With Prejudice to None: A Few-Shot, Multilingual Transfer Learning Approach to Detect Social Bias in Low Resource Languages

Nihar Ranjan Sahoo, Niteesh Mallela, Pushpak Bhattacharyya
CFILT, Indian Institute of Technology Bombay, India
{nihar, niteesh, pb @cse.iitb.ac.in}

Abstract

Warning: *This paper contains offensive material by way of examples and case studies which is unavoidable due to the nature of the work.*

In this paper, we describe our work on social bias detection in a low-resource multilingual setting in which the languages are from two very divergent families- Indo-European (English, Hindi, and Italian) and Altaic (Korean). Currently, the majority of the social bias datasets available are in English and this inhibits progress on social bias detection in low-resource languages. To address this problem, we introduce a new dataset for social bias detection in Hindi and investigate multilingual transfer learning using publicly available English, Italian, and Korean datasets. The Hindi dataset contains $\sim 9k$ social media posts annotated for (i) binary bias labels (bias/neutral), (ii) binary labels for sentiment (positive/negative), (iii) target groups for each bias category, and (iv) rationale for annotated bias labels (a short piece of text). We benchmark our Hindi dataset using different multilingual models, with XLM-R achieving the best performance of 80.8 macro-F1 score. Our results show that the detection of social biases in resource-constrained languages such as Hindi and Korean may be improved with the use of a similar dataset in English. We also show that translating all datasets into English does not work effectively for detecting social bias, since the nuances of source language are lost in translation.

1 Introduction

Social media has become one of the most important sources of information for users in recent years (Twenge and Campbell, 2019). The rise in the use of social media platforms (Kemp, 2020), combined with the ease of access, allows unrestricted flow of content containing social biases, and stereotypes on such sites. Furthermore, inequalities in social-media use exist across countries and regions as a result of various societal norms, cultures, and histories. Thus, prejudice and societal biases vary across cultures.

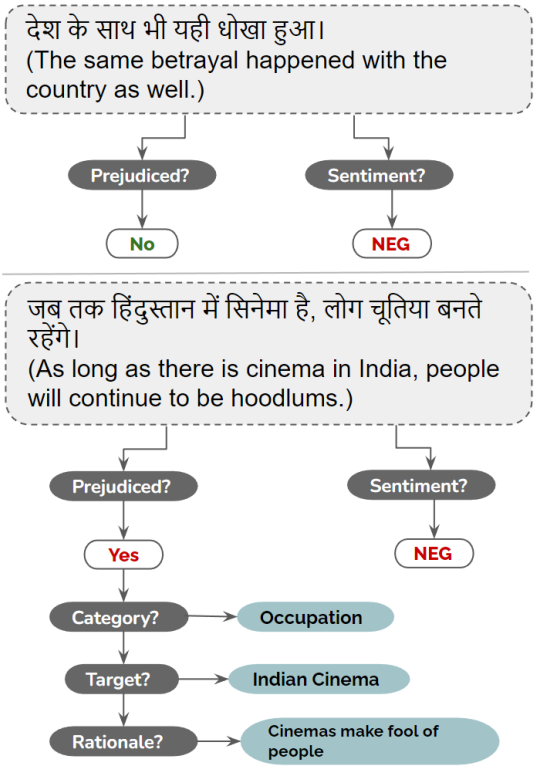

Figure 1: Examples from our Dataset. Each post is annotated with multiple categorical labels. In addition, annotators are asked to write the rationale behind the social bias in the post.

Motivation: There has been a lot of focus on how to identify social bias either in data or model in recent years. This is because it is crucial that the systems we create do not encourage pre-existing prejudices. While there has been some study of this topic, it has largely been limited to *high-resource* languages like English (Dev et al., 2022; Röttger et al., 2022).

Despite the fact that social bias is inextricably linked to cultural and linguistic characteristics of the language, non-English datasets (Lauscher et al., 2020; Kurpicz-Briki, 2020; Liang et al., 2020) are

Findings of the Association for Computational Linguistics: ACL 2023, pages 13316–13330

limited, hindering the development of social bias identification in other languages. In this work, we are expanding the bias detection task from english language to non-english languages. The results show that large datasets are not always required to develop efficient methods for identifying social bias in these resource-constrained languages.

We intend to address the identification of Social Bias in the Hindi language. With 602 million active speakers[1], Hindi is the world's third most spoken language. Despite the fact that a sizable proportion of these folks choose to communicate online in Devanagari[2] (script for Hindi), there has been essentially no research on social bias detection in Hindi. Kumar et al. (2021a) focuses on social bias detection in code-mixed and transliterated Hindi language and (Bhatt et al., 2022) releases a social bias benchmarking dataset using Indian English. Although both of these datasets are focusing on social biases in the Indian context, none of them are in the Hindi language.

In this paper, we present our manually annotated dataset for Social Bias detection in Hindi. We have annotated $\sim 9k$ online social media posts with multiple categorical labels. We empirically investigate the impact of languages from two different language families on the downstream task of social bias detection. Further, we show that for bias detection, translating all datasets into English does not perform well.

Our contributions are:

- Identification of social bias in text across four languages (e.g., Hindi, English, Italian, and Korean) using multilingual transfer learning.
- A **new social bias detection dataset** in Hindi with $\sim 9k$ instances, along with an accompanying annotation guideline which will be a valuable resource for researchers studying social bias detection in low-resource languages.
- Baseline experiments as useful benchmarks for future research on social bias detection in Hindi and other languages.

The rest of the paper is organized as follows: Related works are discussed in section 2. Section 3 gives insight into our dataset, terminologies, and annotation process. The methodologies and experiments are discussed in Section 4. Detailed error analysis is presented in Section 5 followed by the concluding remarks, and the discussion on future works in Section 6.

[1] https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers

[2] https://en.wikipedia.org/wiki/Languages_of_India

2 Related Works

The presence of social bias in language representations is mostly caused by the undesired and skewed associations within the training data. Given the growing social effect of NLP applications, studying these undesired relationships is paramount (Bender and Friedman, 2018; Crawford, 2017). The initial attempts to tackle this issue focused on measuring and mitigating gender biases from word embeddings (Bolukbasi et al., 2016; Caliskan et al., 2017; Zhao et al., 2017; Garg et al., 2018; Sun et al., 2019). Additionally, multiple works have explored solutions to identify race, and religion bias in word embeddings (Manzini et al., 2019). Many subsequent works (May et al., 2019; Zhao et al., 2019; Kurita et al., 2019) have also focused on contextualized language representation from models like BERT for bias detection.

