Cultivating Fair and Accurate Knowledge: Exploring Toxicity Detection, Bias Mitigation, and Fact Verification in Multilingual Wikipedia

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computer Science and Engineering
by Research

by

Ankita Maity
2021701017
ankita.maity@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
March 2024

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "**Cultivating Fair and Accurate Knowledge: Exploring Toxicity Detection, Bias Mitigation, and Fact Verification in Multilingual Wikipedia**" by **Ankita Maity**, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date	Advisor: Prof. Vasudeva Varma
Date	Co-advisor: Dr. Manish Gupta

To my parents and fluffy ball of fur Shijju for supporting me throughout.

Acknowledgments

I would like to thank everyone who has been part of this journey. I grew a lot during my time at IIIT Hyderabad pursuing MS by Research, and that growth was largely due to others inspiring me every step of the way.

I am very grateful to Professor Vasudeva Varma for letting me work on a topic close to my heart, for providing invaluable guidance every time I got stuck or was unable to see the road forward, and for showing us that presenting our work well is as vital as the work itself.

I was very fortunate to have Dr. Manish Gupta as my co-advisor. I was always inspired by his work ethic, which in turn pushed me to work harder than I believed I could. I thank him for the weekly (and sometimes even more frequent!) discussions, which turned an abstract problem statement into concrete results. Manish sir taught me how to do research and, more importantly, that doing research can be fun.

The IRE Lab was my home during my time at IIIT Hyderabad, and I am very thankful to all lab members for making me feel part of the team, for collaborating on research projects, for providing feedback on numerous ideas during lab meetings, and of course, for the numerous treats.

I am grateful to Bhavyajeet and Sagar, the 'lab manager' and 'lab HR', respectively, for collaborating on SemEval shared tasks and also helping me out whenever I needed it. Shivansh helped out with both the Fact Verification and NPOV projects (especially with refining the reinforcement learning part), and his ideas have made both projects better. Aditya was helpful throughout, and his suggestions have helped me out many times. I am thankful to Gokul for introducing me to IndicWiki when I joined, and I also learned a lot from TAing alongside him.

I am also grateful to other lab members like Dhaval, Pavan, Rudra, Tushar, Anubhav, Nirmal, Harshit, Manav and Vijay sir. Thank you for collaborating on research projects, thank you for throwing surprise birthday parties, and overall, thank you for making my time at IIIT memorable.

This list can never be complete, and I am grateful for all the research-intensive interactions I have had during my time at IIIT, which have made me a better researcher and a better person.

Abstract

Wikipedia is one of the primary sources of encyclopedic content online. It is one of the most widely read websites. It is also regarded as a quality data source in many machine-learning pipelines. To maintain the high quality of its articles, Wikipedia has three core content policies, one of which is "Neutral Point of View (NPOV)". This policy is a set of principles, including "avoiding stating opinions as facts" and "preferring nonjudgmental language." Whenever we refer to "bias", we refer to it within these guidelines.

This work studies how to enhance the quality of the Indian language Wikipedia articles. We looked at existing work on dataset curation from English Wikipedia for bias detection and tried replicating that for Indian languages. We discuss the hurdles faced in this process and discuss translation (along with various quality checks to reduce noise) as a viable alternative.

Much of this thesis is dedicated to automatically detecting whether a sentence can be called biased and trying to remove the bias if so. Bias detection is challenging because certain words lead to bias if written in some contexts while not in others. For bias detection in Indian languages, we perform binary classification using MuRIL, InfoXLM and mDeBERTa in zero-shot, monolingual and multilingual settings. For human evaluation, we note how this is a subjective task and disagreement among annotators is expected. Thus, we also experiment with different settings like loss functions specific for subjective tasks and include anonymized annotator-specific information to help us understand the level of disagreement.

For bias mitigation, we perform style transfer using IndicBART, mT0 and mT5. These models provide strong baseline results for the novel multilingual tasks. We study how different text generation metrics may or may not be able to capture the quality of debiasing and how to evaluate our models best.

Reinforcement learning offers a way to fix the problems observed in the debiased results of the style transfer module. Also, it helps us combine the classification and style transfer modules. We formulate three reward functions specific to our debiasing task and study the results of training fully/partially with these rewards compared to vanilla mT5.

Yet another way of improving the quality of the Indian language Wikipedias is to verify the accuracy and reliability of the information presented. All material in Wikipedia must be attributable to a reliable, published source. Thus, we try to identify if the information in a sentence is factually correct or needs a citation. In contrast to previous work, we do this at the fact level instead of the sentence level for more accurate results.

Thus, these measures will enhance the quality of the Indian language Wikipedia articles and increase its credibility as the largest source of free, fair, and accurate information.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this thesis, we explore ways of enhancing the quality of articles in multilingual Wikipedia through automated detection and removal of subjective bias, and verification of factual content. This chapter explains the need for such a system and gives an overview of our contributions in this space.

1.1 Motivation

Wikipedia is one of the primary sources of encyclopedic content online. It is one of the most widely read websites. It is also regarded as a quality data source in many machine-learning pipelines. Thus, it is of great importance that the content on this website is written in a neutral, encyclopedic tone, free from subjective bias, and that the information presented is factually correct. Considering Wikipedia's (1) volume and diversity of content, (2) frequent updates, and (3) large and diverse user base, automated ways to enhance the quality of its articles are essential. This can assist editors in making changes more effectively and prevent inaccurate information or dilution of information on one of the most important websites in the world. Particularly, lower article quality and fewer editors of Indian language Wikipedia pages make such a system indispensable.

1.2 Problem Description

To maintain the high quality of its articles, Wikipedia has three core content policies: neutral point of view, verifiability, and no original research [1]. These are described as:

- **Neutral point of view (NPOV)** – All Wikipedia articles and other encyclopedic content must be written from a neutral point of view, representing significant views fairly, proportionately and without bias.

[1] `https://en.wikipedia.org/wiki/Wikipedia:Core_content_policies`

- **Verifiability** – Material challenged or likely to be challenged, and all quotations, must be attributed to a reliable, published source. In Wikipedia, verifiability means that people reading and editing the encyclopedia can check that information comes from a reliable source.

- **No original research** – Wikipedia does not publish original thought: all material in Wikipedia must be attributable to a reliable, published source. Articles may not contain any new analysis or synthesis of published material that serves to advance a position not clearly advanced by the sources.

These policies, together, determine the type and quality of material acceptable in Wikipedia articles. This work mainly focuses on the NPOV and Verifiability problems in the Indian language Wikipedias. We define NPOV more specifically with the following set of principles. Throughout this thesis, whenever we refer to "bias", we are referring to it within the scope of the following encyclopedic guidelines [2].

- **Avoid stating opinions as facts.** Usually, articles will contain information about the significant opinions that have been expressed about their subjects. However, these opinions should not be stated in Wikipedia's voice. Rather, they should be attributed in the text to particular sources, or where justified, described as widespread views, etc. For example, an article should not state that "genocide is an evil action" but may state that "genocide has been described by John So-and-so as the epitome of human evil."

- **Avoid stating seriously contested assertions as facts.** If different reliable sources make conflicting assertions about a matter, treat these assertions as opinions rather than facts, and do not present them as direct statements.

- **Avoid stating facts as opinions.** Uncontested and uncontroversial factual assertions made by reliable sources should normally be directly stated in Wikipedia's voice. Unless a topic specifically deals with a disagreement over otherwise uncontested information, there is no need for specific attribution for the assertion, although it is helpful to add a reference link to the source in support of verifiability. Further, the passage should not be worded in any way that makes it appear to be contested.

- **Prefer nonjudgmental language.** A neutral point of view neither sympathizes with nor disparages its subject (or what reliable sources say about the subject), although this must sometimes be balanced against clarity. Present opinions and conflicting findings in a disinterested tone. Do not editorialize. When editorial bias towards one particular point of view can be detected the article needs to be fixed. The only bias that should be evident is the bias attributed to the source.

- **Indicate the relative prominence of opposing views.** Ensure that the reporting of different views on a subject adequately reflects the relative levels of support for those views and that it does not

[2] https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view

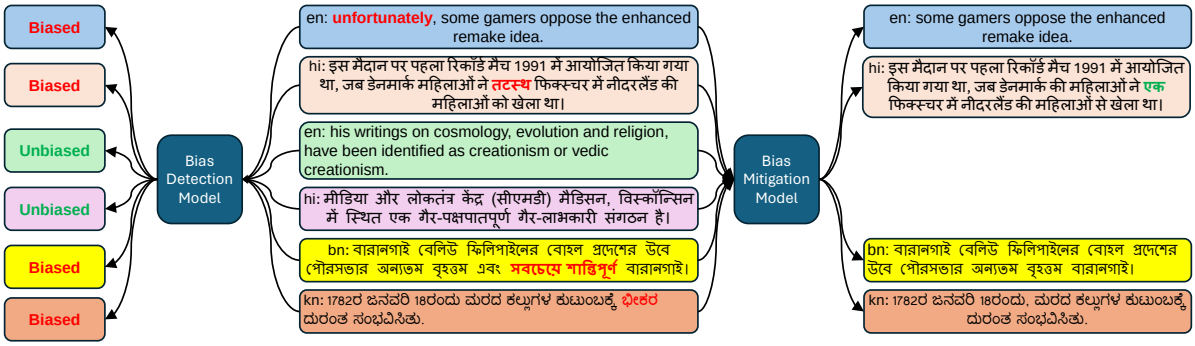

Figure 1.1 Bias detection and mitigation examples from MWIKIBIAS dataset.

give a false impression of parity, or give undue weight to a particular view. For example, to state that "According to Simon Wiesenthal, the Holocaust was a program of extermination of the Jewish people in Germany, but David Irving disputes this analysis" would be to give apparent parity between the supermajority view and a tiny minority view by assigning each to a single activist in the field.

In this work, we study how to detect sentences that violate the NPOV guidelines and convert them to more neutral sentences in Indian languages. Bias detection is challenging because certain words lead to bias if they are written in some contexts, while not in other contexts. This ambiguity arises due to language's inherent complexity and flexibility, which can lead to different interpretations of the same expression by different individuals. Additionally, subjective tasks can lead to disagreements between annotators with different perspectives or interpretations of the same text.

For **bias detection**, we perform binary classification using MuRIL [23], InfoXLM [10] and mDeBERTa [19] in zero-shot, monolingual and multilingual settings. Additionally, we focus on developing methods able to capture agreements/disagreements rather than focusing on developing the "best" model as defined by a single metric like F1. Since a "truth" cannot be assumed, "soft" evaluation is the primary form of evaluating performances, i.e. an evaluation that considers how well the model's probabilities reflect the level of agreement among annotators.

Bias mitigation is challenging because of subjectivity and context-dependence, and the models need to strike a good balance between fairness and content preservation. For bias mitigation, we perform style transfer using IndicBART [12], mT0 [39] and mT5 [58]. These models provide strong baseline results for the novel multilingual tasks. Further, using a reinforcement learning-based approach, we augment these models with rewards that help to learn the target style better.

For **verifiability**, while there have been efforts at identifying if the information in a sentence is factually correct or needs a citation, most of these approaches are monolingual and only present for high-resource languages. Furthermore, these solutions work on the granularity of a sentence. Complex sentences from Wikipedia articles can be made up of multiple facts. In such cases, the correctness of each of these facts can be more helpful than the correctness of the sentence as a whole. For this, we need

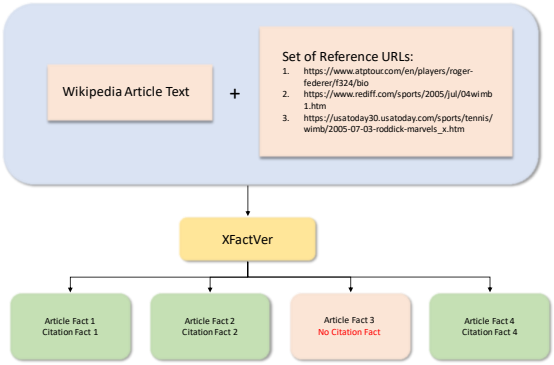

Figure 1.2 Example of the cross-lingual fact extraction and verification problem.

to have specific information about the availability of citations for each fact. Thus, it becomes necessary first to extract factual information from the sentences and then predict the label for each of those. Figure 1.2 shows an example of the XFactVer problem.

The pipeline for cross-lingually extracting factual information can also be used for multiple purposes, like automatically populating knowledge graphs such as Wikidata or utilising natural language text from multiple sources to create a common knowledge graph. Once the facts are extracted, we pass each of the facts along with semantically selected sentences from the reference through a classifier pipeline, which predicts if the citations support the fact or if the fact is in need of further citation. Such a pipeline can be used for automatically citing text on the low-resource editions of Wikipedia and reducing the manual efforts needed to identify sentences needing citations.

1.3 Main Contributions

Overall, we make the following contributions in this thesis:

- We propose multilingual bias detection, mitigation and fact verification (with granularity at the fact level) for Indian languages.
- We contribute two novel datasets, MWIKIBIAS and MWNC, to multilingual natural language generation (NLG) community. Across 8 languages, they contain ∼568K and ∼78K samples for bias detection and mitigation respectively. We also make XFACTVER, a cross-lingual dataset for fact extraction and verification public.
- We rigorously experiment with multiple transformer-based models and training setups to contribute a set of baselines for seven Indian languages for the dual problems for bias detection and mitigation (Figure 1.1) and show how n-gram based metrics are not suitable for evaluating this task.
- We experiment with different settings like loss functions specific for subjective tasks and include anonymized annotator-specific information to help us understand the level of disagreement. We perform an in-depth analysis of the performance discrepancy of these different modelling choices.
- We show the viability of deep reinforcement learning in achieving the desired style of our outputs for Indian languages.
- We propose a pipeline for automated cross-lingual fact extraction and verification, with the granularity at the fact level instead of the sentence level.

1.4 Thesis Outline

This thesis is structured as follows:

- **Chapter 1** (this chapter) explained the need for article quality enhancements in multilingual Wikipedia by means of bias mitigation and fact verification and gave an overview of our contributions in this space.
- **Chapter 2** gives an overview of the current state of the art in bias detection, mitigation and fact verification. Most of the works referred to deal with monolingual data, but we include multilingual works wherever available.
- **Chapter 3** provides a description of the data used for our experiments. It describes the various approaches we took in order to get a good-quality dataset and the challenges we faced while doing so. It should be noted that while the datasets used for bias mitigation and fact verification were datasets we created, for bias detection specifically, we used additional datasets already available for a shared task in order to compare our approaches to other participating teams.

- **Chapter 4** lists the approaches we took to detect whether a sentence is biased. We experimented with English, Arabic and 7 Indian languages for this task.
- **Chapter 5** describes the style transfer module and provides an analysis of the kind of errors faced. We use both automated and human evaluation in order to do this. We design reward functions for reinforcement learning to mitigate the problems observed during style transfer and train our baseline models in many different settings to understand the effects of these rewards.
- **Chapter 6** suggests an additional method of enhancing the quality of an encyclopedic article - by verifying that the facts listed in it are correct. We describe an automated pipeline of fact extraction and verification in this chapter.
- **Chapter 7** reiterates the key takeaways and suggests future work needed in this critical area.

The **appendix** lists some additional works performed, such as detecting human values behind arguments, detecting misogyny, as well as translation based augmentation in multilingual tweet analysis.

Chapter 2

Related Work

In this section, we discuss related work on detecting bias in Wikipedia, work done on style transfer, and work on the two stages of our approach - fact extraction and verification. We also discuss work done in incorporating multilingualism in all these aspects.

2.1 Bias Detection

Detecting various forms of bias in text has been a well-studied problem, particularly for English. The earliest work we referred to that tried to set baselines for this task using Wikipedia was Recasens et al. [48]. This work was expanded on by Pryzant et al. [46] and Zhong et al. [60], whose datasets we have also used for our task. They contribute the original WNC and WIKIBIAS datasets, respectively, from which we have derived our mWNC and mWIKIBIAS datasets.

Hube and Fetahu [22] propose a supervised classification approach, which relies on an automatically created lexicon of biased words and other syntactical and semantic characteristics of biased statements. However, this approach relies on external resources like Conservapedia, which are not readily available for multilingual settings.

There has been work on trying to detect specific kinds of biases, such as promotional tone [14], puffery [7], political bias [15], and gender and racial bias [17]. However, these are not sufficient to address POV bias, which is more general and can thus arise from sources external to this. Our work adds to this line of research and aims to provide a more generalized view than these.

