese-Taylor[1,2], Yutaka Matsuo[1]
[1]The University of Tokyo
[2]National Institute of Advanced Industrial Science and Technology
{iffatmaab, emarrese, matsuo}@weblab.t.u-tokyo.ac.jp

Abstract

In this paper we present a thorough investigation of automatic bias recognition on BASIL, a dataset of political news which has been annotated with different kinds of biases. We begin by unveiling several inconsistencies in prior work using this dataset, showing that most approaches focus only on certain task formulations while ignoring others, and also failing to report important evaluation details. We provide a comprehensive categorization of these approaches, as well as a more uniform and clear set of evaluation metrics. We argue about the importance of the missing formulations and also propose the novel task of simultaneously detecting different kinds of biases in news. In our work, we tackle bias on six different BASIL classification tasks in a unified manner. Eventually, we introduce a simple yet effective approach based on data augmentation and preprocessing which is generic and works very well across models and task formulations, allowing us to obtain state-of-the-art results. We also perform ablation studies on some tasks to quantify the strength of data augmentation and preprocessing, and find that they correlate positively on all bias tasks.

1 Introduction

News outlets have expanded to become one of the most influential and prominent platforms within mass media. News articles play a central role in transforming individual and public opinion (Hamborg et al., 2019). Public opinion of journalists influence viewers to become biased towards a particular issue. The consequence of media bias is massive, and raises questions about the credibility of news. For example, Bernhardt et al. (2008) reported that media bias led to the election of the wrong candidate, and according to Wolton (2019), voters are always well-informed with unbiased rather than biased media outlets. Moreover, biased media outlets have also been found to provide less information, which reduces voter's welfare Duggan and Martinelli (2011).

Misinformation has enormous potential in changing the individual and public beliefs, expectations, or desired conclusions. The harmful biases contained in news media require rigorous analysis to be detected and quantified, but once addressing these issues can improve the quality of research process to maximize the accuracy and credibility of research results (Johnson et al., 2020). To enhance transparency and reliability in promoting accurate information, it is important to realize bias in machine learning methods and how humans perceive bias (Sun et al., 2019).

Seminal work by Fan et al. (2019) has been a key contribution in bias detection in news, with the introduction of the BASIL dataset. Critically, to the best of our knowledge, this dataset is the first to be annotated with different kinds of bias. This is due to the fact that, as pointed out by Fan et al. (2019), some kinds of bias can only be analyzed in a broader context, because a given sentence becomes ambiguous in isolation at any given time. In this sense, their proposal was to differentiate *informational bias* from *lexical bias*. On one hand, the former is usually presented using speculative, imperative, and tangential clauses that convey information in a factual and neutral way (van den Berg and Markert, 2020) to sway readers' opinions towards news entities (Guo and Zhu, 2022a), therefore depending mostly on the context. On the other hand, lexical bias instead depends on linguistic attributes like word choice and syntax and originates from content realization or how things or events are expressed, and is generally not depending on the context (Greene and Resnik, 2009; Hube and Fetahu, 2019; Iyyer et al., 2014; Yano et al., 2010; Recasens et al., 2013). Table 1 shows examples of informational and lexical bias on New York Times (NYT) news (2nyt; source: nyt, event: 2) on the BASIL dataset. The three sentences report

Proceedings of the Sixth Fact Extraction and VERification Workshop (FEVER), pages 66–77

Bias	Polarity	Sentence	Index #
Informational	Neg	**The president again suggested that he should win the Nobel Peace Prize, and he reviewed which conservative commentators had been supportive of him, while dismissing Ann Coulter, who has not.**	10
Lexical	Neg	Sounding alternately **defensive and aggrieved**, Mr. Trump explained his failure to secure wall funding during his first two years in office when Republicans controlled both houses of Congress by saying, I was a little new to the job.	11
Neutral	*None*	He blamed certain people, a particular one, for not having pushed this faster, a clear reference to former Speaker Paul D. Ryan of Wisconsin, a Republican.	12

Table 1: Examples extracted from BASIL dataset, from New York Times (NYT) section and discussing the same event, showing how informational and lexical bias manifest. In the examples, text spans annotated with informational bias and lexical bias are highlighted in bold, and we refer to sentences annotated with no bias as 'Neutral'.

on Donald Trump as the main event. As seen, informational bias covers the article context broadly, often depending on the complete sentence, whereas lexical bias is expressed in polarized words as highlighted.

Though the development of datasets such as BASIL has brought an interesting new paradigm to look at bias in news, thereby drawing the attention of several members of the research community and leading to abundant prior work, we note that these studies mostly focus on informational bias only (van den Berg and Markert, 2020; Guo and Zhu, 2022b; Fan et al., 2019; Lei et al., 2022), with lexical bias studied solely by Fan et al. (2019). While we surmise that this could be partially due to the fact that lexical bias appears comparatively less frequently in news articles, making the automatic detection of both kinds of bias simultaneously difficult, we think this should be no reason to focus on either and argue that detecting lexical bias from informational bias should be equally significant. In this context, our take is well-aligned with previous work by Zhou and Bansal (2020), who highlighted the uniqueness of lexical bias and demonstrated its importance through several experiments.