More recently, many datasets (Nangia et al., 2020; Sap et al., 2020) have been created to measure social biases like gender, race, profession, religion, age, and so on in language models. Blodgett et al. (2021) has reported that these datasets lack clear definitions and have ambiguities and inconsistencies in annotations. Researchers have also investigated the presence of biases in models trained for various NLP tasks like machine translation (Stanovsky et al., 2019; Savoldi et al., 2021), question answering (Li et al., 2020), and coreference resolution (Webster et al., 2018).

There have been a lot of notable efforts towards identifying data bias in the problem of hate speech and offensive languages detection(Waseem and Hovy, 2016; Davidson et al., 2019; Sap et al., 2019; Mozafari et al., 2020). Borkan et al. (2019) has discussed unintended bias in hate speech detection models for identity terms like Islam, lesbian, bisexual, etc. Recent studies have also investigated the usefulness of counter-factual data augmentation (Dixon et al., 2018; Nozza et al., 2019; Sahoo et al., 2022; de Vassimon Manela et al., 2021) to reduce the effect of unintended bias in these tasks.

However, most of the bias detection and mitigation research is in English and has focused on western culture. Few recent works have explored the is-

Post	Label			Rationale
	Bias	Target Group	Category	
मेरे देश के हिन्दु बहुत निराले है। कुछ तो पक्के राम भक्त है और कुछ बाबर के साले है 🙏जय श्री राम 🙏 (Hindus of my country are very different. Some are staunch Ram devotees and some are the brother-in-law of Babur 🙏Jai Shri Ram 🙏)	Yes	hindus	Religion	talking about the hindus in negative manner
कश्मीर में कहां है शान्ती ????????? दलाल मीडिया दिखाता क्यो नही???? बोलेगा। (Where is peace in Kashmir????????? Why doesn't the broker media show this????)	Yes	Journalist	Occupation	reporters do not show right information on Kashmir
आजम खान की बेटी पर ससूर ने किया रेप। अब ये उठता है कि होने वाला बच्चा आजम खान के रियल दमाद को क्या बोलेगा। (Azam Khan's daughter was raped by father-in-law. Now it arises that the child to be born is the real son-in-law of Azam Khan.)	No			
मै भाजपा का कार्यकर्ता नहीं हूं,लेकिन फिर भी मोदी जी मेरा मान है,योगी जी सम्मान है,और अमित जी अभिमान है,और मै अपने मान, सम्मान,अभिमान को कभी झुकने नहीं दूंगा। (I am not a BJP worker, but still Modi ji is my honor, Yogi ji is honor, and Amit ji is pride, and I will never let my honor, honor, pride.)	Yes	Modi ji,Amit shah,Yogi ji	Personal Favour	favouritism towards political personalities
बनारस हिंदू विश्वविद्यालय से एक और छात्र लापता हो गया है। इससे पहले लापता हुए छात्र का भी अभी तक कोई सुराग नहीं लगा है (Another student from Banaras Hindu University has gone missing. Before this, there is no clue of the student who went missing.)	No			

Figure 2: Examples of social media posts in our Dataset, along with annotations.

sue of social bias in languages such as Arabic, Italian, Spanish, French, and Korean (Lauscher et al., 2020; Sanguinetti et al., 2020; Zhou et al., 2019; Kurpicz-Briki, 2020; Moon et al., 2020). There are very few research works towards tackling this challenge on Indian context. Pujari et al. (2019) explores binary gender bias in Hindi languages, and Gupta et al. (2021) investigates gender bias in Hindi-English machine translation using different fairness metrics. Sambasivan et al. (2021) analyzes and discusses multiple dimensions of algorithmic fairness in India. Through a detailed qualitative study, the authors suggest seven potential dimensions of algorithmic unfairness in India: Caste, Gender, Religion, Ability, Class, Sexual Orientation, and Ethnicity. Gangula et al. (2019) made available a dataset to identify bias towards political parties in Telugu. Kumar et al. (2021b) released a multilingual dataset in four languages like Hindi, Bangla, Meitei, and Indian English.

Multilingual models acquire cross-lingual knowledge through the sharing of layers that allow for the alignment of representations across languages (Wang et al., 2019). In general, the low-resource languages in a multilingual framework benefit from the existence of other languages (Liu et al., 2020). Lees et al. (2020) explores the multilingual transfer learning between English and Italian for hate speech and stereotype detection tasks. We study the effectiveness of multilingual transfer learning (also in few-shot setting) for four different languages.

In the following section, we provide a comprehensive discussion of our annotated dataset, delving into its details, the definitions of each categorical label, accompanied by relevant examples, and an overview of the annotation process.

3 Hindi Social Bias Dataset

We have constructed this dataset[3] intending to explore the identity-related social prejudices and stereotyping in Hindi from social media platforms such as Instagram, Facebook, Twitter, etc. The major part of the dataset was initially developed as part of CONSTRAINT-2021[4] (Bhardwaj et al., 2020) which focuses on hostility detection in low-resource regional languages. We augment this dataset with 994 more twitter posts which are scraped using Twitter API[5]. We scraped the tweets using some keywords relating to the Indian peninsula, such as, *adivasi, dalit, dowry, child labor, casteism, farmer-protest, muslim, islam, hindu, hinduism, article15, article370, jain, poverty, sikh*. For scraping we use both English and corresponding hindi keywords. After collecting all the tweets, we use Language Identification Models to filter Hindi tweets. After that we again filer the tweets based on the likes and retweet count. The tweets with minimum 100 likes and retweets are used for annotation purposes. The tweets were collected between January 1, 2021 and November 30, 2022.

We follow the hierarchical annotation scheme by (Zampieri et al., 2019; Singh et al., 2022) and annotate each post for (1) the presence of social bias in it, (2) the category of social bias, (3) specific identity group(s) if the post holds prejudice against/towards it, and (4) sentiment for each post. We also ask the annotators to ground their judgment by mentioning the rationale behind the social bias specified in the post.

