Universiteit Leiden
The Netherlands

Bachelor DSAI

Mutually Intelligible, Separate Outcomes:

Linguistic Bias Between Hindi and Urdu in Sentiment Classification

Raaed Naveed Afzal Khan
s3624838

Supervisors:

Dr. Flor Miriam Plaza del Arco & Dr. Natalia Amat Lefort

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

Abstract

The growing implementations of AI systems across various degrees of societies provide many avenues for growth. The implementation of these systems often comes hand in hand with the inclusion of various societal biases. Studies have shown varying degrees of linguistic bias in English; degraded performance has been shown in particular for low-resource registers. This thesis aims to study the performance of encoder-based models in a sentiment analysis polarity classification task between Hindi and Urdu - two distinct, yet strongly similar and mutually intelligible languages in order to screen for the presence of linguistic bias.

This is researched through analysing the difference in classification performance on a dataset of 6,999 samples of Urdu and Hindi. Models are fine-tuned for the classification task, and then performance metrics are compared across languages, and the difference is compared across models. These metrics are paired with a suite of bias measurements contained in the AI Fairness 360 toolkit. The Average odds difference, disparate impact ratio, demographic parity ratio and statistical parity difference scores are used to determine whether the differences in performance constitute linguistic bias.

This thesis finds there to be a substantial difference in model performance across language groups, with the language-specific models exhibiting larger gaps than the multilingual baseline. According to the fairness framework and the error analysis, these differences are consistent with the presence of linguistic bias in the models' treatment of the two languages. However, because the Urdu and Hindi corpora differ in domain (reviews from news sites versus product and media reviews) and in romanisation method, the observed gaps cannot be fully disentangled from domain and romanisation approaches - the error analysis also indicates that all three factors contribute. The findings therefore indicate that speakers of one language group may receive differential treatment in sentiment analysis tasks, signalling a risk of allocational harms, while also demonstrating the need for parallel, domain-controlled test datasets to confirm the extent of the bias present without domain confound.

Contents

1 Introduction **1**
- 1.1 Research Questions . . . 2
- 1.2 Thesis Overview . . . 2

2 Background & Related Work **2**
- 2.1 Sentiment Analysis . . . 2
- 2.2 Bias in NLP . . . 3
  - 2.2.1 Multilingual Bias . . . 5
- 2.3 Hindi vs Urdu - "Hindustani" . . . 7
  - 2.3.1 Romanisation . . . 9
  - 2.3.2 Linguistic Imperialism . . . 9

3 Methodology **11**
- 3.1 Data . . . 11
  - 3.1.1 Collection . . . 11
  - 3.1.2 Processing . . . 12
  - 3.1.3 Description . . . 13
- 3.2 Models . . . 14
  - 3.2.1 BERT_Roman_Urdu . . . 14
  - 3.2.2 MuRIL . . . 15
  - 3.2.3 mBERT . . . 15
- 3.3 Fairness . . . 16

4 Results **18**
- 4.1 Classification Performance . . . 18
- 4.2 Error Analysis . . . 22
- 4.3 Bias . . . 28

5 Discussion **30**
- 5.1 Interpretations . . . 30
- 5.2 Limitations . . . 32
- 5.3 Further Research . . . 33
- 5.4 Conclusions . . . 34

References **39**

List of Tables

1 Example Sentences in Urdu . . . 11
2 Example Sentences in Hindi . . . 12
3 Fine-tuning Hyperparameters for BERT_Roman_Urdu . . . 14
4 Fine-tuning Hyperparameters for MuRIL . . . 15
5 Fine-tuning Hyperparameters for mBERT . . . 16

6 * 18
7 Prediction Distribution by Input Language Statistics 21
8 Misclassifications of BERT_Roman_Urdu in Urdu 22
9 Sample misclassifications of BERT_Roman_Urdu in Hindi 23
10 Sample misclassifications of mBERT in Urdu 24
11 Sample misclassifications of mBERT in Hindi 25
12 Sample misclassifications of MuRIL in Urdu 26
13 Sample misclassifications of MuRIL in Hindi 27
14 * 28

List of Figures

1 Comparison of Devanagari and Nastaliq scripts with Romanisation alongside 9
2 Distribution of Polarities across Languages 13
3 Comparison of Accuracy by Model and Language 19
4 Comparison of Macro F1 by Model and Language 19
5 Confusion Matrices by Model and Language 20
6 Prediction Rates Across Model by Language 21

1 Introduction

Recently, we have seen the rise in use of Natural Language Processing (NLP) systems to a capacity where they are now entrenched across society. These systems allow namely for the cognition of human speech and text by computer intelligence systems. A subset of these tasks are those of sentiment analysis based natures. First introduced in literature by Pang and Lee in 2002 [1], sentiment analysis refers to the computational analysis of people's moods, opinions, feelings or broader sentimental polarity through language processing. Today, sentiment analysis systems have become ubiquitous in informing decisions across multiple various fields; examples have been found across healthcare [2], marketing [3], governance [4] and innumerable other domains.

Despite the benefits of the increased technological opportunities afforded by the increase in adaptations of such systems, they are implemented at jeopardy of exacerbating societal biases [5]. The ubiquity of their uses adds to the severity of the issue, and research into the matter shows that there are a variety of different factors underlining the bias problem pernicious to current models. The studies of Bolukbasi et al. and Baste et al. have shown that models encode societal biases into their training data [6] [7], and the study of Kiritchenko and Mohammad shows that sentiment analysis systems embody these biases as well - for these to propagate further downstream represents an opportunity for an increase in systemic inequality, an outcome that responsible AI implementations should aim to prevent.

The forms of biases found in current applications are truly innumerable. Haozhe et al. tasked Large Language Models (LLMs) with selecting candidates to hire varying only between masculine and feminine variants of names pertaining to European American, African American and Hispanic American identities, and found them to make biased judgements [8]; the study of Curto et al. into pre-trained word embeddings has shown that current AI systems corroborate descriptions of poverty with negative emotional descriptions [9], a discriminatory pattern of judgement corroborated by Curry et al., who found that NLP performance degraded proportionally to socio-linguistic register pertinent to class [10]. Studies of several popular LLMs have shown in fact that they are unable to properly comprehend non-standard dialects of English, and when recreating them often respond in ways that consist of stereotyping and demeaning responses [11].

In multilingual contexts, the waters become murkier. The majority of research in computational linguistics as a whole, as well as into sentiment analysis systems is conducted upon models trained primarily for English, or for use in primarily English speaking settings [12]. It should be a common goal that the implementation of AI systems should be carried out in such a manner that is not discriminative of users' linguistic background. Here, this is taken to mean that the experience said user derives should not be degraded as a result of the language they are speaking. Unfortunately, this is not at all the case. Research has shown that performance degrades in multilingual scenarios [13] [11]. This essentially results in a scenario where users must either suffer degraded performance in their own language, or shift to the dominant language. In either scenario, this represents linguistic imperialism, [14] carried out through the incapacity of these systems.

For speakers of under-resourced languages, this represents the dual burden of the simple unfairness of lacking proper resources in their language of choice compounded with a dearth of research that may shed light on potential issues underlying its implementations, particularly those that may result in potential allocational harms such as biased judgements. This thesis aims to study the presence of linguistic bias in the treatment of two languages that share a colloquial register similar enough to the standard of near total mutual intelligibility. Representing a speech community of

c. 700 million people, Hindi and Urdu are the lingua francae of large swathes, as well as official languages of India and Pakistan respectively. Despite this, the aforementioned relative absence of research into other languages continues into Hindi and Urdu, particularly in cross-lingual contexts. Therefore, this thesis contributes a novel assessment of the presence of linguistic bias potentially experienced by speakers of either language.

This thesis will therefore attempt to ascertain the existence of any bias in the way encoder-based models handle Hindi and Urdu in a sentiment analysis task. This research hopes to capture the differences in performance, and ascertain whether these indicate the presence of linguistic bias. This thesis thus also seeks to inform future work on the nature of bias in NLP, particularly in languages with less available resources and literature. In order to achieve this objective, this thesis will aim to answer the following research questions.

1.1 Research Questions

- RQ1: Does sentiment classification performance vary between Hindi and Urdu across encoder-based models?
- RQ2: Do these differences in treatment across model between languages represent linguistic bias?

1.2 Thesis Overview

The following sections of this thesis are structured according to the following outline: In Chapter 2, an overview of the background relating to the research, and some key related work in this field is given, as well as the definition of bias used in this experiment; In Chapter 3, the methodology utilised to tackle this issue is outlined, an explanation of the experimental variables captured to answer the research question is given, as well as an explanation of the measurement of bias; Chapter 4 contains the results of the experiment conducted, as well as an explanation of the misclassifications pertinent to this study, and the results of the fairness evaluation; and Chapter 5 concludes with addressing the implications of the results given in the preceding chapters, an overview of a number of possible limitations to the findings of this thesis, and a collection of avenues for extension to this research.

2 Background & Related Work

This thesis aims to investigate the biases present in Hindi-Urdu NLP models through considering their performance in sentiment analysis, particularly in multilingual contexts. In this section, background of the task of sentiment analysis is given, alongside a selection of related works on the topic of bias in NLP, as well as the definition of bias used for this study. This section concludes with a background of the languages being studied, as well as an explanation of the choice of romanisation.

2.1 Sentiment Analysis

Sentiment analysis is a field of NLP which refers to the computation of people's opinions, moods, emotions and sentiments through the analysis of speech and text. Early work in the field included de-

termining the polarity of a sentence (positive, negative or neutral) [1], where later developments have resulted in affording us a greater computational understanding of more complex communicational patterns [15] [16] [17].

The growth in the diverse applications of NLP systems is also visible in those of sentiment analysis systems. Studies into the use of sentiment analysis in healthcare found that patient reviews and social media posts could be analysed to monitor for signs of adverse reactions to medication, or in helping hospitals identify opportunities to improve their services, or as support in the creation of personalised health plans, etc. [2]. Sentiment analysis systems in the business world find usage in predicting and understanding the market, be it by predicting which deal to market, or by helping understand the trends at the moment [3]. Approaches utilising sentiment analysis have also been used by governments in automating their processing of the feedback given by civilians to state services [4]. It is here that the relevance of the task should become evident, amidst various applications of these systems, the ability of prejudiced decision-making to cause harmful downstream impacts grows, thereby necessitating the study of any factors that may result in biased decision making.

Sentiment analysis systems are unfortunately no strangers to the same biases we see in everyday society. Terminology differentiating identity groups has shown to adversely impact classification decisions, [18] [19]. Studies on word embeddings have demonstrated that they encode biases across gender and identity axes [6] [20], [7], but these studies tackle a computational method that is faster falling out of mainstream usage. Research into successive methodologies such as contextualised word embeddings have revealed however that the problems seem not to be assuaged, with differing degrees of societal biases still being encoded [21] [7] [22].

This is particularly concerning when considering that most of the developments across sentiment analysis have been in English, and current approaches to the problem in multilingual contexts have been known to involve translating source material into English, or using English language tools such as SentiWordNet in order to evaluate content in other languages [23]. Studies have shown that the degrees of inequalities found in English sentiment analysis systems translate with a greater extent when compared in other languages, raising questions on the feasibility of such an approach [24]. For methodologies that take a purely native approach to dealing with this problem, the toll of limited resource availability and computational bottlenecks for developing populations provide serious structural limitations [25]. The frontiers of multilingual sentiment analysis are vast, and the presence of bias within them remains near entirely unmapped.

2.2 Bias in NLP

Bias used here is a disproportionate inclination or preference towards one group of people, ideology or language inter alia. This study investigates the extent to which two closely related languages are treated differently in sentiment analysis systems. In order to measure this, a variant of the definition of bias utilised by Jentzsch and Turan is used [19]. For the purposes of their study, they defined bias through terms of harm, where they "consider it harmful if a classifier that is trained to distinguish positive and negative movie reviews prefers performers and film characters of one gender over another". Therefore, a model in the context of this research is said to display bias if it treats one language with preferential treatment to another. For example, a biased model could present in its (presumed) tendency to predict a positive sentiment more often for samples in Urdu than in Hindi, vice-versa, or indeed any other disproportionate outcome. Similarly, a model might

indicate signs of linguistic bias according to the given definition if it achieved better classification accuracy for samples of a particular language, e.g. Hindi samples achieving a higher classification accuracy than Urdu samples. This thesis also utilises a quartet of tests from the AI Fairness 360 toolkit in order to better gauge these differences, although the scope of these tests remains limited to the latter example of difference.

Persons should not have to suffer degraded or incorrect performance as a result of their spoken language, and to investigate this is the first necessary step in mitigating this harm. Sentiment Analysis tools have been used by several state, non-governmental and social organisations geared towards speakers of Hindustani, and for this to carry on carrying the weight of unfair treatment exacted upon them as consequence of improperly oriented systems can not continue to pass, and indeed can result in harmful outcomes if it comes to pass [19].

