Preface

The rapid growth of social media has led to an unprecedented volume of user-generated content, characterized by the frequent use of code-mixing, where speakers alternate between two or more languages within a single conversation or even within a sentence. This phenomenon presents unique challenges for natural language processing (NLP) systems, which are typically designed for monolingual texts. Consequently, there is an urgent need for advanced text processing techniques tailored to handle code-mixed data effectively. This dissertation addresses this gap by exploring various applications of code-mixed text processing, including word-level language identification, sentiment analysis, hate speech detection, and information retrieval.

First, we focus on the fundamental task of language identification at the word level within code-mixed sentences. This research compares non-contextual input representations (Word2Vec, GloVe, FastText) with contextual input representation (BERT) for automatically identifying languages in code-mixed texts. Employing a deep learning framework that leverages pre-trained models for embedding, we evaluate our approach across six datasets encompassing three language pairs: Bengali-English, Hindi-English, and Spanish-English. Our Bi-LSTM model, built on top of BERT representations, emerges as the best-performing model, significantly outperforming classical approaches.

Second, we investigate sentiment analysis in code-mixed data from three Dravidian language pairs (Malayalam-English, Tamil-English, and Kannada-English), sourced from social media (YouTube) comments. The text is categorized into five sentiment

classes. We emphasize the importance of accurate language identification to enhance sentiment analysis performance. Our hierarchical model, incorporating a language identification module and mBERT, demonstrates improved performance in terms of weighted average F_1 scores, suggesting that addressing sentiment analysis as a multi-level problem offers notable advantages.

Our third work tackles the challenge of detecting hate speech and offensive content within Hindi-English code-mixed social media conversations. We decompose the task into binary and multi-class classifications, employing deep learning models such as BERT and an ensemble model comprising a fine-tuned mBERT and a sentence transformer. By adopting distinct feature extraction strategies for posts, comments, and replies, we effectively manage the linguistic diversity and variable text lengths inherent in our dataset, leading to enhanced classification accuracy.

Finally, we delve into information retrieval from Bengali-English code-mixed data, proposing a language-identification-based solution with query expansion using phonetic encoding. This approach yields promising results, outperforming baseline models. We advocate for the development of a code-mixed stop words list to optimize information retrieval. Our research highlights the necessity of domain-specific strategies, especially in the dynamic landscape of social media.

In summary, this dissertation contributes to the burgeoning field of code-mixed text processing by addressing key challenges and proposing innovative solutions across multiple applications. Our findings lay a robust foundation for future research and development in this critical area.