[3]Code and dataset are available at: https://github.com/sahoonihar/Hindi_Social_Bias

[4]https://constraint-shared-task-2021.github.io/

[5]https://developer.twitter.com/en/docs/twitter-api

3.1 Terminologies

Social Bias: People typically have predetermined opinion or prejudice towards others who do not belong to their social group. A social bias is considered as the preferences for or against individuals or groups based on their social identities such as religion, race, political affiliation, profession, etc (Hammersley and Gomm, 1997). In this work, we focus on religion bias, political bias, occupation bias, and person-directed statements. The definitions of each category are as follows:

- **Religion Bias:** Bias towards/against persons or groups based on their religion or religious beliefs (Muralidhar, 2021) such as Christianity, Islam, Hinduism, etc. The religion against which the prejudice is aimed is the target word. Example:

भारतीय पुलिस द्वारा उत्तर प्रदेश में मुस्लिमों की तबाही
{Bhaartiya police dwara uttar pradesh mein muslimon ki tabaahi}
(Destruction of Muslims by Indian Police in Uttar Pradesh)

- **Political Bias:** Preconceived statement directed towards or against persons or groups based on their political beliefs. In India, the major political parties with various philosophies are the BJP, Congress, and Shiv Sena, among others. Example:

जामिया कालेज में चहेरा छुपाकर घुसे पुलिस की वर्दी में आरएसएस बीजेपी के गुंडे
{Jamiya college mein chehra chhupakar ghuse police ki vardi mein RSS BJP ke gunde}
(RSS BJP goons in police uniform entered Jamia College hiding their faces)

- **Personal Attack:** The social media posts that contain biased remarks made against renowned personalities. Personal attacks also include verbal abuse, insults, or threats directed at the individual (Vidgen et al., 2021). Example:

राहुल गाँधी को राष्ट्रीय मुर्ख घोषित करना चाहिए 😀 सहमत हो भाइयो? 🤔
{Rahul gandhi ko raashtriya murkha ghoshit karna chahiye 😀 sehmat ho bhaiyon? 🤔 }
(Rahul Gandhi should be declared a national fool 😀 do you agree brother? 🤔)

- **Personal Favor:** It shows favoritism towards famous individuals. There are many instances of bias towards politicians and celebrities from the entertainment industry in our dataset. Example:

पीएम मोदी को भगवान की तरह मानते हैं मेरठ के दिव्यांग जुड़वां भाई
{PM Modi ko bhagwan ki tarah maante hein Meerut ke divyaang judwa bhai}
(Specially-abled twin brothers of Meerut treat PM Modi like a God)

- **Occupation Bias:** Prejudice towards individuals on the basis of their professions. It also displays the preconceived belief against any vocation. Example:

जब तक हिंदुस्तान में सिनेमा है, लोग चूतिया बनते रहेंगे।
{Jab tak Hindustan mein cinema hai, log chutiya bante rahenge.}
(As long as there is cinema in India, people will continue to be hoodlums.)

- **Caste Bias:** Caste bias demonstrates societal injustice by highlighting caste inequalities (Sambasivan et al., 2021). Example:

हिन्दू राष्ट्र असल में ब्राह्मण राष्ट्र ही है मित्र। मुसलमान तो बहाना हैं, दलित-पिछड़े-आदिवासी असल निशाना हैं।
{Hindu raashtra asal mein braahman raashtra hi hai mitra. Musalmaan tou bahana hai, dalit-pichhde-aadivasi asal nishana hai.}
(The Hindu nation is actually a Brahmin nation friend. Muslims are excuses, Dalit-backward-tribals are real targets.)

- **Other Biases:** We also annotate for the other biases like race bias, class bias, etc. Discrimination based on economic background is referred to as class prejudice. Race bias is referred to as favoritism for a group of people on basis of their dialect, color, or region. We do not evaluate these categories individually due to their marginal presence in the dataset. Example:

मुंबई में लगे पाकिस्तान जिंदाबाद के नारे वह भी मंत्री जी के सामने
{Mumbai mein lage pakistan zindabaad ke naare veh bhi mantri ji ke saamne}
(Pakistan Zindabad slogans raised in Mumbai in front of the minister)

3.2 Annotation Process

Because the task is so complicated, we decided to engage three specialized annotators with understanding of Indian history, culture, and politics rather than crowd-sourcing. Each annotator determines if there is prejudice against any identity, such

as religion, race, or person-directed statements, given a social media post. If the post is labeled as biased, annotators have to annotate for bias categories and targets. For biased posts, annotators have to further mention the rationale behind the underlying bias in the form of free text. Finally, for each post, the annotators are asked to provide the also label for sentiment (Positive/Negative). The hierarchical annotation approach is depicted in Figure 1.

Acknowledging the difficulty of the task, we provide a detailed guideline and questionnaire set. The questionnaire set contains multiple two-choice (yes/no) questions for each categorical variables of our task. The Inter Annotator Agreement (IAA) was calculated using Krippendorff's alpha (Krippendorff, 2011). The IAA (α) for bias label, sentiment are 0.662 and 0.72 respectively, which shows good agreement among annotators.

To get the gold label, the data discrepancies were resolved through adjudication. Figure 2 shows some samples annotations from the dataset. We discuss more details of annotation process and guidelines in **Appendix C**.

3.3 Annotator Demographics and Treatment

All the three annotators were trained and selected through extensive one-on-one discussions. We paid very reasonable salary to all of the them for doing the annotations. They went through few days of initial training where they would annotate many examples which would then be validated by an expert and were communicated properly about any wrong annotations during training. As there are potential negative side effects of annotating such biased and sensitive posts, we used to have regular discussion sessions with them to make sure they are not excessively exposed to the harmful contents. All the annotators were Indian female and were of age between 27 to 42. One of the annotator has master degree in computer applications. Other two annotators have master degree in linguistics. The expert was an Indian female with post-graduation degree in sociology.

3.4 Data Statistics

The final dataset contains 9154 instances, of which 2300 posts are labelled as biased and 2203 posts as positive. We divided the dataset into 70:10:20 for train set, validation set, and test set, ensuring a uniform distribution of each bias category in each set. The training set contains 6388 posts, whereas validation and test sets each include 901 and 1863 posts respectively.