2.2 Disagreements in Subjective Tasks

The 2nd edition of the Learning with Disagreements (Le-Wi-Di) task was held in SemEval 2023, while the first version was held in SemEval 2021 [54]. A survey paper by Uma et al. [56] was also released, which identified several NLP and CV tasks for which the gold-standard idealisation has been shown not to hold. It used them to analyse the extensive literature on learning from data possibly containing disagreements.

Akhtar et al. [2] who introduced the HS-Brexit dataset, trained different classifiers for each annotator, and then took an ensemble of classifiers to detect abusive language.

In the case of multiclass problems (for example, classifying different kinds of hate speech instead of simply distinguishing between hate speech vs non-hate speech), there have been efforts to frame it as which class is harder to classify instead of which text belongs to which class [45]. For our work, we stick to binary classification.

We worked with datasets that had annotated labels. However, for tasks without annotated labels to calculate soft loss, augmentation techniques like mixup, as shown by [59], could be used to distribute the probability mass over more than one class.

2.3 Style Transfer and Reinforcement Learning

Various methods to transfer the style of text from one form to another have been proposed recently. Most rely on the availability of a parallel dataset, though there has been some work on unsupervised approaches to style transfer as well. For example, unsupervised approaches have been used by Dale et al. [13] for text detoxification and by Krishna et al. [28] for formality transfer.

For this work, we follow a supervised approach for multilingual style transfer by machine translating the English resource. This is similar to the best-performing approaches of Lai et al. [30], who try four different settings: (1) pseudo-parallel data in the target language via machine translating the English resource; (2) non-parallel style data in the target language; (3) no style data in the target language; (4) no parallel data at all, to find that the first method performs the best.

To enhance the performance of our baseline models, we augment them with rewards. Similar techniques can also be used to de-bias the outputs from large language models, as shown by Liu et al. [36].

2.4 Multilingual Wikipedia

There has also been limited work on expanding quality checks on Wikipedia to multilingual settings - for example, Aleksandrova et al. [3] work on bias detection for Bulgarian and French, but their method requires a collection of language-specific NPOV tags, making it difficult to extend to other languages. Yet another way of enhancing the quality of multilingual Wikipedia is by automatically detecting vandalism, as done by Trokhymovych et al. [53].

2.5 Fact Extraction

Structured fact extraction from unstructured textual data is a widely studied problem. Two Indian languages - Hindi and Telugu have been covered in a prior work [26]. We extend this to four other Indian

languages while avoiding translation. Our work is most similar to [50] - we experiment with other fact extraction methods and extend their work to include verification as well.

2.6 Fact Verification

Prior work on fact verification has centred around the FEVER (Fact Extraction and VERification) benchmark [42, 27]. FEVER consists of 185,445 claims generated by altering sentences extracted from Wikipedia and subsequently verified without knowledge of the sentence they were derived from. The claims are classified as Supported, Refuted or NotEnoughInfo. For the first two classes, the annotators also recorded the sentence(s) forming the necessary evidence for their judgment. However, most prior work on fact verification is monolingual and works on sentence level instead of fact level [51, 21].

A popular method for verification is adding individual facts to a knowledge graph [40] and then using various knowledge graph comparison methods to compare and verify facts across the two graphs [61, 38, 9].

Chapter 3

Data

This chapter describes the different datasets we used for each of our tasks. For the NPOV task, we outline the various ways we tried to create our dataset from the Indian language Wikipedia and their shortcomings, leading us to translate the English data collected from the English Wikipedia with various quality checks in place for translation errors. For fact verification, we describe how we created our XFactVer dataset from two existing datasets, XAlign and XWikiRef. The Learning With Disagreements task already had a dataset created specifically for the task with predefined train/val/test splits for easy comparison on the leaderboard, so we briefly outline how the data was created and the specific labels in the data we used for our task.

3.1 Data for Bias Detection and Mitigation in Indian Languages

We required a dataset with consistent, substantial, parallel data points for bias mitigation. However, we found that most datasets for bias-related topics are domain-specific and represent a small part of what Wikipedia-related biases are about. Thus, datasets that failed to capture the broad range of topics that fall under the NPOV definition could not be used.

Wiki Neutrality Corpus (WNC) [46] and WIKIBIAS [60] corpus were created by looking for NPOV-related tags in the edit history of the English Wikipedia dumps. Both datasets have parallel sentence structures. We tried to replicate the data curation pipeline of these datasets, as detailed below. However, they did not work well with low-resource Indian languages due to a lack of consistency in tag usage for edits in the revision history of the Indian language version of Wikipedia.

A number of approaches went into trying to create a dataset from scratch:

- **Using only language independent tags**: Multilingual Wikipedia dumps have two kinds of tags: those that are language-dependent (consisting of words that are specific to a particular language) and those that are language-independent (tags like POV and NPOV, which occur in all language versions of Wikipedia). We tried to see if the language-independent tags were enough to collect enough data from Indian Wikipedia dumps. Each such tag is used to render a specific text on Wikipedia related to neutrality bias. For example, {{NPOV}} and {{POV}} give the rendered

text as “The neutrality of this article is disputed” while {{NPOV language}} gives us “The neutrality of the style of writing in this article is questioned”.

We tried looking at popular templates that other papers have used and manually trying to find even more such language-independent templates. We used a diachronic retrieval method for testing. This means that between the time that the template has been added and subsequently removed, the text should have been changed to reflect the bias being removed.

However, number of sentences extracted was low in this case. For example, only 116 sentences for Hindi and 1473 sentences for Italian could be collected. Also, since the datasets extracted were not large, we assessed the quality through human evaluation. It was observed that sentences are fragmented, and the tag is used out of context in many sentence pairs.

- **Using source code tags as well as comment tags**: We wanted to see if comment tags on Wikipedia, which we can directly scrape, could help us increase our dataset size. However, we found that the comment tags were generally not NPOV-related but used for other purposes [1]. Hence, we decided to concentrate on source code tags only (the tags we directly get from Wikipedia dumps) instead of searching through editor comments on Wikipedia.

- **Using domain-independent sources known to be biased**: This is similar to Hube and Fetahu’s work on detecting biased statements in Wikipedia [22]. By using an external resource like Conservapedia, which is a Wiki shaped according to right-conservative ideas, including strong criticism and attacks, especially on liberal politics and members of the Democratic Party of the United States, we can get contrastive statements. However, Conservapedia is English only. While domain-specific sources exist in regional languages, they would not generalise well to the broad range of topics Wikipedia covers. Further, a parallel dataset would be challenging to obtain since there isn’t a direct correlation between the source article and the Wikipedia article. For example, sentences from an opinion piece about a news event may not correlate sufficiently well with the Wikipedia article on the event to consider them corresponding biased and unbiased sentence pairs.

- **Using language specific tags**: As mentioned in the first point, using only language-independent tags does not give us a large number of sentences. So, we had to consider language-specific tags as well. To do this, we constructed an initial seed list of English NPOV-related tags. Then, for languages we did not know, we tried to find tags that co-occur in articles with these English tags. If these resultant tags were related to bias, they were added to the seed list and the co-occurrence script was rerun to get more tags until saturation.

  Our assumption here was that a biased article might contain both English and multilingual NPOV tags. Unfortunately, this is not the case. Although multilingual tags exist, they do not co-occur with English tags and thus cannot be found in this manner.

[1] https://en.wikipedia.org/wiki/Special:Tags

Dataset	Split	Size
mWNC	train	258.79k
	val	14.38k
	test	14.38k
mWIKIBIAS	train	252k
	val	14k
	test	14k

Table 3.1 Classification dataset statistics.

Dataset	Split	Size
mWNC	train	25k
	val	7.19k
	test	7.19k
mWIKIBIAS	train	25k
	val	7k
	test	7k

Table 3.2 Style transfer/RL dataset statistics.

- **Direct translation**: We translate the Wiki Neutrality [46] and WIKIBIAS [60] corpus into target languages of Hindi (hi), Marathi (mr), Bengali (bn), Gujarati (gu), Tamil (ta), Telugu (te) and Kannada (kn). However, the translation missed subtleties of text (for example, IndicTrans gave the same translation for both the biased and the unbiased part of the sentence pair for quite a few such sentence pairs). To get proper translations, we had to restrict the sentences to a maximum length and filter out sentence pairs with the same input and target sentence. This resulted in approximately 12 per cent of the Wiki Neutrality corpus being removed, so we could not use the same train-test split as the authors.

Thus, finally, due to the difficulties mentioned above, we had to translate the English datasets despite the various problems associated with translation. Thus, although the data we worked with was not directly from the Indian language Wikipedia, we observed that **even the translated sentences were of higher quality than the data obtained from Indian Wikipedias**. And despite the filtering restrictions we put in place to check translation quality, the **resulting number of sentences was much higher** than we would have gotten otherwise, thus enabling us to finetune advanced transformer-based architectures.

We translated the datasets using IndicTrans [47] to make our own MWIKIBIAS and MWNC datasets. For WNC, we used the 'biased-full' part, and for WIKIBIAS, we used the 'tag_binary' part. The noisier part of WIKIBIAS was almost a parallel dataset with 210888 source sentences and 211503 target sentences. We took Jaccard similarity (between source and target sentence) with a threshold $>$0.75 to get a parallel dataset from this.

We added regex filtration (URLs, phone numbers, punctuations, email IDs, etc.) and translation-specific filtration to reduce noise introduced by translation. We filtered out all sentence pairs where the translated source text and the translated target text were the same for at least one of our target languages. This occurred because the source and target texts were very similar, often differing by only a single word. We noticed that translation also caused repetition of words and filtered out sentence pairs where such repetition occurs.

We then divided the filtered dataset into a train/val/test split of 90/5/5 for our style transfer experiments. Thus, our style transfer dataset is a parallel dataset formed in this manner. Since we do not require a parallel dataset for classification, we labelled every source sentence (the sentence before the 'NPOV'-

	Task	Language	Size	Disaggregated labels	Pool annotators	Additional information
HS-Brexit	Hate speech detection	English	Train/Dev: 952 Test: 168	6	6	Aggressiveness, Offensiveness
ArMIS	Misogyny and sexism detection	Arabic	Train/Dev: 798 Test: 145	3	3	
MD-Agreement	Offensive language detection	English	Train/Dev: 7696 Test: 3057	5	>800	Domain

Table 3.3 Overview of the datasets used for Le-Wi-Di.

related tag was removed) as biased and every target sentence (the sentence after the removal of the 'NPOV' tag) as unbiased and then joined the source and target parts of our parallel dataset, leading the classification dataset to be twice the size, as shown in Tables 3.1 and 3.2. The number of samples in each language for both datasets is consistent. Some examples of biased and unbiased sentences in our dataset are shown in figs. 5.1 and 5.2.

3.2 Data for the Learning with Disagreements Task

Natural language expressions, such as sentences and phrases, can often have multiple possible interpretations depending on the **context** in which they are used. This ambiguity arises due to language's inherent complexity and flexibility, which can lead to **different interpretations of the same expression by different individuals**. Additionally, subjective tasks can lead to **disagreements between annotators with different perspectives** or interpretations of the same text.

MWIKIBIAS and MWNC datasets are used for binary classification where we assume that editors' judgement on whether a piece of text is biased or not, based on whether they added/removed the NPOV tags, is sound. However, this is an inherently subjective task, and editors do not univocally agree on a sentence as being biased or unbiased. On Wikipedia, sometimes this disagreement manifests as edit wars [2].

To help understand how to better deal with this issue, we participated in the SemEval-2023 task 11 Learning With Disagreements (Le-Wi-Di) [32]. This focuses entirely on similarly subjective tasks, where training with aggregated labels makes much less sense. In this task, we worked with three (textual) datasets with different characteristics in terms of languages (English and Arabic), tasks (misogyny, hate

[2] https://en.wikipedia.org/wiki/Wikipedia:Edit_warring

	Example	Individual Annotations	Soft labels [0,1]	Hard label
HS-BREXIT (Hate Speech)	*It's an invasion of soldiers not a migration of refugees. url*	0,0,0,1,1,0	[0.67,0.33]	0
	user user London still has the muslim mayor. Get rid of him, and we'll come to visit. #Brexit	0,1,0,1,1,1	[0.33,0.67]	1
MD-AGREEMENT (Offensive language)	*This is why so many people think Germans are the worst tourist.*	0,0,1,1,1	[0.4,0.6]	1
	Don't be afraid of Covid. Don't let it dominate your life. The freaking president of the United States. #TrumpCovid-Hoax #TrumpLied200KDied	1,1,0,1,0	[0.4,0.6]	1

Figure 3.1 Example sentences in the LEWIDI task.

speech, offensiveness detection) and annotations' methodology (experts, specific demographic groups, AMT-crowd). We leverage this additional information in order to get more accurate estimates of each annotator's annotation (results are given in detail in section 4.2.2).

All the datasets provide a multiplicity of labels for each instance. The focus is on developing methods able to capture agreements/disagreements rather than focusing on developing the best model. Since a "truth" cannot be assumed, "soft" evaluation is the primary form of evaluating performances, i.e. an evaluation that considers how well the model's probabilities reflect the level of agreement among annotators.

The three datasets we worked with for this task all deal with Twitter data - **HS-Brexit** [2], **ArMIS** [4] and **MultiDomain Agreement** [33] dataset. While HS-Brexit and MultiDomain Agreement deal with English tweets, ArMIS deals with Arabic tweets. Details of these datasets are given in Table 3.3, and a few example sentences can be viewed in Figure 3.1.

The **"HS-Brexit" dataset** consists of 1,120 English tweets collected with keywords related to immigration and Brexit. The dataset was annotated with hate speech (in particular xenophobia and islamophobia), aggressiveness, offensiveness, and stereotype by six annotators belonging to two distinct groups: a target group of three Muslim immigrants in the UK and a control group of three other individuals. All the annotations are binary, and the dataset is unbalanced towards the negative class across all four dimensions: between 7% of instances annotated with the positive class for aggressiveness and 18% for offensiveness. An analysis of the disaggregated annotation revealed interesting patterns in this dataset. In particular, in all cases of total disagreement between the two groups, the target group indicated the presence of hate, and the control group indicated its absence, but never the other way round.

The **"ArMIS" dataset** is a dataset of Arabic tweets with binary labels created to study the effect of sexism judgments of bias - particularly where judges stand on the axis from conservative to liberal. The data was annotated by three people, one self-identifying as a conservative male, one as a moderate

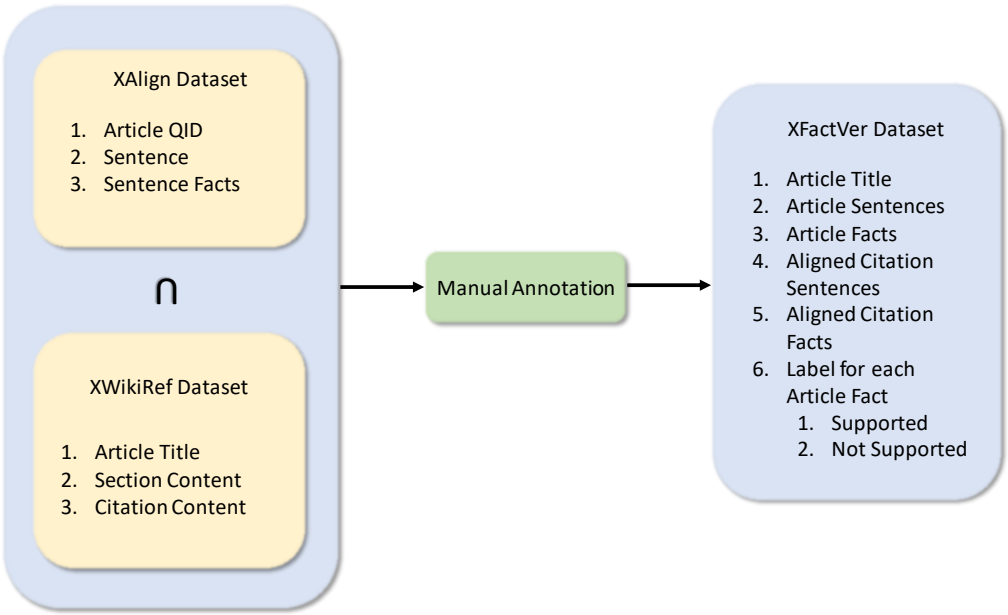

Figure 3.2 Components of the XFACTVER dataset.

female, and the last as a liberal female. The annotators labelled the tweets for sexism using the AMI guidelines from Anzovino et al [6].