There is undeniably ample research on the nature of biases present in NLP systems. Kiritchenko and Mohammad [18] conducted bias analysis on sentiment analysis systems through constructing an "Equity Evaluation Corpus" - a large collection of sentences consisting of pairing emotionally charged vocabulary paired with either African-American or European American names, or noun phrases signifying gender (e.g. I had a heartbreaking conversation with *my aunt* or *my uncle*), and tasked 219 sentiment analysis systems with measuring the intensity of a select catalogue of sentiments was present in the sentences within the corpus. The results displayed widespread symptoms of bias, i.a. assigning sentiments of "anger", "fear" and "sadness" with higher intensity to African-American names, and "joy" and "happiness" more often for European-American names. The authors also found significant signs of gender bias in statements in which emotionally charged vocabulary was used in conjunction with gendered terms, finding that the majority of studied models assigned higher sentiment intensity predictions across female exemplars.

Jentzsch and Turan [19] studied 63 sentiment classification systems comprised of 7 BERT models with 9 training conditions, measuring gender bias through the extent to which the intensity of predicted emotion varies between different samples across gendered names. Of the studied configurations, 57 showed significant biases, of which 16 "preferred" male terms over female, whereas the rest of the biased cohort displayed various signs of misogyny. From the two aforementioned studies, it may be remarked that contemporary models show preferential treatment of certain groups over others.

Bolukbasi et al. [6] in the development of de-biasing algorithms for word embeddings, discover that these encode gender biases present in society into the systemic representations used by NLP systems, studying the associations between word embeddings and gender, finding strong evidence of direct bias in word embeddings and indirect bias in word embeddings, where for example a sexist analogy like man is to woman as computer programmer is to home-maker could be expected [6]. These findings are corroborated by the work of Caliskan et al., who introduce the WEAT, a testing framework to identify bias in word embeddings, showing that socially sensitive terms such as ethnic or gendered names are associated to pleasant or unpleasant attributes, an example being their chosen systems propensity to provide pleasant associations to European American names more often than to African American ones [20].

The work of May et al. embellishes on the work of Caliskan et al., extending the WEAT framework to sentence level encoders, developing the SEAT testing framework and applying it across a catalogue of encoder-based models, finding significant gender biases present in BERT with regards to the double bind stereotype, and the flowers/insect task from the Caliskan testing suite [26].

The work of Kurita et al. shows that models utilising contextualised word embeddings such

as those in this thesis exhibit signs of bias as demonstrated in prior works by tasking a BERT model with the Gendered Pronoun Resolution task, where a pronoun-containing expression is to be paired with the entity uttering the expression. Their work shows that BERT struggles in performing co-reference resolution for female candidates, reflecting the aforementioned findings [21]. The findings of Basta et al., who similarly studied the presence of gender bias in contextualised word embedding based models using methodology stemming from that of Bolukbasi et al. found that while the extent of direct bias is to a lesser degree - a measure of 0.03 as opposed to 0.08 found in the study on static word embeddings - the presence of bias has only somewhat been diminished, as opposed to justly mitigated. Furthermore, the turn to contextualised word embeddings does result in stereotyping words relating to identities clustering more closely with gendered terminology [7]. The studies of Fleisig et al. reveal in fact that when responding to minority groups, LLMs that make use of contextualised word embeddings often respond in stereotyping manners, reflecting the evident harms created by this phenomenon [11].

The body of work on the nature of bias in current NLP systems demonstrates the simple, yet malignant fact that our usage of these systems, through propagating socially learned biases in downstream tasks, further enforces the degree of unfairness in our society, as well as that the preferential treatment displayed above may stem from the manner in which models learn societal biases. An important consideration to be made of the aforementioned studies is that the given work has focused on English language tasks, where the conclusions drawn from this may not necessarily generalise to the linguistic and cultural connotations of other languages. Thus, some treatment of studies analysing the presence of bias in multilingual contexts is given.

2.2.1 Multilingual Bias

As prior mentioned, of the body of literature concerning bias in NLP, a remarkably low proportion of this focuses in nature on the presence of such in multilingual contexts [23]. The research of Câmara et al., following the earlier mentioned work of Kiritchenko and Mohammad, establish multilingual Equity Evaluation Corpora designed to study uni-sectional and intersectional social biases in Spanish and Arabic alongside English, using them across a catalogue of transformer-based models. Their findings indicate diverse signs of bias, English models flag sentences with Arab names as "fearful" more often than Arabic models would flag the same sentence with the same name, models in English and Spanish predict fear more often for women than men, models in Arabic sentences referring to women predict anger less often, amidst other signifying factors. These results indicate that to some extent the encoding of social biases into NLP models discussed above also applies, but these may not be the only factors resulting in biased outcomes.

Similarly, Goldfarb-Tarrant et al. built a counterfactual evaluation corpus in order to analyse the prevalence of gender and racial biases in sentiment analysis models in Chinese, German, English, Japanese, and Spanish - using a methodology considerate of grammatical phenomena inherent to the aforementioned languages. Across all languages, test sets for evaluating gender bias are given, and test sets for evaluating anti-migrant and racial bias for German and Japanese languages are also given [24]. The authors compare the difference in scores for polarity detection across groups in different languages, finding that when controlling for gender bias, models showed that English language models tended to display lesser degrees of bias than others. In controlling for migration based and racial biases, the Japanese and German models were shown to be biased against minority groups, whereas the Spanish models reversed this bias i.e. biased against the majority group.

The work of Gamboa et al. in the field of bias in multilingual contexts provides insights that may put this into perspective. Their research demonstrates that the development of bias-mitigating architecture in NLP suffers from the challenge of dealing with the varying nature of biases present across different languages and cultures. They find that while the presence of some over-arching categories of bias can be generalised across languages, the distinct manner in which they arise vary dependent on culture and language. In the context of this research, especially considering aforementioned work regarding biases encountered in society transferring into computational methodologies, this may present in the form of biased outcome when considering the multilingual context of the study [27].

This consideration regarding the variety in biases across different languages is echoed in the findings of Sahoo et al., who in the process of developing a bias evaluation dataset for Hindi detect the presence of different forms of gender and occupational bias present in different languages based off of cultural connotations (e.g. the caste bias found in Hindi is not replicated in the Korean language, and the connotation of the colour pink with the queer community seems to be something limited to the Anglophone world in the context of their research) [28].

While many studies approaching bias in NLP center solely around gender or identity based constructs, Levy et al. compared gender, race, nationality and religion biases across sentiment analysis systems using mBERT and RoBERTa in English, Chinese, Hebrew, Italian and Spanish. Their findings found that while there was a degree of social biases intrinsic to the models learned representations as consistent with prior research, that biases were not amplified or reduced particularly for any specific language. That being said, the research still found some particularly telling signs of bias, such as: the sentimental judgement for the religion adhered to by the majority of a language's speaker receiving more positive judgements than other religions; nationalities where the given language being studied has official status being given more positive judgements than other nationalities; and most alarmingly - that persons identified with the Black race were adjudged more negatively across all languages. This study should act as a watershed moment for research on bias in NLP, displaying how deeply and across different axes of identity the issue cuts [29].

The research of Smirnov showed by tasking LLMs to provide a description of political events in Ukraine in either Russian or Ukrainian languages how prompt language can influence the nature of the outcome, demonstrating using the chosen example of the military conflict in Ukraine how models in the catalogue would consistently give syntactically different responses in Russian and Ukrainian. This represents differential treatment on behalf of the model with respect to speaker language, and can lead to harmful consequences if these responses are utilised for purposes such as the dissemination of politically charged information or material [30].

As for directly controlling for performance between two languages as is done within this study, the work of Wong and M'hiri compares performance in French and English, the governing language of the researchers' country of residence, finding that Francophone systems outperform their Anglophone equivalents, pointing towards some idea of bias present for speakers of one of the two languages within Canadian environments [31]. The comparison between the outcomes in speakers of two languages studied here may only be exacerbated by the subtle difference of choice in linguistic study here selected.

From the established body of work studying biases in multilingual contexts, it can be derived that while the presence of bias that may be gleaned by referring to the English-based body of literature certainly translates as well, the extent and degree of this does not necessarily have to be the same. In order to understand the nature of biases in other linguistic contexts, further study is warranted,

an ends that this study thus serves as the means to.

2.3 Hindi vs Urdu - "Hindustani"

In light of the aforementioned background on sentiment analysis and work on bias within NLP, as well as in more broader multilingual contexts - some consideration is also warranted to that focused on biases in Hindi-Urdu contexts. Before this, some background to both languages.

What are today Hindi and Urdu are both descendants of a language known today as Khari Boli (standing language), a vernacular that developed in what is today the region around Delhi [32] around the 11th century CE. Traced to its roots, the language evolved from Shauraseni Prakrit, the dramatic register of Prakrit, developing as a lingua franca amongst various regional languages at the time. Over time, the political climate of the region, in combination with the inclusion of vocabulary through other borrowings led the language to become infused with some Persian and Arabic grammatical and lexical structures. Although the general spoken content of the language has remained largely similar, it may be better for the purposes of this study to consider the development of each language separately in order to properly enunciate some differences in their modern equivalents [33].

The linguonym Hindi originates from Hindvi, a reference to the vernacular of the region in poetry of the time, likewise Hindustani from its duplicate, both being Persian terms referring to a broader idea of the region. Urdu likewise finds its origins in the verse of "zaban-e-Urdu-e-Mualla" - the language of the camp of the exalted, a reference to one of the ruling dynasties of the time. At the behest of the turns of history, much extant literature remains in the Perso-Arabic script [34]. At the behest of colonial history, a movement to develop a language innate to scripts of the region amidst a growing movement of encouraged sectarian divides resulted in the development of a separate register of the language, in the Devanagari script that additionally included Sanskrit derived replacements for previously inherited loanwords, resulted in the formation from the half-written, mostly spoken language of Hindustani to the remarkably distinct languages of Hindi and Urdu.

The events of the 20th century had a remarkably transformative impact on South Asian peoples, a fact certainly reflected within its languages. The joint independences of India and Pakistan, along with the language's historical nature as the lingua franca of the region led both nations to select Hindi and Urdu as respective official languages of each country. Although the divide had formerly been only in name and orthography, their uses across different societies resulted in both languages acquiring influences from diverse other sources, among which include borrowings and vocabulary from other languages in their respective societies. Alongside this development, the growth in sectarian division amongst communities in both societies over the years has also enveloped that between the two languages, causing their usage in some societies to become intrinsically tied to religious identity as well as nationality, where Urdu speakers are associated with the Muslim faith and Hindi with the Hindu, regardless that such binary distinction rarely captures the linguistic complexities of the societies in which they are used. In Hindi and Urdu contexts, this may manifest as linguistic bias towards speakers of one language, meted out as a form of religious based or national identity based discrimination [35]. While this may have resulted in some extent to the differences in linguistic minutiae being present, this serves not as contradiction to the fact of the vernacular dialects of both languages spoken today being to a great extent mutually intelligible to each other, a principle illustrated in Figure 1. For the purposes of this thesis, this allows in a sense the languages to be considered as a particularly low resource register of the other language.

Amidst these shared linguistic structures, the increase in cross-lingual communication afforded by recent developments has resulted in greater cultural exchange than ever, carrying amongst it vital interchange of ideas and information. Amongst this exchange, developments in technology have resulted in computational resources in Hindi and Urdu (albeit spoken or romanised) being available to members of both speech communities.

On to the status of the developments of NLP regarding Hindi and Urdu. The work of Malik et al. measure bias in Hindi language representations, comparing using modified versions of the WEAT and SEAT test, comparing bias measurements across a test set designed to measure bias in the Hindi language and across the test set used in the WEAT test utilised by the study of Caliskan et al. as prior mentioned [20], the authors utilise Hindi translations of semantically neutral templates used in the earlier mentioned work of May et al. [26]. The authors measure for bias across gender, caste and religious identities. The findings report significant gender bias in Hindi language representations, and the findings for caste and religion based bias are even more shocking. The representations identifying terms identifying lower caste persons are strongly associated with negative terminology, and likewise for upper caste. The authors consideration of religion contrasts religious last names and terms with adjectives, finding discrimination towards Muslim groups [36].

Gupta et al. [37] developed test datasets in the Hindi language to examine the prevalence of gender bias across select LLMs in Hindi, using these to solve a co-reference resolution task, finding remarkable signs of bias, among other examples - performance suffered for female entities, to the selected models were more likely to assign emotional reactions even when male entities were clearly designated as having expressed them, and models often responded in stereotyping manners, bias that the authors hypothesise originate within the training data.

Kumar et al. [38] developed a bias detection framework for Indian-centric languages, with the results of their research conducted with the use of this displaying various forms and degrees of biases present across the models selected for this research in Hindi. The authors found that religious bias towards minority religions in Indian society was a common issue throughout the selected model catalogue. Having considered these findings, it is also important to note studies like this are few and far between, representing the dearth of scholarship on the presence of bias in the South Asian languages [39].

The prior studies represent a smaller subset of an already minuscule amount tackling bias in the computation of the languages spoken in South Asia. Most research centred on bias in these contexts focuses more on using NLP methodologies to diagnose and detect biases present in society [40]. Beyond the prior mentioned work, there is relatively minimal attention regarding the degree to which the biases these models are being used to detect are themselves present within system architectures. In the absence of research designed to diagnose the extent of social prejudices encoded in system applications, we run the serious risk of continuing our development with these biases going unchecked.