The majority of the biased instances come from the religion, political, and personal attack categories. This is possibly because the major source of the dataset was BBCHindi, social media posts related various news articles. The two most frequent targets for political class are the BJP and the Congress, which reflects the political affiliation among Indians. Similarly, most posts on religion bias target Hindus and Muslims, reflecting the Hindu-Muslim strife in India.

4 Experiments

This section describes all the experimental configurations. The training methodology is detailed in the **Appendix A**. Our experiments focus on predicting the presence of bias and its categories at the sentence level. We investigate multilingual transfer learning to measure the extent of task generalization across languages.

Language	# Train	# Valid	# Test	Categories
Hindi	6388 (25%)	901	1863	religion, political, occupation, *etc.*
English	6166 (50%)	600	1692	gender, race, religion, profession
Italian	6824 (44%)	500	1263	religion, race
Korean	5810 (34%)	505	1581	gender, others

Table 1: Statistics of all datasets used for experiments. Percentage of biased instances in each training dataset is shown in bracket.

Dataset: To explore multi-lingual transfer learning we use our annotated dataset in Hindi and publicly available English (Nadeem et al., 2020), Italian (Sanguinetti et al., 2020), Korean (Moon et al., 2020) datasets. English dataset (Stereoset) has posts collected after curating a set of target terms from Wikidata triplets. For each target term, there are three associated sentences corresponding to stereotypical, anti-stereotypical, and unrelated associations. We disregard the anti-stereotype associations as many of them lack relevancy and veracity (Blodgett et al., 2021). The dataset was created to assess bias in four categories: gender, race, religion, and profession. As the entire dataset is not publicly available, we use a portion of it that is. The Korean dataset was constructed using the comments from entertainment, news platforms. The dataset has two major bias labels: gender bias and other bias, which takes into account prejudice towards various attributes such as political affiliation, age, and religion. The Italian dataset is an expansion of an Italian hatespeech dataset that has been

annotated for the existence of stereotypes towards Muslims, Roma, and immigrants. Table 1 depicts the distribution of train, test, and validation sets across all four datasets.

4.1 Baselines

Along with random class and majority class baselines, we use Logistic Regression (LR) and Support Vector Machines (SVM)(Hearst et al., 1998) as baselines. For SVM, we experimented with different kernels and C-values. Linear kernel with C-value of 5 performs the best for the binary bias prediction task.

For LR and SVM baselines, we experimented with TF-IDF features (for bigrams and trigrams) and features from transformer based model (XLM-RoBERTa). In SVM and LR, the class weight parameter is set to *balanced* allowing the model to discover the appropriate weights for imbalance classes.

4.2 Mono-lingual models

We looked into two well-known multilingual pre-trained language models such as m-BERT[6] (Devlin et al., 2018), XLM-R[7] (Conneau et al., 2019). As we aim to compare and study four different languages in a single framework, we did not use any langugae specific transformer models like KoBERT, indic-BERT, *etc.* We fine-tune each model on supervised datasets and use a fully connected layer on top of each of the language model to get two outputs (for binary bias prediction). This we call as monolingual fine-tuning. The best performance is reported based on the macro-F1 Score of test set results with tuned hyperparameters (refer B). Table 2 shows the results of all the fine-tuned multilingual models when tested on the in-domain data (e.g: testing Hindi data on the model trained using Hindi train set). XLM-R[8]outperforms mBERT for all four languages. As a result, we only use XLM-R for all other experiments. We call these monolingual models as *XLM_L*, where *L* can be one among *ENG, HI, KOR, and IT*.

4.3 Multi-lingual transfer learning

We investigate multilingual transfer learning (MTL) to determine how successfully training can be transferred[9] from one language to another. In table 4, we show the results of zero-shot bias detection (direct inference) for target language, as well as the performance improvements after sequential fine-tune of the model using target language. In sequential fine-tuning step, we continuously fine-tune the source language models using the target language. We call these multilingual models as *XLM_S_L*, where both *S* and *L* can be one among *ENG, HI, KOR, and IT*. Both *S* and *L* can not be same.

4.4 MTL based on Translation

For this study, we translate all the non-English datasets into English using Google translate[10] api. As there are abudant of resources (datasets and models) already available for English, a general approach is to do classification followed by English translation. We investigate the effectiveness of this approach for bias detection using Hindi, Italian and Korean datasets. Similar to previous approach, we perform both zero-shot and sequential fine-tuning for translated datasets and report the results in table 5.

Language	Model	Metrics			
		Accuracy	Precision	Recall	F1
Hindi	m-BERT	83.6	75.3	80.0	77.1
	XLM-RoBERTa	83.3	75.6	82.8	**77.9**
English	m-BERT	90.2	90.5	90.3	90.2
	XLM-RoBERTa	95.2	95.8	95.5	**95.8**
Italian	m-BERT	73.4	73.9	74.0	73.3
	XLM-RoBERTa	75.1	75.8	75.8	**75.2**
Korean	m-BERT	74.8	72.5	73.2	72.8
	XLM-RoBERTa	76.6	74.2	73.9	**74.1**

Table 2: Results of different multi-lingual models using in-domain dataset. The first column reflects the language used to train and test models. The top performances among models are in **bold**.

Model	Metrics			
	Accuracy	Precision	Recall	F1
Majority Class	74.7	37.7	50.0	42.7
Random Class	51.7	51.4	51.9	48.2
LR + TF	82.3	74.5	60.2	61.4
LR + XF	76.9	69.7	77.2	71.1
SVM + TF	80.6	69.3	61.0	63.7
SVM + XF	76.1	67.9	74.1	69.3
XLM-RoBERTa	**83.3**	**75.6**	**82.8**	**77.9**

Table 3: Performance comparison for all baseline models with XLM-R model on Hindi dataset. TF: Tf-Idf features, XF: XLM-RoBERTa features.

[6] https://huggingface.co/bert-base-multilingual-cased

[7] https://huggingface.co/xlm-roberta-base

[8] XLM and XLM-R are used as abbreviation of XLM-RoBERTa.