The **"MultiDomain Agreement" dataset** of around 10,000 English tweets from three domains (BlackLivesMatter, Election2020, Covid-19). Five annotators via AMT annotated each tweet for offensiveness. Particular focus was put on pre-selecting tweets to be annotated that are likely to lead to disagreement. Indeed, almost 30% of the dataset has then been annotated with a two vs three annotators disagreement, while another 30% of the dataset has an agreement of one vs four judgments.

3.3 Data for Fact Verification in Indian Languages

Most existing work on fact verification utilizes the FEVER benchmark and is focused on sentence-level verification of facts. We worked with fact verification at the factoid level in Indian languages. For this, we constructed the XFactVer dataset. We used two existing datasets, the XAlign [1] and the XWikiRef [52] datasets. The XAlign dataset contains sentences from Indian language Wikipedia articles from the persons domain along with aligned facts from Wikidata. The XWikiRef dataset contains articles in Indian languages along with text from their references. We extract the intersection of these two datasets by getting the entities which are present in both XAlign and XWikiRef. To get the common entities,

Language	Articles	Sentences	Facts
Bengali	11,468	53,522	106,165
Odia	1,635	7,601	13,035
English	4,715	17,326	39,540
Punjabi	3,491	12,324	25,758
Tamil	6,003	21,937	38,100
Hindi	5,796	20,277	40,062
Total	**33,108**	**132,987**	**262,660**

Table 3.4 XFACTVER dataset statistics for each of the languages.

we needed to match titles in XWikiRef to Wikidata QIDs in XAlign. We used WikiMapper [3] to get the article title from the QID but found that exact matches between article titles give us very few data points. Removing spaces and punctuations gives us many more article matches, leading to the statistics shown in Table 3.4. We work on six languages - Bengali (bn), Odia (or), English (en), Punjabi (pa), Tamil (ta) and Hindi (hi), which are languages present in both XAlign and XWikiRef.

We extract the top 10 sentences from the reference text for all the article sentences in the dataset. We do this by checking semantic similarity between (article text, reference text) as (question, answer) pairs [4]. In order to construct the golden test data for every sentence, we manually annotate each fact. Four annotators (Computer Science students) annotated the test set, with all the six languages being split equally between them. The manual annotators provide two possible labels - either the fact is supported with respect to the reference sentences, or it isn't supported. Using this approach, we construct the XFactVer dataset. The constructed golden test dataset contains a sentence from the Indian language Wikipedia article, context from citations, manually aligned facts, and a manually annotated label. Figure 3.2 describes the components of the dataset and shows how the intersection was taken.

[3] https://github.com/jcklie/wikimapper

[4] https://huggingface.co/SeyedAli/Multilingual-Text-Semantic-Search-Siamese-BERT-V1

3.4 Relation Between the Three Datasets

mWNC and mWIKIBIAS are parallel datasets used for binary classification and style transfer. They include only the binary biased/not biased label. In addition to the above, the three datasets for Le-Wi-Di also include labels given by each human annotator. Thus, HS-Brexit includes six labels for each sentence, ArMIS includes three labels per sentence, and MD-Agreement uses five labels per sentence. Taking the majority label would have given us a dataset that was very similar to the original WNC and WIKIBIAS corpus.

Both mWNC/mWIKIBIAS and XFactVer have Wikipedia sentences in Indian languages but differ in the intent of curating such encyclopedic sentences. While the former datasets have labels for whether the sentence is biased or not, the latter labels whether the facts present in the article sentence are supported or not by the references. Thus, both can be used for different binary classification tasks, as shown in the following sections.

Chapter 4

Bias Detection

In this chapter, we aim to detect whether a given sentence has a neutrality bias or not. This task is challenging due to its subjective nature, and the multilingual aspect of the datasets we used makes it even more challenging.

Wikipedia has a guide for editors on what words are likely to induce bias and thus should be avoided[1]. Some examples include:

- **Peacock terms:** Words such as these are often used without attribution to promote the subject of an article while neither imparting nor summarizing verifiable information. For example, words like legendary, best, great, acclaimed, iconic, visionary, etc.

- **Weasel words:** These are words and phrases aimed at creating an impression that something specific and meaningful has been said when, in fact, only a vague or ambiguous claim has been communicated. A common form of weasel wording is through vague attribution, where a statement is dressed with authority yet has no substantial basis. For example, some people say, many scholars state, it is believed/regarded/considered, many are of the opinion, most feel, experts declare, it is often reported, etc.

- Other expressions that are flattering, disparaging, vague, clichéd, or endorsing of a particular viewpoint.

Here, we aim to learn a classifier which is able to detect words of this nature in a sentence. We describe in detail our main classification experiments for bias and toxicity detection. The first part of this chapter contains details on binary classification on MWIKIBIAS and MWNC datasets where we choose our best-performing model to be the one with the highest F1 scores. Later, we move to softer evaluation metrics like minimising the cross entropy to take into account disagreements between annotators for subjective tasks like this one.

[1] https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Words_to_watch

4.1 Binary Classification over MWIKIBIAS and MWNC datasets

For our baseline modelling approaches, we rely heavily on multilingual transformer-based models, which are trained on the MWIKIBIAS and MWNC datasets. The experiments below were done in the following settings:

- **Zero-shot setting:** We train only on English and test on each of the other languages.
- **Monolingual setting:** We train and test one language at a time.
- **Multilingual setting:** All eight languages are used for training at once. This results in a single checkpoint that we then test each language with.

For classification, we learn classifiers based on encoder-only models like InfoXLM [10], MuRIL [23], and mDeBERTa [19] with a twin linear layer setup to detect whether a sentence is biased. We took the best classifier to be the one with the highest macro F1 score. To keep comparisons fair, we use the base versions of all three models, each having 12 layers and 768 hidden states.

4.1.1 Overall Baseline Performance

For classification, as shown in Tables 4.1 and 4.2, mDeBERTa and MuRIL, both trained in a multilingual setting, exhibit the strongest performance, with mDeBERTa slightly outperforming MuRIL. We observed that the monolingual models tend to identify the sentences without bias better than the multilingual versions. However, they are worse at identifying sentences with bias (higher true negatives but lower true positives). Overall, considering macro F1 scores, multilingual models outperform monolingual models, which in turn outperform zero-shot approaches. We also show the detailed results for multilingual training in Table 4.4, since these are our strongest classification models.

Since **mDeBERTa trained in a multilingual way** is our best-performing model, and we use it later to evaluate the style transfer results, we also check the average logit scores of this model to see if it is able to distinguish biased text from unbiased text. We find that unbiased text gives an average score of 0.3243, while biased text gives a score of 0.6012. Thus, there is an appreciable gap between these scores, confirming its suitability to check the debiasing part.

4.1.1.1 Implementation Details

We finetuned our classification models on the entire mWNC and mWIKIBIAS datasets. Our validation and test set splits remain the same for our style transfer and reinforcement learning experiments, but we reduce our training data to 25 per cent for efficiency.

For MuRIL and InfoXLM, we use a learning rate of 1e-6, weight decay of 0.001, and dropout of 0.1. We trained for 15 epochs using a batch size of 320 and mixed precision training. For mDeBERTa, we use a learning rate 2e-5 with a weight decay of 0.01, keeping the other parameters the same.

	Accuracy	Precision	Recall	F1	MCC
MuRIL (zero shot)	0.593	0.615	0.593	0.572	0.206
MuRIL (monolingual)	0.627	0.652	0.627	0.610	0.277
MuRIL (multilingual)	0.651	0.663	0.651	0.644	0.314
InfoXLM (zero shot)	0.593	0.597	0.593	0.588	0.190
InfoXLM (monolingual)	0.610	0.621	0.610	0.600	0.230
InfoXLM (multilingual)	0.634	0.645	0.634	0.626	0.279
mDeBERTa (zero shot)	0.610	0.616	0.610	0.605	0.226
mDeBERTa (monolingual)	0.648	0.656	0.648	0.643	0.304
mDeBERTa (multilingual)	0.651	0.656	0.651	**0.648**	0.308

Table 4.1 Classification baseline results for MWIKIBIAS.

	Accuracy	Precision	Recall	F1	MCC
MuRIL (zero shot)	0.630	0.640	0.630	0.624	0.270
MuRIL (monolingual)	0.648	0.660	0.648	0.642	0.307
MuRIL (multilingual)	0.667	0.672	0.667	0.665	0.340
InfoXLM (zero shot)	0.621	0.628	0.621	0.615	0.249
InfoXLM (monolingual)	0.632	0.637	0.632	0.627	0.269
InfoXLM (multilingual)	0.655	0.658	0.655	0.653	0.312
mDeBERTa (zero shot)	0.630	0.640	0.630	0.623	0.270
mDeBERTa (monolingual)	0.666	0.669	0.666	0.664	0.335
mDeBERTa (multilingual)	0.670	0.670	0.670	**0.669**	0.340

Table 4.2 Classification baseline results for MWNC.

We use a batch size of 12 for the style transfer baseline experiments and train for 10 epochs, using early stopping with a patience of 3. We use Adafactor optimiser with a learning rate of 1e-3 for mT5 and mT0 and AdamW optimiser with a learning rate of 1e-4 for IndicBART. All models use a weight decay of 0.01.

For our reinforcement learning experiments, Adafactor optimiser with a learning rate of 1e-4 is used for mT5 and mT0. All other parameters remain the same as the style transfer baselines.

All models were run on 4 NVIDIA V100 GPUs having 32GB of RAM.

	bn	en	gu	hi	kn	mr	ta	te	avg
mWIKIBIAS	0.778	0.711	0.777	0.780	0.787	0.776	0.786	0.790	**0.773**
mWNC	0.758	0.767	0.742	0.741	0.749	0.740	0.751	0.753	**0.750**

Table 4.3 Classifier accuracy using our best classifier (target copy).

		MuRIL				InfoXLM				mDeBERTa			
		Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
mWikiBias	bn	64.55	**65.65**	64.55	63.92	62.17	63.38	62.17	61.30	**64.62**	65.15	**64.62**	**64.31**
	en	73.69	**74.74**	73.69	73.40	72.89	73.74	72.89	72.65	**74.19**	74.57	**74.19**	**74.09**
	gu	63.77	**64.93**	63.77	63.06	62.03	63.25	62.03	61.13	**63.91**	64.35	**63.91**	**63.63**
	hi	**65.31**	**66.37**	**65.31**	64.73	63.54	64.60	63.54	62.86	65.01	65.34	65.01	**64.82**
	kn	**64.33**	**65.63**	**64.33**	63.57	62.48	63.50	62.48	61.76	63.96	64.49	63.96	**63.64**
	mr	**62.74**	**63.98**	**62.74**	61.89	61.47	62.52	61.47	60.65	62.26	62.71	62.26	**61.92**
	ta	63.05	64.47	63.05	62.12	61.99	63.23	61.99	61.07	**63.80**	**64.55**	**63.80**	**63.33**
	te	**63.46**	**64.90**	**63.46**	62.56	60.81	62.17	60.81	59.69	63.34	63.99	63.34	**62.91**
	avg	65.11	**66.33**	65.11	64.41	63.42	64.55	63.42	62.64	**65.14**	65.64	**65.14**	**64.83**
mWNC	bn	**66.75**	**67.34**	**66.75**	**66.47**	65.01	65.32	65.01	64.83	66.46	66.53	66.46	66.42
	en	71.08	71.43	71.08	70.96	71.57	71.66	71.57	71.54	**72.92**	**72.92**	**72.92**	**72.91**
	gu	66.00	**66.48**	66.00	65.76	64.33	64.63	64.33	64.15	**66.42**	66.46	**66.42**	**66.40**
	hi	67.13	**67.53**	67.13	66.95	66.28	66.44	66.28	66.20	**67.45**	67.47	**67.45**	**67.44**
	kn	66.40	**66.92**	66.40	66.14	64.77	65.04	64.77	64.61	**66.55**	66.61	**66.55**	**66.52**
	mr	**64.90**	**65.48**	**64.90**	**64.57**	63.70	63.94	63.70	63.54	64.44	64.56	64.44	64.37
	ta	**65.70**	**66.29**	**65.70**	65.38	64.36	64.69	64.36	64.15	65.68	65.83	65.68	**65.60**
	te	**65.78**	**66.43**	**65.78**	65.44	63.93	64.30	63.93	63.69	65.75	65.87	65.75	**65.68**
	avg	66.72	**67.24**	66.72	66.46	65.49	65.75	65.49	65.34	**66.96**	67.03	**66.96**	**66.92**

Table 4.4 Detailed language-wise bias detection results for multilingual setup.

4.1.2 Experiments Using Contrastive Learning

Since mDeBERTa trained in a multilingual way was our best baseline, we tried to see if using contrastive learning would further increase its performance. Contrastive learning aims to pull together an anchor and a "positive" sample in embedding space and push apart the anchor from many "negative" samples. This would thus increase the gap between probabilities of positive to negative samples. We tried using this in two ways: pairwise and triplet loss.

4.1.2.1 Pairwise loss

This creates a criterion that measures the loss given inputs $x1, x2$ and a label y (containing 1 or -1).

If $y = 1$ then it assumed the first input should be ranked higher (have a larger value) than the second input, and vice-versa for $y = -1$.

The loss function for each pair of samples in the mini-batch is:

$$\text{loss}(x1, x2, y) = \max(0, -y * (x1 - x2) + \text{ margin})$$

We calculated pairwise loss in two ways: by taking the negative sample as the corresponding unbiased sentence and with a random unbiased sentence. Taking the corresponding unbiased sentence as a negative

	mDeBERTa on mWIKIBIAS					mDeBERTa on mWNC				
Metric	**Acc**	**P**	**R**	**F1**	**MCC**	**Acc**	**P**	**R**	**F1**	**MCC**
bn	0.608	0.648	0.608	0.58	0.253	0.6	0.641	0.6	0.569	0.238
en	0.635	0.65	0.635	0.626	0.285	0.623	0.662	0.623	0.598	0.282
gu	0.609	0.645	0.609	0.583	0.252	0.601	0.642	0.601	0.57	0.24
hi	0.615	0.646	0.615	0.592	0.259	0.604	0.644	0.604	0.574	0.245
kn	0.605	0.639	0.605	0.579	0.241	0.597	0.635	0.597	0.567	0.229
mr	0.596	0.64	0.596	0.562	0.232	0.589	0.635	0.589	0.55	0.219
ta	0.596	0.637	0.596	0.564	0.23	0.592	0.638	0.592	0.555	0.225
te	0.594	0.638	0.594	0.559	0.227	0.588	0.635	0.588	0.549	0.218
avg	0.607	0.643	0.607	**0.581**	0.247	0.599	0.642	0.599	**0.567**	0.237

Table 4.5 Results of contrastive learning using pairwise loss in a multilingual training setup.

sample worked better than taking a random unbiased sentence. **This is due to the corresponding text being a harder negative.** Also, based on validation loss, we set the margin to 0.4. Results are given in Table 4.5.

However, our best F1 scores were 0.581 for mWIKIBIAS and 0.587 for mWNC, which was below our baseline performance. Hence, we did not proceed with using pairwise loss.

4.1.2.2 Triplet loss

This creates a criterion that measures the triplet loss given an input tensors $x1, x2, x3$ and a margin with a value greater than 0 . This is used for measuring a relative similarity between samples. A triplet is composed by a, p and n (i.e., anchor, positive examples and negative examples respectively).

The loss function for each sample in the mini-batch is:

$$L(a, p, n) = \max\left\{d\left(a_i, p_i\right) - d\left(a_i, n_i\right) + \text{margin}, 0\right\}$$

where

$$d\left(x_i, y_i\right) = \left\|\mathbf{x}_i - \mathbf{y}_i\right\|_p$$

The norm is calculated using the specified p value and a small constant ε is added for numerical stability. While p and n were given as biased and corresponding unbiased sentences, the anchor a was a random sentence that could be biased or unbiased. If it were a biased sentence, we would decrease its similarity to the unbiased sample and increase its similarity to the biased sample and vice versa for unbiased anchor sentences.

However, this performed worse than even the pairwise loss. Thus, we could not use it for our final experiments to judge the style transfer outputs.

While these are empirical results, and we cannot say what definitively caused our contrastive learning experiments not to do better than our baselines, we suspect that it is in part because our source and target sentences were very similar to each other. Thus, the model was unable to contrast well between their representations.