Concerning sentiment analysis in Hindi and Urdu, research up to the present has been more exploratory in nature, geared primarily around establishing first principles as opposed to evaluating fairness [25] [40]. The work of Ashraf et al. tackles this problem by utilising a BERT-based methodology for sentiment analysis in Urdu, outperforming the accuracy of prior models by 26.09% and F-1 score by 25.87% [41]. Works such as this demonstrate that while the adoption of sentiment analysis systems in Urdu have transpired, the lack of studies analysing their fairness indicate that it does so uncontrolled for fairness. As the development of AI systems is felt on a more global scale by the increased connectivity of the modern era, and the increased adoption of AI systems amidst

dissimilar increases in the rate of digital literacy and understanding of these systems is a concern that has been noticed in prior research, particularly in the developing world, it is important that some measure of equity be extended as well to this field, as it should not be acceptable that this dearth of research should proceed unfilled.

2.3.1 Romanisation

A somewhat core feature of this research is the romanisation of Hindi/Urdu text. The reasons for this can be motivated by the following. The prior established history and nature of the two languages have resulted in a contemporary situation of the two languages such that when either Urdu or Hindi is spoken, the degree of mutual intelligibility between speakers of either language is to such an extent that without very overt dialectic giveaways or prior information, speakers of one language may struggle to tell apart those of the other. This has also been somewhat visualised in Figure 1. As shown in this example, despite the differences in orthography - Nastaliq being written right-to-left as opposed to Devanagari being left-to-right, as well as both languages sharing wildly different rules for the transcription of vowels, and composition of word shapes - the lexical content of both languages is remarkably similar when romanised.

Native Script	Romanised
I went to the market today English - Latin Script	~~
मै आज बाज़ार गया था Hindi - Devanagari Script	main aaj bazar gaya tha
میں آج بازار گیا تھا Urdu - Nastaliq Script	mein aaj bazar gaya tha.

Figure 1: Comparison of Devanagari and Nastaliq scripts with Romanisation alongside

In order to best replicate this experience of cross linguistic communication that often underlines such uses of both languages, data was at first sought, and then later algorithmically romanised so as to mask the differences in orthography between both languages that would not be present in the more general use cases of either language in romanised format.

This romanisation did introduce some limitations to the findings of this research, a matter that is further addressed in Chapter 5.2.

2.3.2 Linguistic Imperialism

Some attention in the context of this thesis is also warranted to the topic of linguistic imperialism, which refers to the dominance of one particular language in a specific context thereby resulting in the diminished use and possible eventual extinction of marginalised tongues and forms of expression [14].

Much of the Internet is in English, this is both a symptom and a necessary consequence of the main cause of the issue [42] [12], the assimilation of global culture into an Anglo-amalgam. Especially in

the current geopolitical climate, the implicit suppression of foreign peoples through the introduction of roadblocks that effectively inhibit the usage of tools designed to aid and ease their lives and workflows, resulting in a situation where they are forced to utilise said tools in manners not in line with their intended designation but rather in ways that befit the English speaking population and their desires and narratives. In all senses the capacity of non-English speakers to fully benefit from these systems is minimised. Beyond this, this is a myopic view of the development of computational linguistics, one that supposes that the linguistic features of English are in some way universal, thereby ignorant of other means of transferring and storing information.

Not only does this present the obvious ethical issues, this also translates poorly for the development of AI systems [43], which only furthers the amount of inequalities their usage procures. Thereby, in an already biased landscape, furthering scholarship of the prevalence of, and of the understanding of the natures of issues plaguing AI systems in languages other than the global "standard" serves as a means to prevent the growth of these inequalities, and to work towards providing a more fair distribution of resources of languages in computation. In having considered the relative lack of scholarship and resources available upon and in Hindi and Urdu, this can be seen as a necessary step in the alleviation and prevention of the continuance of this phenomenon.

3 Methodology

In this section, an explanation of the methodology taken in order to fulfil the objective of the research is given, alongside a description of the experimental setup used to obtain the results of which, wherein an explanation of bias measurement is also given. This explanation includes formally the collection, subsequent processing and treatment of data, selection of and processing done to the model catalogue, and the experimental setup and bias measurement.

3.1 Data

3.1.1 Collection

The dataset used for this study is a self-agglomerated collection of texts sourced from several pre-existing datasets in Hindi and Urdu [44] [45] [46] [47] [48] [49] [50]. All datasets used in this thesis are openly accessible. In order to maximise the comparability of results across both languages, care was taken to accumulate datasets in Hindi and Urdu of similar sizes. The Urdu language had a large collection of reader reviews and comments from several Pakistani newspapers, which were labelled as either positive, neutral, or negative [49]. This dataset consisted of samples that were already in romanised format, and the authors of the dataset disclosed that these reviews were also written in romanised Urdu, implying a representative set of the romanised Urdu of its speaking population. This dataset served for the Urdu corpus of this thesis.

In Hindi, the procurement of datasets containing similar quantities of material proved counter-intuitively to be more challenging than its Urdu counter endeavour. This resulted in the final corpus being stitched together from several sources. Initially, sources were sought out that contained romanised Hindi text, but the limited presence of these resulted in the final approach containing texts that were sourced in the Devanagari script, which added an additional romanisation step to the research methodology. The final Hindi corpus consists of user reviews sourced from IMDb and Amazon. Due to the increase in amount of datasets compiled into the final Hindi corpus, there was a degree of variation in labelling conventions, which resulted in the addition of a label sanitisation step into data processing. Table 1 contains some examples for each class in Urdu, and Table 2 likewise contains a collection of examples in Hindi.

Table 1: Example Sentences in Urdu

Sentence	**Label**
1 aur mazedar app	Positive
mera vote pti ke sath hai .kyon ka mery khayal sy wo hiis mulk ko bacha sakte ha	Positive
zarafat ke makhsoos andaz ny un ke khatut ko dilchasp nasar nigari bana deya	Positive
emergency ward mein inho ne aik nurse ka role nibhaya	Neutral
ye movie abe indian movie ke samne zero hai	Negative
ab mat bolna modi ne karaya ha sb kuch	Negative

The Urdu sentences, top to bottom can be translated as follows:

- 1 au... - One more fun app

Table 2: Example Sentences in Hindi

Sentence	**Label**
unki jugalbandi rang laati hai.	Positive
keemat ke layik.	Positive
main subha yoga karta hun.	Neutral
kahani mein dama nahin hone ke bavajud dilchasp kirdaar.	Neutral
mujhe yeh aavaz pareshaan karati hai.	Negative
lambe samay tak attki rahi.	Negative

- mera vo... - I support PTI, because I think *they* are the ones who can save this country
- zarafat... - Their humorous style has made their threats interesting.
- Emerg... - She took on the role of a nurse in the emergency ward.
- Ye mov... - This movie is nothing in comparison to an Indian one.
- Ab ma... - Don't go saying Modi did all of this now.

The Hindi sentences, top to bottom can be roughly translated as follows:

- unki ju... - Their Jugalbandi really lights up the atmosphere
- keema... - Worth its price
- main su... - I do yoga in the morning
- kahani me... - Seems refreshing even though he isn't in the story at all.
- mujhe ye... - Hearing this noise (or voice) really worries me.
- lambe sa... - She remained there (in an annoying manner) for too long

The nature of the romanisation, namely the inconsistencies in spelling conventions, and the loss of some orthographic cues of sounds did raise some limitations, that have further been addressed in Chapter 5.2.

3.1.2 Processing

The processing of the collected datasets commenced with romanising the Hindi texts, a task accomplished with relative ease using the **Aksharamukha** library. A slight issue this raised in having this romanisation be consistent with the standard used in prior research and in the Urdu dataset was the transliteration scheme containing several diacritics, and transliterating all words to end with vowel sounds, where most human based transliterations would end with consonants. A further step thus involved modifying these romanisations so that they maintained a similar lexical quality as examples of the Urdu corpora. This was done through modifying diacritics and sorting through words to remove excessive vowel endings without any loss of syntactic value. Through

comparing with a dataset of romanised Hindi words, suitable replacements for these were identified and then algorithmically applied so that the final Hindi corpus was romanised to a similar standard as its Urdu corpus. While this has induced an amount of noise to the findings of this research, care was taken in the proceedings of this research to ensure that these replacements remain true to the common standards used in contemporaneous use examples.

At this point, it became apparent that the corpus contained: many samples that were too short in length to be of any meaningful value; samples that were either written in English with Hindustani grammatical rules, or in English altogether, and samples so strewn with lexical errors that they were rendered indecipherable. In order to cleanse the corpus of the systematic noise incurred by these, the following were removed: instances of less than 3 words, instances where the proportion of English words exceeded half of the sentence, and sentences that contained those mistakes that were so commonly repeated and easy to spot that their removal was made possible.

This was followed up by sanitisation of labels within the Hindi collection. Due to the Hindi collection of datasets amassing in number to match the quantity of samples in the Urdu dataset, the collected datasets ubiquitously had varying label conventions: some numeric, and others varying with case and length. This was mitigated by remapping all examples to standardised label names. The data further proved to be unequal in nature due to containing more samples of Urdu than of Hindi, as well as differences in the distribution of polarities across both languages. In order to maintain a degree of comparability in classification results across languages, the data was further stratified so as to balance between both quantity and class distribution. Since the models used would also require fine-tuning for the task of this thesis, this stratification also had to be addressed when portioning data for these purposes. As a result of this, a set of 600 samples was selected from both corpora in such a manner that maintained equal class distribution to the test set.

3.1.3 Description

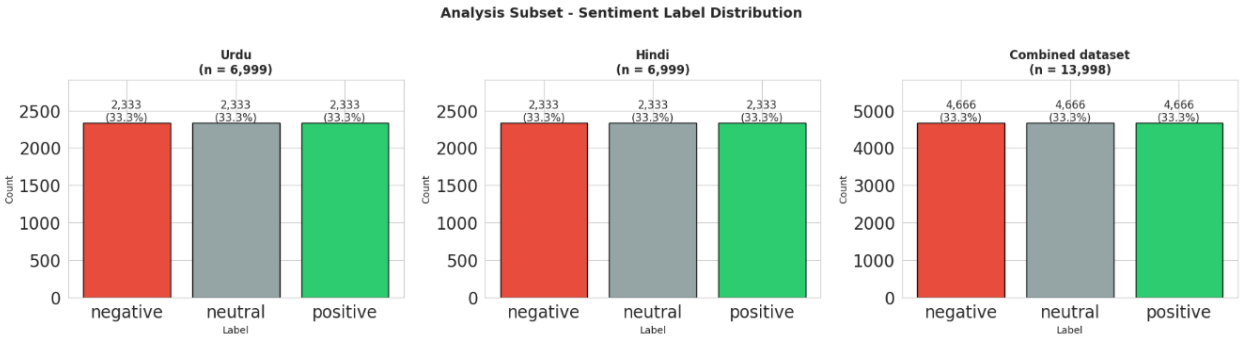

Figure 2: Distribution of Polarities across Languages

The final dataset consists of 6,999 samples each of Urdu and Hindi. This distributes into 2333 negative, neutral, and positive samples across both languages. In Figure 2 a visualisation of the distribution of polarities across languages is given. The lengths of texts in the corpus vary, with texts in Urdu ranging from 3-363 words in length, with a mean review length of 14.7 and median of 10. Hindi had some considerably longer outliers, resulting in a range of 3-1462 words. The majority of reviews were much shorter in length however, with a median of 9 words and a mean of 23.0.

3.2 Models

This thesis uses three models for assessing the performance of sentiment polarity classification. Multilingual BERT, or mBERT is used as a baseline model by which the difference between models across languages is compared across the language specific models chosen for this thesis. MuRIL (Multilingual Representations for Indian languages) is used as a model trained on Indian languages including romanised Hindi, that serves as a model designed to test the performance in Hindi language models. BERT_Roman_Urdu is one of the few available Urdu models for the task that are also trained on romanised text, and is thus used as the model by which the performance of Urdu language models is measured.

The chosen models required fine-tuning for the task, which was carried out using a portion of the corpora held out of the final task set. Each model designated for a specific language was fine-tuned using only samples of its training language (n=1998), with mBERT receiving a uniformly split distribution across languages in its fine-tuning data.

3.2.1 BERT_Roman_Urdu

BERT_Roman_Urdu is a BERT-based model trained on the Urdu language in both Nastaliq and romanised scripts. The model had to be fitted with a novel classification head, as it was trained primarily for binary classification tasks, this feature allowed it to be adapted to handle the ternary nature of the task. The procedure was geared to optimise performance on the task, ensuring that there were no structural limitations in the final outcomes that would be analysed.