[9] We look into multilingual transfer for only the social bias detection task.

[10] https://py-googletrans.readthedocs.io/en/latest/

Target Lang.→ Source Lang.↓	Direct Inference				Sequential Finetune			
	English	Hindi	Italian	Korean	English	Hindi	Italian	Korean
English (XLM_ENG)	**95.8**	59.2	50.5	48.9	-	80.7	74.6	**75.4**
Hindi (XLM_HI)	44.2	**77.9**	58.4	66.7	95.5	-	**76.5**	74.5
Italian (XLM_IT)	43.9	61.1	**75.2**	62.6	94.9	**80.8**	-	73.8
Korean (XLM_KOR)	50.2	62.9	39.7	**74.1**	95.3	80.4	75.7	-

Table 4: Comparison of monolingual fine-tuning vs multilingual fine-tuning for all datasets. Four source language models, XLM_ENG, XLM_HI, XLM_IT and XLM_KOR are the fine-tuned XLM-R models on English, Hindi, Italian and Korean datasets respectively. Last four columns correspond to sequential fine-tuning of all datasets using source language models. Best F1-scores are shown in **bold**.

4.5 Few-shot MTL

The creation of dataset for culture-specific social bias detection is very time-consuming and expensive. To tackle this issue, we explore the multilingual transfer learning in a few-shot setting. We evaluate the performances using *XLM_L* and *XLM_ENG_L* models. For fine-tuning, we use few examples (represented as N) from training sets of Hindi (*HI*), Italian (*IT*) and Korean (*KOR*) languages. We use following values of N : 25, 50, 100, 200, 400, 800, 1600. We randomly sample equal number of instances from both neutral and bias classes of the three datasets. We repeat this experiment five times for each N to report (table 6) mean and standard deviation of macro F1-scores and for plotting the 95% confidence interval in figure 3.

Configuration↓	Inference Language		
	Hindi	Italian	Korean
Zero-shot-Tr	56.6	54.6	49.1
XLM_L-Tr	75.7	74.3	69.2
XLM_ENG_L-Tr	75.8	73.9	68.6
Best model	77.9	75.2	74.1

Table 5: F1 scores on English-translated test set of each language. **Zero-shot-Tr:** direct inference on XLM_ENG model. **XLM_L-Tr:** XLM is fine-tuned on English translation of respective datasets. **XLM_ENG_L-Tr:** sequential fine-tune of XLM-ENG model using translated datasets. **Best model:** Scores using in-domain dataset (Table 2).

5 Results and Analysis

Table 2 shows how both mBERT and XLM-R models performed on Hindi, English, Italian and Korean datasets. XLM-R performs better on all four datasets (macro-F1 of 77.9, 95.8, 75.2, 74.1 respectively on Hindi, English, Italian and Korean datasets). The English dataset used for experimental purposes is balanced for bias and neutral classes and was created following some templates. It is not scraped from any social media platforms. On the other hand, Korean, Italian, and Hindi datasets are scraped from respective social media platforms; they have more natural and long sentences as compared to English datasets. Due to this, there are human errors (Grammatical, spelling, syntactical, and pragmatic errors) and convoluted constructions in datasets other than the English dataset. Both models perform best on the English dataset, possibly because the English dataset is less complex and balanced as compared to other datasets. As XLM-R consistently performs better for all languages, we use XLM-R for all other experiments.

Table 3 shows the comparison among all baseline models for on the Hindi dataset for bias prediction. For both LR and SVM, the features extracted from XLM-R model works better than Tf-Idf features. However, fine-tuning XLM-R model on the Hindi dataset gives the best performance of 77.9 macro-F1.

Table 4 shows performances of all the multilingual transfer learning experiments. In zero-shot setting (direct inference), there is very poor knowledge transfer between English and any other dataset. However, Hindi performs decently on *XLM_IT* and *XLM_KOR* models in zero shot setting. This is due to the fact that English dataset is template based dataset and other three datasets are annotated using social media comments. Also, we show that all the models perform better after fine-tuning them with training set of target language. The hypothesis is that the source language models trained on monolingual data provides better initialization for multilingual fine-tuning. Multilingual fine-tuning using Hindi data performs better over monolingual fine-tuning (macro F1 of 77.9) for every source language model. The Korean dataset performs best when ENG model is used as base model. Italian dataset performs best when the HI model

