BOOTSTRAPPING SENTIMENT ANALYSIS AND MITIGATING BIAS IN LOW-RESOURCE LANGUAGES

A thesis submitted in partial fulfillment

for the Award of Degree of

Doctor of Philosophy

by

Satyam Dwivedi

Department of Humanistic Studies

Indian Institute of Technology (BHU)

Varanasi - 221005

Roll Number: 15191502

Year of Submission: 2025

CERTIFICATE

It is certified that the work contained in the thesis titled "**Sentiment Analysis and Mitigating Bias in Low Resource Languages**" by **Satyam Dwivedi** has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

It is further certified that the student has fulfilled all the requirements of Comprehensive Examination, Candidacy and SOTA for the award of PhD Degree.

Sanjukta 19/3/25

Signature of the Supervisor

Affiliation

सहयुक्त आचार्य/Associate Professor
मानविकी अध्ययन विभाग/Department of Humanistic Studies
भारतीय प्रौद्योगिकी संस्थान/Indian Institute of Technology
(काशी हिन्दू विश्वविद्यालय)/(Banaras Hindu University)
वाराणसी-२२१००५ (उ०प्र०)/Varanasi-221005 (U.P.)

DECLARATION BY THE CANDIDATE

I, **Satyam Dwivedi**, certify that the work embodied in this thesis is my own bona fide work and carried out by me under the supervision of **Dr. Sanjukta Ghosh** from 2016 to 2025, at the **Department of Humanistic Studies, Indian Institute of Technology (BHU)** Varanasi. The matter embodied in this thesis has not been submitted for the award of any other degree/diploma. I declare that I have faithfully acknowledged and given credits to the researchers wherever their works have been cited in my work in this thesis. I further declare that I have not willfully copied any other's work, paragraphs, text, data, results, etc., reported in journals, books, magazines, reports dissertations, theses, etc., or available at websites and have not included them in this thesis and have not cited as my own work.

Satyam Dwivedi

Date: March 19, 2025 **Signature of the Student**

Place: Varanasi, India Satyam Dwivedi

CERTIFICATE BY THE SUPERVISOR

It is certified that the above statement made by the student is correct to the best of my knowledge.

Sanjukta 19/3/25

Signature of the Supervisor

Affiliation

सहयुक्त आचार्य/Associate Professor
मानवतावादी अध्ययन विभाग/Department of Humanistic Studies
भारतीय प्रौद्योगिकी संस्थान/Indian Institute of Technology
(काशी हिन्दू विश्वविद्यालय)/(Banaras Hindu University)
वाराणसी-२२१००५ (उ०प्र०)/Varanasi-221005 (U.P.)

19/03/2025

Signature of the Head of Department

Head
for Department of Humanistic Studies
Indian Institute of Technology
(Banaras Hindu University)
VARANASI-221005 (U.P.)

COPYRIGHT TRANSFER CERTIFICATE

Title of the Thesis: **Sentiment Analysis and Mitigating Bias in Low Resource Languages**

Name of the Student: **Satyam Dwivedi**

COPYRIGHT TRANSFER

The undersigned hereby assigns to the Indian Institute of Technology (BHU) Varanasi all rights under copyright that may exist in and for the above thesis submitted for the award of the PhD degree.

Satyam Dwivedi

Date: March 19, 2025 **Signature of the Student**

Place: Varanasi, India Satyam Dwivedi

Note: However, the author may reproduce or authorize others to reproduce material extracted verbatim from the thesis or derivative of the thesis for author's personal use provided that the source and the Institute's copyright notice are indicated.

त्वदीयं वस्तु गोविन्द तुभ्यमेव समर्पये।

O Govinda! What is Yours, I offer back to You.

सदाशिवसमारम्भाम् शङ्करचार्यमध्यमाम्।
अस्मदाचार्यपर्यन्ताम् वन्दे गुरुपरम्पराम्॥

**I bow to the lineage of teachers, beginning with Lord Sadāśiva,
passing through Ādi Śaṅkarācārya, and continuing up to my own preceptor.**

Acknowledgement

When I look back, my life feels like a series of plots and subplots—woven together by people, places, and situations, each forming a unique recipe of effort, experience, and outcomes. Different chapters unfolded across different geographies, each a distinct box of time and energy focused on specific goals. Among them all, this research journey stands out as the most intense—filled with the highest concentration of dramatic events, focused efforts, ups, and downs. It has been a box like no other, and completing it was only possible because of the collective contributions of everyone who walked this path with me.