The model was fine-tuned using the Huggingface Trainer API. Tokenisation was handled using the model's native tokeniser, with maximum length set to 256 tokens. The fine-tuning data was split into training and validation sets using a 85-15 distribution, resulting in 1698 training samples and 300 validation. The model was trained exclusively on the Urdu language set, running for 3 epochs with a learning rate of 2×10^{-5}, and batch size of 16. An overview of hyperparameters for BERT_Roman_Urdu are given in Table 3

Table 3: Fine-tuning Hyperparameters for BERT_Roman_Urdu

Parameter	**Value**
Base Model	BERT-Base-Roman-Urdu
Optimizer	AdamW
Learning Rate	2x10^{-}5
Number of Epochs	3
Batch Size	16
Weight Decay	0.01
Warmup Ratio	0.1
Evaluation Strategy	Per Epoch
Metric	Accuracy

3.2.2 MuRIL

Multilingual representations for Indian Languages (MuRIL) is a BERT-based model trained on 17 Indian languages [51]. Trained for such tasks in the Hindi language, this model afforded the additional benefit of being trained on romanised Hindi samples as well, allowed for its use in this thesis. Additionally of note, Urdu is amongst the languages that MuRIL was trained on, which adds an additional dimension in which the cross-lingual comparison may be conducted.

The model was fine-tuned for the purposes of this task on solely the Hindi language, with training and validation split similarly being of 85% training samples to 15% validation samples, which once again resulted in 1698-300 samples being reserved for each. The fine-tuning ran for 4 epochs, with a learning rate of 3×10^{-5}, and batch size of 32. An overview of the hyper-parameters for the fine-tuning of MuRIL are given in Table 4

Table 4: Fine-tuning Hyperparameters for MuRIL

Parameter	**Value**
Base Model	MuRIL-Base-Cased
Optimizer	AdamW
Learning Rate	3x10⁻5
Number of Epochs	4
Batch Size	32
Weight Decay	0.01
Warmup Ratio	0.1
Evaluation Strategy	Per Epoch
Metric	Accuracy

3.2.3 mBERT

Multilingual BERT (mBERT) is a BERT based model pre-trained on 104 languages, of which Hindi and Urdu are both included [52]. While both languages are pre-trained in their native script, prior research approaches have demonstrated the ability of mBERT to adequately process romanised text with appropriate fine-tuning. For this purpose, it is fine-tuned with a set of 1998 samples, of which 999 are in Urdu and 999 in Hindi, the distribution of which into training and validation samples should come as no surprise at 1698 and 300 samples respectively.

The fine-tuning consisted of both Hindi and Urdu language samples, with the max token length once again set to 256 tokens. The training and validation split remains the same as detailed above, with a split of 85-15% resulting in 1698-300 samples being assorted. The batch size is set consistent with the process of MuRIL at 32, and the learning rate is also set to 3×10^{-5}. F1-Score is the optimised metric, and the learning rate is set to differ with a cosine rate. An overview of hyper-parameters for mBERT are given in Table 5

This thesis uses an adaptation of the methodology of Wong and M'hiri [31], and Jentzsch and Turan [19], where the differences in the models classification of a dataset of reviews into polarity valuations of positive, neutral and negative are compared across different languages, and the presence of biases within the results are identified with use of the methodologies described in Chapter 3.3. The accuracies, or the rate of correct predictions, and Macro-F1 scores - the mean of

Table 5: Fine-tuning Hyperparameters for mBERT

Parameter	Value
Base Model	multilingual-BERT
Optimizer	AdamW
Learning Rate	3x10^{-}5
Number of Epochs	4
Batch Size	32
Weight Decay	0.01
Warmup Ratio	0.1
Evaluation Strategy	Per Epoch
Metric	F1-Score

precision and recall rates, where precision refers to the rate at which the model correctly identifies the polarity of a given text compared to all identifications of that polarity within the sample, and recall refers to the rate at which the model correctly identifies polarities compared to all instances of them within the sample - are computed for Hindi and Urdu language test sets for each chosen model, and the difference in results are compared across language sets.

Each model classified the test set of each language separately, in order to maintain direct comparability between groups and avoid introducing unnecessary noise through transliteration variations and other similar factors. In order to best capture the difference in results across each language, the difference between statistics of accuracy and F1-scores are compared across models. Classification results are compared across language between models, and the difference between the two is compared with the differences observed in the multilingual model to identify if there is significant preferential treatment of language in either model. Both models were tasked with classifying sentiments across a test set of 6,999 samples in each language.

Additionally, this thesis considers the classification distribution of the results. The study also measures the difference in predicted polarity distribution as compared to ground truth distributions, and addresses whether these differences can be said to be indicative of biased performance. Here the variable being measured is the proportion of the models predicted polarity in comparison with the real polarity, which has been held to be near uniform as explained in 3.1.3.

3.3 Fairness

The bias measurement carried out in this thesis is conducted using the AI Fairness 360 toolkit [53], a bias detection and mitigation framework designed to measure fairness between two separate groups. In this thesis, the two groups are defined as the Urdu language and the Hindi language respectively. The methodology used did require some adaptation in order to be used for the context of this research, as the AIF360 toolkit is based around binary classification scenarios. In order to consider the difference in classification performance as a binary classification problem between favourable decisions, this thesis defines a correct classification as a favourable outcome, and incorrect classification as unfavourable.

The extent to which differences in classification results constitute linguistic bias was evaluated using a standard cohort of fairness metrics in the AIF360 toolkit. The Statistical Parity Difference, or

the SPD score. The SPD score represents the difference in the likelihood of obtaining a "favourable" decision dependant on group identity. It is calculated by taking the difference of the conditional likelihood of obtaining a favourable decision based off of which group one is a member of. Since this thesis considers the difference in classification results across languages, a favourable decision is here defined as a true positive (i.e. a correctly identified polarity). A SPD score of 0 implies the most equal treatment, where both languages have equal odds of correct classification. Similarly the Disparate Impact ratio measures the ratio of a favourable outcome pertinent to group membership, serving as a means of comparing the rate at which correct classifications differ between languages.

The Average Odds difference evaluates the difference in true positive and false positive rates between groups. In the context of this study, true positives are defined as cases in which the model is able to correctly identify the polarity of a given text. Due to the presence of a medial class, a false positive is in this classification scenario seen as a "flipped" prediction, that is to say flagging negative as positive, or vice versa. The Demographic Parity ratio calculates the likelihood of group identity influencing favourable outcome. For the purposes of this thesis, the difference in performance results were said to be bias if samples of either language group had a significant likelihood of receiving higher classification performance dependent on language.

Additionally, it is important to consider that while these metrics compare only the differences found in correct classifications, they do not contain all the subtleties of bias that may be found in the models treatments of different language groups. Since the AI Fairness 360 toolkit is designed primarily to capture the difference in favourable label outcome, in the context of this thesis this results in it deriving a comparison only in classification accuracy. In order to assess the extent of which the differences in model predictions constitute linguistic bias between model treatment beyond this factor, an analysis of erroneous classifications was conducted so as to better determine if the found differences can truly be said to indicate the presence of linguistic bias as opposed to a reflection of domain difference in the dataset.

4 Results

In this section, the results of the classification task are given, through firstly the performance metrics of the polarity classification task across all three models and the degree of variation between each models performance, the bias measurements outlined in 3.3, and an analysis of common error patterns found in the models results is given.

4.1 Classification Performance

The accuracies and macro-F1 scores across each model and language are given in Table 6. Additionally, a comparison of accuracy across the languages for each model is given in Figure 3, and F1 across languages and model in Figure 4

Table 6: *
Accuracy and F1-score per Model and Language

Model	**Overall**		**Hindi**		**Urdu**	
	Accuracy	**Macro F1**	**Accuracy**	**Macro F1**	**Accuracy**	**Macro F1**
BERT_Roman_Urdu	47.38%	47.38%	37.64%	37.41%	57.10%	57.15%
mBERT	57.43%	57.35%	62.13%	62.12%	52.72%	52.62%
MuRIL	52.74%	51.69%	66.58%	66.49%	38.99%	34.76%

BERT_Roman_Urdu model had an overall accuracy of 47.38% across the entire corpus, and a F1-score of 47.38% as well. mBERT and MuRIL models performed slightly better in comparison, coming out with accuracies of 57.43% and 52.74%, and F1-scores of 57.35% and 51.69% respectively. Across the Hindi corpus, BERT_Roman_Urdu model had an accuracy of 37.64%, and a F1-score of 37.41%. This was the lowest across the catalogue. The multilingual BERT model had an accuracy and a F1-score of 62.13% and 62.12% respectively, and MuRIL outperformed all other models with accuracy and F1 scores of 66.58% and 66.49%. Across the Urdu corpus, a marked degradation of performance is visible across the model catalogue with exception to the Urdu model, which still barely outperforms the rest. The Urdu model had the highest accuracy score of 57.10%, and F1-score of 57.15%. The mBERT model also displays a decrease in performance when compared with Hindi, with an accuracy of 52.72% and F1-score of 52.62%. MuRIL achieves the poorest performance in the Urdu language, with classification accuracy of 38.99% and F1-score of 34.76%.

Across the Hindi and Urdu languages, there is a statistically significant difference in classification performances. BERT_Roman_Urdu has a difference of 19.47 percentage points in favour of Urdu in accuracy and 19.74 percentage points in F1-score between the Urdu and Hindi languages. mBERT has a difference of 9.41 percentage points in accuracy and 9.50 percentage points in F1-score between Hindi and Urdu, performing slightly better in the Hindi language. MuRIL has a difference of 27.59 percentage points in accuracy, and a difference of 31.73 percentage points in F1-score between Hindi and Urdu, by far the largest gap between the languages. Both BERT_Roman_Urdu and MuRIL thus display a much greater degree of variation in their treatment of both languages than the multilingual baseline.

The confusion matrix of each model by language is given in Figure 5. In Urdu, BERT_Roman_Urdu correctly predicted negative sentiments 56.15% of the time. Of the incorrect negative predictions, the

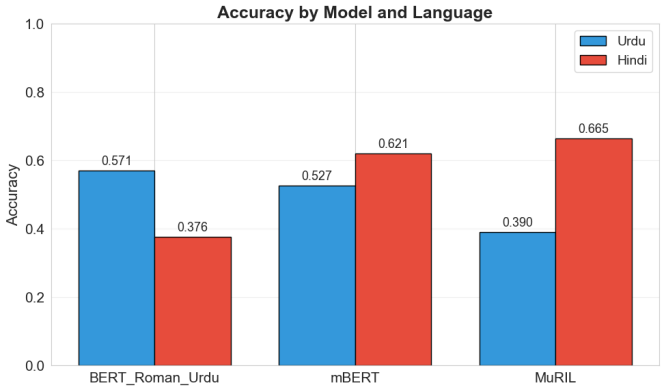

Figure 3: Comparison of Accuracy by Model and Language

model predicted the output to be neutral for 25.93% of the samples, and positive for the remaining 17.92%. The model achieved the highest correct classification accuracy for neutral samples, of 58.64% of neutral samples. Of the incorrect neutral classifications, 22.16% were classed negative and 19.20% positive. The model performed similarly for negative samples as positive samples, catching 56.54%, and misclassifying 26.70% as neutral and 16.76% as negative, mirroring the distribution of classifications assigned to positive samples. The multilingual BERT model outperformed the

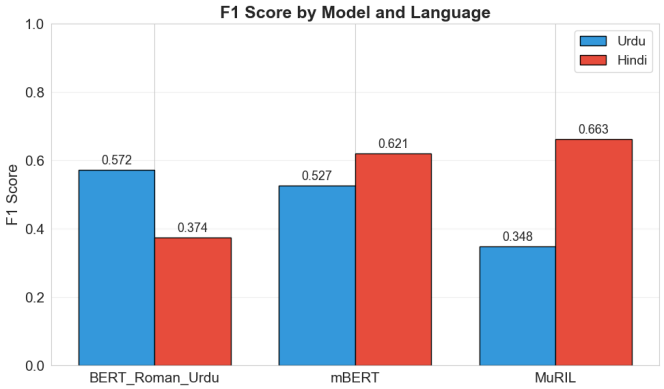

Figure 4: Comparison of Macro F1 by Model and Language

catalogue in predicting negative samples, catching negative samples 60.87% of the time, and classifying the missed samples as neutral 23.15% of the time, and as positive the remaining 15.95%. Neutral samples were correctly classified at a rate of 47.92%. The misclassification rates reveal a slight lean towards negative classifications, coming out 32.79% of the time as opposed to 19.29% of neutral samples being classified as positive. For true positives, the model was correct 49.38% of the time, classifying 22.42% of the remnant positive samples as neutral, and the remaining 28.20% of positive samples were misclassified as negative.