4.1.3 Using ChatGPT API

We also tried prompting GPT-4 using the API to try to see how it performs in Indian languages. We used few shot prompting using examples taken from Wikipedia quizzes to help new users understand what the NPOV policy is and is not[2]. The final version of the prompt used was the following:

```
Please classify the following sentence as either violating Wikipedia's
Neutral Point of View Policy or not. Note that the sentence is in {LANG},
but you should still try to classify it as best as possible. The policy
is not violated if all the significant views published by reliable
sources on a topic are represented fairly, proportionately, and, as far
as possible, without editorial bias. Thus, a neutral sentence should
avoid stating opinions as facts, avoid stating seriously contested
assertions as facts, avoid stating facts as opinions, prefer nonjudgmental
language, and indicate the relative prominence of opposing views.
Thus, you must perform the following task:
1. Provide your response in a JSON format containing two keys, "reason"
and "label".
2. You should explain your reason in the "reason" field. You can use
{LANG} or any other language to do this.
3. The "label" should be assigned '1' if it violates the policy and
assigned '0' if it is a neutral {LANG} sentence and does not violate the
policy.
4. Do not provide any additional information except the JSON.
```

While the above prompt performed well on the few sentences we tested it on before running it on a quarter of the entire test set, it did not do well on the test set. From the "reason" field, ChatGPT generally tried to translate the sentence to English by itself before attempting to justify whether it was neutral or not. However, in this process, the main issues observed were:

- It could not translate many of the sentences and mostly gave up, saying it could not understand the sentence.

[2] https://en.wikipedia.org/wiki/Help:Introduction_to_policies_and_guidelines/neutrality_quiz

- The translation happened correctly, but it could not detect the bias properly and went with the label 0 (unbiased) much more than the label 1 (biased). This led to a really poor F1 score.

Thus, we decided not to use ChatGPT for the later stages of our work.

4.2 Using Soft Labels to Measure Annotator Disagreement

This section describes another binary classification task. However, there is one addition. All the datasets provide a **multiplicity of labels for each instance** instead of a single biased/unbiased label. The focus is on developing methods able to capture agreements/disagreements rather than focusing on developing the best model. Since a "truth" cannot be assumed, "soft" evaluation is the primary form of evaluating performances, i.e. an evaluation that considers how well the model's probabilities reflect the level of agreement among annotators.

4.2.1 Methodology

Since the datasets we used have both textual information and external information, we needed to combine this information to derive meaningful insights. This combination can be done in various ways - from simply concatenating the information to using the attention-based summation of the information.

The main models we used were BERTweet [41] for HS-Brexit and MD-Agreement, and AraBERT [5] for ArMIS. LM-based text embeddings were common across all datasets, but other embeddings we used varied based on the datasets. For HS-Brexit, we used embeddings for aggressiveness and offensiveness information. For MD-Agreement, we first concatenated the tweet's domain information to the tweet's text. Also, since there were over 800 annotators for MD-Agreement, we needed additional embeddings to capture this. For this, we used one-hot vectors and let the model learn information about the annotators based on their annotations. In our experiments, we concatenate the embeddings and then use attention-based mixing as a combining module. We use auxiliary losses to improve model performance. Figure 4.1 shows a high-level overview of this structure.

Initially, we use hard labels (for the submission) but later also experiment with soft labels and apply softmax over the logits produced by the classifier. [55] found that soft-loss training systematically outperforms gold training when the objective is to achieve a model whose output mimics most closely the distribution of labels produced by the annotators. We compare the effects of using soft loss training with respect to hard loss training on the given datasets, as explained in the results section.

For the final experiment, we improve our mixers and add more complexity to the model. We experiment with multi-head attention-based mixing for this. The final embedding is obtained after three layers of multi-head attention-based mixing followed by feed-forward layers. Since, for ArMIS, no additional information was present, this dataset was excluded from this experiment.

Implementation details: For our experiments, we use Adam optimizer with a learning rate of 1e-6 and cyclicLR scheduler with triangular2 mode. We train the model for 30 epochs with a batch size of 16.

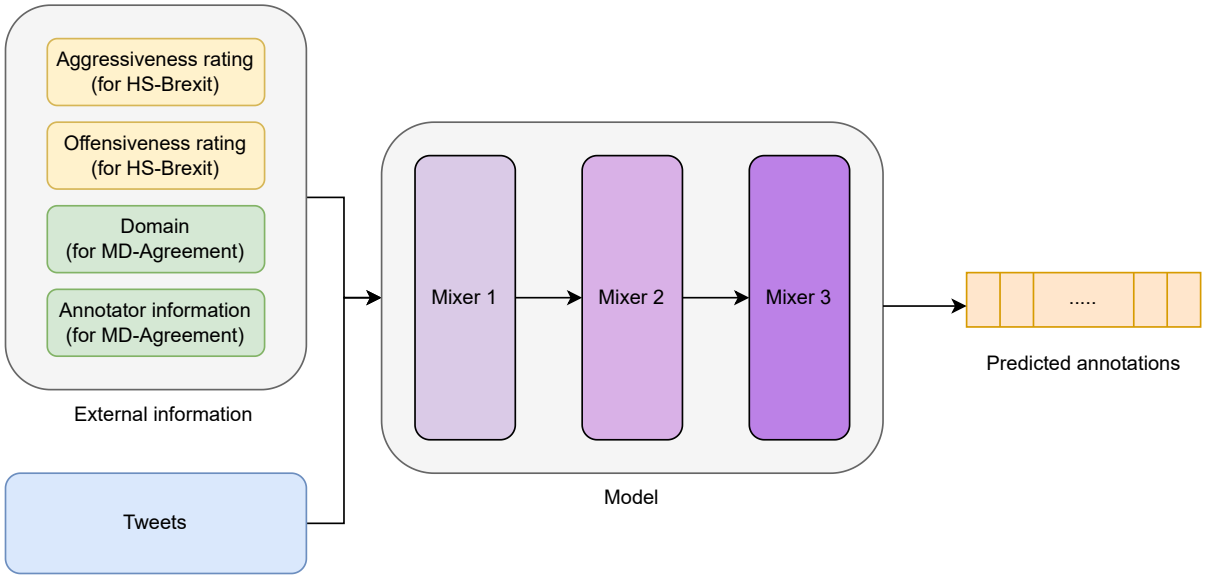

Figure 4.1 A high level overview of our model.

4.2.2 Results

4.2.2.1 Overall Performance

We summarize the results from our experiments in Table 4.6. Cross entropy was used as the primary evaluation metric, but we also show micro F1 scores alongside cross-entropy. Using hard loss gives the results that were submitted for the competition, and we compare that with our other experiments.

The addition of soft loss most helped MD-Agreement results, but the results were mixed for HS-Brexit and ArMIS. In fact, the best performance of ArMIS for cross-entropy came from hard loss. This may be because ArMIS used no additional information besides text and had the least number of annotators (just three) to distribute the probability mass.

Our architectural improvements, which included designing better mixers, gave better cross entropy and micro F1 results for both HS-Brexit and MD-Agreement datasets. **The best results for cross entropy for test sets of these two datasets resulted from this.**

4.2.2.2 Error Analysis

Some tweets have a larger amount of disagreement than others. Two cases are of particular interest to us. We wanted to check how many obvious cases (annotators agreed with 75 per cent certainty over the class the tweet belonged to) our system was missing.

We also wanted to check how many less obvious cases (there was less than 35 per cent agreement between annotators) our system could predict correctly.

Testing strategy	HS-Brexit				ArMIS				MD-Agreement			
	val		test		val		test		val		test	
	CE	F1	CE	F1	CE	F1	CE	F1	CE	F1	CE	F1
Majority baseline	2.71	0.89	5.62	0.89	8.23	0.60	8.91	0.57	7.74	0.65	7.38	0.67
Hard loss	0.47	0.86	0.75	0.84	4.55	0.57	**4.01**	0.58	7.50	0.51	9.92	0.42
Soft loss	0.65	0.88	1.07	0.86	3.82	0.58	4.70	0.56	6.42	0.57	8.73	0.50
Better mixers with multi-head attention	0.58	0.88	**0.58**	0.84	-	-	-	-	3.40	0.59	**3.70**	0.58

Table 4.6 Results for cross entropy and micro F1 across the three datasets.

For the **obvious cases our system was missing**,

- HS-Brexit had 26 predictions wrong in total, out of which 10 were obvious cases.
- ArMIS had 61 predictions wrong in total, out of which 37 were obvious cases.
- MD-Agreement had 1275 predictions wrong in total, out of which 877 were obvious cases.

For the **less obvious cases our system was able to predict correctly**,

- HS-Brexit had 33 less obvious instances, out of which our system correctly predicted 17 instances.
- ArMIS had 53 less obvious instances, out of which our system correctly predicted 29 instances.
- MD-Agreement had 856 less obvious instances, out of which our system correctly predicted 458 instances.

Thus, on average, our model was able to predict 53.24 per cent of controversial/less obvious cases correctly, which seems promising. However, more work is needed since 55.97 per cent of incorrect predictions were obvious cases.

4.3 Summary

In this chapter, we saw different ways of detecting whether text is biased in any way or is a neutral text that belongs in an encyclopedia. For Indian languages, we experimented with zero-shot, monolingual and multilingual setups, while for English, we also experimented with a few other datasets of a similar nature. For this task, we used hard labels like F1 scores and later experimented with soft labels in which we applied softmax over the logits produced by the classifier. We also compared the effects of using soft loss training with respect to hard loss training.

Chapter 5

Bias Mitigation

Bias detection enables us to find which texts are biased through various classification approaches. While unbiased texts can be left as is, the sentences detected as having some form of bias need to be debiased to maintain a neutral tone throughout the article being considered. We consider style transfer as a way to perform this debiasing. Text style transfer is a text generation problem that involves modifying the style of a given text while preserving its underlying content. The goal is to transform the linguistic characteristics of the text, such as formality, tone, or sentiment, from one style to another while ensuring that the core meaning or information remains unchanged.

For instance, in transforming biased sentences to unbiased ones, the style transfer task focuses on changing the tone and language expressions associated with bias without altering the factual content. This chapter outlines the experimental details of this process for our MWIKIBIAS and MWNC datasets.

5.1 Style Transfer Baselines

5.1.1 Models Tested

For style transfer (Figure 5.1), we fine-tune the following multilingual encoder-decoder transformer-based models over the MWIKIBIAS and MWNC parallel corpus to perform debiasing. We use the small versions of all three models for our experiments.

- mT5 (google/mt5-small) [58]
- IndicBART (ai4bharat/IndicBART) [12]
- mT0 (bigscience/mt0-small) [39]

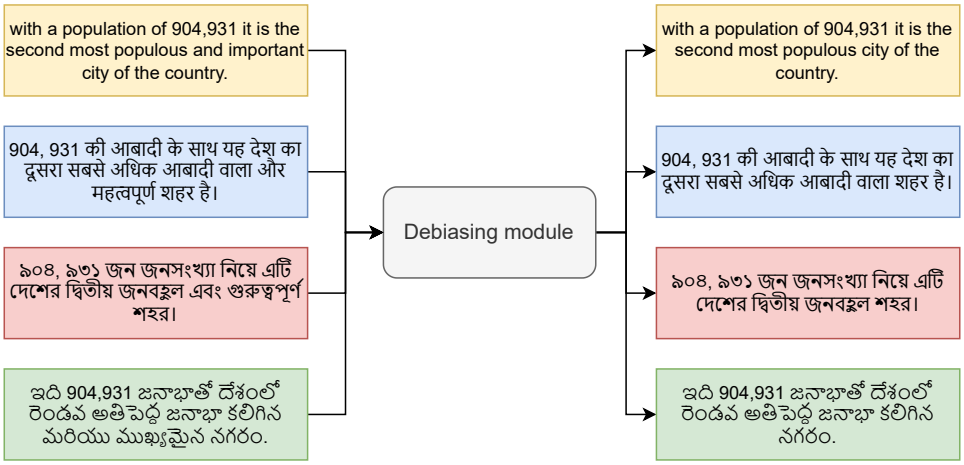

Figure 5.1 The debiasing module.

5.1.2 Choice of Metrics

The effectiveness of bias mitigation models should be evaluated broadly on two aspects: match with ground truth and debiasing accuracy. For measuring match with groundtruth unbiased sentences, we use standard NLG metrics like BLEU, METEOR, chrF and BERT-Score.

Classifier accuracies using our best-performing classifier (mDeBERTa model trained in a multilingual setting) were considered for evaluation in addition to content preservation metrics like BLEU scores. To do this, we passed the outputs of our style transfer module to the classifier and measured the accuracy, assuming the ground truth to be unbiased. Thus, this gives us an estimate of how well our style transfer module was able to debias the text. For setting a threshold on how high we could expect our accuracy values to be using this classifier, we also pass the target text part of our style transfer dataset, which we know to be unbiased, to the same classifier and measure the accuracy values (reported as target copy values) in Table 4.3.

A model can easily obtain a high match with ground truth by simply copying words from the input (since the source and the target sentences, in this case, are very similar). Similarly, a model can easily obtain a high accuracy score by predicting a constant highly unbiased sentence independent of the input. A good model should be able to strike a favourable tradeoff between the two aspects. Among the four metrics for computing the match, **BERT-Score** has been shown to be the most reliable in NLG literature because it captures semantic match rather than just a syntactic match. Thus, we evaluated the best style transfer model to be that with the **highest value of the content preservation metrics as well as the highest classifier accuracy**.

	BLEU	METEOR	chrF	BERTScore	Classifier accuracy
IndicBART (monolingual)	63.67	75.87	80.04	91.58	0.59
IndicBART (multilingual)	46.32	64.62	68.94	88.47	0.59
mT0 (monolingual)	61.57	77.05	80.84	93.24	0.62
mT0 (multilingual)	60.86	77.04	80.89	93.20	0.63
mT5 (monolingual)	58.81	76.74	80.23	92.97	0.59
mT5 (multilingual)	63.26	77.41	81.39	**93.40**	**0.64**

Table 5.1 Style transfer baseline results for MWIKIBIAS.

	BLEU	METEOR	chrF	BERTScore	Classifier accuracy
IndicBART (monolingual)	54.98	69.25	75.27	90.99	0.56
IndicBART (multilingual)	17.58	59.67	61.15	85.54	0.54
mT0 (monolingual)	53.09	70.01	75.75	91.27	0.59
mT0 (multilingual)	55.23	70.61	76.54	**91.50**	**0.60**
mT5 (monolingual)	55.39	70.28	76.22	91.36	0.57
mT5 (multilingual)	55.27	70.46	76.41	91.47	0.59

Table 5.2 Style transfer baseline results for MWNC.

5.1.3 Results and Analysis

Results for style transfer are given in Tables 5.1 and 5.2 (these results are averaged across all eight languages, but we also show the full results in Tables A.7 and A.8). We observe that **the models leave a large percentage of sentences unchanged for the debiasing task.** In many cases, the input and output sentences are very similar, so it may be **hard for the models to figure out the correct part of the sentence to change.** This contributes to the difficulty of the task. The high value of n-gram based content preservation metrics like BLEU/METEOR/chrF only tells part of the story because some of the highest values are obtained for models that do not change the biased sentences in any way.

We take the best model to be the one with the highest classifier accuracy and the highest value of BERTScore. **Since these two metrics correlate, we can choose a single best model for each of our datasets**. Thus, the best model for mWIKIBIAS is mT5, trained in a multilingual way, and the best model for mWNC is mT0, trained in a multilingual way. Further experiments using reinforcement learning were thus performed using these two experimental settings.

Broadly, multilingual models outperform monolingual counterparts. And as expected, both models work best for English.

5.1.4 Human Evaluation

We asked 4 Computer Science bachelors students with language expertise to evaluate the generated outputs (mT5 multilingual for MWIKIBIAS and mT0 multilingual for MWNC) on 3 criteria, each

Lang.	MWIKIBIAS			MWNC		
	Fluency (↑)	Bias (↓)	Meaning (↑)	Fluency (↑)	Bias (↓)	Meaning (↑)
bn	4.42	3.12	4.79	3.94	2.68	4.80
en	**4.92**	2.72	**4.84**	**4.86**	**2.40**	**4.92**
hi	4.20	3.20	4.76	4.60	2.64	**4.92**
te	4.40	**2.50**	4.81	3.88	2.45	4.75

Table 5.3 Human evaluation results

on a scale of 1 to 5: fluency, whether the bias is reduced and whether the meaning is preserved when compared to input. This was done for 50 samples per language for both datasets. Table 5.3 shows that automated evaluation correlates well with human judgment, with English predictions showing the best results. MWNC is easier for the models to debias than MWIKIBIAS. The model outputs were generally fluent and had similar content as the input text. However, a wider variance in bias mitigation abilities was observed for the 3 Indian languages tested compared to English. Ambiguity in bias assessment and noise in the reference text made $\sim$20% of the samples challenging for human annotators.