In light of this issue, in our work, we propose approaches to detect both informational and lexical bias simultaneously. Our holistic view on bias detection enables us to reveal important trade-offs between informational and lexical bias, and we perform a sensitivity analysis of such trade-offs in various task formulations. Based on our findings, we propose a set of data augmentation techniques which we combine with deep learning models to improve performance.

Furthermore, our focus on both bias detection problems leads us to propose more clear and consistent task formulations. Specifically, we note out several inconsistencies in the reporting of performance on previous work, and provide a framework to improve uniformity and clarity, and to avoid problems we found in prior work. Consequently, we highlighted incoherent task formulations of bias classification tasks, absence of a significant task of detecting both informational and lexical bias simultaneously, and missing evaluation metrics of bias interpreted labels. Our findings are connected to a related issue which has been brought to attention by van den Berg and Markert (2020), showing that sentence-based splits can introduce data leakage since labeled sentences from the same or similar articles can appear in training and test subsets, resulting in overestimation of the predicting accuracy. Our proposed framework not only enables us to adequately compare our results with previous work, which we often find lacks structure in this sense, but also helps pave the way for future research involving the BASIL dataset. In summary, our work provides the following contributions:

1. We unveil and remove inconsistencies in prior work, providing uniform and more clear evaluation metrics with various task formulations on BASIL.

2. To the best of our knowledge, we are the first to propose the task of simultaneously detecting both informational and lexical bias, and

distinguish them from neutral examples. We propose models to tackle this new task and establish its importance in overall bias detection problem setting.

3. We propose an approach based on data augmentation and preprocessing which is generic and works well across a selection of models and settings in bias detection in news, allowing us to obtain state-of-the-art results on BASIL.

2 Related work

There has been a growing interest in the investigation of linguistic information presented by neural models (Liu et al., 2019a). A study by Jia et al. (2019) worked on the bias and distortion of online commentary information based on online reviews, and illustrate online review components significance over increased reputation and false reviews. Hovy and Prabhumoye (2021) study five sources of bias including data, annotation process, input representations, models, and research design where research design is the most difficult to detect because it requires systematic analysis and subjectivity of human perceptions.

The work on BASIL dataset in the literature is not exhaustive to the best of our knowledge. BASIL dataset provided by Fan et al. (2019) use BERT and RoBERTa for informational and lexical bias detection while treating sentences in isolation, whereas informational bias is also explored with different types of contexts like textual, article, event and domain using BiLSTM's, BERT and Event Context-Inclusive Model (EvCIM) inspired by Context Aware Model of Papalampidi et al. (2019), we will omit a detailed explanation of these models, referring readers to (van den Berg and Markert, 2020). Fan et al. (2019) demonstrate that informational bias is more challenging to detect due to its dependence on content selection as compared to lexical bias. Another study by Lee et al. (2021) introduce a general purpose misinformation UnifiedM2 model for bias detection in BASIL and handle tasks like fake news, clickbait and rumors. Contrastive learning and Graph Attention Network in the MultiCTX (Multi-level ConTeXt) model uses triplets of BASIL to detect informational bias as proposed by Guo and Zhu (2022b). Similarly, bias sentence identification is also studied by Lei et al. (2022) through local and global discourse structures using RoBERTa for addressing bias. We compare our results with the aforementioned current state-of-the-art methods to investigate the significance of our proposed approach.

Another topic relevant to our work is data augmentation. In general, data augmentation in NLP is at an emerging stage compared to Computer Vision (Shorten et al., 2021; Shleifer, 2019). Among data augmentation techniques in NLP one that has recently shown excellent results in the context of machine translation is backtranslation (Sennrich et al., 2015). The idea is to pair monolingual data with an automatic backtranslation generated by a given model or external system, which led to considerable performance improvements on neural machine translation, showing that even for small amounts of in domain monolingual data, backtranslation is effective for domain adaptation. More recently, Edunov et al. (2018) scale backtranslation to hundreds of millions of monolingual sentences and achieve substantial improvements. The same approach is more recently been successfully adapted to other task. For example, Shleifer (2019) focus on sentiment classification and use backtranslation on only 500 examples, obtaining an 8.1% improvement over the augmentation-free baseline based on ULMFit (Howard and Ruder, 2018) on the IMDB dataset. Our work is similar to this, but we adapt the approach for bias detection.

Finally, our work is related to stemming, the process of linguistic normalization where variant forms of a word are converted to a stem or root word. There are different methods for stemming, but most of them follow a rule-based or linguistic approach, each one with its own advantages and limitations (Jivani et al., 2011). Relevant for us is the fact that stemming has been shown to add semantic value in feature selection, as for example Biba and Gjati (2014) proved that stemming of composite words greatly improves classification of fake news. Moreover, Mahendra et al. (2021) showed that cleaning and stemming resulted in the greatest model performance on the medical domain for the task in mortality prediction on ICU (Intensive Care Unit) patients. We refer readers to a thorough survey of stemmers spanning over the past 50 years by Singh and Gupta (2016).