MuRIL was by far the best model in the catalogue at detecting positives in the Urdu language, with true positive rate of 65.20%. Of the misclassified samples, a meagre 8.44% were classified as neutral, and the remaining 26.36% as negative. Of neutral samples, a mere 8.92% of samples were correctly classified as neutral, with 38.32% of samples being classified as negative and 52.75% as positive. The trend for predicting the majority of samples in the positive class for the Urdu

language continued for the negative samples, where a true negative rate of 42.86% was compounded with a false positive rate of 49.81%. The remaining 7.33% of negative samples were misclassified as neutral. Across the Hindi dataset, BERT_Roman_Urdu performs poorly on negative samples,

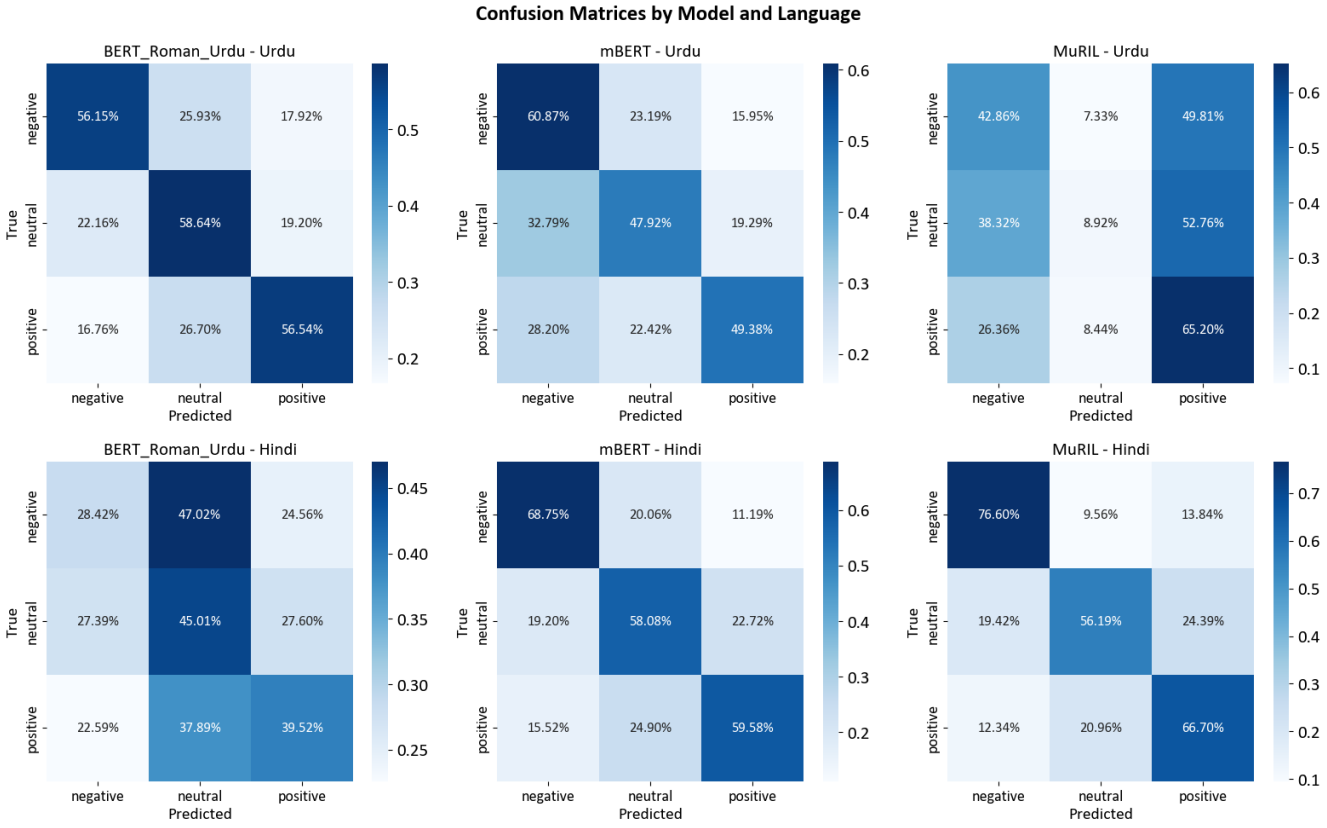

Figure 5: Confusion Matrices by Model and Language

correctly classifying only 28.42% of negative samples. The majority of the misclassifications are neutral, accounting for 47.02% of samples, with the remaining 24.56% being false positives. In neutral samples, the model correctly predicts neutrality in 45.01% of neutral samples, classifying 27.39% as negative and 27.60% as positive. The model has a true positive rate of 39.52%, with 37.89% of positive samples being classified as neutral, and 22.59% as negative.

The multilingual BERT model performed better on negative samples in Hindi than Urdu, with correct classification rate of 68.75%, the true negative rate also being the metric in which it had the best score across the Hindi corpus. 20.06% of negative samples were misclassified as neutral, and the model marked 11.19% of negative samples as positive. For neutral samples, the model correctly classified 58.08% of neutral samples as such. The remaining 19.20% and 22.72% it misclassified as negative and positive respectively. The model had a true positive rate of 59.59%, with 24.90% of positives being marked as neutral and 15.52% as negative.

MuRIL has by far the best true negative rate for Hindi samples at 76.60%, with 9.56% of negative samples misclassified as neutral and 13.84% as positive. Of the neutral samples, 56.19% were correctly classified, with 24.39% of neutral samples being classified as positive and 19.42% as negative. 23.67% of misses were classed as positive, and 20.14% as negative. For negative samples, MuRIL correctly predicted 66.70% of negative samples, marking 20.96% of the rest as neutral and

12.34% as positive.

In order to understand the different ways in which models treat different language groups, this study also analyses the distribution of polarities assigned by the model. These are visualised in form of a direct comparison in Figure 6, and in Table 7 the statistics of each distribution are given. BERT_Roman_Urdu assigned 26.13% of samples as negative, 43.31% as neutral and 30.56%

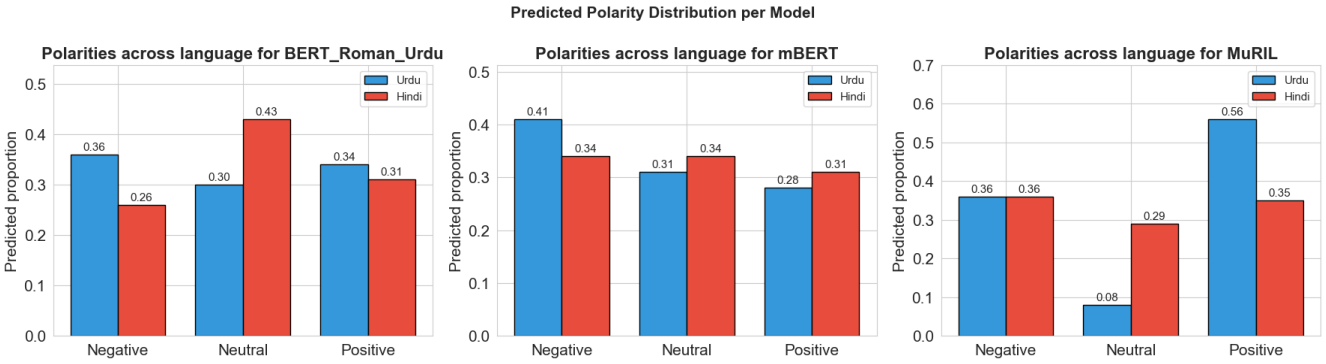

Figure 6: Prediction Rates Across Model by Language

as positive for the Hindi dataset, and 31.69% negative, 37.09% neutral, and 31.22% positive for the Urdu Dataset. mBERT assigned 34.49% negative, 34.35% neutral and 31.16% positive for the Hindi dataset, and 40.62% of samples as negative, 31.18% as neutral and 28.20% as positive for the Urdu dataset. Finally, MuRIL predicted 36.12% of samples as negative, 28.90% as neutral, and 34.98% as positive for the Hindi dataset, and 35.85% as negative, merely 8.23% as neutral and an overwhelming 55.92% of samples as positive across the Urdu dataset. These polarity distributions display some interesting characteristics that may be interpreted as biases similar to those found in the studies of Goldfarb-Tarrant et al. [24].

Table 7: Prediction Distribution by Input Language Statistics

Model	**Language**	**Prediction Distribution (%)**			**Chi-Square Test**	
		Negative	**Neutral**	**Positive**	χ^2	p-**value**
BERT_Roman_Urdu	Hindi	26.13	43.31	30.56	71.51	2.96×10^{-16}*
	Urdu	31.69	37.09	31.22		
mBERT	Hindi	34.49	34.35	31.16	232.28	3.65×10^{-51}*
	Urdu	40.62	31.18	28.20		
MuRIL	Hindi	36.12	28.90	34.98	1143.50	4.91×10^{-249}*
	Urdu	35.85	8.23	55.92		

From these differences, each model displays its own form of preferential treatment. BERT_Roman_Urdu predicts samples as neutral 6.32 percentage points more in Hindi than in Urdu, which is in stark contrast to MuRIL, which barely predicts neutral sentiments for the Hindi corpus at all. BERT_Roman_Urdu predicts positive sentiments 4.43% of the time more often in Hindi than in Urdu, whereas MuRIL predicts positive sentiments for the majority of samples in the Urdu language, with a staggering difference of 20.94 percentage points between languages.

mBERT also maintains the pattern of predicting negative emotional polarities more often than its monolingual counterparts. For both Hindi and Urdu languages, negative sentiment is predicted as the majority class. The model comes closer to the true distribution on Hindi than in Urdu, coming the closest to mirroring the true distribution of the input polarities across the test corpora in this language.

4.2 Error Analysis

A way to gain insight into whether these results constitute bias is to investigate the nature of the errors. Some misclassifications are therefore promptly treated as follows. Examples of misclassifications for Urdu are given as follows: BERT_Roman_Urdu in Table 8, mBERT in Table 10 and MuRIL in Table 12. Examples of Hindi misclassifications are as follows: BERT_Roman_Urdu in Table 9, mBERT in Table 11 and MuRIL in Table 13.

Table 8: Misclassifications of BERT_Roman_Urdu in Urdu

Sentence	**True Label**	**Predicted Label**
asa lagta hai jaise yai gana hi mohabbat hai	Positive	Neutral
khuda ky liay pm sab hath jorr ky rekuest hai ky vazirin or fujion ki ayashion ko kam kardain	Neutral	Negative
sab jhoot h.	Positive	Negative
astaffirullah allah pak lanat kra and ghark kra amin sum amin	Negative	Positive
Isko mall rd kay khaday main phenk do	Negative	Neutral

The example misclassifications of BERT_Roman_Urdu in Urdu, top to bottom, can be translated as follows

- Asa lag... - This song feels like the definition of love
- khuda ky... - For gods sake, I'm begging this prime minister to decrease the degree of concessions awarded to ministers and military.
- sab jhoot... - Its all a lie.
- astaffir... - God forgive me, may god smite and drown them, amen.
- Isko mall... - Throw (indeterminate pronoun) in a pothole on Mall Road.

BERT_Roman_Urdu classified positive samples as neutral 623 times in the Urdu language. Out of a total of 3002 misclassifications, this resulted in 20.8% of errors arising through this misclassification pattern. The next most common pattern had a small difference of 19 fewer errors, being negative answers being misclassified as neutral. The examples in Table 8 display some commonalities of these misclassifications. The statements all contain emotive judgements, but are structured in such a format as to follow the sentence structure of objective claims. It is possible from these findings to hypothesise that the model assigns neutral judgements to these sentences based not on the content of the vocabulary, but the manners in which they are structured.

In terms of samples that were actually neutral, the model misclassified neutral samples as negative 517 times, and as positive 448 times. This hints at a slight bias towards negativity. Reading neutral examples misclassified as negative or positive, a clear pattern is seen where references to figures of state or military, even in neutral contexts result in the model outputting negative polarity. Some other neutral examples misclassified as negative include "hahahahahaha head of army vah", or "Imran Khan Sher e Pakistan" - statements that translate to "(laughter), wow, the head of army" or "Imran Khan the lion of Pakistan". However, an interesting counterexample is a pattern discovered where references to the current ruling party of Pakistan and its members are classified as positive by the model - with sentences like "shehbaz sharif wazir e azam" or "maryam nawaz h hmari cm", which translate to "Shehbaz Sharif the prime minister" and "Maryam Nawaz is our chief minister" received positive predictions as opposed to neutral, or negative in line with the prior pattern.

BERT_Roman_Urdu had the least false positives and negatives in comparison to other errors. The model incorrectly flagged 419 negatives as positives, and 391 positives as negatives. Some key terms that often led to the model outputting false positives are the mentions of celebrities and religion. Sentence four of the examples is a heavily charged sentiment that presumably was flagged as being positive due to mentions of religion, something that seems to be a commonality among false positives. The false negatives tend to often be shorter length sentences. A commonality amongst these samples is that most positive samples misclassified as negative tend to contain sarcasm, or terms like "katil" or "bc", which directly translate to "killer" and a South Asian variant of profanity often used as a term of endearment. It can be that the model is ill-equipped to handle sarcasm and non-standard use of language.

Table 9: Sample misclassifications of BERT_Roman_Urdu in Hindi

Sentence	**True Label**	**Predicted Label**
is teblet ke disple par ungaliyon ke nisan bahut asani se para jate hain	Negative	Positive
thik thak hai	Negative	Neutral
upyog karne ke bad imandar samiksa	Negative	Positive
lohe ka baksa	Positive	Negative
iski 200mah baitri ek bar ke carj men karib 12-13 ghante cali.	Neutral	Negative

The example misclassifications of BERT_Roman_Urdu in Hindi, top to bottom, can be translated as follows

- Is teble... - The display of this tablet becomes stained with fingermarks very easily
- thik tha... - It's perfectly fine.
- upyog ka... - Developed into an in depth review after usage
- loha... - A box of steel!
- iski 2... - Its 200 MAh battery ran for about 12-13 hours after just one charge.