5.2 Using Reinforcement Learning to Augment Style Transfer

5.2.1 Methodology

Reinforcement learning offers a way to fix the problems observed in the debiased results given by the style transfer module. We formulate three reward functions specific to our task of generating debiased multilingual text.

We needed to tell the model not simply to copy the sentences from the input to the output. But for this, the model needed to be able to identify which part of the sentence to change. This is how the change-based reward was designed. The embedding-based reward is a standard reward used in tasks similar to this, and it also follows naturally as a consequence of using the change-based reward. During hyperparameter tuning, we observed that these two rewards correlated well with each other. On the other hand, the KL-divergence based reward helps us combine the classification and style transfer modules. We use these rewards similarly to Lai et al. [29].

- **Change based reward:** Since we notice from our baseline experiments that the models tend to leave a large percentage of sentences unchanged, we add this reward to maximise the number of unbiased words in the generated text. For this, we assume biased words to be words that were removed from the input text in order to produce the reference text. Correspondingly, unbiased words are words that are not present in the difference between the input and reference texts. For calculating the reward, we normalise by dividing by the total number of words generated by the model.

Language	Input text	Reference text	Generated text
Bengali	জেমা লুইস ওয়ার্ড (জন্ম নভেম্বর ৩, ১৯৮৭) একজন অস্ট্রেলীয় সুপারমডেল এবং অভিনেত্রী।	জেমা লুইস ওয়ার্ড (জন্ম নভেম্বর ৩, ১৯৮৭) একজন অস্ট্রেলীয় মডেল এবং অভিনেত্রী।	জেমা লুইস ওয়ার্ড (জন্ম নভেম্বর ৩, ১৯৮৭) একজন অস্ট্রেলীয় মডেল এবং অভিনেত্রী।
English	pauline stainer (born 1941) is an acclaimed english poet.	pauline stainer (born 1941) is an english poet.	pauline stainer (born 1941) is an english poet.
Gujarati	નિકોલસ કોપરનિકસ, પોલિશ ખગોળશાસ્ત્રી અને ગણિતજ્ઞ	નિકોલસ કોપરનિકસ, ખગોળશાસ્ત્રી અને ગણિતજ્ઞ	નિકોલસ કોપરનિકસ, ખગોળશાસ્ત્રી અને ગણિતજ્ઞ
Hindi	डाल्गेटी बे में दो प्रतिष्ठित प्राथमिक विद्यालय हैं: डाल्गेटी बे प्राथमिक विद्यालय और डोनीब्रिसल प्राथमिक विद्यालय।	डाल्गेटी बे में दो प्राथमिक विद्यालय हैं: डाल्गेटी बे प्राथमिक विद्यालय और डोनीब्रिसल प्राथमिक विद्यालय।	डाल्गेटी बे में दो प्राथमिक विद्यालय हैं: डाल्गेटी बे प्राथमिक विद्यालय और डोनीब्रिसल प्राथमिक विद्यालय।
Kannada	ಎಲಿಜಾಬೆಥನ್, ಎಡ್ವರ್ಡಿಯನ್, ಜಾರ್ಜಿಯನ್ ಮತ್ತು (ಎಲ್ಲಕ್ಕಿಂತ ಹೆಚ್ಚು ಪ್ರಸಿದ್ಧವಾದ) ವಿಕ್ಟೋರಿಯನ್ ಇವುಗಳಿಗೆ ಉದಾಹರಣೆಗಳಾಗಿವೆ.	ಎಲಿಜಾಬೆಥನ್, ಎಡ್ವರ್ಡಿಯನ್, ಜಾರ್ಜಿಯನ್ ಮತ್ತು ವಿಕ್ಟೋರಿಯನ್ ಇವುಗಳಿಗೆ ಉದಾಹರಣೆಗಳಾಗಿವೆ.	ಎಲಿಜಾಬೆಥನ್, ಎಡ್ವರ್ಡಿಯನ್, ಜಾರ್ಜಿಯನ್ ಮತ್ತು ವಿಕ್ಟೋರಿಯನ್ ಇವುಗಳಿಗೆ ಉದಾಹರಣೆಗಳಾಗಿವೆ.
Marathi	दुर्दैवाने, पास्कलच्या तरुण मुलाला एक भयंकर आजार झाला.	पास्कलच्या तरुण मुलाला एक भयंकर आजार झाला.	पास्कलच्या तरुण मुलाला एक भयंकर आजार झाला.
Tamil	டயானா ஸ்டார்கோவா (Diana Starkova) ஒரு துருக்கிய-உக்ரைனிய சூப்பர் மாடல், ஆட்டோ பந்தய ஓட்டுநர் மற்றும் முன்னாள் அழகு ராணி ஆவார்.	டயானா ஸ்டார்கோவா (Diana Starkova) ஒரு துருக்கிய-உக்ரேனிய மாடல், ஆட்டோ பந்தய ஓட்டுநர் மற்றும் முன்னாள் அழகு ராணி ஆவார்.	டயானா ஸ்டார்கோவா (Diana Starkova) ஒரு துருக்கிய-உக்ரேனிய மாடல், ஆட்டோ பந்தய ஓட்டுநர் மற்றும் முன்னாள் அழகு ராணி ஆவார்.
Telugu	న్యూ యార్క్ నగరం యొక్క డైమ్ మార్పు మరొక ప్రియమైన సంప్రదాయం.	న్యూ యార్క్ నగరం యొక్క డైమ్ మార్పు మరొక సంప్రదాయం.	న్యూ యార్క్ నగరం యొక్క డైమ్ మార్పు మరొక సంప్రదాయం.

Figure 5.2 Predictions for MWNC dataset generated by mT0-small with reinforcement learning.

- **Embedding based reward:** Using only the change-based reward may lead the models to repeat words when they generate text, as long as the generated words are unbiased. Thus, we also simultaneously aim to maximise the cosine similarity between the reference text and the generated text using Sentence Transformers [49] [1].

- **KL-divergence based reward:** This is based on using our best-performing classifier checkpoint (mDeBERTa trained in a multilingual setting) on the reference and generated texts. After getting the log probabilities, we use this output to calculate KL-divergence.

Embedding based reward is used for content preservation, while the change based and KL-divergence based rewards help with style transfer intensity. The rewards are used for policy learning. The policy gradient is

$$\nabla_{\phi} J(\phi) = E\left[R \cdot \nabla_{\phi} \log\left(P\left(\boldsymbol{y}^{s} \mid \boldsymbol{x}; \phi\right)\right)\right]$$

where R represents the above three rewards, $\boldsymbol{y}^{s}$ is sampled from the distribution of model outputs at each decoding time step, and ϕ are the parameters of the model. The overall objectives for ϕ are the cross entropy loss of the base model and the policy gradient of the different rewards. In addition to calculating the metrics used for the baselines, we also perform human evaluation using the answers given by GPT-3.5 [8] as references.

[1] https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v2

For reinforcement learning, we train the models in two different settings. One is to use RL-based training right from the first epoch. However, our best results were obtained from partial training - where we trained mT0 and mT5 using regular cross entropy for three epochs and then used RL-based training for the remaining seven epochs. Some example outputs from this are given in Figure 5.2.

Strategy	**mT5 multilingual on mWIKIBIAS**					**mT0 multilingual on mWNC**				
Metric	**B**	**M**	**C**	**BS**	**Acc**	**B**	**M**	**C**	**BS**	**Acc**
bn	62.94	75.16	79.42	92.65	0.55	54.41	68.78	75.19	90.45	0.38
en	86.80	92.77	91.64	98.44	0.60	79.42	89.15	88.84	97.68	0.35
gu	63.57	76.00	79.04	92.72	0.53	55.04	69.97	74.86	90.55	0.36
hi	71.85	82.09	81.97	93.98	0.55	63.46	77.26	78.14	91.94	0.37
kn	61.94	74.96	81.18	93.09	0.54	54.69	69.71	78.03	91.22	0.36
mr	60.15	72.76	77.53	92.01	0.56	51.46	66.00	73.11	89.67	0.40
ta	55.91	69.07	77.40	91.49	0.58	49.03	63.35	74.34	89.44	0.39
te	60.42	72.63	78.60	92.08	0.56	52.37	67.07	75.19	90.15	0.39
avg	65.45	76.93	80.85	**93.31**	**0.56**	57.49	71.41	77.21	**91.39**	**0.38**

Table 5.4 Results of RL. B=BLEU, M=METEOR, C=chrF, BS=BERTScore, Acc=classifier accuracy.

Strategy	**mT5 multilingual on mWIKIBIAS**					**mT0 multilingual on mWNC**				
Metric	**B**	**M**	**C**	**BS**	**Acc**	**B**	**M**	**C**	**BS**	**Acc**
bn	63.15	75.67	80.07	92.83	0.57	55.26	68.31	74.94	90.72	0.53
en	86.52	92.72	91.72	98.33	0.63	82.01	88.06	87.95	97.62	0.59
gu	63.52	76.52	79.78	92.91	0.55	55.89	69.43	74.60	90.82	0.51
hi	72.00	82.81	82.85	94.21	0.57	65.00	76.62	77.79	92.21	0.53
kn	61.87	75.67	82.07	93.30	0.56	55.30	69.00	77.60	91.46	0.51
mr	60.44	73.45	78.33	92.18	0.57	52.16	65.35	72.69	89.85	0.54
ta	56.58	69.91	78.43	91.69	0.59	48.98	61.99	73.29	89.49	0.55
te	60.37	73.15	79.35	92.27	0.58	60.62	73.41	79.49	92.30	0.53
avg	65.56	77.49	81.58	**93.47**	**0.58**	59.40	71.52	77.29	**91.81**	**0.54**

Table 5.5 Results of RL partial training. B=BLEU, M=METEOR, C=chrF, BS=BERTScore, Acc=classifier accuracy.

5.2.2 Results and Analysis

Tables 5.4 and 5.5 show the results of using reinforcement learning on our best style transfer baselines. For mT5, we kept the reward coefficients to e = 0.3, u = 0.15 and c = 0.15 for both full and partial training based on loss observed on the validation set (where 'e' represents the embedding-based reward, 'u' represents the change based reward, and 'c' represents the KL-divergence reward using our best performing classifier). For mT0, we kept e = 0.15, u = 0.15, and c = 0.15.

For the same reasons outlined above, we consider BERTScore and classifier accuracy as our primary methods for evaluation. We observe that when averaged across all eight languages, partial training (regular cross entropy for three epochs and then RL-based training for the remaining seven epochs) gives much better results than just using reinforcement learning from the beginning. However, they do not outperform our baseline results, even though we tried various combinations of rewards across different hyperparameter settings.

We did observe some of the metrics outperforming our baselines. Still, after performing a two-tailed test, we got a p-value greater than 0.05, thus proving that our improvements were **not statistically significant**. Thus, using only reinforcement learning was outperformed by partial training, but the baselines outperformed partial training as well. Hence, we concluded that the rewards did not help much, whether used individually or in combination and did not proceed further with this line of training our models.

5.3 Summary

This chapter explored ways to convert multilingual biased text to unbiased text through various style transfer and reinforcement learning techniques. We designed reward functions for reinforcement learning to mitigate the problems observed during style transfer and trained our baseline models in many different settings to understand the effects of these rewards. We also analysed the kinds of errors faced by using both automated and human evaluation.

Chapter 6

Fact Verification as an Additional Quality Check

This chapter explains our two-stage approach to multilingual fact verification. Fact verification is a significant problem, especially now, given the spread of misinformation in today's digital age and when LLMs are known to hallucinate information. It can be used in multiple different contexts when we have text that we do not know whether to rely on or not, and we have some amount of evidence to compare it with.

In factual texts online, whether Bing chat answers or encyclopedic articles, we automatically use citations to verify a given claim's factual correctness. In our specific use case, we extracted and verified factoids from Wikipedia using each article's references.

6.1 End-to-End Pipeline

The automated pipeline for fact-level verification is constructed in the following two phases: fact extraction and fact verification. The proposed solution operates entirely in a cross-lingual setting, where t**he article text and the references can be in any language**. The pipeline integrates several natural language processing techniques to extract the relevant facts from the input sources. The extracted facts are then verified against the references, leveraging the semantic relationships between the facts and the reference sources. Figure 6.1 gives a diagrammatic overview of this process.

6.1.1 Fact Extraction

The task of cross-lingual fact extraction (CLFE) involves extracting English facts from the natural language text of multiple low-resource Indian languages. For this task, we propose two methods. The results for them are shown in Table 6.1.

- The first approach formulates this problem as a **text-to-text task** and finetunes a pre-trained mT5 model on our XFactVer dataset for extracting the English facts.
- For our second method, to check the **viability of LLMs** for this task, we prompt GPT-4 to extract facts in English from the multilingual sentences. We used few shot prompting by using examples

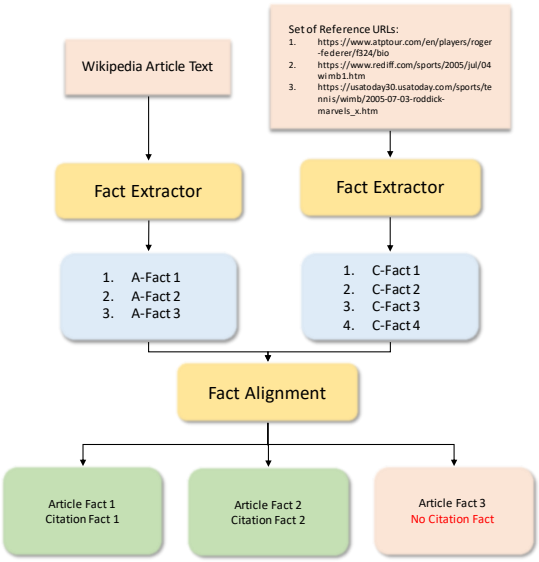

Figure 6.1 Pipeline for automated fact extraction and verification.

from our validation set and generated outputs for our entire test set in this way. The prompt used was:

```
You must extract all facts in English from the following {LANG}
sentence.  A fact consists of a relation and tail entity present
in the sentence.  Return the extracted facts in the form of a
list of lists.
```

Yet another approach is by extracting the possible entities from a sentence and then using classification methods to extract the relation. However, we did not try this since it has been shown to be inferior to the finetuning mT5 approach [50].

6.1.2 Fact Verification

Once we have two sets of facts, one from the article and one from the references, we need to check if the article fact is supported by its reference facts or if there isn't enough information to say so. We also wanted to check if there were contradictions between the article's fact and its references, but we found that in our dataset, there were hardly any instances of contradictions. Hence, we chose to stick with two classes only for the verification part.

- One initial approach we tried was looking for facts with common relation from both the article sentence and the references. We then match the tail to see if it supports the fact. If no fact with a

Model	Metric	bn	or	en	pa	ta	hi	avg
mT5-small	ROUGE-L	0.838	**0.711**	**0.768**	**0.692**	**0.842**	**0.854**	**0.784**
	BERTScore	0.890	**0.860**	**0.883**	**0.865**	**0.924**	**0.932**	**0.893**
GPT-4	ROUGE-L	**0.902**	0.600	0.656	0.601	0.766	0.596	0.687
	BERTScore	**0.954**	0.822	0.868	0.847	0.902	0.833	0.871

Table 6.1 Language wise fact extraction results.

	bn	or	en	pa	ta	hi	avg
Accuracy	66.59	**70.52**	61.90	60.39	66.43	57.76	63.93

Table 6.2 Language wise fact verification results.

matching relation is found in the reference, then we predict that the fact lacks citation. This is a heuristics-based approach and did not give us good results; hence, we did not proceed with it.

- Our best-performing fact verification approach utilizes passing each fact along with the corresponding reference sentences from a classifier. Since the reference text can be very long, we tokenize the reference text into sentences and then use LABSE [16] to find the top 10 semantically similar sentences from the reference to the article sentence. The results from this is shown in Table 6.2.

6.1.3 Implementation Details

For fact extraction using mT5, we use the mT5-small model having 8 encoder and 8 decoder layers. We use Adam optimizer with a learning rate of 2e-5 and train the model for 10 epochs with a batch size of 4. For fact verification, in particular, the threshold to determine semantic similarity was kept at 0.7.