3 Proposed Approach

The Bias Annotation Spans on the Informational Level (BASIL) dataset, provided by Fan et al. (2019), is based on three news sources, i.e., Huff-

ington Post (HPO), Fox News (FOX), and New York Times (NYT) from 2010 to 2019, containing 300 news articles with 100 triplets of news articles taken from each news source. BASIL contains both binary sentence classification labels of informational and lexical bias together with span level annotations for token classification. An isolated sentence is biased if it contains at least one bias span. There are 7,977 sentences in BASIL having 1,249 sentences with informational and 478 sentences with lexical bias (Fan et al., 2019). BASIL has more prevalence of informational bias than lexical bias. It covers news articles of reasonable time i.e., 10 years, representing conservative from FOX, neutral from NYT, and liberal from HPO, respectively.

Since informational bias is presented in a factual and neutral way, there is increasing prevalence of informational bias in news media, as evidenced by Fan et al. (2019). This factual reporting of informational bias makes its prediction more challenging on a sentence level. On the other hand, lexical bias is reportedly easier to capture because of its non-contextual nature, depending mostly on word choice (Chen et al., 2020). For example, a study on Tagalog-speaking Filipino pre-school children by Devanadera and Alieto (2019) showed that lexical bias is related to lexical inventories like nouns, verbs and adjectives produced by young children, with nouns as the leading or dominant lexical bias among children in their narrative production. Despite these facts, the detection of both types of bias remains a difficult and challenging task, as evidenced by the study on BASIL performed by Chen et al. (2020), which includes approaches for the automatic detection of both types of bias (although not distinguishing them).

BASIL has 7,977 sentences, from which 1,249 and 478 are labeled as containing with informational bias and lexical bias, respectively. Following the formulation of Fan et al. (2019), this leads to 6,250 sentences with no bias. After careful consideration of models proposed in prior work utilizing BASIL, we note inconsistencies in the task formulation across papers, which are derived from the way in which these labels are interpreted and used.

In order to shed light into this issue, we begin by organizing prior work and defining the notation we will utilize in the rest of the manuscript, for the sake of simplicity. We refer to sentences annotated with no bias as **'NEU'**, sentences annotated with informational bias as **'INF'**, sentences annotated with lexical bias as **'LEX'**. Additionally, we refer to the combination of neutral sentences with and sentences with lexical bias as **'OTH'**, while 'BIAS' refers to the combination of samples that have both informational and lexical bias.

LABELS	Baseline					
	Fan	**Berg**	**Lee**	**Guo**	**Lei**	**Ours**
INF v/s LEX	-	-	-	-	-	✓
INF v/s NEU	✓	-	-	-	-	✓
INF v/s OTH	-	✓	-	✓	✓	✓
LEX v/s NEU	✓	-	-	-	-	✓
BIAS v/s NEU	-	-	✓	-	-	✓
INF v/s/ LEX v/s NEU	-	-	-	-	-	✓

Table 2: Comparison of formulations proposed by prior work on bias detection using BASIL, showing the different class combinations adopted for training. In the table, work by Fan et al. (2019) is denoted as 'Fan', van den Berg and Markert (2020) as 'Berg', (Lee et al., 2021) as 'Lee', Guo and Zhu (2022b) as 'Guo', and Lei et al. (2022) as 'Lei'.

Using the above definitions, we proceeded to analyze the settings proposed by previous work by Fan et al. (2019), van den Berg and Markert (2020), Lee et al. (2021), Guo and Zhu (2022b), and Lei et al. (2022) which we summarize in Table 2, above. We can see that there is a clear disparity in the way in which the bias detection tasks are approached. We further observe that detection of informational bias strikes as being the main focus in literature so far. We surmise that this issue may be due to the fact that lexical bias samples on BASIL appear to be considerably fewer than informational bias. However, we think that detection of lexical bias (Zhou and Bansal, 2020) should be at least as important as informational bias. While previous results focus mainly on the INF-related settings, we propose to experiment on classifying LEX bias from INF as, depicted in the first row of Table 2. To the best of our knowledge, no prior work on BASIL tried to make such a distinction. The relevance of our approach is evidenced by Chen et al. (2020), who showed that the detection of both INF and LEX bias together is difficult, with bias detection becoming harder at the article level.

In light of these observations, our work shows the results of several experiments we perform to compare and distinguish different bias in BASIL, using various settings. To further study the relationship between labels, we also propose a novel three-way classification approach to directly differ-

entiate informational, lexical and neutral samples (INF/ LEX/ NEU), respectively.