Similar in Hindi as in Urdu, samples being misclassified as neutral made up the bulk of BERT_Roman_Urdu's 4364 errors. Negative samples being classified as neutral occurred 1097 times, and positive samples as neutral 883 times. The examples shown above highlight some key patterns found in these errors. As noticed across the Urdu dataset, sentences that were phrased in an objective manner were often classified as neutral, regardless of syntactic content.

Of actually neutral samples, the model had a slight propensity for errantly predicting positive samples, with 646 misclassifications fitting this pattern and 637 being negative. The model here displayed signs of what can be described as the opposite pattern, marking sentences that syntactically seemed to be following the format of subjective judgements yet did not necessarily have any particular lean as negative or positive based off of the emotive content of words found therein. The work of Chen et al. discusses how sentiment analysis systems tasked with polarity classification sometimes encode polarity association bias by incorrectly flagging neutral samples as having emotive value due to vocabulary contained within them, a phenomenon that may be at hand here [54].

False positives and negatives were once again the least common direction of errors, with BERT_Roman_Urdu having 573 false positives and 528 false negatives in Hindi. This is more than across the Urdu corpus, albeit with a similar difference in error percentage. Some commonalities that incurred false positives were the presence of vocabulary generally used in very positive contexts in the Urdu language. For example, while the example sentence is complaining that their display stains with ease, it could easily be misinterpreted as a remark of satisfaction at the ease at which a product fulfils a purpose. The false negatives, however, paint a more interesting picture. The misclassified samples are filled with vocabulary that is not commonly present in Urdu language contexts, displaying as sure a sign as any of linguistic bias.

Table 10: Sample misclassifications of mBERT in Urdu

Sentence	**True Label**	**Predicted Label**
insallah pmln hi kamyab ho ji	Positive	Negative
mjy pakistanion ki dance ki samajh nahi ati	Neutral	Negative
hanji, unearthed but kept his mouth shut till nov bas karo begyarto, tum hamaray leader nahi ho	Positive	Neutral
lannati log hai	Negative	Neutral
video per yai sab bakvas achi lagte hai . chalan ke sath mar bhi khani padega	Negative	Positive

The example misclassifications of mBERT in Urdu, can be translated as follows

- insa... - Inshallah PML-N shall succeed
- mjy pak... - I just can't understand the dances of Pakistanis.
- hanji... - Oh right, unearthed but kept his mouth shut till November, just quit it you shameless lot, you aren't our leader!
- lanna... - They are shameful people!
- video p... - This kind of nonsense looks good in videos, they need a fine and a beating as well

mBERT broke with the trend of the plurality of misclassifications being classed as neutral. The most common errors were made out as negative, classifying 1424 neutral samples as negative. mBERT also had a remarkably amount of positive examples classified as negative, making this mistake 1191 times. The model somewhat reversed the political trend found in BERT_Roman_Urdu, as well as that relating to religion. Of the positive and neutral samples classified as negative, many contained references to Islamic or Pakistani political figures. The first sample misclassification reflects this, as PML-N is the current administrative force in Pakistan.

mBERT classified neutral sentiments as positive 484 times, and very rarely did it ever classify negative sentiments as positive, only 288 times. The neutral sentences classified as positive were rife with idiom, sarcasm and use of English vocabulary in Urdu lexical structures. It may be reasonable to chalk up a degree of these errors to noise present in the data. Of the negative component, the models propensity to classify sentiments relating to Pakistani political parties as negative seemed not to extend to that of figures related to it. The mention of Nawaz Sharif, the former prime minister of the country in negative contexts consistently resulted in mBERT errantly predicting positive sentiments for the given sentence.

Of errantly predicting neutrality, the model classified positive samples as neutral 304 times and negative as neutral 239 times. The positive sentences classified as neutral seemed to reverse the pattern noticed in the neutral sentences classified as positive. Additionally, the model seemed to struggle with properly adjudging predicate sentences. Statements such as "Acha gana hai" - "It's a good song" were classified as neutral despite being clearly positive. The misclassifications of negative sentences into neutral followed similar patterns as their converses, with short predicate sentences often being misclassified, and sentences with large English content proportions also judged as neutral.

Table 11: Sample misclassifications of mBERT in Hindi

Sentence	**True Label**	**Predicted Label**
isse yah kebal ek hi samay men taiblet ko carj karne ke sath - sath eksatarnal hard draiv ko bhi urja de sakta hai	Neutral	Positive
is prastuti ne mujhe prerit kiya	Positive	Negative
leather strip ke saath yah zenwatch behad hi classy aur stylish hai	Positive	Neutral
yah nirnay puri tarah se galat aur asanvednsil tha.	Negative	Positive
kimat ke layak	Positive	Neutral

The example misclassifications of mBERT in Hindi, can be translated as follows

- isse... - This allows the cable to charge the tablet and power an external hard drive at the same time.
- is pras... - This presentation inspired me.
- leathe... - With the leather strip, this Zenwatch is immeasurably stylish and classy!
- yah nir... - This decision was completely wrong and insensitive.!

- kimat ke layak... - Worth its price

mBERT erred most in classifying neutral samples in Hindi, classifying them as positive 884 times and negative 646 times. Analysis of the positive classifications revealed that this was due to the description of features seeming like the provision of praise. Analysis of the negative judgements revealed a similar pattern, where the product reviews section of the Hindi database resulted in a measure of neutral reviews being pigeonholed into possessing emotional value. The propensity of mBERT to predict emotion in neutral cases can be seen as consistent with the findings of Chen et al, who demonstrate how emotionally charged vocabulary can result in the misclassification of neutral samples.

mBERT misclassified 522 positive samples as negative, and 372 positive samples as neutral. Of its errors in the positive class, the word "lekin" featured in 133 samples predicted as negative. This word translates roughly as a comparative between two ideas where the following idea is held in contradiction to its precedent. Of the 487 negative samples it classifies as positive, the word "lekin" is present in 93 of them. Further analysis of the presence of "lekin" in both classes reveals that it is indeed contained in sentences where sentiment related to one label is shortly contradicted by the other. The presence of these errors implies mBERT's sensitivity to balance contradictory information in the Hindi dataset. The remaining 314 errors consisted of negative samples misclassified as neutral, where the pattern discovered matched that across other models so far in the catalogue of classifying statements phrased in an objective-sounding manner as neutral regardless of emotive content.

Table 12: Sample misclassifications of MuRIL in Urdu

Sentence	**True Label**	**Predicted Label**
kon kon india se yai song sun raha hai i am from india katil fav singer rahat fatih ali khan	Neutral	Positive
katil intne izzat sach main btaye koe kam hai kya	Neutral	Negative
yai jab hakomat main a jatay han to america ky samny inke avay ni nikalti. katil	Negative	Positive
in sab ky bavajud halat ki kashedgi barkarar rahe or yai kashedgi us vakat uroj per pahoch gai jab benazir ky chotay bhai mir murtaza bhutto ko unke apni bahen ky ahde hakomat main ek intahai mashkok police makblay main sath satheon samat halak kardeya gaya	Negative	Neutral
eid main sab theak hojaiga	Positive	Negative

The example misclassifications of MuRIL in Urdu, can be translated as follows

- kon ko... - Who else is listening to this song from India! I'm from India My favourite singer is Rahat Fateh Ali Khan.
- katil... - You killer, be honest with such honour are these the kind of actions to be doing?
- yai jab h... - When they come into government, they show their incompetence in the face of American pressure, killers!

- in sa... - Despite all this, the tension of the situation continued and this tension reached its peak when Benazir's younger brother, Mir Murtaza Bhutto, was killed along with his brother in an extreme Mashkok police mausoleum during the rule of his own sister.!
- eid me... - Everything will be alright by Eid

MuRIL had a total of 4365 misclassifications across the Urdu dataset. A whopping 3132 of these misclassifications came through errantly classifying samples as positive. 1575 neutral examples were classified as positive, and 1558 negative ones. The nature of these misclassifications reveal some interesting patterns. A commonality shared by both misclassified subsets is that they contain large quantities of vocabulary that is derived from sources that are today found more commonly in Hindi poetic literature than in daily use, therefore potentially resulting in a positive skew of found text. Also common within this error subset are sentences strewn with religious terminology. Many negative prayers are predicted as positive, presumably either due to referring to god or due to being prayers in nature. Similarly, neutral descriptions of political or cultural figures are often predicted as positive, once again displaying some measure of counter-intuitively directioned biases.

The model assigned negative polarities at a relatively lower rate, something also visible in the error amounts. 552 neutral samples were misclassified as negative, and the model interpreted positive samples as negative 344 times. The word 'drama' appears in 53 neutral samples misclassified as negative. In the context of these reviews, drama likely refers to Pakistani drama shows, a subtlety MuRIL may here be failing to pick up on. The false negatives also revealed some insight into the errors the model makes. 48 of the 344 samples misclassified contained the term "katil", which translates directly to killer but is also used metaphorically as a superlative. In fact the usage of this word showed up in most contraindicative predictions in the Urdu dataset, signifying MuRIL's poor understanding of this metaphor.

MuRIL's aversion to assigning neutral polarity judgements in the Urdu language was reflected in the amount of erroneous neutral classifications. Only 169 positive and 167 negative samples were misclassified as neutral.

Table 13: Sample misclassifications of MuRIL in Hindi

Sentence	**True Label**	**Predicted Label**
main rozana apne bagice men phul dekhata hun	Neutral	Positive
utarte samay charge karana	Neutral	Negative
yah accha hai lekin upyog ke bad dant men dard hua	Neutral	Negative
apko pavar aur ritarn ki ko kam men lane ke lie inhen thora jyada pres karna hota hai	Negative	Positive
is muly sima par yah bahut accha hai lekin ismen keval ek samasya hai	Positive	Negative

The example misclassifications of MuRIL in Hindi, can be translated as follows

- men roz... - I see flowers in my garden every day.
- utarte... - Charge it as soon as you get off.
- yah acc... - This is good, but my teeth started to hurt after using it.

- apko pa... - You've got to apply some extra force to the power and return keys in order to make them work.
- is muly... - At the price point its great, but there's juust one issue

In Hindi, MuRIL had a total of 2460 misclassifications. The model misclassified neutral examples as positive 762 times, and negative ones 477 times. Across both Hindi and Urdu datasets, the model consistently displays the propensity to predict the positive class both more often than ground truth, and than all other classes. The misclassifications sorted as positive showed some interesting patterns. A great number of the neutral batch of these consisted of sentences that described the daily routines, or events transpiring in the lives of reviewers and thus while not inherently emotionally valenced in nature, did contain such vocabulary. Of the negative subset, the term "bahut" - very - occurred in 76 instances, signifying that the model displaying the propensity to classify superlatives positively, even erroneously.

Breaking from its performance across the Urdu dataset, MuRIL predicted more samples to be neutral in the Hindi corpus, albeit at times still erroneously. 417 positive samples were classified as neutral, and 322 of the negative samples misclassified were classed as neutral. These misclassifications were haphazard in nature, and provided no insights of note to this thesis.

As displayed in the confusion matrix in Figure 5, MuRIL achieved the best classification performance across Hindi in the negative class. This is thus reflected in negative misclassifications being the least occurring of the lot. Neutral samples were misclassified as negative 319 times, and positive samples in a mere 163 occurrences.

The differences displayed in rates across the catalogue can be said to constitute signs of differential treatment across language groups between the models. In addition to this, analysis of their aetiology reveals that while some common patterns in model errors cut across the catalogue, the individual natures of each models propensity to err varies more on factors within the language and model.

4.3 Bias

The aforementioned classification results were used in conjunction with the AIF360 toolkit as defined in Chapter 3.3 in order to compute the Disparate Impact Ratio, Statistical Parity Difference, Average Odds Difference and Demographic Parity Ratio. It is important to consider that in terms of differential treatment between models, values are computed between two distinct groups defined such that Group 0 refers to the Urdu language and Group 1 to the Hindi language. In this case values less than 0 (1 in the case of the DI ratio) are said to indicate bias towards the Urdu language, and values higher (than 1 in the case of the DI ratio) indicate bias towards the Hindi language.

Table 14: *
Bias metrics across each model

Metric	**BERT_Roman_Urdu**	**mBERT**	**MuRIL**
Disparate Impact Ratio	0.6592	1.1786	1.7054
Statistical Parity Difference	−0.1946	0.0942	0.2750
Average Odds Difference	−0.0486	0.0235	0.0688
Demographic Parity Ratio	0.6214	0.8241	0.6489

The ideal value for the Disparate Impact ratio is 1.0, implying perfectly equal treatment between groups. The standard for acceptable treatment varies, but is generally set to around 0.8-1.2. Both Bert_Roman_Urdu and MuRIL have Disparate Impact ratios outside this threshold, coming in at 0.66 and 1.70. Both values imply a level of bias. BERT_Roman_Urdu's score implies that a favourable judgement is more likely for speakers of the Urdu language, and MuRIL's score implies the same for Hindi. The score of mBERT, at 1.17 still implies some level of differential treatment, and is on the more laissez faire end of fairness approaches.

