## **d4data/bias-detection-model**

## 

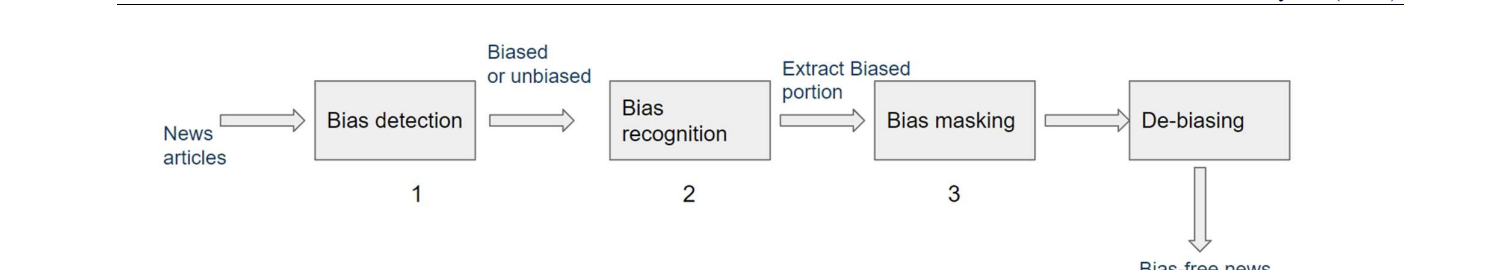
## **About the Model**

An English sequence classification model, trained on MBAD Dataset to detect bias and fairness in sentences (news articles). This model was built on top of distilbert-base-uncased model and trained for 30 epochs with a batch size of 16, a learning rate of 5e-5, and a maximum sequence length of 512.

| **Train Accuracy** | **Validation Accuracy** | **Train loss** | **Test loss** |
| --- | --- | --- | --- |
| 76.97 | 62.00 | 0.45 | 0.96 |

news articles

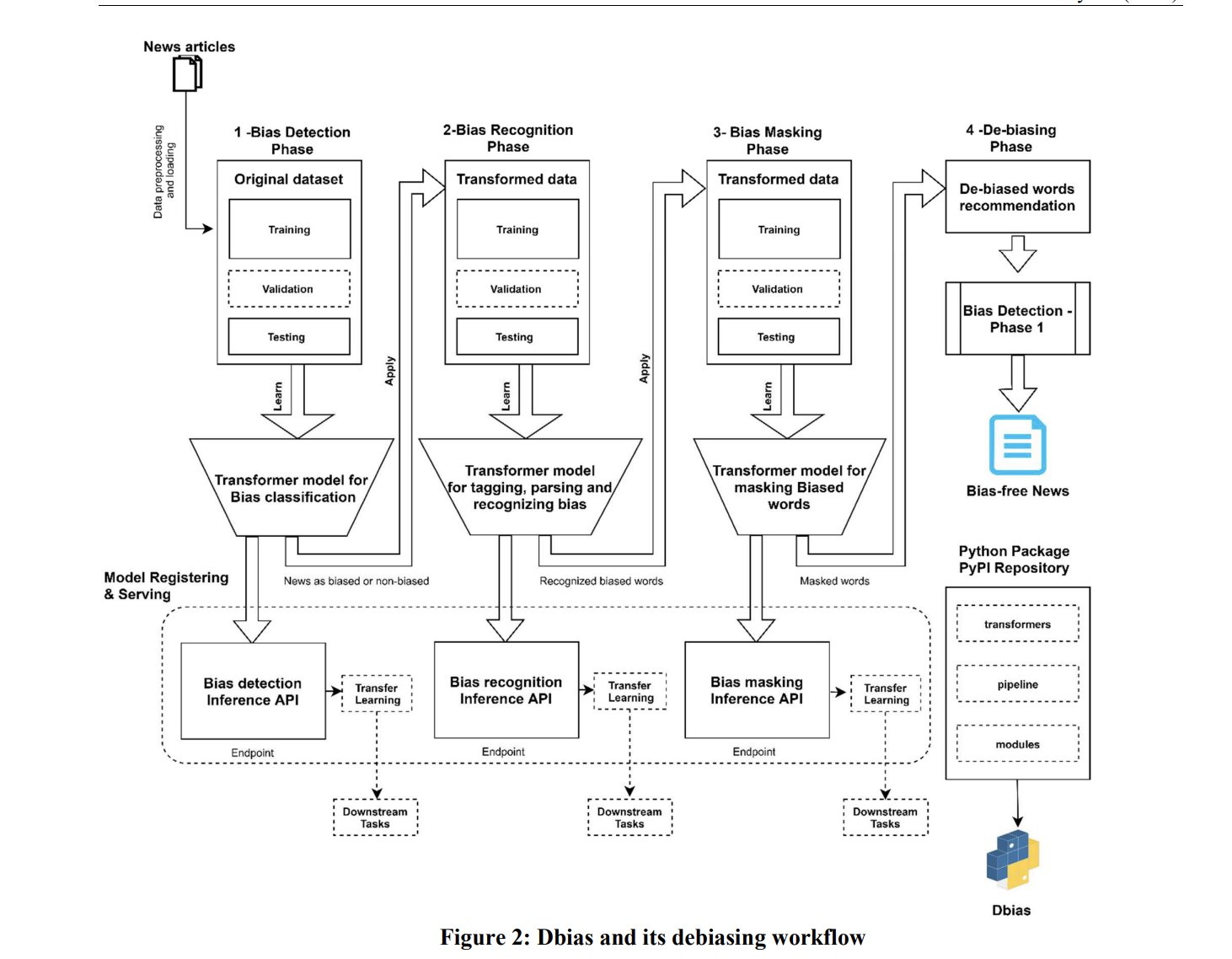
| **Feature** | **Output** |
| --- | --- |
| Text Debiasing | Returns debiased news recommendations with bias probability |
| Bias Classification | Classifies whether a news article is biased or not with probability |
| Bias Words/Phrases Recognition | Extract Biased words or phrases from the news fragment |
| Bias masking | Returns the news fragment with biased words masked out |