As a result of our extensive study, we note that in many cases, performance gaps are due to lack of annotated examples for a given label, with this being particularly the case for lexical bias. In light of this issue, in this paper, we also adapt backtranslation as a mechanism for selective data augmentation in bias detection and use Google Translate python API for backtranslation. Our approach is inspired by previous work of Ma and Li (2020), who proposed a Chinese text data augmentation based on back-translation to generate a corpus to enrich the lexical features of text data, and reported increased performance on text classification tasks, especially when training on smaller datasets. Back-translation is also a commonplace to improve the performance on machine translation tasks (Miyabe and Yoshino, 2015). Our idea is also motivated by seminal work by Mikolov et al. (2013) who achieved 90% precision for translation of words between English and Spanish and found word vectors of both languages contain similar geometric arrangements. We used Google Translate python API which is accessible free under the MIT License.

Finally, we note that data preparation plays a significant role in machine learning, especially for natural language processing tasks (Marinov and Efremov, 2019). Inspired by the work by Zainol et al. (2018) and Ladani and Desai (2020), we **preprocess** sentences in the BASIL dataset to remove noise. Our proposed text preprocessing strategy involves two steps, first the removal of special characters and stop words i.e., words like 'after', 'before', 'the', 'some', 'all', 'such', and second reduction of words into their grammatical root or stem. In the context of bias classification task, as well shown in Section 5 we found that reducing the number of dimensions in terms of space (Rakholia and Saini, 2016) by removing most common words or words that normally carry no meaning has significant impact on bias detection.

4 Experimental Framework

In this section, we discuss the experimental framework including the setup, baselines, implementation details and give details of the models we train.

4.1 Setup

Backtranslation is performed separately on lexical and informational bias samples with one extra example per original (one half original and other half translated), hence doubling the size of lexical bias sentences to 956 samples and informational bias sentences to 2,498 samples respectively. Naturally, our backtranslation-based augmentation is performed only on the training set, with validation and testing always containing the original data samples

Our backtranslation approach is applied on the specific label of interest. For example, for the INF v/s LEX task, augmentation is only performed on lexical bias samples; while in INF v/s LEX v/s NEU task augmentation is performed on both lexical and informational bias samples, whereas no augmentation is performed on neutral sentences in all of of our experiments including this. INF v/s LEX v/s NEU task represent bias detection of whole BASIL corpus data, with neutral samples included. In the task INF v/s OTH, augmentation is only performed on informational bias sentences, where OTH represents the combination of lexical bias and neutral sentences.

We note that other kinds of splits exist for BASIL. Specifically, we find "story splits" by van den Berg and Markert (2020), also exist, where context is created by integrating events and articles. However, our work focuses on sentence classification without access to more context, and thus we report results on the traditional splits to retain consistency. Similarly, our testing data has no augmented example.

4.2 Baselines

Several deep learning approaches for the detection of bias in media have been proposed in previous work. Most of this work focuses on the detection of informational bias, and is based on the fine-tuning of large pre-trained models. In this paper, we concretely consider the informational bias detection approach by Fan et al. (2019), who proposed BERT-based (Devlin et al., 2018) approach and TF-IDF for this.

We also consider several models proposed by van den Berg and Markert (2020), which are used to detect informational bias in different ways. For starters, we compare with their BERT-based model as well as with their RoBERTa-based model Liu et al. (2019b). We also consider WinSCC (windowed Sequential Sentence Classification) which is a variant of SSC Cohan et al. (2019), ArtCIM (Article Context-Inclusive Model) and EvCIM (Event Context-Inclusive Model). We note that ArtCIM

Model	Aug.	Preproc.	INF/ LEX			INF/ LEX / NEU				INF/ OTH		
			Acc	F1-score		Acc	F1-score			Acc	F1-score	
				INF	LEX		INF	LEX	NEU		INF	OTH
BERT	-	-	74.46%	0.573	0.342	79.57%	0.383	0.194	0.880	76.55%	0.404	0.881
	-	✓	76.69%	0.687	0.534	77.78%	0.377	0.309	0.895	80.14%	0.423	0.871
	✓	-	81.37%	0.667	**0.691**	77.56%	**0.432**	**0.417**	0.864	81.53%	0.463	0.868
	✓	✓	**83.97%**	**0.712**	0.678	**81.54%**	0.429	0.401	**0.881**	**83.86%**	**0.507**	**0.899**
LSTM	-	-	72.63%	0.512	0.348	74.88%	0.301	0.209	0.856	73.93%	0.311	0.855
	-	✓	70.24%	0.594	0.432	71.34%	0.277	0.187	0.869	76.09%	0.332	0.823
	✓	-	70.01%	0.677	0.655	73.76%	0.319	**0.456**	**0.873**	75.34%	0.360	**0.875**
	✓	✓	**75.34%**	**0.692**	**0.671**	**75.56%**	**0.325**	0.450	0.851	**78.89%**	**0.381**	0.868
SVM	✓	✓	70.98%	0.491	0.795	74.23%	0.212	0.346	0.877	81.72%	0.178	0.890

Table 3: Results of our ablation study to understand the performance impact of our proposed backtranslation and data preprocessing approaches on three task formulations. In the Table, 'Aug.' denotes the usage of our augmentation techniques, while 'Preproc.' denotes models that included our preprocessing appraoch. We note that no prior work has done for the detection of the first two tasks we consider (INF v/s LEX and INF v/s LEX v/s NEU).