The Statistical Parity difference is a metric ideally at 0, where a variation of 0.1 is generally seen as acceptable. Once again, both models have test metrics outside of this threshold, with BERT_Roman_Urdu showing a bias towards Urdu at -0.19, and MuRIL towards Hindi at 0.27. mBERT was once again near the edge of the acceptable boundary, with a SPD score of 0.094. Both models therefore once again display signs of biased treatment towards their language at similar rates.

Similarly, the Average Odds Difference is ideally 0, with variation between positive and negative -0.1 In this case, all models are within this threshold. the score of BERT_Roman_Urdu being -0.04, with mBERT having a score of 0.023, and MuRIL having a score of 0.068.

The Demographic Parity difference is a metric ideally at 1, where a score of 1 indicates equal outcome regardless of group. The standard threshold is judging fair performance above 0.8. BERT_Roman_Urdu had a DPD score of 0.62, and MuRIL of 0.64 - therefore indicating the presence of bias in both models. mBERT was also quite close to the edge of the threshold, and its DPD score can show to indicate some signs of bias

The differences in classification performance are evidently clear in Table 6, and the models scores for these in the aforementioned fairness metrics derived from these strongly indicate the presence of linguistic bias in the models treatments of Hindi and Urdu. Out of four chosen fairness scores, three indicate the presence of bias in both BERT_Roman_Urdu and MuRIL. The means by which these vary by model and the patterns that are shown in the data are further treated in Chapter 5.1.

5 Discussion

In this section, the findings of the results of the experimental methodology as prior given in this thesis are discussed in order to ascertain the nature of their implications. This section also contains an overview of the structural limitations of this research, and possible avenues future research may take that have here been discovered.

5.1 Interpretations

The answer to RQ1 can be considered by analysing found differences in the results themselves. The results displayed different dimensions of sentiment classification performance across the model catalogue. Models achieved vastly different rates for accuracy and macro-F1 scores across Hindi and Urdu corpora, and displayed markedly different natures of classification decisions. The Urdu model displayed degraded performance across samples in the Hindi language, a fact that was vice versa replicated in the Hindi model. The degree of variation between the two languages in both Urdu and Hindi models was shown to be greater than that experienced in the multilingual baseline.

Analysis of the confusion matrices further demonstrates the extent of the differences between language groups across the model catalogue. The Hindi model was shown to possess a proclivity to classify sentiments from the Urdu dataset positively, regardless of true class identity. Alongside this MuRIL showed complete aversion to predicting sentiments in the neutral class in Urdu, a phenomenon that it did not replicate within the Hindi dataset. BERT_Roman_Urdu exhibiting similar classification behaviour with the neutral class, albeit its classification performance was absent of any under-represented class as in MuRIL. Another insight revealed amidst the comparison of classification metrics and confusion matrices across models was the performance of mBERT achieving higher rates of correct predictions in Hindi than in Urdu across all classes. Based off of these findings, it is possible to answer RQ1 by thus confirming that the performance of encoder-based models in Hindi and Urdu varies between languages.

In order to answer RQ2, namely to address whether the found differences can be said to constitute linguistic bias, the fairness of classification performance across language groups was measured using a quartet of AI Fairness tests, in which both BERT_Roman_Urdu and MuRIL had scores indicating bias in a variety of tests. The DI ratio, SPD and DPD scores for both models strongly indicated the presence of linguistic bias towards either language in both models. These indicated for BERT_Roman_Urdu that Hindi speaking users of this model face a provably worse outcome in sentiment analysis tasks than Urdu speaking users. The converse was also indicated for MuRIL, highlighting the extent to which this gap of unfair treatment between the two languages needs yet to be bridged. These findings therefore demonstrate that the found differences in classification performances can be said to constitute linguistic bias in model treatment of language. However, the presence of data originating from slightly different domains in the dataset does somewhat confound the validity of these differences being able to indicate the presence of bias, and therefore these findings are compared with the misclassifications to see if the patterns there indicate whether these found biases are a result of difference in domain or if they can truly be said to indicate bias.

Analysis of model misclassifications also revealed fascinating patterns of note. BERT_Roman_Urdu struggled with identifying neutral statements that were formatted as objective judgements or facts in both Hindi and Urdu. The model also displayed some signs of biases not directly targeted in this thesis in its misclassification patterns in both Urdu and Hindi. The model displayed some

signs of political biases and religious biases in Urdu, tending to misclassify samples referring to certain figures as positive, and some as negative. These biases reflect to some extent the sentiments found in broader Pakistani society - and while their presence is explicable - in consideration of prior research into the biases of NLP, merit further investigation [6] [19] [28]. The model also showed some extent of positive bias towards religious groups, where mention of religious terms spurred multiple false positives despite the content of those instances being instead imprecations. The model displayed the propensity to err in predicting neutral sentiments much more often in Hindi than Urdu. The errant misclassified samples revealed BERT_Roman_Urdu's tendency to anchor judgements off of the presence of words often found in other contexts in Urdu. Additionally, short sentences often threw the model off, explaining a large degree of the neutral predictions.

The misclassifications of mBERT reveal some interesting patterns. The majority of misclassifications in Urdu are made by the model predicting negative sentiments. Of these misclassifications, in a collection of cases once again the mention of political figures results in the presence of negative label outcome. In Hindi, the majority of errors arises instead through misclassifying sentiments positively. These differences also reveal interesting implications for outcomes even in models designated as being suited primarily for multilingual uses.

The misclassifications of MuRIL shed some light on the propensity of the model to predict positive sentiments in the Urdu language. The presence of vocabulary that may be considered niche in Hindi contexts, as well as the use of metaphors that do not translate as perfectly to Hindi speakers resulted in a non-trivial portion of positive misclassifications, but many positive misclassifications followed a trend observed in BERT_Roman_Urdu of predicting the polarity of samples containing religious terms as positive. Analysis of the Hindi subset indicated the models propensity to enforce a judgement on otherwise neutral sample text. This characteristic somewhat contextualises the models lack of neutral predictions across the Urdu dataset. In summation, the findings of error analysis across all three models indicate that while some of the patterns of the differences found are explainable due to noise in the datasets and model pitfalls not derived through bias, there are still a remarkable nature of differences in model treatment of each language group that arises due to subtle factors such as word register and idiom usage.

The fairness metrics in this study define favourable outcome as a correct classification and group membership as input language; under this definition the bias measures are closely tied to the differences in accuracy, and so embellish on the nature of that gap, they do not provide an additional source of testing for the presence of linguistic bias and do not in and of themselves account for the differences in domain. The Urdu and Hindi corpora are not parallel: they differ in domain and in the manner of romanisation. The error analysis shows that all three potential causes — differential treatment, topic mismatch, and transliteration noise — are active in the observed misclassifications. This thesis therefore finds the answer to RQ2 as follows: that while the observed differences appear consistent with linguistic bias, and at least partly attributable to it, this comes with the caveat that the present design is not able to fully separate bias from domain and romanisation confounds. These findings imply that the linguistic content used by Urdu or Hindi speaking users of NLP may experience differential treatment in sentiment analysis tasks using these models. These differences may result in allocational harms when used for downstream tasks, and therefore these results show the necessity of mitigating measures of these biases, as well as further research to confirm the extent to which this bias is free from domain confounds.

The implications of these findings stem from the applications of sentiment analysis systems as a whole. As prior demonstrated, we are seeing an increase in possibilities for these across various

fields: included but not limited to governance, healthcare and marketing. These systems underline key elements of everyday life and it is thus essential that their constituent parts are fair. The ability for users to receive fair treatment regardless of linguistic background is necessary. The differences between languages found across models discussed shortly prior paints a picture of the opposite scenario for speakers of Hindi and Urdu.

This unequal treatment can result in alarming consequences. In order to better illustrate this statement, let us consider some example applications. Recent developments have resulted in certain states adopting policies of monitoring the digital presence of prospective entrants into their borders. In this context, use of either MuRIL or BERT_Roman_Urdu would result in provably different outcomes for speakers of either other language, something that may adversely impact the fairness of admission they may otherwise receive. Similarly, usage of these models within state lines may result in speakers from minority groups receiving unfair treatment in measuring indication of public services, response to natural or engineered disasters, or in determining public opinion regarding policy proposals. Usage of these models in healthcare carries the same risk of automated systems missing out on the emotional needs or opinions of speakers of the opposite language. These outcomes require stringent mitigation, in order best to alleviate these adverse results. Assuming the degree of bias is consistent across a test set designed to control for domain and romanisation confounds, this thesis finds that usage of these models in situations in which both Hindi and Urdu speech is considered may result in a biased outcome.

5.2 Limitations

The core of the limitations that may constrain the generalisability of this research arise from the nature of the data, the fine-tuning procedure taken for the task, as well as the training of the models themselves. Among some key limitations experienced within the body of this research arose within the datasets of this research. Since the experience of romanised Hindi and Urdu speakers was sought to be captured, datasets were collected with a preference for romanised text. This proved to be easier for Urdu than Hindi. The quantity of samples of the few initially collected Urdu datasets far exceeded that of the many Hindi datasets that had been accumulated in order to attempt at a fair distribution. Due to structural limitations in sourcing romanised Hindi texts, the inclusion of texts in the Devanagari script was necessitated.

As a result of this, the additional step of transliterating these texts and the algorithmic handling of the romanised Hindi text in order to maintain a similar standard to vernacular use, as seen in the Urdu dataset resulted in the detraction of a layer of representativeness of the Hindi dataset. However, this controlled romanisation was conducted in such a manner as to maintain the greatest degree of fidelity between linguistic group representations. The representativeness of self-romanised texts that the Urdu dataset enjoys does not however come without its own pitfalls. Since the romanisation was user-applied across all instances, it does not conform to a common standard, resulting in the presence of several different variations of the same word being used in different scenarios in the dataset. An additional limitation this romanisation incurs is the loss of some orthographic cues e.g. the voiceless retroflex plosive is written as a T, which is the same as the English voiceless alveolar plosive t. This may result in some words being misconstrued, but is a source of noise across both datasets.

Additionally, while the collection of texts in both final processed corpora originated from user reviews marked as either positive, neutral or negative, their exact origins vary from Urdu having

reader reviews sourced solely from news sites, whereas the Hindi corpus contains reviews of products, cinema, personalities and other such topics. The difference in topics does result in some degree of the context being different across datasets, as well as there being a difference in the emotive nature of each languages data corpus.

In order for the models to be prepared for the classification task, they required fine-tuning. While care was taken to ensure that this was done on a partition of the data not used in the final task, and that the distribution of labels somewhat matched the test set distribution, this was still done on data romanised to the aforementioned self and algorithmically performed standard. This does raise some questions on the generalisability of the findings, and further research on the matter is warranted in order to address this issue. Even after fine-tuning, the models received low accurate accuracy in the other language. While testing of this on different subsets of Hindi and Urdu data revealed this to be a consistent feature of either model's architecture, it is possible that improper training of the model may result in performance disparities. However, this is a concept that is captured within this thesis' conception of linguistic bias.

Additionally, the use of AIF360 as the bias detection framework for this thesis resulted in the loss of some information regarding the nature of bias. This occurred namely due to the AIF360 framework resulting in a binary outcome (Fair/Unfair), which loses a degree of information into the nature of the misclassification that was used to derive this outcome e.g. BERT_Roman_Urdu's proclivity and MuRIL's aversion to assigning the neutral sentiment, or other model specific prediction flips. While this is somewhat addressed in the qualitative segment of the error analysis, it remains still as insight lost in the final quantitative judgement of bias using this framework.

A limitation that may somewhat limit the generalisation of this research to more broader contexts is the presence of solely one model per language in comparing between each language. While this arises somewhat out of the lack of passable models trained for Urdu or Hindi (or both), the caveat that these biases may be limited specifically to these models is one that should be addressed in further research.

5.3 Further Research

Future work may extend upon this research by comparing for differences on a more similar test set across both languages, where sentences are created using a template that uses language-specific terms per each group, using a variation of the SEAT framework introduced by May et al. [26]. Even if this methodology is not followed, research centred on replicating these findings on a more symmetric test set would serve as valuable extensions of the findings of this thesis.

The findings of this thesis may also be embellished upon with the adaptation of the framework of Camara et al, by building a test set in both Hindi and Urdu languages that may be used to test for bias in model treatment via variations in responses to samples that contain vocabulary with direct counterparts in both languages. Additionally, the error analysis of BERT_Roman_Urdu and MuRIL revealed reason to suspect some degree of religious and identity based biases - this research could thus be expanded by including a dimension of including terms associated with different identity groups in both societies to by investigating potential variations in these the presence of these biases may be either confirmed or denied.

