Text Processing on Code-Mixed Social Media Data

कोड-मिश्रित सोशल मीडिया डेटा पर टेक्स्ट प्रोसेसिंग

Thesis submitted in partial fulfillment
for the Award of Degree

Doctor of Philosophy

by

Supriya Chanda

सुप्रिय चन्द

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY
(BANARAS HINDU UNIVERSITY)
VARANASI - 221005

Roll No. **18071008** Year **2024**

Chapter 8

Conclusions and Future Work

This dissertation presents several foundational contributions to enhance text processing in code-mixed social media data. The research encompasses various applications, ranging from word-level language identification to sentence-level tasks such as sentiment analysis, hate speech detection, and information retrieval in code-mixed texts. A range of deep learning techniques are employed on existing datasets to assess the efficacy of current systems across various tasks. Also, a novel dataset for code-mixed information retrieval is created, and several retrieval related experiments are conducted.

A primary focus of this research is on language identification at the word level, a crucial task for processing code-mixed datasets. Significant contributions include the comparison of non-contextual input representations (Word2Vec, GloVe, FastText) with contextual input representation (BERT) in automatically identifying languages within code-mixed data. The proposed approach utilizes a deep learning framework that leverages pre-trained models for embedding, providing a pioneering solution to this challenge. Evaluation of the model spans six distinct datasets, covering three language pairs: Bengali-English, Hindi-English, and Spanish-English. The results indicate that our Bi-LSTM model on top of BERT neural representations of code-mixed data is the best-performing model.

This thesis also explores sentiment analysis for code-mixed data from Dravidian language pairs (Malayalam-English, Tamil-English, and Kannada-English), sourced from social media comments. The text is categorized into five classes: Positive, Negative, Mixed Feelings,

Unknown State, and Not in Language. The research highlights the importance of language identification in enhancing sentiment analysis performance. The study demonstrates that a hierarchical model with an LID and mBERT module improves weighted average F_1 scores.

In the domain of hate speech detection, this research addresses the challenge of identifying hate speech and offensive content within Hindi-English code-mixed conversations on social media. The task is subdivided into binary and multi-class classifications. Various deep learning models, including BERT, are employed, and an ensemble model comprising a fine-tuned mBERT model and a sentence transformer is proposed. The investigation reveals that the strategic approach of distinct feature extraction from posts, comments, and replies effectively addresses the linguistic diversity within the dataset.

Finally, this dissertation delves into the challenges of information retrieval from Bengali-English code-mixed data. A language-identification-based solution with query expansion employing different phonetic algorithms is proposed, showing promising results that outperform the baseline. The research advocates for the development of a code-mixed stop words list to bolster the performance of information retrieval systems and critiques the conventional approach of employing term frequency as a metric for term importance.

8.1 Future Directions

While significant progress has been made, there are numerous avenues for future research to further enhance text processing in code-mixed social media data.

Low-Resourced Language Pairs: More work is needed on low-resourced language pairs. Studying these pairs will provide more insightful and exciting findings. For instance, Tulu, a language primarily spoken in Karnataka, India, is similar to Kannada. Addressing code-mixing between Tulu-Kannada-English will present a more challenging task.

Sentiment Analysis: Future work should focus on creating error-free datasets to experimentally validate the hypothesis that language identification should precede sentiment analysis. Additionally, addressing the class imbalance problem and investigating the impact of distinct levels of code-mixing in sentences on downstream tasks will contribute valuable

insights.

Hate Speech Detection: This study only focuses on Hindi-English data. Future research should examine other language pairs with a reasonable amount of data. Additionally, a directed graph-based approach could be more effective for understanding the contextual relationship between posts, comments, and replies.

Information Retrieval: The dataset for code-mixed information retrieval currently comprises only Bengali-English language pairs and data from a single social media domain. Future work should explore other language pairs and domains. Investigating the ratio of retrieval data from code-mixed and monolingual sources and training the DeepCT model with datasets in different languages will provide a broader understanding of retrieval effectiveness.

Code-mixing is a pervasive phenomenon, and text processing on code-mixed social media data is an emerging topic of interest in both academia and industry. This thesis has explored several aspects of this field, laying the groundwork for future exploration and advancements.