Model	INF / NEU		BIAS / NEU		LEX / NEU		INF / OTH	
	Acc	INF F1	Acc	BIAS F1	Acc	LEX F1	Acc	INF F1
TF-IDF (Fan et al., 2019)	-	26.02	-	-	-	-	-	-
BERT (Fan et al., 2019)	-	43.27	-	-	-	31.49	-	-
RoBERTa (Lee et al., 2021)	-	-	72.80	65.50	-	-	-	-
UnifiedM2 (Lee et al., 2021)	-	-	81.00	**70.20**	-	-	-	-
BERT (van den Berg and Markert, 2020)	-	-	-	-	-	-	-	38.26
RoBERTa (van den Berg and Markert, 2020)	-	-	-	-	-	-	-	49.89
WinSSC (van den Berg and Markert, 2020)	-	-	-	-	-	-	-	38.67
ArtCIM (van den Berg and Markert, 2020)	-	-	-	-	-	-	-	42.80
EvCIM (van den Berg and Markert, 2020)	-	-	-	-	-	-	-	44.10
MultiCTX Guo and Zhu (2022b)	-	-	-	-	-	-	-	46.08
RoBERTa Lei et al. (2022)	-	-	-	-	-	-	-	46.47
BERT (Ours)	**87.00**	**49.60**	**82.34**	69.00	**95.70**	**62.30**	**83.86**	**50.70**

Table 4: Comparison with previous work on four bias tasks using our BERT model. Results of our BERT on INF v/s NEU with augmentation of only informational bias, BIAS v/s NEU with augmentation of both informational and lexical bias, LEX v/s NEU with augmentation of only lexical bias, and INF v/s OTH with augmentation of only informational bias, respectively. BERT (Ours) report averaged results of three seed runs on all bias tasks. In the Table, 'Acc' denotes accuracy and only one prior work on task BIAS v/s NEU (Lee et al., 2021) reported accuracy.

and EvCIM integrate article and event context respectively and use BiLSTM's (Hochreiter and Schmidhuber, 1997) for bias detection. We also consider MultiCTX, as proposed by Guo and Zhu (2022b), which is based on contrastive learning on triplets sampled from different articles. We compare another model by Lei et al. (2022) built on RoBERTa that incorporates global functional discourse structure and local rhetorical discourse relations for detecting bias.

We further compare our models with the binary classification of informational and lexical bias with a RoBERTa model fine-tuned by Lee et al. (2021), and with their proposed UnifiedM2 model. This is a comprehensive misinformation detection model that was trained on the concatenation of multiple misinformation domains into a single unified setup.

4.3 Implementation Details

Our models are mainly based on BERT, in particular the implementation from HuggingFace (Face, 2021). We fine-tune BERT on our augmented and preprocesed data, using learning rate of 5×10^{-5}, and batch sizes of 16 or 32. The maximum epoch count in our experiments reaches up to 15.

We additionally use a model based on a LSTM that receives the same data which is already preprocessed as described previously. For this LSTM, we use a hidden size dimension of 200, with embeddings based on GloVe (Pennington et al., 2014). Finally, we also propose a simple approach based on a linear kernel SVM where we pass preprocessed

augmented training sentences to the model.

For all of our experiments, we use 80/10/10 split with non-overlapping samples for train-validation-test, respectively. We reported the average performance of our models using three seed runs in our experiments.

5 Results & Discussion

We begin by testing the impact of our proposed data augmentation and preprocessing approaches by running ablation studies on three task settings. We first consider the newly-introduced tasks of INF/ LEX and INF/ LEX/ NEU, in addition to the more standard INF/ OTH setting, which is generic and of particular interest in the prior work. We measure the impact of each component of our proposed work by repeating the experiments and dividing our ablation test into two steps. To study and measure the significance of our proposed components, we first experiment only with data augmentation (denoted as 'Aug.') which involves the backtranslation of samples with bias, which is followed by our preprocessing step (denoted as 'Preproc.') which includes data cleaning, removal of stop words and stemming. The evaluation criteria we used is accuracy as 'Acc' and F1-micro score separately on each class of the bias i.e, INF, LEX, NEU, and OTH as explained earlier.

From our literature review, as discussed in Section 3 we found that many potentially relevant task formulations were missing. We surmise this could be due to the scarcity problem of lexical bias samples in BASIL, with interest focusing on only informational bias. We therefore propose to tackle all relevant task formulations.

Concretely, as to the best of our knowledge there is no prior work done to classify lexical bias from informational, we believe we are the first to identify and report experiments on the tasks of INF/ LEX as binary classification, as well as INF/ LEX/ NEU, use a multi-class classification.