6.2 Results

The results are summarized in Tables 6.1 and 6.2.

6.2.1 Fact Extraction

From Table 6.1, we observe that ROUGE-L and BERTScore correlate well, and thus, either metric can be used to find our best-performing model. Other than Bengali, mT5-small outperforms GPT-4 in all the languages, with the best results observed for Hindi. Thus, fine-tuning a much smaller model outperforms a SOTA model used with few shot prompting, even for English.

6.2.2 Fact Verification

From Table 6.2, we see that our proposed system achieves an average accuracy of 63.93%. **It can be observed that the system does not suffer from a language divide.** Even extremely low-resource languages like Odia and Punjabi perform very close to or higher than the average, while higher-resource languages like English perform worse than average.

6.3 Summary

In this chapter, we looked at an additional method of enhancing the quality of an encyclopedic article: by verifying that the facts listed in it are correct. We described a completely automated pipeline of cross-lingual fact extraction and verification. Our results highlighted the viability of this method for extremely low-resource languages.

Chapter 7

Conclusion

This thesis explored different aspects of critical problems, such as detecting and mitigating biased and toxic content and verifying facts in multilingual Wikipedia. We made all of the code used for the results presented here public. We contributed three datasets in Indian languages, MWNC, MWIKIBIAS, and XFACTVER, to verify our pipeline or those devised in the future. We list some challenges faced and key insights and takeaways from working on these problems below:

7.1 Insights from Dataset Creation

Getting data for all the aspects of the problems we wanted to study was challenging, particularly due to the current quality of articles and the general unstructured organisation of Indian language Wikipedias. We were forced to translate data from English Wikipedia for bias detection rather than rely directly on the Indian Wikipedias. Even for fact verification, we were unable to find much data for contradictions, i.e. when the reference article directly contradicted what was present in the Wikipedia page, and thus could not frame it as a more conventional NLI problem, as originally intended.

7.2 Insights from Bias Detection Experiments

We presented baseline results using standard Transformer based models for bias detection. When considering hard metrics for evaluation like F1 scores, more advanced methods like contrastive learning utilizing pairwise/triplet loss were surprisingly unhelpful for this task. Another key takeaway was the inability of the current popular ChatGPT API to evaluate a very subjective task like this in a multilingual setting.

We also explored the need for alternate ways to evaluate a model's performance on subjective tasks instead of just using F1 scores. When there is no clear answer that all annotators can agree with, it is crucial to consider the factor of their disagreement over just getting a single label for a data point. Our results highlight the benefits of using soft loss over hard loss for such controversial cases. We also find

that using better ways to combine multiple channels of information, which can potentially help us model the annotators and predict their choices, can lead to the best results.

However, the deep learning models of today are primarily encouraged to focus on hard evaluation scores like F1 and disregard the noise in the data, which leads to excellent results in constrained lab environments but fails in real-world scenarios. Finding more ways to incorporate the subjectivity of real-world data and people's opinions could help make these models more robust and generalizable.

7.3 Insights from Style Transfer and Reinforcement Learning Experiments

We presented baseline results using standard Transformer based models for style transfer as well as experimented with reinforcement learning based methods which used detection-based scores to enhance generation. A key challenge faced in this part came from the fact that the current Transformer based models struggle with identifying which part of the sentence to change when dealing with input and reference data that is highly similar. The key finding here was that neither full nor partial training using reinforcement learning outperformed our standard baselines trained with just cross-entropy loss. And none of the methods described fully alleviates the problem of models just copying input text directly as the output.

7.4 Insights from Fact Verification Experiments

We proposed the task of cross-lingual fact extraction and verification and contributed relevant baselines for the same. Instead of the conventional sentence-level approach, we did this at the fact level to get more accurate results for complex sentences present in Wikipedia textual data. Surprisingly, we find lower-resource Indian languages to perform comparably, or in a few cases, even better than English, all without relying on translation. Our results highlighted the viability of this method for extremely low-resource languages.

7.5 Future Work

We also outline some future work that could be undertaken in this area:

- For bias detection, we can explore more approaches like ensemble learning with our current models or even more domain-specific models like those used to detect promotional tone, puffery, political bias, and gender and racial bias to see how they compare with the more general NPOV detection.
- We can explore other ways of using reinforcement learning. One way could be to obtain manually annotated word lists for biased and unbiased words in each of our target languages and use that for

the embedding-based reward instead of automatically attempting to infer this information from the text.

- Further work on fact verification can utilize Set Transformers [31] to augment the mT5-based generation, which is particularly useful in our case since the extracted facts are permutation invariant.
- For fact verification, we can also try to construct knowledge graphs from the triples obtained from the article and reference sentences. Thus, this would then be the problem of comparing two knowledge graphs, for which various graph-based algorithms have been proposed, like CG-MuAlign [61].

Overall, this thesis has significantly contributed to enhancing the article quality of multilingual Wikipedia. However, there is still some scope to modify or extend the components described here, as outlined above, so there is scope for further research in this area.

Appendix A

Full Results and Some Other Applications

The appendix provides two main sets of information: the complete classification and style transfer results instead of the summarized data and a few NLP tasks undertaken that were not directly related to the main contribution of the thesis.

A.1 Complete Detailed Results of Classification and Style Transfer

Here, we give the tables for the classification and style transfer baselines. The tables A.1 to A.6 give the full language-wise results for our classification baselines, while tables A.7 and A.8 give the full language-wise results for our style transfer baselines.

Model	Metric	bn	gu	hi	kn	mr	ta	te	avg
MuRIL	**Accuracy**	0.598	0.596	0.603	0.596	0.584	0.595	0.577	0.593
	Precision	0.621	0.608	0.646	0.607	0.615	0.608	0.603	0.615
	Recall	0.598	0.596	0.603	0.596	0.584	0.595	0.577	0.593
	F1	0.577	0.583	0.571	0.585	0.554	0.582	0.550	0.572
	MCC	0.218	0.204	0.245	0.203	0.197	0.202	0.178	0.206
InfoXLM	**Accuracy**	0.592	0.588	0.610	0.593	0.588	0.595	0.583	0.593
	Precision	0.594	0.592	0.617	0.602	0.591	0.599	0.586	0.597
	Recall	0.592	0.588	0.610	0.593	0.588	0.595	0.583	0.593
	F1	0.590	0.584	0.603	0.583	0.585	0.592	0.580	0.588
	MCC	0.185	0.180	0.227	0.195	0.180	0.194	0.170	0.190
mDeBERTa	**Accuracy**	0.615	0.602	0.627	0.617	0.599	0.606	0.603	0.610
	Precision	0.619	0.607	0.636	0.624	0.605	0.614	0.609	0.616
	Recall	0.615	0.602	0.627	0.617	0.599	0.606	0.603	0.610
	F1	0.612	0.598	0.621	0.612	0.592	0.599	0.598	0.605
	MCC	0.234	0.209	0.263	0.241	0.203	0.220	0.212	0.226

Table A.1 Zero-shot results on MWIKIBIAS.

Model	Metric	bn	en	gu	hi	kn	mr	ta	te	avg
MuRIL	Accuracy	0.618	0.715	0.611	0.626	0.620	0.604	0.610	0.611	0.627
	Precision	0.643	0.738	0.637	0.650	0.645	0.629	0.633	0.637	0.652
	Recall	0.618	0.715	0.611	0.626	0.620	0.604	0.610	0.611	0.627
	F1	0.600	0.707	0.592	0.610	0.602	0.584	0.591	0.591	0.610
	MCC	0.260	0.452	0.247	0.275	0.263	0.231	0.241	0.246	0.277
InfoXLM	Accuracy	0.590	0.706	0.594	0.616	0.606	0.586	0.592	0.589	0.610
	Precision	0.603	0.718	0.611	0.627	0.613	0.593	0.594	0.605	0.621
	Recall	0.590	0.706	0.594	0.616	0.606	0.586	0.592	0.589	0.610
	F1	0.577	0.702	0.577	0.607	0.599	0.577	0.589	0.574	0.600
	MCC	0.193	0.424	0.204	0.242	0.219	0.179	0.185	0.193	0.230
mDeBERTa	Accuracy	0.644	0.743	0.634	0.646	0.635	0.624	0.631	0.628	0.648
	Precision	0.653	0.747	0.646	0.648	0.647	0.630	0.645	0.633	0.656
	Recall	0.644	0.743	0.634	0.646	0.635	0.624	0.631	0.628	0.648
	F1	0.639	0.742	0.625	0.645	0.627	0.620	0.622	0.624	0.643
	MCC	0.297	0.490	0.280	0.294	0.282	0.254	0.276	0.261	0.304

Table A.2 Monolingual classification results on MWIKIBIAS.

Model	Metric	bn	en	gu	hi	kn	mr	ta	te	avg
MuRIL	Accuracy	0.646	0.737	0.638	0.653	0.643	0.627	0.631	0.635	0.651
	Precision	0.656	0.747	0.649	0.664	0.656	0.640	0.645	0.649	0.663
	Recall	0.646	0.737	0.638	0.653	0.643	0.627	0.631	0.635	0.651
	F1	0.639	0.734	0.631	0.647	0.636	0.619	0.621	0.626	0.644
	MCC	0.302	0.484	0.287	0.317	0.299	0.267	0.275	0.283	0.314
InfoXLM	Accuracy	0.622	0.729	0.620	0.635	0.625	0.615	0.620	0.608	0.634
	Precision	0.634	0.737	0.633	0.646	0.635	0.625	0.632	0.622	0.645
	Recall	0.622	0.729	0.620	0.635	0.625	0.615	0.620	0.608	0.634
	F1	0.613	0.726	0.611	0.629	0.618	0.606	0.611	0.597	0.626
	MCC	0.255	0.466	0.253	0.281	0.260	0.240	0.252	0.229	0.279
mDeBERTa	Accuracy	0.646	0.742	0.639	0.650	0.640	0.623	0.638	0.633	0.651
	Precision	0.651	0.746	0.643	0.653	0.645	0.627	0.646	0.640	0.656
	Recall	0.646	0.742	0.639	0.650	0.640	0.623	0.638	0.633	0.651
	F1	0.643	0.741	0.636	0.648	0.636	0.619	0.633	0.629	0.648
	MCC	0.298	0.488	0.282	0.303	0.284	0.250	0.283	0.273	0.308

Table A.3 Multilingual classification results on MWIKIBIAS.

Model	Metric	bn	gu	hi	kn	mr	ta	te	avg
MuRIL	**Accuracy**	0.637	0.631	0.646	0.631	0.621	0.628	0.619	0.630
	Precision	0.643	0.634	0.652	0.639	0.640	0.636	0.636	0.640
	Recall	0.637	0.631	0.646	0.631	0.621	0.628	0.619	0.630
	F1	0.633	0.629	0.642	0.626	0.607	0.623	0.607	0.624
	MCC	0.280	0.265	0.298	0.270	0.260	0.264	0.254	0.270
InfoXLM	**Accuracy**	0.618	0.621	0.643	0.618	0.617	0.621	0.609	0.621
	Precision	0.623	0.626	0.648	0.630	0.621	0.632	0.617	0.628
	Recall	0.618	0.621	0.643	0.618	0.617	0.621	0.609	0.621
	F1	0.615	0.617	0.640	0.609	0.613	0.612	0.601	0.615
	MCC	0.242	0.246	0.291	0.248	0.237	0.252	0.226	0.249
mDeBERTa	**Accuracy**	0.637	0.627	0.645	0.632	0.617	0.630	0.625	0.630
	Precision	0.645	0.635	0.654	0.642	0.628	0.642	0.636	0.640
	Recall	0.637	0.627	0.645	0.632	0.617	0.630	0.625	0.630
	F1	0.632	0.621	0.639	0.625	0.609	0.621	0.617	0.623
	MCC	0.281	0.262	0.299	0.273	0.245	0.272	0.260	0.270

Table A.4 Zero-shot results on MWNC.

Model	Metric	bn	en	gu	hi	kn	mr	ta	te	avg
MuRIL	**Accuracy**	0.651	0.689	0.639	0.657	0.644	0.630	0.637	0.637	0.648
	Precision	0.662	0.700	0.660	0.665	0.652	0.645	0.642	0.650	0.660
	Recall	0.651	0.689	0.639	0.657	0.644	0.630	0.637	0.637	0.648
	F1	0.645	0.685	0.628	0.652	0.639	0.620	0.634	0.630	0.642
	MCC	0.313	0.389	0.298	0.322	0.296	0.275	0.279	0.287	0.307
InfoXLM	**Accuracy**	0.625	0.695	0.621	0.642	0.624	0.613	0.618	0.616	0.632
	Precision	0.630	0.697	0.626	0.650	0.628	0.620	0.624	0.621	0.637
	Recall	0.625	0.695	0.621	0.642	0.624	0.613	0.618	0.616	0.632
	F1	0.621	0.694	0.616	0.637	0.621	0.607	0.613	0.611	0.627
	MCC	0.254	0.392	0.247	0.291	0.252	0.234	0.242	0.237	0.269
mDeBERTa	**Accuracy**	0.662	0.735	0.657	0.670	0.658	0.642	0.652	0.651	0.666
	Precision	0.668	0.736	0.658	0.674	0.661	0.647	0.656	0.654	0.669
	Recall	0.662	0.735	0.657	0.670	0.658	0.642	0.652	0.651	0.666
	F1	0.659	0.734	0.657	0.669	0.657	0.639	0.649	0.649	0.664
	MCC	0.330	0.471	0.315	0.344	0.320	0.288	0.308	0.304	0.335

Table A.5 Monolingual classification results on MWNC.

Model	Metric	bn	en	gu	hi	kn	mr	ta	te	avg
MuRIL	**Accuracy**	0.668	0.711	0.660	0.671	0.664	0.649	0.657	0.658	0.667
	Precision	0.673	0.714	0.665	0.675	0.669	0.655	0.663	0.664	0.672
	Recall	0.668	0.711	0.660	0.671	0.664	0.649	0.657	0.658	0.667
	F1	0.665	0.710	0.658	0.670	0.661	0.646	0.654	0.654	0.665
	MCC	0.341	0.425	0.325	0.347	0.333	0.304	0.320	0.322	0.340
InfoXLM	**Accuracy**	0.650	0.716	0.643	0.663	0.648	0.637	0.644	0.639	0.655
	Precision	0.653	0.717	0.646	0.664	0.650	0.639	0.647	0.643	0.658
	Recall	0.650	0.716	0.643	0.663	0.648	0.637	0.644	0.639	0.655
	F1	0.648	0.715	0.641	0.662	0.646	0.635	0.642	0.637	0.653
	MCC	0.303	0.432	0.290	0.327	0.298	0.276	0.290	0.282	0.312
mDeBERTa	**Accuracy**	0.665	0.729	0.664	0.675	0.665	0.644	0.657	0.657	0.670
	Precision	0.665	0.729	0.665	0.675	0.666	0.646	0.658	0.659	0.670
	Recall	0.665	0.729	0.664	0.675	0.665	0.644	0.657	0.657	0.670
	F1	0.664	0.729	0.664	0.674	0.665	0.644	0.656	0.657	0.669
	MCC	0.330	0.458	0.329	0.349	0.332	0.290	0.315	0.316	0.340

Table A.6 Multilingual classification results on MWNC.

