

Towards Detection of Subjective Bias using Contextualized Word Embeddings

Tanvi Dadu*
NSIT Delhi
tanvid.co.16@nsit.net.in

Kartikey Pant*
IIIT Hyderabad
kartikey.pant@research.iiit.ac.in

Radhika Mamidi
IIIT Hyderabad
radhika.mamidi@iiit.ac.in

ABSTRACT

Subjective bias detection is critical for applications like propaganda detection, content recommendation, sentiment analysis, and bias neutralization. This bias is introduced in natural language via inflammatory words and phrases, casting doubt over facts, and presupposing the truth. In this work, we perform comprehensive experiments for detecting subjective bias using BERT-based models on the Wiki Neutrality Corpus(WNC). The dataset consists of 360k labeled instances, from Wikipedia edits that remove various instances of the bias. We further propose BERT-based ensembles that outperform state-of-the-art methods like $BERT_{large}$ by a margin of 5.6 F1 score.

1 INTRODUCTION

In natural language, subjectivity refers to the aspects of communication used to express opinions, evaluations, and speculations[7], often influenced by one's emotional state and viewpoints. Writers and editors of texts like news and textbooks try to avoid the use of biased language, yet subjective bias is pervasive in these texts. More than 56% of Americans believe that news sources do not report the news objectively¹, thus implying the prevalence of the bias. Therefore, when presenting factual information, it becomes necessary to differentiate subjective language from objective language.

There has been considerable work on capturing subjectivity using text-classification models ranging from linguistic-feature-based models[5] to finetuned pre-trained word embeddings like BERT[4]. The detection of bias-inducing words in a Wikipedia statement was explored in [5]. The authors propose the "Neutral Point of View" (NPOV) corpus made using Wikipedia revision history, containing Wikipedia edits that are specifically designed to remove subjective bias. They use logistic regression with linguistic features, including factive verbs, hedges, and subjective intensifiers to detect bias-inducing words. In [4], the authors extend this work by mitigating subjective bias after detecting bias-inducing words using a BERT-based model. However, they primarily focused on detecting and mitigating subjective bias for single-word edits. We extend their work by incorporating multi-word edits by detecting bias at the sentence level. We further use their version of the NPOV corpus called Wiki Neutrality Corpus(WNC) for this work.

The task of detecting sentences containing subjective bias rather than individual words inducing the bias has been explored in [2]. However, they conduct majority of their experiments in controlled settings, limiting the type of articles from which the revisions were extracted. Their attempt to test their models in a general setting is dwarfed by the fact that they used revisions from a single Wikipedia

article resulting in just 100 instances to evaluate their proposed models robustly. Consequently, we perform our experiments in the complete WNC corpus, which consists of 423,823 revisions in Wikipedia marked by its editors over a period of 15 years, to simulate a more general setting for the bias.

In this work, we investigate the application of BERT-based models for the task of subjective language detection². We explore various BERT-based models, including *BERT*, *RoBERTa*, *ALBERT*, with their *base* and *large* specifications along with their native classifiers. We propose an ensemble model exploiting predictions from these models using multiple ensembling techniques. We show that our model outperforms the baselines by a margin of 5.6 of F1 score and 5.95% of Accuracy.

2 BASELINES AND APPROACH

In this section, we outline baseline models like $BERT_{large}$. We further propose three approaches: optimized BERT-based models, distilled pretrained models, and the use of ensemble methods for the task of subjectivity detection.

2.1 Baselines

- (1) **FastText**[3]: It uses bag of words and bag of n-grams as features for text classification, capturing partial information about the local word order efficiently.
- (2) **BiLSTM**: Unlike feedforward neural networks, recurrent neural networks like BiLSTMs use memory based on history information to learn long-distance features and then predict the output. We use a two-layer BiLSTM architecture with GloVe word embeddings as a strong RNN baseline.
- (3) **BERT** [1]: It is a contextualized word representation model that uses bidirectional transformers, pretrained on a large 3.3B word corpus. We use the $BERT_{large}$ model finetuned on the training dataset.

2.2 Proposed Approaches

- (1) **Optimized BERT-based models**: We use BERT-based models optimized as in [8] and [9], pretrained on a dataset as large as twelve times as compared to $BERT_{large}$, with bigger batches, and longer sequences. *ALBERT*, introduced in [9], uses factorized embedding parameterization and cross-layer parameter sharing for parameter reduction. These optimizations have led both the models to outperform $BERT_{large}$ in various benchmarking tests, like *GLUE* for text classification and *SQuAD* for Question Answering.
- (2) **Distilled BERT-based models**: Secondly, we propose to use distilled BERT-based models, as introduced in [6]. They are

*The first two authors contributed equally to the work.

¹<https://news.gallup.com/opinion/gallup/235796/americans-misinformation-bias-inaccuracy-news.aspx>

²Made available at <https://github.com/tanvidadu/Subjective-Bias-Detection>

smaller general-purpose language representation model, pre-trained by leveraging distillation knowledge. This results in significantly smaller and faster models with performance comparable to their undistilled versions. We finetune these pretrained distilled models on the training corpus to efficiently detect subjectivity.

- (3) **BERT-based ensemble models:** Lastly, we use the weighted-average ensembling technique to exploit the predictions made by different variations of the above models. Ensembling methodology entails engendering a predictive model by utilizing predictions from multiple models in order to improve Accuracy and F1, decrease variance, and bias. We experiment with variations of *RoBERTa_{large}*, *ALBERT_{xxlarge.v2}*, *DistilRoBERTa* and *BERT* and outline selected combinations in Table 1.