I am deeply grateful to my advisor, Dr. Sanjukta Ghosh, for her unwavering support, patience, and persistence throughout this journey. The freedom she allowed me created the space I needed to explore ideas, grow, and bring out the best in myself. This work would not have been possible without her guidance.

A special thanks to Dr. Anil Thakur for his tremendous support on countless occasions and for the occasional but profound wisdom he shared—ranging from reflections on the self and society to education, educator, and everything in between. Those conversations opened new perspectives for me and taught me to look beyond the surface. One of his witty remarks that has stayed with me is: "Only those who have a PhD line on their palm complete this journey." I may not know much about palmistry, but it seems that somehow, I developed that line after all.

I am grateful to Dr. R. K. Mishra, Dr. Ajit Kumar Mishra, Dr. Sukomal Pal, Dr. Chandan Upadhyay, and Dr. Rajaram Shukla for their support, valuable insights, and rigorous evaluation of

my work. Their constructive feedback and guidance have played a crucial role in shaping this journey.

To my parents and my brother, my mirror image—your constant emotional and mental support through every high and low gave me the courage and strength to navigate this demanding journey.

To my better half—thank you for being my anchor, my reminder, and my courage. Your presence gave me the strength to see this journey through to its end.

Lastly, my heartfelt gratitude to friends and peers who made this journey lighter and brighter with shared laughter, countless memories, and support. Along the way, I built bonds beyond bloodlines—relationships that I will always hold close as my most cherished treasures.

Satyam Dwivedi

March 2025, Varanasi

Preface

In the fast-evolving domain of Artificial Intelligence (AI), Natural Language Understanding (NLU) is instrumental in enabling seamless human-machine interaction. With advancements in Machine Learning (ML) techniques such as Transfer Learning and Task-Efficient Fine-Tuning, along with the rise of transformer-based architectures like Large Language Models (LLMs) demonstrating emergent capabilities, AI systems are shown to efficiently handle tasks and languages with little to no prior exposure. This progress has driven both academia and industry to adopt and productionize off-the-shelf models trained on high-resource languages, such as English, for applications in less widely studied languages like Hindi, Thai, and Vietnamese among others. As these Language Models (LMs) become more deeply integrated into everyday applications, it is crucial to identify and address their inherent challenges to ensure their equitable accessibility and adaptability, particularly in low-resource linguistic settings. This thesis examines NLU applications with a specific focus on Sentiment Analysis and Bias Identification and Mitigation techniques in low-resource languages. The research explores strategies for bootstrapping Sentiment Analysis with finite data, including developing models from scratch and adapting existing ML models. It further examines the linguistic, cultural, and ethical challenges of using off-the-shelf models to low-resource or nearly unseen languages, where limited data availability may hinder ML models' effectiveness and adaptability. A key contribution of this study is assessing the extent to which NLU models trained on high-resource languages can be effectively

adapted to Hindi and other Indian languages, revealing their capabilities, limitations, and inherent biases.

The motivation for this research stems from the researcher's prior experience in bootstrapping Machine Translation tools for low-resource languages, involving key processes such as data curation, synthetic data generation, and subsequently the development of both research and production-level ML models. A major challenge in this process these days is the heavy reliance on pre-trained LMs, which are primarily trained on high-resource languages and inherently influenced by their cultural contexts. These models often struggle to generalize effectively to languages with distinct syntactico-semantic structures, socio-cultural nuances, and contextual interpretations, leading to misrepresentation and underperformance in low-resource linguistic settings. Moreover, biases embedded in these models, including gender bias, societal stereotypes, and cultural misrepresentations, pose serious challenges to fairness, inclusivity, and ethical usage. Addressing these issues requires a comprehensive investigation into NLU applications, alongside the development of Bias Identification and Mitigation strategies specifically designed for low-resource languages, ensuring that these models are both linguistically adaptable and socially responsible.

The research first delves into Sentiment Analysis, exploring its various levels, tasks, and methodologies. It provides a structured analysis of different approaches, including:

1. Lexicon-based methods, which rely on predefined sentiment dictionaries,
2. Corpus-based approaches, which leverage annotated datasets for sentiment classification, and
3. ML techniques, including both traditional models and modern deep-learning architectures.