Strategy	Model	Metric	bn	en	gu	hi	kn	mr	ta	te	avg
Monolingual	IndicBART	BLEU	56.98	87.55	61.72	70.98	61.66	57.83	55.65	56.98	63.67
		METEOR	73.84	93.51	75.32	81.88	75.08	68.72	68.69	69.89	75.87
		chrF	78.09	92.35	78.47	81.86	81.42	74.83	76.97	76.33	80.04
		BERTScore	91.72	98.59	91.17	93.74	92.52	85.96	89.80	89.11	91.58
		Classifier accuracy	0.54	0.44	0.52	0.54	0.54	0.53	0.54	0.55	0.53
	mT0	BLEU	60.69	87.87	56.01	66.96	59.89	57.48	48.77	54.87	61.57
		METEOR	75.40	93.62	76.31	81.97	75.12	73.01	68.36	72.62	77.05
		chrF	79.62	92.51	79.10	81.76	81.47	77.89	75.94	78.44	80.84
		BERTScore	92.66	98.61	92.67	93.97	93.08	91.98	91.00	91.93	93.24
		Classifier accuracy	0.55	0.64	0.52	0.56	0.55	0.54	0.57	0.57	0.56
	mT5	BLEU	59.46	87.71	61.39	62.04	58.44	55.45	54.90	31.12	58.81
		METEOR	75.50	93.70	76.05	81.96	75.22	72.81	68.41	70.25	76.74
		chrF	79.57	92.59	79.10	81.31	81.36	77.52	76.16	74.20	80.23
		BERTScore	92.56	98.62	92.68	93.57	92.93	91.95	91.22	90.25	92.97
		Classifier accuracy	0.52	0.61	0.53	0.49	0.52	0.54	0.59	0.51	0.54
Multilingual	IndicBART	BLEU	35.32	83.34	47.10	56.58	39.85	34.98	16.54	56.88	46.32
		METEOR	62.32	91.28	66.36	73.85	58.47	57.31	36.58	70.82	64.62
		chrF	64.89	90.29	70.51	73.80	67.18	62.55	45.11	77.17	68.94
		BERTScore	86.88	98.18	89.56	90.78	87.83	86.13	76.77	91.61	88.47
		Classifier accuracy	0.55	0.46	0.55	0.54	0.55	0.55	0.57	0.58	0.54
	mT0	BLEU	59.91	83.90	57.87	68.40	55.91	58.10	51.38	51.39	60.86
		METEOR	75.45	91.95	75.90	82.61	75.25	73.21	69.43	72.48	77.04
		chrF	79.61	90.74	78.98	82.37	81.39	78.05	77.62	78.32	80.89
		BERTScore	92.67	97.96	92.67	94.05	93.04	92.05	91.32	91.87	93.20
		Classifier accuracy	0.57	0.62	0.56	0.56	0.56	0.57	0.57	0.57	0.57
	mT5	BLEU	60.60	86.02	61.35	69.36	60.84	58.19	53.03	56.72	63.26
		METEOR	75.82	92.68	76.39	82.87	75.63	73.25	69.70	72.97	77.41
		chrF	80.03	91.63	79.54	82.76	82.05	78.11	78.15	78.86	81.39
		BERTScore	92.81	98.30	92.85	94.16	93.28	92.12	91.57	92.11	93.40
		Classifier accuracy	0.56	0.65	0.57	0.56	0.58	0.58	0.59	0.58	0.58

Table A.7 Full style transfer baseline results on MWIKIBIAS.

Strategy	Model	Metric	bn	en	gu	hi	kn	mr	ta	te	avg
Monolingual	IndicBART	BLEU	52.35	82.24	51.10	63.01	53.52	50.37	39.13	48.12	54.98
		METEOR	65.04	87.70	68.79	76.39	66.75	64.41	59.98	64.92	69.25
		chrF	72.02	87.29	73.77	77.37	75.56	71.64	71.19	73.28	75.27
		BERTScore	89.84	97.71	90.33	91.89	90.85	89.47	88.05	89.81	90.99
		Classifier accuracy	0.51	0.56	0.45	0.47	0.51	0.49	0.52	0.49	0.50
	mT0	BLEU	54.08	82.81	50.87	62.37	54.30	50.66	36.56	33.06	53.09
		METEOR	67.70	88.53	69.24	76.39	67.33	64.68	61.43	64.75	70.01
		chrF	74.33	88.29	74.15	77.41	76.07	72.10	71.82	71.81	75.75
		BERTScore	90.50	97.85	90.59	92.07	91.12	89.63	89.10	89.27	91.27
		Classifier accuracy	0.52	0.58	0.49	0.51	0.55	0.53	0.52	0.52	0.53
	mT5	BLEU	54.60	82.64	50.69	57.91	55.09	50.76	45.75	45.64	55.39
		METEOR	67.74	88.93	68.90	76.38	68.05	64.93	61.88	65.42	70.28
		chrF	74.38	88.72	73.80	77.08	76.78	72.29	73.08	73.60	76.22
		BERTScore	90.53	97.88	90.47	91.81	91.23	89.65	89.28	90.03	91.36
		Classifier accuracy	0.52	0.55	0.47	0.48	0.51	0.51	0.49	0.52	0.51
Multilingual	IndicBART	BLEU	14.61	33.82	11.16	21.12	20.43	8.73	8.01	22.77	17.58
		METEOR	56.88	78.65	56.30	66.04	60.58	51.71	46.29	60.88	59.67
		chrF	58.47	75.42	55.25	62.86	67.16	52.18	50.93	66.93	61.15
		BERTScore	84.20	95.10	82.83	86.28	87.62	81.39	79.32	87.57	85.54
		Classifier accuracy	0.55	0.54	0.52	0.53	0.57	0.53	0.52	0.59	0.54
	mT0	BLEU	54.76	79.06	55.47	63.12	54.69	50.24	35.66	48.84	55.23
		METEOR	68.48	87.56	69.56	76.81	68.88	65.44	61.91	66.21	70.61
		chrF	75.10	87.38	74.73	77.85	77.55	72.65	72.51	74.55	76.54
		BERTScore	90.72	97.42	90.79	92.19	91.40	89.77	89.37	90.31	91.50
		Classifier accuracy	0.54	0.61	0.51	0.53	0.51	0.52	0.55	0.53	0.54
	mT5	BLEU	54.01	81.86	53.72	60.88	53.89	51.48	38.80	47.54	55.27
		METEOR	68.29	88.23	69.17	76.47	68.73	65.22	61.73	65.86	70.46
		chrF	74.85	88.05	74.26	77.38	77.44	72.62	72.62	74.07	76.41
		BERTScore	90.64	97.73	90.68	92.07	91.37	89.81	89.36	90.13	91.47
		Classifier accuracy	0.53	0.58	0.51	0.52	0.51	0.53	0.55	0.52	0.53

Table A.8 Full style transfer baseline results on MWNC.

A.2 Other NLP Tasks Using Transformers

This section explores different avenues to solve various challenges faced in problems involving text as input. These tasks were done during the course of this thesis as a part of various shared tasks. We will be discussing our work on three different problems:

- Explainable Detection of Online Sexism.
- Translation-Based Augmentation in Multilingual Tweet Analysis.
- Detecting Human Values Behind Arguments.

The tasks all use transformers to operate on textual data but vary significantly on how they are used to get optimal results.

A.2.1 Explainable Detection of Online Sexism

In Chapter 4, we saw different ways of detecting whether the text is biased in any way or is a neutral text that belongs in an encyclopedia. We used hard labels like F1 scores and later experimented with soft labels in which we applied softmax over the logits produced by the classifier. Both these tasks were binary classification tasks. In this section, we also show some experiments with hierarchical classification, where our focus lies on multi-level training techniques for bias detection, specifically targeting online misogyny and sexism.

We use the Explainable Detection of Online Sexism dataset for this task. This dataset was presented as part of the SemEval-2023 Task 10 [25] and sourced from social platforms like Reddit[1] and Gab[2]. The dataset presents three interconnected subtasks (designated A, B, C) that progressively classify sexist remarks. Task A involves a straightforward binary distinction between sexist and non-sexist posts. Moving to Task B, it classifies these identified sexist posts into four specific sexist categories. Lastly, Task C offers a detailed breakdown by categorizing these posts into 11 distinct forms or "detailed sexism vectors". We participated in all these subtasks and documented our outcomes.

To benefit from the structured relationship among these subtasks, creating a taxonomy for sexist content labeling, our primary focus lies on multi-level training techniques. This approach involves leveraging insights from a higher-level task to enhance performance on a lower-level task. We explored five distinct transformer architectures for this purpose. Additionally, we incorporated domain-adaptive pretraining as an initial step before multi-level training, tailoring the models to the specific nuances of this dataset. Given the skewed distribution of classes in Tasks B and C, we used the focal loss [35] method.

We specify the methodologies employed and present our findings, showcasing the efficacy of domain-adaptive pretraining for Task A, multi-level training for Task B, and focal loss for Task C. Tables A.9, A.10, and A.11 illustrate these results. The most promising outcomes, based on macro-F1 scores, were

[1] `https://www.reddit.com/`

[2] `https://gab.com/`

Strategy	Model	F1	Precision	Recall	Accuracy
Basic finetuning	BBU	81.62	81.97	82.47	86.80
	RB	83.22	83.94	82.58	87.90
	DB	83.69	83.69	83.69	88.00
	DBV3	82.51	82.69	82.32	87.20
	RL	85.73	**87.11**	84.56	89.85
Pretraining → finetuning	BBU	81.93	83.02	81.00	87.10
	RB	83.86	84.56	83.22	88.35
	DB	82.40	85.02	80.52	87.85
	DBV3	83.93	84.15	83.72	88.25
	RL	**85.93**	86.90	**85.09**	**89.90**

Table A.9 Performance of our models on the validation set of Task A.

derived from the RoBERTa-large pre-trained model, and we show these results using the validation split of the dataset.

A.2.2 Translation-Based Augmentation in Multilingual Tweet Analysis

Social media platforms have become prevalent channels for communication, attracting millions who produce and disseminate content regularly. Twitter, in particular, stands out as a favored medium for brief messages or tweets that convey a spectrum of emotions, viewpoints, and feelings[34] [11]. These tweets offer crucial insights into diverse societal and political matters, positioning them as indispensable tools for scholars and decision-makers alike.

The concept of intimacy in language denotes the extent of emotional closeness or familiarity shared among individuals, often evident in their word choices, tone, and contextual communication[57]. This level of closeness can fluctuate based on interpersonal relationships, profoundly influencing the outcomes of interpersonal exchanges.

Recently, delving into the analysis of intimacy within social media content has garnered attention.[43] Such scrutiny sheds light on facets of human conduct like network formation, information dissemination, and online community dynamics. Nevertheless, probing intimacy within social media texts poses challenges, demanding advanced techniques to handle and decode vast amounts of unstructured textual data. Assessing intimacy across multiple languages adds complexity due to linguistic intricacies and cross-cultural nuances.

The objective of the task [44] was to evaluate the intimacy of tweets across 10 distinct languages, presented as scores ranging from 1 to 5. We harness domain-specific attributes and pre-trained models tailored to discern intimacy nuances in tweets. Additionally, we employ a translation-centric data enhancement approach, significantly boosting scores for unfamiliar languages. This study also delves into various modifications we made to our approach, shedding light on each component's impact. As depicted in Figure A.1, our approach secured third position for uncharted languages, registering a Pearson's r

Strategy	Model	F1	Precision	Recall	Accuracy
Basic finetuning: CE	BBU	58.91	62.39	56.68	62.14
	RB	63.10	64.84	63.85	65.23
	DB	68.33	67.07	69.90	69.34
	DBV3	61.43	61.71	62.13	63.79
	RL	69.41	69.19	69.71	70.99
Basic finetuning: FOCAL	BBU	58.60	60.43	57.21	62.14
	RB	64.02	65.36	62.90	64.81
	DB	67.79	67.53	68.21	69.14
	DBV3	62.72	62.03	63.86	62.55
	RL	69.67	69.11	71.47	**71.60**
Pretraining → finetuning: FOCAL	BBU	60.91	59.97	62.00	61.32
	RB	57.30	57.17	57.49	57.61
	DB	62.13	60.33	64.80	62.55
	DBV3	56.98	56.16	58.69	56.17
	RL	69.57	67.24	**73.39**	69.55
Pretraining → Task A → finetuning: FOCAL	BBU	66.33	65.94	67.28	67.28
	RB	62.87	64.92	61.36	62.96
	DB	67.05	65.60	68.87	68.93
	DBV3	62.53	64.19	61.21	64.20
	RL	**69.96**	**69.24**	70.98	70.99

Table A.10 Performance of our models on the validation set of Task B.

Strategy	Model	F1	Precision	Recall	Accuracy
Basic finetuning: CE	BBU	33.07	33.51	34.43	54.12
	RB	38.48	37.91	39.57	55.76
	DB	44.90	54.68	43.10	56.58
	DBV3	34.97	35.68	36.09	55.14
	RL	50.35	52.34	49.31	61.32
Basic finetuning: FOCAL	BBU	36.22	34.97	39.27	47.33
	RB	43.05	42.48	44.34	50.21
	DB	49.79	51.25	49.17	55.35
	DBV3	41.23	40.28	43.95	49.59
	RL	**54.62**	**56.33**	**53.79**	**61.32**
Pretraining → finetuning: FOCAL	BBU	37.32	41.34	38.19	47.33
	RB	39.83	38.87	42.07	47.33
	DB	51.16	51.90	52.19	58.23
	DBV3	41.38	40.21	43.67	50.62
	RL	51.67	55.95	50.97	60.08
Pretraining → Task A → Task B → finetuning: FOCAL	BBU	40.68	48.61	41.04	51.44
	RB	43.15	44.22	43.00	53.50
	DB	49.97	51.04	49.77	57.41
	DBV3	39.68	42.80	39.64	51.23
	RL	51.15	51.98	50.80	59.67

Table A.11 Performance of our models on the validation set of Task C.

score of 0.485, and ranked tenth overall with a Pearson's r score of 0.592. Comprehensive outcomes are presented in Table A.12.

Ablation		pretrained model				filtering	emojis		translation		others
	submitted system	distill-bert	mbert	TwHIN-bert	xlm-T	no cleaning	no emoji	emoji-2-text	no trans	trans test	no trans no emoji
English	0.706	0.602	0.636	0.688	0.706	0.704	**0.723**	0.706	0.704	0.702	0.715
Spanish	0.725	0.604	0.622	**0.727**	0.711	0.720	0.705	0.709	0.694	0.680	0.678
Portuguese	0.648	0.514	0.545	0.606	**0.676**	0.671	0.674	0.668	0.645	0.652	0.645
Italian	**0.727**	0.590	0.558	0.710	0.698	0.694	0.690	0.692	0.694	0.709	0.695
French	0.628	0.559	0.580	0.631	0.675	0.681	0.674	0.674	0.681	0.680	**0.692**
Chinese	0.698	0.666	0.664	0.721	0.714	0.717	0.720	**0.729**	0.720	0.677	0.708
Hindi	0.203	0.176	0.174	0.189	**0.217**	0.160	0.184	0.184	0.200	0.235	0.206
Dutch	0.591	0.488	0.487	0.567	**0.630**	0.608	**0.611**	0.603	0.602	0.604	0.592
Korean	0.307	0.277	0.269	**0.404**	0.358	0.306	0.372	0.359	0.322	0.319	0.374
Arabic	**0.644**	0.395	0.365	0.637	0.605	0.572	0.647	0.623	0.653	0.604	0.628
Seen Lang	0.684	0.591	0.612	0.687	0.704	**0.707**	0.699	0.702	0.694	0.684	0.693
Unseen Lang	0.485	0.410	0.384	**0.516**	0.477	0.471	0.484	0.477	0.434	0.367	0.420
Overall	0.592	0.512	0.510	**0.605**	0.602	0.601	0.601	0.600	0.573	0.535	0.570

Table A.12 Results from our submitted system and related ablations.

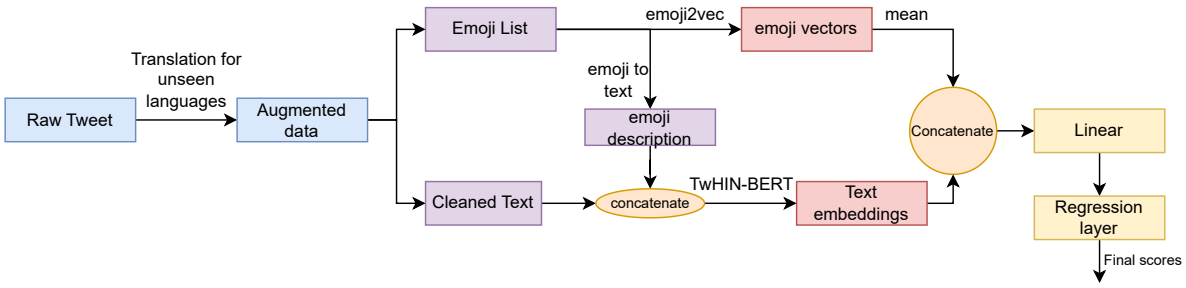

Figure A.1 Pipeline for the proposed system.

A.2.3 Detecting Human Values Behind Arguments

Humans frequently arrive at varied conclusions even when presented with the same premise. This divergence often stems from their underlying values. Identifying the values behind the arguments helps understand the argument itself. Downstream tasks, such as generating arguments in favor or against, can gain from such value discernment. In this study[24], our objective is to pinpoint 20 distinct value categories within a specified premise, stance, and conclusion triplet illustrated in figure A.2. The dataset is sourced from four distinct cultural regions and annotated by [37].