The Dbias can detect biases in any kind of text. The only requisite is to train the model on the domainspecific data.

Transfer learning techniques: Transfer learning is a technique to transfer the knowledge contained in larger, different but related source domain to a target domain [54]. The goal is to improve the performance of target domain with the existing knowledge of the source domain. Bidirectional Encoder Representations from Transformers (BERT)

The task of identifying a named entity (a real-world object or concept) in unstructured text and then classifying the entity into a standard category is known as named entity recognition [58]. Mehrabi et al. 2020 [32] used named entities to determine whether female names are more frequently tagged as non-person than male names. Some other researchers [59, 60] use the named entities to identify biases based on occupation, race, and demographics. These named entity recognition models usually recognize the biased entities from the data, without mitigating them.



19233

0045473f40ec42a2bd2ca0ee35df0b75

52078298e0974ccb9d7badd5bd25299f

67

2020-08-12 06:11:13

Non-biased

Entirely factual

67

6

19235

0045473f40ec42a2bd2ca0ee35df0b75

5ee3b6054e8c4ee0a371033ef805edb9

67

2020-08-12 06:11:39

Biased

terrorists

Expresses writer’s opinion

67

This is a collective pipeline comprises of 3 Transformer models to de-bias/reduce amount of bias in news articles. The three models are:

* An English sequence classification model, trained on the [MBIC Dataset](https://github.com/Media-Bias-Group/Neural-Media-Bias-Detection-Using-Distant-Supervision-With-BABE), to detect bias and fairness in sentences (news articles). This model was built on top of distilbert-base-uncased model and trained for 30 epochs with a batch size of 16, a learning rate of 5e-5, and a maximum sequence length of 512.
* An Entity Recognition model, which is is trained on MBIC Dataset to recognize the biased word/phrases in a sentence. This model was built on top of roberta-base offered by Spacy transformers.
* A Masked Language model, which is a Pretrained model on English language using a masked language modeling (MLM) objective.