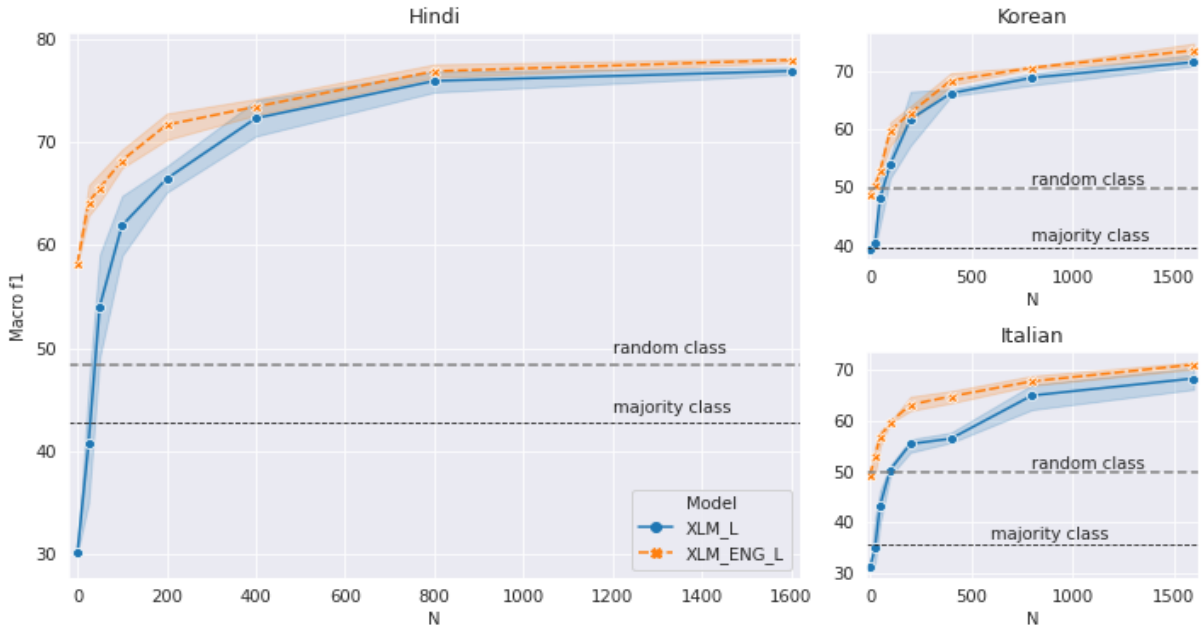

Figure 3: Macro *F1* scores on the test set of three target languages *Hindi*, *Korean* and *Italian* for different values of *N*, the number of training examples in the few-shot setting. The label *XLM_L* represents the monolingual fine-tuning of XLM with the data of a target language *L* (*Hindi/Korean/Italian*; call this *L*-pretraining). *XLM_ENG_L*, on the other hand, represents sequential fine-tuning, first with *ENG* data and then with *L* data. Notice the impact of sequential pre-training. GIVEN a desired *F1*-score, the data requirement reduces compared to *L*-pretraining, and GIVEN a fixed amount of training data, the *F1*-score is pushed up. F1 scores for all the values of N are mentioned in **Appendix D** (table 6).

is used as base model and vice versa. In general, multilingual fine-tuning outperforms monolingual fine-tuning across languages for bias detection.

Hindi also performs well on English base model data due to a good overlap of religion, occupation, and race biases in both datasets. Korean dataset has majorly gender bias instances along with other biases like political affiliation, religion, race, *etc.* Only the English dataset has significant instances of gender bias in addition to Korean. This improves the performance of the Korean dataset when English is used as the base model. Due to category overlap between the two datasets, such as political affiliation, religion, and race, the Korean dataset also performs well (macro F1 of 74.5) when Hindi is used as the base model. When measured using the XLM-R model, the average perplexity of the English, Hindi, Italian, and Korean datasets are respectively, 77.05, 85.02, 103.23, 145.57. From perplexity scores and the performances of the monolingual model, it is evident that the Korean dataset is complex in nature, and the gain in performance in the multilingual model for the Korean dataset can be attributed to the learning from the source language (English or Hindi).

The Italian dataset has a higher percentage of religion biases than race biases, and both the Italian and Hindi datasets were gathered from their respective social media platforms. The English dataset was derived from a wiki corpus, whereas the Korean dataset was derived from news articles. As a result, both Hindi and Italian dataset help each other.

The results of multilingual few-shot experiments are shown in figure 3 and table 6 (Appendix). When fine-tuned on XLM_L, we can attain an F1-score of at least 70 utilising ~ 350 training instances for the Hindi dataset. However, we need only ~ 150 instances to achieve similar F1 score when fine-tuned using XLM_ENG_L model. This behaviour is also observed in Korean and Italian datasets. Furthermore, for all values of N, multilingual few-shot fine-tuning (sequential fine-tune) performs better than monolingual fine-tuning for all three languages. *When the target-language data is limited, there is considerable benefit accruing from an initial round of fine-tuning using English data.*

However, the correlation between the amount of target language fine-tuning data and the improvement in model performance is inversely proportional. The marginal gains of increasing *N* decline sharply across models and target language test sets.

For example, for Hindi, XLM_L improves by 32 macro-F1 from $N = 25$ to 400, and by just 4 from $N = 400$ to 1600. Similar trend is also observed for XLM_ENG_L model.

Why translations can not be used directly? One fundamental question is whether we can utilise publicly accessible English datasets to forecast bias in datasets in other languages by translating them to English. Table 5 shows that zero-shot inference using an English translation of a Hindi dataset yields a macro-F1 of 56.6, much lower than the highest F1 score of 80.8. The trend is similar for Italian and Korean datasets also. One leading cause might be the loss of meaning following translation. Another possibility is that present translation algorithms are incapable of interpreting region-specific slur phrases correctly, as mentioned in the original text. Interestingly, even though the translation is correct the English translation has low frequency of occurrence attenuating its influencing power. The findings from the XLM_L-Tr and XLM_ENG_L-Tr combinations in table 5 support both of these interpretations. Even after fine-tuning the XLM or ENG basic model, the results are still subpar on certain datasets. Also, mismatches in bias categories can contribute to poor generalisation.

6 Conclusion and Future Work

We present a comprehensive dataset of $\sim 9k$ Hindi posts with multiple annotations: social bias and their categories, the sentiment of the post, the target group, and the rationales of the bias in the post. We demonstrate the capability of multilingual transfer learning using our dataset and publicly available English, Italian, and Korean datasets. Multilingual fine-tuning (sequential fine-tune) is found to be effective for Hindi, Italian, and Korean datasets, in the sense of reducing data requirements given a performance level or increasing performance level, given a fixed amount of data. Our results show that irrespective of the language family (we have dealt with Indo-European and the Altaic family here), the bias detection task benefits from multilingual sequential fine-tuning. Using few-shot experiments we show that only a small amount of target-language fine-tuning data is required to achieve strong performance and initial fine-tuning on English data can ameliorate data requirement. We report benchmarks on our dataset for the bias detection task in 4 languages. We plan to investigate the effect of multi-task training, for example, bias-and-sentiment, bias-and-explanation, and so on.

7 Acknowledgements

We would like to thank the anonymous reviewers as well as the ACL action editors. Their insightful comments helped us improve the current version of the paper. Additionally, we would like to thank Manisha, Rashmi, Sandhya Singh for their contributions to data annotation and useful comments.

Ethics Statement

Our work aims at capturing various social biases in Hindi social media posts and demonstrates the annotation quality on biases in one of existing dataset. We briefly discuss the annotation guidelines given to the annotators for the task. Also, studies of social biases come with ethical concerns of risks in deployment (Ullmann and Tomalin, 2020). As these biased news articles or social media posts can create potentially harm to any user or community, it is required to conduct this kind of research to detect them. If done with precautions, such research can be quite helpful in automatic flagging of users and news firms creating such contents.

Researchers working the problem of social bias detection on any form of text would benefit from the dataset we have collated and from the inferences we got from multiple training strategies.