A major focus is placed on the challenges associated with Sentiment Analysis in low-resource languages, including ambiguity, sarcasm, negation, multipolarity, and lack of annotated data. The study evaluates how different rule-based and statistical models perform in sentiment classification tasks across multiple Indian languages, with a particular focus on Hindi and other low-resource Indian languages. Through extensive experimentation, the research uncovers language-specific patterns and limitations, offering novel insights into how NLU models can be optimized for sentiment analysis in linguistically diverse and resource-scarce environments.

The research extends its focus beyond sentiment analysis to examine Bias Identification and Mitigation in NLU models, a growing concern in AI. The study conducts two experimental investigations:

1. Evaluating Bias in Small-Scale LMs– Testing tools such as Flair, TextBlob, and Vader for gender and societal biases, analyzing how they interpret and process linguistic inputs in low-resource languages.
2. Assessing Bias in LLMs– Investigating how state-of-the-art LLMs such as BARD, GPT, and LLAMA generate biased outputs when used in Hindi and other Indian languages.

Through qualitative and quantitative evaluations, the study highlights systematic biases present in these models, particularly in gender representation, cultural depictions, and implicit stereotypes. The research also explores novel bias mitigation techniques, including:

1. Template-driven synthetic data generation, used to counteract model biases by creating balanced datasets,
2. Prompt engineering and in-context learning, which involve restructuring input and incorporating on-the-go contextual information to minimize biased outputs, and

3. Explainability methods, which help interpret model predictions and identify underlying biases.

The experiments reveal that bias identification and mitigation techniques can significantly improve fairness, but also underscore the complex trade-offs between bias reduction, model performance, and language adaptability. These findings contribute to the broader discussion of Responsible AI and the necessity of context-aware bias detection frameworks for low-resource languages.

The research culminates in a synthesis of key findings, presenting guidelines for ethical NLU model development and proposing advancements in bias detection methodologies. It underscores the necessity of multidisciplinary collaborations to ensure AI systems are inclusive and contextually aware. Additionally, the study outlines future research directions, particularly in bias mitigation strategies for underrepresented linguistic communities.

List of Figures

List of Tables

WX Transcription Notation

Vowels

अ	आ	इ	ई	उ
A	A	i	I	u
ऊ	ए	ऐ	ओ	औ
U	e	E	o	O
ऋ	ॠ	ऌ	ं	ः
q	Q	L	M	H

Consonants

क्	ख्	ग्	घ्	ङ्
k	K	g	G	**f**
च्	छ्	ज्	झ्	ञ्
c	C	j	J	**F**
ट्	ठ्	ड्	ढ्	ण्
t	T	d	D	N
त्	थ्	द्	ध्	न्
w	W	x	X	n
प्	फ्	ब्	भ्	म्
p	P	b	B	m
य्	र्	ल्	व्	
y	r	l	v	
श्	ष्	स्	ह्	
S	R	s	h	

List of Abbreviations

Abbreviation	Full-form
ADJ	Adjective
ADV	Adverb
AI	Artificial Intelligence
ANOVA	Analysis of Variance
BERT	Bidirectional Encoder Representations from Transformers
COCA	Corpus of Contemporary American English
F1	Harmonic Mean of Precision and Recall
FN	False Negative
FP	False Positive
FST	Finite State Transducer
GPT	Generative Pre-trained Transformer
GPU	Graphics Processing Unit
ICL	In-Context Learning
LHS	Left-Hand Side
LIME	Local Interpretable Model-agnostic Explanations
LLM	Large Language Model
LM	Language Model
LR	Lexical Rules
ML	Machine Learning
N	Noun (Syntactic), Negative (Accuracy Metrics)
NEG	Negations
NER	Named Entity Recognition
NLU	Natural Language Understanding
NP	Noun Phrase
P	Pre/Post-position (Syntactic), Positive (Accuracy Metrics)
PE	Prompt Engineering
PFT	Pre-trained Fine-Tuning
PNG	Person Number Gender
PoS	Parts of Speech
PRON	Pronoun
RegEx	Regular Expression
RHS	Right-Hand Side
RQ	Research Question

SC	Strong Construction
SFT	Supervised Fine-Tuning
SHAP	SHapley Additive exPlanations
SI	Subjectivity Index
SLM	Small Language Model
SoTA	State-of-the-Art
STEM	Science, Technology, Engineering, Mathematics
TN	True Negative
TP	True Positive
TPU	Tensor Processing Unit
V	Verb
VADER	Valence Aware Dictionary and sEntiment Reasoner
VAUX	Auxiliary Verb
VP	Verb Phrase
WC	Weak Construction
WSD	Word Sense Disambiguation