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.