3 EXPERIMENTS

3.1 Dataset and Experimental Settings

We perform our experiments on the *WNC* dataset open-sourced by the authors of [4]. It consists of aligned pre and post neutralized sentences made by Wikipedia editors under the neutral point of view. It contains 180k biased sentences, and their neutral counterparts crawled from 423, 823 Wikipedia revisions between 2004 and 2019. We randomly shuffled these sentences and split this dataset into two parts in a 90 : 10 Train-Test split and perform the evaluation on the held-out test dataset.

For all BERT-based models, we use a learning rate of $2 * 10^{-5}$, a maximum sequence length of 50, and a weight decay of 0.01 while finetuning the model. We use FastText’s recently open-sourced automatic hyperparameter optimization functionality while training the model. For the BiLSTM baseline, we use a dropout of 0.05 along with a recurrent dropout of 0.2 in two 64 unit sized stacked BiLSTMs, using softmax activation layer as the final dense layer.

3.2 Experimental Results

Table 1 shows the performance of different models on the *WNC* corpus evaluated on the following four metrics: Precision, Recall, F1, and Accuracy. Our proposed methodology, the use of finetuned optimized BERT based models, and BERT-based ensemble models outperform the baselines for all the metrics.

Among the optimized BERT based models, *RoBERTa_{large}* outperforms all other non-ensemble models and the baselines for all metrics. It further achieves a maximum recall of 0.681 for all the proposed models. We note that *DistilRoBERTa*, a distilled model, performs competitively, achieving 69.69% accuracy, and 0.672 F1 score. This observation shows that distilled pretrained models can replace their undistilled counterparts in a low-computing environment.

We further observe that ensemble models perform better than optimized BERT-based models and distilled pretrained models. Our proposed ensemble comprising of *RoBERTa_{large}*, *ALBERT_{xxlarge.v2}*, *DistilRoBERTa* and *BERT* outperforms all the proposed models obtaining 0.704 F1 score, 0.733 precision, and 71.61% Accuracy.

	Models/Metrics	Precision	Recall	F1	Acc
Baselines	FastText	0.613	0.612	0.613	61.24%
	BiLSTM+GloVe	0.648	0.647	0.648	64.76%
	<i>BERT_{large}</i>	0.681	0.587	0.631	65.66%
Single Model	<i>ALBERT_{xxlarge.v2}</i>	0.667	0.579	0.620	64.56%
	DistillBERT	0.731	0.608	0.664	69.28%
	DistillRoBERTa	0.730	0.623	0.672	69.69%
	<i>RoBERTa_{large}</i>	0.723	0.681	0.702	71.09%
Ensemble model	<i>BERT_{Ensemble}</i> +DistillBERT	0.731	0.610	0.665	69.36%
	<i>RoBERTa_{Ensemble}</i>	0.732	0.679	0.704	71.57%
	RoBERTa+ALBERT				
	+DistillRoBERTa+BERT	0.733	0.677	0.704	71.61%

Table 1: Experimental Results for the Subjectivity Detection Task

4 CONCLUSION

In this paper, we investigated BERT-based architectures for sentence level subjective bias detection. We perform our experiments on a general Wikipedia corpus consisting of more than 360k pre and post subjective bias neutralized sentences. We found our proposed architectures to outperform the existing baselines significantly. BERT-based ensemble consisting of *RoBERTa*, *ALBERT*, *DistillRoBERTa*, and *BERT* led to the highest F1 and Accuracy. In the future, we would like to explore document-level detection of subjective bias, multi-word mitigation of the bias, applications of detecting the bias in recommendation systems.

REFERENCES

- [1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. <http://arxiv.org/abs/1810.04805>
- [2] Christoph Hube and Besnik Fetahu. 2019. Neural Based Statement Classification for Biased Language. In *12th ACM International Conference on Web Search and Data Mining (WSDM)*.
- [3] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of Tricks for Efficient Text Classification. <http://arxiv.org/abs/1607.01759> cite arxiv:1607.01759.
- [4] Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2019. Automatically Neutralizing Subjective Bias in Text. arXiv:arXiv:1911.09709
- [5] Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic Models for Analyzing and Detecting Biased Language. In *ACL (1)*. The Association for Computer Linguistics, 1650–1659. <http://dblp.uni-trier.de/db/conf/acl/acl2013-1.html#RecasensDJ13>
- [6] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:arXiv:1910.01108
- [7] Janyce Wiebe. 2002. Tracking Point of View in Narrative. *Computational Linguistics* 20 (07 2002).
- [8] Naman Goyal Jingfei Du Mandar Joshi Danqi Chen Omer Levy Mike Lewis Luke Zettlemoyer Veselin Stoyanov Yinhan Liu, Myle Ott. 2020. RoBERTa: A Robustly Optimized BERT Pretraining Approach. In *Submitted to International Conference on Learning Representations*. <https://openreview.net/forum?id=SyxS0T4tvS> under review.
- [9] Sebastian Goodman Kevin Gimpel Piyush Sharma Radu Soricut Zhenzhong Lan, Mingda Chen. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *Submitted to International Conference on Learning Representations*. <https://openreview.net/forum?id=H1eA7AEtvS> under review.