The generalisability of this research can either be strengthened or denied by a further study that carries out this comparison between a wider catalogue of models trained for both languages, that may better establish the presence and degree of the found bias within this thesis.

The research body forming the background of this thesis reveals a severe lack of studies of bias in the Urdu language altogether. Further studies are therefore warranted on investigating the extent to which proven biases in other language's NLP approaches extend also to the Urdu language. Similarly, there is a severe dearth of equity evaluation testing suites available for the Urdu language that have been developed in English, and in other languages. Further work designed to develop this is a valuable necessity in the pursuit of fairer AI systems.

5.4 Conclusions

This thesis aimed to capture the degree of linguistic bias present in sentiment analysis models trained in Hindi and Urdu. In order to do this, the thesis aimed to investigate whether or not performance in a polarity classification task varied between models, and whether these differences indicated signs of linguistic. In order to investigate this, this thesis aimed two answer two research questions, the first question aimed at capturing whether sentiment classification models achieve different performance across the Hindi and Urdu language, and the latter on whether these differences constitute linguistic bias. RQ1 was answered in the presence of differential treatment across the findings, with further analysis of classification performance and fairness testing indicating that found differences were indicative of linguistic bias between language groups in each language. However, further analysis into the findings also revealed that the degree of bias discovered could not be fully separated from domain and transliteration based confounds in the test dataset, thereby answering RQ2.

The presence of this bias indicates that under the current situation of computational resources in use in society, the usage of those in Hindi and Urdu contexts carry with them the presence of unfair outcomes for the speech community of each language across communities. Usage of sentiment analysis systems has been applied across various fields, by state and non-state actors. The findings of this thesis indicate that such uses of these models in Hindi or Urdu speaking contexts may result in biased outcomes for cross-lingual speakers. However, the extent to which this holds applicable should be isolated by determining if the findings are replicated across a domain controlled dataset romanised in a standard manner across both languages.

References

[1] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, (Barcelona, Spain), pp. 271–278, July 2004.

[2] A. Zunic, P. Corcoran, and I. Spasic, "Sentiment analysis in health and well-being: systematic review," *JMIR medical informatics*, vol. 8, no. 1, p. e16023, 2020.

[3] J. Berger, A. Humphreys, S. Ludwig, W. W. Moe, O. Netzer, and D. A. Schweidel, "Uniting the tribes: Using text for marketing insight," *Journal of Marketing*, vol. 84, no. 1, pp. 1–25, 2020.

[4] R. M. Rosa, J. D. de Souza, and C. d. N. Freitas, "Development and deployment of sentiment analysis ai on citizens' feedback in goiás," *Conference on Digital Government Research*, vol. 26, May 2025.

[5] J. A. van Dijk, "Digital divide research, achievements and shortcomings," *Poetics*, vol. 34, no. 4, pp. 221–235, 2006. The digital divide in the twenty-first century.

[6] T. Bolukbasi, K. Chang, J. Y. Zou, V. Saligrama, and A. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," *CoRR*, vol. abs/1607.06520, 2016.

[7] C. Basta, M. R. Costa-jussà, and N. Casas, "Evaluating the underlying gender bias in contextualized word embeddings," in *Proceedings of the First Workshop on Gender Bias in Natural Language Processing* (M. R. Costa-jussà, C. Hardmeier, W. Radford, and K. Webster, eds.), (Florence, Italy), pp. 33–39, Association for Computational Linguistics, Aug. 2019.

[8] H. An, C. Acquaye, C. Wang, Z. Li, and R. Rudinger, "Do large language models discriminate in hiring decisions on the basis of race, ethnicity, and gender?," *ACL Anthology*, 2024.

[9] G. Curto, M. F. Jojoa Acosta, F. Comim, and B. Garcia-Zapirain, "Are ai systems biased against the poor? a machine learning analysis using word2vec and glove embeddings," *AI society*, vol. 39, no. 2, pp. 617–632, 2024.

[10] A. C. Curry, G. Attanasio, Z. Talat, and D. Hovy, "Classist tools: Social class correlates with performance in nlp," *ACL Anthology*, 2024.

[11] E. Fleisig, G. Smith, M. Bossi, I. Rustagi, X. Yin, and D. Klein, "Linguistic bias in chatgpt: Language models reinforce dialect discrimination," *ACL Anthology*, 2024.

[12] E. M. Bender, "Linguistically naïve != language independent: Why NLP needs linguistic typology," in *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?* (T. Baldwin and V. Kordoni, eds.), (Athens, Greece), pp. 26–32, Association for Computational Linguistics, Mar. 2009.

[13] N. Sharma, K. Murray, and Z. Xiao, "Faux polyglot: A study on information disparity in multilingual large language models," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (L. Chiruzzo, A. Ritter, and L. Wang, eds.), (Albuquerque, New Mexico), pp. 8090–8107, Association for Computational Linguistics, Apr. 2025.

[14] R. Phillipson, "Linguistic imperialism / r. phillipson.," 12 2018.

[15] B. Liu, *Sentiment analysis : mining opinions, sentiments, and emotions*. Studies in natural language processing, Cambridge: Cambridge University Press, second edition. ed., 2020.

[16] L. Lei and D. Liu, *Conducting sentiment analysis*. Cambridge elements. Elements in corpus linguistics, Cambridge: Cambridge University Press, 1st ed. ed., 2021.

[17] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational linguistics*, vol. 37, no. 2, pp. 267–307, 2011.

[18] S. Kiritchenko and S. M. Mohammad, "Examining gender and race bias in two hundred sentiment analysis systems," *ACL Anthology*, 2018.

[19] S. Jentzsch and C. Turan, "Gender bias in BERT - measuring and analysing biases through sentiment rating in a realistic downstream classification task," in *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)* (C. Hardmeier, C. Basta, M. R. Costa-jussà, G. Stanovsky, and H. Gonen, eds.), (Seattle, Washington), pp. 184–199, Association for Computational Linguistics, July 2022.

[20] A. C. Islam, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora necessarily contain human biases," *CoRR*, vol. abs/1608.07187, 2016.

[21] K. Kurita, N. Vyas, A. Pareek, A. W. Black, and Y. Tsvetkov, "Measuring bias in contextualized word representations," in *Proceedings of the First Workshop on Gender Bias in Natural Language Processing* (M. R. Costa-jussà, C. Hardmeier, W. Radford, and K. Webster, eds.), (Florence, Italy), pp. 166–172, Association for Computational Linguistics, Aug. 2019.

[22] A. Câmara, N. Taneja, T. Azad, E. Allaway, and R. Zemel, "Mapping the multilingual margins: Intersectional biases of sentiment analysis systems in English, Spanish, and Arabic," in *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion* (B. R. Chakravarthi, B. Bharathi, J. P. McCrae, M. Zarrouk, K. Bali, and P. Buitelaar, eds.), (Dublin, Ireland), pp. 90–106, Association for Computational Linguistics, May 2022.

[23] K. Dashtipour, S. Poria, A. Hussain, E. Cambria, A. Y. A. Hawalah, A. Gelbukh, and Q. Zhou, "Multilingual sentiment analysis: State of the art and independent comparison of techniques," *Cognit. Comput.*, vol. 8, pp. 757–771, June 2016.

[24] S. Goldfarb-Tarrant, A. Lopez, R. Blanco, and D. Marcheggiani, "Bias beyond English: Counterfactual tests for bias in sentiment analysis in four languages," in *Findings of the Association for Computational Linguistics: ACL 2023* (A. Rogers, J. Boyd-Graber, and N. Okazaki, eds.), (Toronto, Canada), pp. 4458–4468, Association for Computational Linguistics, July 2023.

[25] I. U. Khan, A. Khan, W. Khan, M. M. Su'ud, M. M. Alam, F. Subhan, and M. Z. Asghar, "A review of urdu sentiment analysis with multilingual perspective: A case of urdu and roman urdu language," *Computers*, vol. 11, no. 1, 2022.

[26] C. May, A. Wang, S. Bordia, S. R. Bowman, and R. Rudinger, "On measuring social biases in sentence encoders," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), (Minneapolis, Minnesota), pp. 622–628, Association for Computational Linguistics, June 2019.

[27] L. C. L. Gamboa, Y. Feng, and M. Lee, "Social bias in multilingual language models: A survey," Sept. 2025.

[28] N. Sahoo, N. Mallela, and P. Bhattacharyya, "With prejudice to none: A few-shot, multilingual transfer learning approach to detect social bias in low resource languages," in *Findings of the Association for Computational Linguistics: ACL 2023* (A. Rogers, J. Boyd-Graber, and N. Okazaki, eds.), (Toronto, Canada), pp. 13316–13330, Association for Computational Linguistics, July 2023.

[29] S. Levy, N. John, L. Liu, Y. Vyas, J. Ma, Y. Fujinuma, M. Ballesteros, V. Castelli, and D. Roth, "Comparing biases and the impact of multilingual training across multiple languages," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (H. Bouamor, J. Pino, and K. Bali, eds.), (Singapore), pp. 10260–10280, Association for Computational Linguistics, Dec. 2023.

[30] O. Smirnov, "The language you ask in: Language-conditioned ideological divergence in LLM analysis of contested political documents," Feb. 2026.

[31] E. Wong and F. M'hiri, "Analyzing language bias between french and english in conventional multilingual sentiment analysis models," 04 2026.

[32] S. Mody, *The Making of Modern Hindi: Literary Authority in Colonial North India*. 10 2018.

[33] R. Delacy, S. Ahmed, and L. P. P. (Firm), *Hindi, Urdu & Bengali*. LP Phrasebook Guides, Lonely Planet, 2011.

[34] C. R. King, *One Language, Two Scripts: The Hindi Movement in Nineteenth Century North India*. Bombay: Oxford University Press, 1994.

[35] A. Ranjan, "Language, religion, and identity: Hindi and urdu in colonial and post-colonial india," *India Review*, vol. 21, no. 3, pp. 286–306, 2022.

[36] V. Malik, S. Dev, A. Nishi, N. Peng, and K.-W. Chang, "Socially aware bias measurements for Hindi language representations," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, eds.), (Seattle, United States), pp. 1041–1052, Association for Computational Linguistics, July 2022.

[37] I. Gupta, I. Joshi, A. Dey, and T. Parikh, ""since lawyers are males..": Examining implicit gender bias in hindi language generation by LLMs," in *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, (New York, NY, USA), pp. 3254–3264, ACM, June 2025.

[38] K. Kumar, S. Mendke, R. Parihar, S. Mayya, S. Venkatesh, and S. G. Koolagudi, "DBNLP: detecting bias in natural language processing system for india-centric languages," *Int. J. Inf. Technol.*, vol. 17, pp. 3291–3306, July 2025.

[39] N. Sambasivan, E. Arnesen, B. Hutchinson, T. Doshi, and V. Prabhakaran, "Re-imagining algorithmic fairness in india and beyond," 2021.

[40] S. Agrawal, K. Gupta, D. Gautam, and R. Mamidi, "Towards detecting political bias in Hindi news articles," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop* (S. Louvan, A. Madotto, and B. Madureira, eds.), (Dublin, Ireland), pp. 239–244, Association for Computational Linguistics, May 2022.

[41] M. R. Ashraf, Y. Jana, Q. Umer, M. A. Jaffar, S. Chung, and W. Y. Ramay, "Bert-based sentiment analysis for low-resourced languages: A case study of urdu language," *IEEE Access*, vol. 11, pp. 110245–110259, 2023.

[42] B. Danet and S. C. Herring, *The multilingual Internet : language, culture, and communication online*. Oxford ;: Oxford University Press, 2007.

[43] R. Ta and N. T. Lee, "How language gaps constrain generative ai development," 2023.

[44] P. R. Shetty, "Hindi sentiment dataset," 2024.

[45] Udrasht, "Hindi sentiment analysis corpus from amazon reviews."

[46] K. Pitroda, "Iit patna movie reviews hindi sentiment analysis," Aug 2023.

[47] M. Irfan, "Mirfan899/urdu: Collection of urdu datasets for pos, ner, sentiment, summarization and nlp tasks.."

[48] Ahsan, "Roman urdu dataset," Mar 2020.

[49] M. Ahmed, "Roman urdu word and reviews dataset.," Apr 2024.

[50] Aimlab, "Aimlab/sentiment-analysis-roman-urdu at main."

[51] S. Khanuja, D. Bansal, S. Mehtani, S. Khosla, A. Dey, B. Gopalan, D. K. Margam, P. Aggarwal, R. T. Nagipogu, S. Dave, S. Gupta, S. C. B. Gali, V. Subramanian, and P. Talukdar, "Muril: Multilingual representations for indian languages," 2021.

[52] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[53] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, "Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," 2018.

[54] J. Chen, S. Karimi, D. Molla, and C. Paris, "To labor is not to suffer: Exploration of polarity association bias in LLMs for sentiment analysis," in *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics* (K. Inui, S. Sakti, H. Wang, D. F. Wong, P. Bhattacharyya, B. Banerjee, A. Ekbal, T. Chakraborty, and D. P. Singh, eds.), (Mumbai, India), pp. 70–78, The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, Dec. 2025.