Limitations

The most notable limitation of our work is the lack of external context. Consideration of external contexts that may be relevant for the classification task in our current models, such as the profile bio, user gender, post history, current and past political scenarios of the concerned region, and so on, might prove beneficial for the results in this field. Our research now focuses majorly on only six types of social biases rather than all conceivable degrees of prejudice. We also focused on utilising Hindi, English, Korean, and Italian in our study, and the Hindi dataset is primarily from the Indian context. The limited scope of concern can be further explored with our presented experiments to prove to be fruitful for a wider range of audiences by covering datasets bias annotations pertaining to other low-resource languages. We show the effectiveness of few-shot transfer learning using language models with relatively fewer parameters as compared to recent state-of-the-art language models.

References

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Mohit Bhardwaj, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2020. Hostility detection dataset in hindi.

Shaily Bhatt, Sunipa Dev, Partha Talukdar, Shachi Dave, and Vinodkumar Prabhakaran. 2022. Re-contextualizing fairness in NLP: The case of India. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping norwegian salmon: an inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*.

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.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. WWW '19.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Kate Crawford. 2017. The trouble with bias.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. *CoRR*, abs/1905.12516.

Daniel de Vassimon Manela, David Errington, Thomas Fisher, Boris van Breugel, and Pasquale Minervini. 2021. Stereotype and skew: Quantifying gender bias in pre-trained and fine-tuned language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2232–2242, Online.

Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2022. On measures of biases and harms in NLP. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 246–267.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73.

Rama Rohit Reddy Gangula, Suma Reddy Duggenpudi, and Radhika Mamidi. 2019. Detecting political bias in news articles using headline attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Gauri Gupta, Krithika Ramesh, and Sanjay Singh. 2021. Evaluating gender bias in hindi-english machine translation. *CoRR*, abs/2106.08680.

M. Hammersley and R. Gomm. 1997. Bias in social research. *Sociological Research Online*, 2(1):7–19.

M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28.

Simon Kemp. 2020. Digital 2020: 3.8 billion people use social media.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Ritesh Kumar, Enakshi Nandi, Laishram Niranjana Devi, Shyam Ratan, Siddharth Singh, Akash Bhagat, and Yogesh Dawer. 2021a. The comma dataset v0.2: Annotating aggression and bias in multilingual social media discourse.

Ritesh Kumar, Enakshi Nandi, Laishram Niranjana Devi, Shyam Ratan, Siddharth Singh, Akash Bhagat, and Yogesh Dawer. 2021b. The comma dataset V0.2: annotating aggression and bias in multilingual social media discourse. *CoRR*, abs/2111.10390.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations.

Mascha Kurpicz-Briki. 2020. Cultural differences in bias? origin and gender bias in pre-trained german and french word embeddings.

Anne Lauscher, Rafik Takieddin, Simone Paolo Ponzetto, and Goran Glavaš. 2020. AraWEAT: Multidimensional analysis of biases in Arabic word embeddings. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*.

Alyssa Lees, Jeffrey Sorensen, and Ian Kivlichan. 2020. *Jigsaw @ AMI and HaSpeeDe2: Fine-Tuning a Pre-Trained Comment-Domain BERT Model*.

Tao Li, Tushar Khot, Daniel Khashabi, Ashish Sabharwal, and Vivek Srikumar. 2020. Unqovering stereotyping biases via underspecified questions. *CoRR*, abs/2010.02428.

Sheng Liang, Philipp Dufter, and Hinrich Schütze. 2020. Monolingual and multilingual reduction of gender bias in contextualized representations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5082–5093, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,*.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Jihyung Moon, Won-Ik Cho, and Junbum Lee. 2020. Beep! korean corpus of online news comments for toxic speech detection. *CoRR*, abs/2005.12503.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on BERT model. *CoRR*, abs/2008.06460.

Deepa Muralidhar. 2021. *Examining Religion Bias in AI Text Generators*, page 273–274.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *CoRR*, abs/2004.09456.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*.

Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. 2019. Unintended bias in misogyny detection. In *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 149–155.

Arun K. Pujari, Ansh Mittal, Anshuman Padhi, Anshul Jain, Mukesh Jadon, and Vikas Kumar. 2019. Debiasing gender biased hindi words with word-embedding. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, ACAI 2019.

Paul Röttger, Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022. Data-efficient strategies for expanding hate speech detection into under-resourced languages.

Nihar Sahoo, Himanshu Gupta, and Pushpak Bhattacharyya. 2022. Detecting unintended social bias in toxic language datasets. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*.

Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-imagining algorithmic fairness in india and beyond. *CoRR*, abs/2101.09995.

Manuela Sanguinetti, Gloria Comandini, Elisa Nuovo, Simona Frenda, Marco Stranisci, Cristina Bosco, Tommaso Caselli, Viviana Patti, and Irene Russo. 2020. Haspeede 2 @ evalita2020: Overview of the evalita 2020 hate speech detection task.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. *ACL*.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *CoRR*, abs/2104.06001.

Sandhya Singh, Prapti Roy, Nihar Sahoo, Niteesh Mallela, Himanshu Gupta, Pushpak Bhattacharyya, Milind Savagaonkar, Nidhi, Roshni Ramnani, Anutosh Maitra, and Shubhashis Sengupta. 2022. Hollywood identity bias dataset: A context oriented bias analysis of movie dialogues.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

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. In *Proceedings of the*

57th Annual Meeting of the Association for Computational Linguistics.

Jean M Twenge and W Keith Campbell. 2019. Media use is linked to lower psychological well-being: Evidence from three datasets. *Psychiatric Quarterly*, 90(2):311–331.

Stefanie Ullmann and Marcus Tomalin. 2020. Quarantining online hate speech: technical and ethical perspectives. *Ethics and Information Technology*, 22(1):69–80.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. Introducing CAD: the contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Zirui Wang, Jiateng Xie, Ruochen Xu, Yiming Yang, Graham Neubig, and Jaime Carbonell. 2019. Cross-lingual alignment vs joint training: A comparative study and a simple unified framework.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the gap: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6(0):605–617.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining gender bias in languages with grammatical gender. *CoRR*, abs/1909.02224.