As shown in Table 3, we found that in these tasks considerable performance improvements is achieved when both data augmentation and preprocessing are applied. For the binary INF/ LEX, we found that just by augmenting lexical bias, we are able to attain excellent performance. An improved accuracy of 83.97% with INF F1-score of 0.712 is achieved as compared to LSTM and SVM when both data preprocessing and augmentation are applied. We also observe that the highest LEX F1-score of 0.691 is comparatively lower than informational bias in the INF/ LEX task. One possible reason for that is the larger class imbalance of lexical bias sentences, which persists even after augmentation of lexical bias samples —1,259 informational v/s 956 lexical bias sentences, respectively. Despite this we note that the F1-score of lexical bias detection still approach considerably high value suggesting lexical bias can be classified from informational bias when computed together. Similarly, we observe that our augmentation approach has a more significant impact in performance compared to our pre-processing technique. When augmentation is applied alone, it improves the LEX F1-score from 0.534 to 0.691 using BERT, and from 0.432 to 0.655 for the LSTM. While our preprocessing also consistently improve performance, the gains are not as dramatic as the ones provided by the data augmentation.

We further note that our three-way classification task INF/ LEX/ NEU follows a similar trend, with significant performance improvements due to our data augmentation technique. A sharp rise in accuracy is observed by inclusion of both augmentation and preprocessing, leading to an accuracy of 81.54% when using BERT.

While we believe our experiments clearly show the effectiveness of our data cleaning and augmentation approaches in detecting all kinds of bias in BASIL, it is also worth pointing out that there are some limitations to what these techniques can do. For example, we note that when we only use our preprocesing, a decline in accuracy from 79.57% to 77.78% and INF F1-score from 0.383 to 0.377 using BERT are observed, while for the LSTM accuracy drops from 74.88% to 71.34% with a drop in INF F1-score from 0.301 to 0.277. We believe a possible reason for this might be similar to what Wendland et al. (2021) observed on a similar experiment, where it was shown that stemming could lead to lower accuracy and F1 scores. Since data preparation plays a significant role in capturing the knowledge gaps of natural language processing tasks, our rigorous ablation analysis of the tasks in Table 3 support the hypothesis that improvements in the detection of both informational and lexical bias at the sentences can be improved by performing data augmentation based on backtranslation and by preprocessing the data using stemming.

From prior work on BASIL as discussed previously and as shown in Table 2, most of the attention

has been given to detecting only informational bias, especially the task of INF/ OTH, namely to classify informational bias from a combination of neutral and lexical bias sentences, which prompts us to consider this task also for our ablation study. Since informational bias is more common and difficult to detect as found by Fan et al. (2019); Chen et al. (2020), we highlight this task for our ablation study to demonstrate the significance of our proposed methods and to provide uniform results while comparing it with other methods in further experimental work. From these results, we observe similar trends compared to the other tasks, again showing the effectiveness of our proposed approach, which leads our BERT-based model to obtain a maximum F1-score of 0.507.

Having established the effectiveness of our proposed techniques, we now move on to compare our models with previous work, as shown in Table 4. Concretely, we compare against our selection of baseline models on four different existing tasks formulations derived from BASIL. For the comparisons in this section, we use our best model for each task i.e., the BERT-based approach combined with both of our proposed components, as evident from Table 3. Similarly, as before, we report accuracy and F1-score of each of the class. However, due to lack of completeness in the evaluation metrics found in previous work, we are only able to compare F1-scores of the baseline models with our proposed approach, except in the case of BIAS/ NEU task, where some previous work also report accuracy.

The first task shown in Table 4, INF/ NEU or namely to detect informational bias from neutral sentences, is only performed by Fan et al. (2019). We could only compare INF F1-score of baseline models with our proposed approach and found an improvement of 15% with final score of 49.60 using our BERT model, which compares against the reported performance of 43.27 F1-score.

In our second binary task, BIAS/ NEU, we use the same model configurations as Lee et al. (2021) with a batch size of 32, a learning rate of 5×10^{-6} with 15 as a maximum epoch count. Here BIAS corresponds to the combination of informational and lexical bias sentences, while NEU symbolizes only neutral sentences, which therefore corresponds to detecting bias v/s no bias sentences irrespective of their type. In this task, the highest BIAS F1-score is reported by UnifiedM2, and though we see that our proposed BERT-based approach does not outperform this model, our results are competitive as we observe we attain a similar performance. Furthermore, we note that our results in this task indicate that small focused data augmentation techniques as ours could be nearly as effective as more complicated training procedures including many tasks, as the approach proposed by UnifiedM2.

To detect lexical bias alone from neutral sentences, we study the task of LEX/ NEU. As shown in Table 4, in this case, the only baseline available is the BERT model proposed by Fan et al. (2019). For this task, we perform data augmentation of only lexical bias sentences, and we can see that BERT-based model outperforms the baseline BERT resulting in almost doubled LEX F1-score improvement, with our model attaining an F1-score of 62.3, while for the baseline BERT the value is just 31.49. We note that although BERT is our best model choice on this setting in terms of F1-score, our LSTM outperforms our BERT in terms of accuracy, obtaining 86.64%.