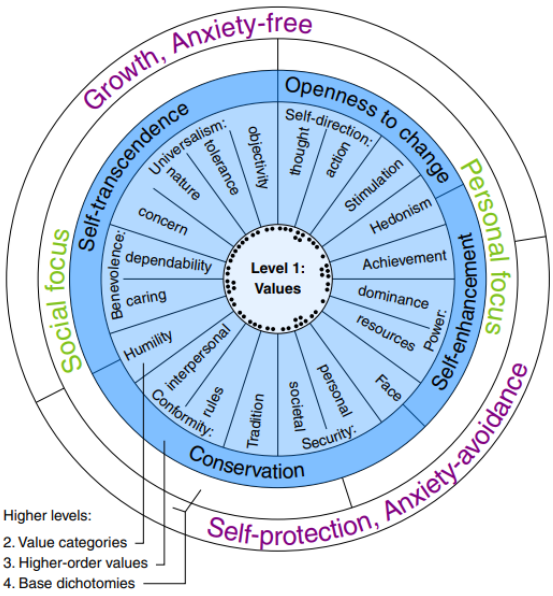

Figure A.2 Values in the data are organized from higher level to lower level.

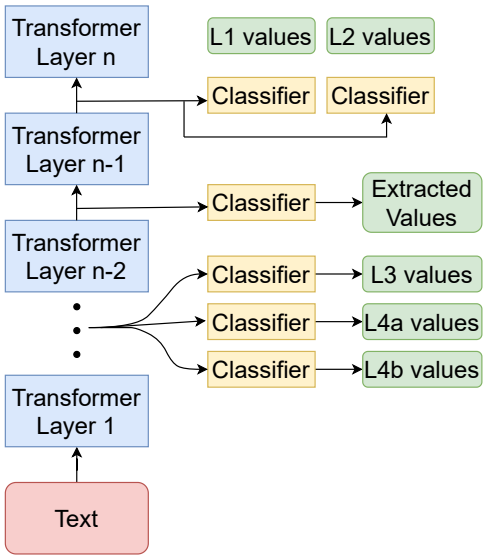

Figure A.3 Using internal hidden states to feed classifiers to exploit the hierarchy in values.

We propose a technique for multi-label classification of premise, stance, and conclusion pairs utilizing encoder-only LMs as the foundational model. We employ DeBERTa[20], a pre-trained language model known for its proficiency across numerous NLP applications, especially classification tasks. Our method involves tokenizing the premise, stance, and conclusion texts using the pre-established tokenizer. These segments are then merged and inputted into the LM, producing a combined text representation. Subsequently, a neural network maps this representation to a predefined set of values. Training leverages a multi-margin loss function, while evaluation metrics are accuracy, precision, recall, and F1 score. This approach can potentially result in a highly accurate and effective NLP model for identifying values in arguments.

A key insight from our research underscores the significance of value hierarchy in discerning them, as depicted in figure A.3. Incorporating five high-level values — Self-direction, Power, Security, Conformity, Benevolence, and Universalism — enhanced the model's efficacy compared to solely relying on the 20 values. Notably, these five values are derived from the broader set of 20. To illustrate, the value Self-direction: action aligns with the broader Self-direction class. Our model surpassed the benchmarks set by the 1-Baseline and Random Baseline in [18] and performed on par with the BERT Baseline.

Related Publications

- **Ankita Maity**, Anubhav Sharma, Rudra Dhar, Tushar Abhishek, Manish Gupta and Vasudeva Varma, "Multilingual Bias Detection and Mitigation for Indian Languages", **WILDRE Workshop at LREC-COLING 2024**.
- Anubhav Sharma*, **Ankita Maity***, Tushar Abhishek, Rudra Dhar, Radhika Mamidi, Manish Gupta and Vasudeva Varma, "Multilingual Bias Detection and Mitigation for Low Resource Languages", **Wiki Workshop 2023**.
- Shivansh Subramanian*, **Ankita Maity***, Aakash Jain*, Bhavyajeet Singh*, Harshit Gupta*, Lakshya Khanna* and Vasudeva Varma, "Cross-Lingual Fact Checking: Automated Extraction and Verification of Information from Wikipedia using References", **ICON 2023 Main Conference**.
- **Ankita Maity**, Pavan Kandru, Bhavyajeet Singh, Kancharla Aditya Hari and Vasudeva Varma, "IREL at SemEval-2023 Task 11: User Conditioned Modelling for Toxicity Detection in Subjective Tasks", **SemEval Workshop 2023**.

Other Publications

- Bhavyajeet Singh, **Ankita Maity**, Pavan Kandru, Kancharla Aditya Hari and Vasudeva Varma, "iREL at SemEval-2023 Task 9: Improving understanding of multilingual Tweets using Translation-Based Augmentation and Domain Adapted Pre-Trained Models", **SemEval Workshop 2023**.
- Pavan Kandru, Bhavyajeet Singh, **Ankita Maity**, Aditya Hari and Vasudeva Varma, "Tenzin-Gyatso at SemEval-2023 Task 4: Identifying Human Values behind Arguments using DeBERTa", **SemEval Workshop 2023**.
- Nirmal Manoj, Sagar Joshi, **Ankita Maity** and Vasudeva Varma, "iREL at SemEval-2023 Task 10: Multi-level Training for Explainable Detection of Online Sexism", **SemEval Workshop 2023**.

*Equal contributions.

Bibliography

[1] T. Abhishek, S. Sagare, B. Singh, A. Sharma, M. Gupta, and V. Varma. Xalign: Cross-lingual fact-to-text alignment and generation for low-resource languages. In *Companion Proceedings of the Web Conference 2022*, WWW '22, page 171–175, New York, NY, USA, 2022. Association for Computing Machinery.

[2] S. Akhtar, V. Basile, and V. Patti. Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection, 2021.

[3] D. Aleksandrova, F. Lareau, and P. A. Ménard. Multilingual sentence-level bias detection in wikipedia. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 42–51, 2019.

[4] D. Almanea and M. Poesio. ArMIS - the Arabic misogyny and sexism corpus with annotator subjective disagreements. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2282–2291, Marseille, France, June 2022. European Language Resources Association.

[5] W. Antoun, F. Baly, and H. Hajj. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France, May 2020. European Language Resource Association.

[6] M. Anzovino, E. Fersini, and P. Rosso. Automatic identification and classification of misogynistic language on twitter. In *Natural Language Processing and Information Systems*, pages 57–64. Springer International Publishing, 2018.

[7] A. Bertsch and S. Bethard. Detection of puffery on the english wikipedia. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 329–333, 2021.

[8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[9] M. Chen, W. Zhang, Y. Zhu, H. Zhou, Z. Yuan, C. Xu, and H. Chen. Meta-knowledge transfer for inductive knowledge graph embedding. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 927–937, New York, NY, USA, 2022. Association for Computing Machinery.

[10] Z. Chi, L. Dong, F. Wei, N. Yang, S. Singhal, W. Wang, X. Song, X.-L. Mao, H.-Y. Huang, and M. Zhou. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, 2021.

[11] G. Cislaru. Emotions in tweets: From instantaneity to preconstruction. *Social Science Information*, 54(4):455–469, 2015.

[12] R. Dabre, H. Shrotriya, A. Kunchukuttan, R. Puduppully, M. M. Khapra, and P. Kumar. Indicbart: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863, 2022.

[13] D. Dale, A. Voronov, D. Dementieva, V. Logacheva, O. Kozlova, N. Semenov, and A. Panchenko. Text detoxification using large pre-trained neural models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7979–7996, 2021.

[14] C. De Kock and A. Vlachos. Leveraging wikipedia article evolution for promotional tone detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5601–5613, 2022.

[15] L. Fan, M. White, E. Sharma, R. Su, P. K. Choubey, R. Huang, and L. Wang. 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, 2019.

[16] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang. Language-agnostic BERT sentence embedding. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[17] A. Field, C. Y. Park, K. Z. Lin, and Y. Tsvetkov. Controlled analyses of social biases in wikipedia bios. In *Proceedings of the ACM Web Conference 2022*, pages 2624–2635, 2022.

[18] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, and M. Potthast. Continuous Integration for Reproducible Shared Tasks with TIRA.io. In *Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023)*, Lecture Notes in Computer Science, Berlin Heidelberg New York, Apr. 2023. Springer.

[19] P. He, J. Gao, and W. Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*, 2022.

[20] P. He, X. Liu, J. Gao, and W. Chen. {DEBERTA}: {DECODING}-{enhanced} {bert} {with} {disentangled} {attention}. In *International Conference on Learning Representations*, 2021.

[21] K.-H. Huang, C. Zhai, and H. Ji. CONCRETE: Improving cross-lingual fact-checking with cross-lingual retrieval. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1024–1035, Gyeongju, Republic of Korea, Oct. 2022. International Committee on Computational Linguistics.

[22] C. Hube and B. Fetahu. Detecting biased statements in wikipedia. In *Companion proceedings of the the web conference 2018*, pages 1779–1786, 2018.

[23] S. Khanuja, D. Bansal, S. Mehtani, S. Khosla, A. Dey, B. Gopalan, D. K. Margam, P. Aggarwal, R. T. Nagipogu, S. Dave, et al. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*, 2021.

[24] J. Kiesel, M. Alshomary, N. Mirzakhmedova, M. Heinrich, N. Handke, H. Wachsmuth, and B. Stein. Semeval-2023 task 4: Valueeval: Identification of human values behind arguments. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada, July 2023. Association for Computational Linguistics.

[25] H. R. Kirk, W. Yin, B. Vidgen, and P. Röttger. SemEval-2023 Task 10: Explainable Detection of Online Sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada, July 2023. Association for Computational Linguistics.

[26] K. Kolluru, M. Mohammed, S. Mittal, S. Chakrabarti, and M. . Alignment-augmented consistent translation for multilingual open information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2502–2517, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[27] A. Krishna, S. Riedel, and A. Vlachos. ProoFVer: Natural logic theorem proving for fact verification. *Transactions of the Association for Computational Linguistics*, 10:1013–1030, 2022.

[28] K. Krishna, J. Wieting, and M. Iyyer. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762, 2020.

[29] H. Lai, A. Toral, and M. Nissim. Thank you bart! rewarding pre-trained models improves formality style transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 484–494, 2021.

[30] H. Lai, A. Toral, and M. Nissim. Multilingual pre-training with language and task adaptation for multilingual text style transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 262–271, 2022.

[31] J. Lee, Y. Lee, J. Kim, A. Kosiorek, S. Choi, and Y. W. Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3744–3753. PMLR, 09–15 Jun 2019.

[32] E. Leonardelli, G. Abercrombie, D. Almanea, V. Basile, T. Fornaciari, B. Plank, V. Rieser, A. Uma, and M. Poesio. SemEval-2023 task 11: Learning with disagreements (LeWiDi). In A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, and E. Sartori, editors, *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318, Toronto, Canada, July 2023. Association for Computational Linguistics.

[33] E. Leonardelli, S. Menini, A. Palmero Aprosio, M. Guerini, and S. Tonelli. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics.

[34] J. S. Y. Liew and H. R. Turtle. Exploring fine-grained emotion detection in tweets. In *Proceedings of the NAACL Student Research Workshop*, pages 73–80, San Diego, California, June 2016. Association for Computational Linguistics.

[35] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[36] R. Liu, C. Jia, J. Wei, G. Xu, L. Wang, and S. Vosoughi. Mitigating political bias in language models through reinforced calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14857–14866, 2021.

[37] N. Mirzakhmedova, J. Kiesel, M. Alshomary, M. Heinrich, N. Handke, X. Cai, B. Valentin, D. Dastgheib, O. Ghahroodi, M. Sadraei, E. Asgari, L. Kawaletz, H. Wachsmuth, and B. Stein. The Touché23-ValueEval Dataset for Identifying Human Values behind Arguments. *CoRR*, abs/2301.13771, 2023.

[38] I. Mondal, Y. Hou, and C. Jochim. End-to-end construction of NLP knowledge graph. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1885–1895, Online, Aug. 2021. Association for Computational Linguistics.

[39] N. Muennighoff, T. Wang, L. Sutawika, A. Roberts, S. Biderman, T. Le Scao, M. S. Bari, S. Shen, Z. X. Yong, H. Schoelkopf, X. Tang, D. Radev, A. F. Aji, K. Almubarak, S. Albanie, Z. Alyafeai, A. Webson, E. Raff, and C. Raffel. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, 2023.

[40] A. Nadgeri, A. Bastos, K. Singh, I. O. Mulang', J. Hoffart, S. Shekarpour, and V. Saraswat. KGPool: Dynamic knowledge graph context selection for relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 535–548, Online, Aug. 2021. Association for Computational Linguistics.

[41] D. Q. Nguyen, T. Vu, and A. Tuan Nguyen. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online, Oct. 2020. Association for Computational Linguistics.

[42] L. Pan, W. Chen, W. Xiong, M.-Y. Kan, and W. Y. Wang. Zero-shot fact verification by claim generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 476–483, Online, Aug. 2021. Association for Computational Linguistics.

[43] J. Pei and D. Jurgens. Quantifying intimacy in language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5307–5326, Online, Nov. 2020. Association for Computational Linguistics.

[44] J. Pei, V. Silva, M. Bos, Y. Liu, L. Neves, D. Jurgens, and F. Barbieri. Semeval 2023 task 9: Multilingual tweet intimacy analysis, 2022.

[45] J. Peterson, R. Battleday, T. Griffiths, and O. Russakovsky. Human uncertainty makes classification more robust. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9616–9625, 2019.

[46] R. Pryzant, R. D. Martinez, N. Dass, S. Kurohashi, D. Jurafsky, and D. Yang. Automatically neutralizing subjective bias in text. In *Proceedings of the aaai conference on artificial intelligence*, volume 34, pages 480–489, 2020.

[47] G. Ramesh, S. Doddapaneni, A. Bheemaraj, M. Jobanputra, A. Raghavan, A. Sharma, S. Sahoo, H. Diddee, J. Mahalakshmi, D. Kakwani, et al. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162, 2022.

[48] M. Recasens, C. Danescu-Niculescu-Mizil, and D. Jurafsky. 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, 2013.

[49] N. Reimers and I. Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, 2020.

[50] B. Singh, S. V. P. K. Kandru, A. Sharma, and V. Varma. Massively multilingual language models for cross lingual fact extraction from low resource Indian languages. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pages 11–18, New Delhi, India, Dec. 2022. Association for Computational Linguistics.

[51] S. Subramanian and K. Lee. Hierarchical Evidence Set Modeling for automated fact extraction and verification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7798–7809, Online, Nov. 2020. Association for Computational Linguistics.

[52] D. Taunk, S. Sagare, A. Patil, S. Subramanian, M. Gupta, and V. Varma. Xwikigen: Cross-lingual summarization for encyclopedic text generation in low resource languages. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 1703–1713, New York, NY, USA, 2023. Association for Computing Machinery.

[53] M. Trokhymovych, M. Aslam, A.-J. Chou, R. Baeza-Yates, and D. Saez-Trumper. Fair multilingual vandalism detection system for wikipedia. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, page 4981–4990, 2023.

[54] A. Uma, T. Fornaciari, A. Dumitrache, T. Miller, J. Chamberlain, B. Plank, E. Simpson, and M. Poesio. SemEval-2021 task 12: Learning with disagreements. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 338–347, Online, Aug. 2021. Association for Computational Linguistics.

[55] A. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, and M. Poesio. A case for soft loss functions. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8:173–177, Oct. 2020.

[56] A. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, and M. Poesio. Learning from disagreement: A survey. *J. Artif. Intell. Res.*, 72:1385–1470, 2021.

[57] L. C. Wynne and A. R. Wynne. The quest for intimacy*. *Journal of Marital and Family Therapy*, 12(4):383–394, 1986.

[58] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, 2021.

[59] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

[60] Y. Zhong, J. Yang, W. Xu, and D. Yang. Wikibias: Detecting multi-span subjective biases in language. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1799–1814, 2021.

[61] Q. Zhu, H. Wei, B. Sisman, D. Zheng, C. Faloutsos, X. L. Dong, and J. Han. Collective multi-type entity alignment between knowledge graphs. In *Proceedings of The Web Conference 2020*, WWW '20, page 2241–2252, New York, NY, USA, 2020. Association for Computing Machinery.