A Approach

Let $D_t = \{(x^i{}_t, y^i{}_t)\}_{i=1}^{N_t}$ be a training dataset with N_t examples, where $y^i{}_t$ is ground truth label for $x^i{}_t$ training instance. Further, let $D_s = \{(x^i{}_s, y^i{}_s)\}_{i=1}^{N_s}$ be the test dataset.

Given a sequence of words $x = \{w_i\}_{i=1}^{n}$ and corresponding target y, where n is the length of sequence x, we encode the input instance using model M. For logistic regression and SVM, the encoding E is the TF-IDF vector corresponding to the input x. For transformer based model, we first tokenize the input x into subword token $T = \{t_i\}_{i=1}^{m}$, where m is the number of subword tokens corresponding to the input x.

Then we feed "$[CLS]T[SEP]$" as input to the transformer encoder and obtain a d_m-dimensional hidden representation $\boldsymbol{h}$ for each input instance. Here, $\boldsymbol{h}$ is the embedding corresponding to $[CLS]$ token of the final layer of the transformer. For the training set, the hidden representation can be represented as $H = \{h_i\}_{i=1}^{N_t}$.

$$\hat{\boldsymbol{y}}_i = \text{softmax}\left(\boldsymbol{W}\boldsymbol{h}_i + \boldsymbol{b}\right) \tag{1}$$

The final hidden representation h_i is fed into the linear layer, which is then followed by a softmax function to generate the predicted label distribution $\hat{\boldsymbol{y}}_i \in \mathbb{R}^{d_y}$ for bias detection or bias category detection task. d_y is two for bias detection task and six for bias category detection. $\boldsymbol{W} \in \mathbb{R}^{d_y \times d_m}$ and $\boldsymbol{b} \in \mathbb{R}^{d_y}$ are trainable parameters along with internal parameters of transformer for transformer based models. We use cross-entropy loss between ground truth labels $\boldsymbol{y}_i$ and the predicted labels $\hat{\boldsymbol{y}}_i$ for each instance i to train the classifier.

$$\mathcal{L} = -\sum_{i=1}^{N_t}\sum_{j=1}^{d_y} \boldsymbol{y}_i^j \log \hat{\boldsymbol{y}}_i^j \tag{2}$$

B Training Details

We fine-tune all the multilingual model for five epochs. Max token length of 128 is used. We also use a dropout layer in our model. We use Adam optimizer and experiment with different learning rates: $1e-05, 2e-05, 3e-05, 4e-05, 5e-05$, different batch sizes: $8, 16, 32$, epsilon = $1e-08$, decay = 0.01, clipnorm = 1.0 were used.

Experiments were run with a single NVIDIA DGX A100 GPU. All of our implementations uses Huggingface's transformer library (Wolf et al., 2020).

C Annotation Details

C.1 Guideline

We share the definitions corresponding to each label in our dataset with annotators. We provide them the possible bias target groups and a detailed questionnaire to reduce the annotation effort. There are multiple questions in the questionnaire corresponding to each bias category. Some of the questions corresponding to each bias category is presented below.

Religion Bias:

- Does the post mock tradition/customs specific to religious group?
- Does the post use a religious slur/cuss word?
- Does the post compare two religion/religious sub-groups based on some attribute?
- Does the post favor or oppose the activities of a religious group?

Political Bias:

- Does the post refer to a political party in a prejudiced manner/mocking/contempt?
- Does the post favor or oppose the activities of a political group?

Personal Attack:

- Does the text has abusive mention against any famous personality (politician/celebrity)?
- Does the post refer to the negative traits of any famous personality(politician/celebrity) which are not factual?

Occupation Bias:

- Does the text refer to a profession in a negative/positive manner?
- Does the post compare people from same profession / different profession based on attributes like wage/skill level/skill set/Position/policy/identity?

Caste Bias:

- Is the text using any slur/cuss words against any caste?
- Is the text associating positive/negative attribute to any caste?

D Other Results

Configuration↓	Inference Language		
	Hindi	Korean	Italian
XLM_L, N-25	40.67 ± 6.48	40.3 ± 0.48	34.8 ± 4.26
XLM_L, N-50	54.0 ± 5.83	48.13 ± 3.30	43.2 ± 2.67
XLM_L, N-100	61.9 ± 3.63	53.9 ± 1.63	50.3 ± 0.83
XLM_L, N-200	66.4 ± 1.54	61.7 ± 3.80	55.4 ± 1.18
XLM_L, N-400	72.31 ± 1.91	66.2 ± 0.50	56.5 ± 0.86
XLM_L, N-800	76.3 ± 0.67	68.8 ± 0.89	65.0 ± 2.06
XLM_L, N-1600	77.1 ± 0.48	71.5 ± 0.94	68.36 ± 1.75
XLM_ENG_L, N-25	64.16 ± 1.12	50.3 ± 0.08	52.9 ± 2.36
XLM_ENG_L, N-50	65.5 ± 1.55	52.76 ± 0.40	56.7 ± 1.28
XLM_ENG_L, N-100	68.24 ± 0.97	59.7 ± 1.81	59.7 ± 0.08
XLM_ENG_L, N-200	71.67 ± 1.48	62.83 ± 1.13	63.19 ± 1.17
XLM_ENG_L, N-400	73.4 ± 1.00	68.4 ± 0.92	64.8 ± 0.91
XLM_ENG_L, N-800	76.7 ± 1.00	70.6 ± 0.18	67.8 ± 0.80
XLM_ENG_L, N-1600	77.8 ± 0.38	73.6 ± 1.07	71.1 ± 0.63

Table 6: Results of multilingual few-shot experiments. Each row represent performance on test set of Hindi, Italian, Korean for different configurations. N-100 shows that 100 examples (50-neutral, 50-biased) are used for few-shot finetune. Mean and std. dev. of macro F1-score for 5 random runs are reported.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations

☑ A2. Did you discuss any potential risks of your work?
Ethics Statement

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract; 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3

☑ B1. Did you cite the creators of artifacts you used?
2; 4

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
2; 4

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Due to the nature of the research work.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
3; C.1

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
3

C ☑ Did you run computational experiments?

5; B; E

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
B

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
B

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
5; E

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
B

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

3; C

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
C

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
C

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
C