Finally, we consider the task of INF/ OTH, where we find most of previous work has focused. Concretely, for this task, we compare our BERT-based model with six previous models from the literature. As can be seen, we find that our model is able to outperform all existing prior work, achieving the best performance in terms of with INF F1-score of 50.9. Our model is followed by the RoBERTa model fine-tuned by (van den Berg and Markert, 2020) with INF F1-score of 49.89, where RoBERTa by Lei et al. (2022) as third with 46.47 and MultiCTX follows in the fourth position with F1-score of only 46.08. We believe this indicates that our approach is also better at recognizing informational bias as a type of misinformation.

Furthermore, we note that for the INF/ OTH task, many of our considered baselines are based on pre-trained models similar to ours, such as BERT or RoBERTa. As our model is based fundamentally on the same deep learning model, we believe these results suggest that our augmentation and pre-processing approaches might work for those models also. This adds to our observations derived from Table 3, where we saw that combining the LSTM with our proposed approach leads to consistent improvements also.

6 Conclusion

This paper presents different techniques of phrasing bias to tackle media bias in new outlets. We propose an approach that relies on current neural network models to capture sentence level biased language. We defined how data augmentation is applicable to less frequent bias in news articles and measure the effect of its performance across different models. Human annotation is costly and conditions where obtaining new misinformation samples is difficult, our approach is significant to such real life cases. Since our proposed approach involves simple feature extraction techniques to tackle a particularly small and unbalanced biased dataset, we believe our work can be used to mitigate bias and improve the quality of the model's predictions in real-world scenarios. We identify some novel tasks in BASIL and our augmentation technique effectively detect informational and lexical bias sentences simultaneously, while also outperforming in other tasks. In our work, we incorporate different methods to process bias and illustrate the importance of our proposed components. A key distinguishing feature of our work is the removal of inconsistencies of prior work in reporting and evaluating bias types of BASIL. Ablation studies are also performed by varying training data in different tasks and our technique suggest significance of each proposed component in different experimental settings. We found the performance improvement of our proposed approach in almost all tasks as compared to several state-of-the-art techniques, hence this proves that our methodological standpoint in using small augmented data is well-aligned in finding informational and lexical bias sentences in different classification tasks. Similarly, our work tries to propose a way of regulating different task formulations of BASIL which are unclear in prior work. We intend to explore context in BASIL news articles as future work, besides trying other feature selection techniques. We believe further parameter optimization and fine-tuning for different task formulations can also improve the results.

Limitations

One major limitation of our work is that we only experimented on an English dataset. While other lexical and syntactic features can be captured by text processing techniques and also backtranslation performed with other or multiple languages can be used to see the effect on performance. Other English news articles may also be useful for analyzing bias, and require further research analysis to verify media bias. Similarly, though our proposed approach works well for detecting bias in BASIL, we provide no evidence to suggest if this will also work on other misinformation-related tasks. The same applies for models other than the ones we tested in this paper, which though includes a broad selection (SVMs, LSTMs and Transformers) is not completely comprehensive.

Ethical Considerations

The interpretation of bias detection results is crucial. For cases, where different political entities are debatable in news media, may mislead the bias detection model and removing such bias require more flexible and tolerating approach while dealing with such entities. Therefore, the results reported in our work highlight the need for mitigating bias and further research is required to investigate the biased influence towards particular issues at various stages of the training model.

Acknowledgements

The authors wish to express gratitude to the funding organisation as this work has been supported by the Mohammed bin Salman Center for Future Science and Technology for Saudi-Japan Vision 2030 at The University of Tokyo (MbSC2030).

References

Dan Bernhardt, Stefan Krasa, and Mattias Polborn. 2008. Political polarization and the electoral effects of media bias. *Journal of Public Economics*, 92(5-6):1092–1104.

Marenglen Biba and Eva Gjati. 2014. Boosting text classification through stemming of composite words. In *Recent Advances in Intelligent Informatics*, pages 185–194. Springer.

Wei-Fan Chen, Khalid Al-Khatib, Benno Stein, and Henning Wachsmuth. 2020. Detecting media bias in news articles using gaussian bias distributions. *arXiv preprint arXiv:2010.10649*.

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Dan Weld. 2019. Pretrained language models for sequential sentence classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics.

Aprillette Devanadera and Ericson Alieto. 2019. Noun bias among tagalog-speaking filipino pre-school children. *Devanadera, A., & Alieto, E.(2019). Lexical Bias among Tagalog-speaking Filipino Pre-school Children. Asian EFL*, 24(4.1):207–225.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

John Duggan and Cesar Martinelli. 2011. A spatial theory of media slant and voter choice. *The Review of Economic Studies*, 78(2):640–666.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Hugging Face. 2021. The ai community building the future. *URL: https://huggingface. co*.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In plain sight: Media bias through the lens of factual reporting. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6343–6349, Hong Kong, China. Association for Computational Linguistics.

Stephan Greene and Philip Resnik. 2009. More than words: Syntactic packaging and implicit sentiment. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 503–511, Boulder, Colorado. Association for Computational Linguistics.

Shijia Guo and Kenny Q. Zhu. 2022a. Modeling multi-level context for informational bias detection by contrastive learning and sentential graph network.

Shijia Guo and Kenny Q Zhu. 2022b. Modeling multi-level context for informational bias detection by contrastive learning and sentential graph network. *arXiv preprint arXiv:2201.10376*.

Felix Hamborg, Karsten Donnay, and Bela Gipp. 2019. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries*, 20(4):391–415.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Christoph Hube and Besnik Fetahu. 2019. Neural based statement classification for biased language. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1113–1122, Baltimore, Maryland. Association for Computational Linguistics.

Qiong Jia, Yue Guo, and Stuart Barnes. 2019. Understanding information bias: The perspective of online review component: An abstract. In *Academy of Marketing Science Annual Conference*, pages 157–158. Springer.

Anjali Ganesh Jivani et al. 2011. A comparative study of stemming algorithms. *Int. J. Comp. Tech. Appl*, 2(6):1930–1938.

Jessica L Johnson, Donna Adkins, and Sheila Chauvin. 2020. A review of the quality indicators of rigor in qualitative research. *American journal of pharmaceutical education*, 84(1).

Dhara J Ladani and Nikita P Desai. 2020. Stopword identification and removal techniques on tc and ir applications: A survey. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 466–472. IEEE.

Nayeon Lee, Belinda Z. Li, Sinong Wang, Pascale Fung, Hao Ma, Wen-tau Yih, and Madian Khabsa. 2021. On unifying misinformation detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5479–5485, Online. Association for Computational Linguistics.

Yuanyuan Lei, Ruihong Huang, Lu Wang, and Nick Beauchamp. 2022. Sentence-level media bias analysis informed by discourse structures. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10040–10050.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual representations.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jun Ma and Langlang Li. 2020. Data augmentation for chinese text classification using back-translation. In *Journal of Physics: Conference Series*, volume 1651, page 012039. IOP Publishing.

Malini Mahendra, Yanting Luo, Hunter Mills, Gundolf Schenk, Atul J Butte, and R Adams Dudley. 2021. Impact of different approaches to preparing notes for analysis with natural language processing on the performance of prediction models in intensive care. *Critical care explorations*, 3(6).

Martin Marinov and Alexander Efremov. 2019. Representing character sequences as sets : A simple and intuitive string encoding algorithm for nlp data cleaning. In *2019 IEEE International Conference on Advanced Scientific Computing (ICASC)*, pages 1–6.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Mai Miyabe and Takashi Yoshino. 2015. Evaluation of the validity of back-translation as a method of assessing the accuracy of machine translation. In *2015 International Conference on Culture and Computing (Culture Computing)*, pages 145–150. IEEE.

Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. 2019. Movie plot analysis via turning point identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1707–1717, Hong Kong, China. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Rajnish M Rakholia and Jatinderkumar R Saini. 2016. Lexical classes based stop words categorization for gujarati language. In *2016 2nd international conference on advances in computing, communication, & automation (ICACCA)(Fall)*, pages 1–5. IEEE.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659, Sofia, Bulgaria. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Sam Shleifer. 2019. Low resource text classification with ulmfit and backtranslation. *arXiv preprint arXiv:1903.09244*.

Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. 2021. Text data augmentation for deep learning. *Journal of big Data*, 8(1):1–34.

Jasmeet Singh and Vishal Gupta. 2016. Text stemming: Approaches, applications, and challenges. *ACM Computing Surveys (CSUR)*, 49(3):1–46.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.

Esther van den Berg and Katja Markert. 2020. Context in informational bias detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6315–6326, Barcelona, Spain (Online). International Committee on Computational Linguistics.

André Wendland, Marco Zenere, and Jörg Niemann. 2021. Introduction to text classification: impact of stemming and comparing tf-idf and count vectorization as feature extraction technique. In *European Conference on Software Process Improvement*, pages 289–300. Springer.

Stephane Wolton. 2019. Are biased media bad for democracy? *American Journal of Political Science*, 63(3):548–562.

Tae Yano, Philip Resnik, and Noah A Smith. 2010. Shedding (a thousand points of) light on biased language. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 152–158.

Zuraini Zainol, Mohd TH Jaymes, and Puteri NE Nohuddin. 2018. Visualurtext: a text analytics tool for unstructured textual data. In *Journal of Physics: Conference Series*, volume 1018, page 012011. IOP Publishing.

Xiang Zhou and Mohit Bansal. 2020. Towards robustifying nli models against lexical dataset biases. *arXiv preprint arXiv:2005.04732*.