Contents

List of Figures **xix**

List of Tables **xxv**

List of Abbreviations **xxvi**

Preface **xxix**

1 Introduction **1**

1.1 What are Social Media? 1
1.2 Rise of Social Media 3
1.3 Code-Mixing 5
1.4 Why CM text processing is important? 7
1.5 Tasks on Code-Mixed data 9
1.5.1 Language Identification (LID) 10
1.5.2 Sentiment Analysis 11
1.5.3 Hate Speech and Offensive Content Identification 12
1.5.4 Information Retrieval 13
1.6 Motivation and Challenges 13
1.7 Dissertation Overview 15
1.8 Research Goals 16
1.9 Contribution 17
1.10 Structure of the thesis 19

2 Background **21**

2.1 What are Languages and Scripts? 21
2.2 Transliteration and Code-Mixing 23
2.3 Are Code-Switching and Code-Mixing same? 24
2.4 Types of Code-Mixing 24

2.5 Measuring the amount of CM in corpus . . . 26
2.6 Phonetic Matching Algorithms . . . 28
2.6.1 Soundex . . . 29
2.6.2 Phonix . . . 29
2.6.3 Hindex . . . 30
2.7 Types of IR . . . 30
2.8 IR Framework . . . 34
2.9 Deep Learning Framework . . . 37
2.10 Evaluation Metrics . . . 43
2.10.1 Classification Tasks . . . 44
2.10.2 IR Tasks . . . 47

3 Literature Review 49
3.1 Literature Review in Language Identification . . . 49
3.2 Literature Review in Sentiment Analysis . . . 54
3.3 Literature Review in Hate speech and Offensive Content Identification 57
3.4 Literature Review in Code-Mixed Information Retrieval . . . 60

4 Code-Mixed Language Identification in Word Level 67
4.1 Problem Statement . . . 67
4.2 Dataset . . . 68
4.2.1 Examples . . . 70
4.3 Experimental Setup . . . 72
4.3.1 Baseline Approaches . . . 73
4.3.2 Proposed Approaches . . . 74
4.4 Results . . . 79
4.5 Discussion . . . 87
4.5.1 Comparison with the baselines . . . 89
4.6 Summary of this Chapter . . . 93

5 Sentiment Analysis on Dravidian Code-Mixed Data 95
5.1 Problem Statement . . . 96
5.2 Datasets . . . 96
5.3 Methodology and Experiment Setup . . . 99
5.3.1 Multilingual BERT or mBERT . . . 99
5.3.2 Language Identification tool by Googletrans . . . 99
5.3.3 Data Pre-processing . . . 100

5.3.4 Methodology for RQ-1 . . . 100
5.3.5 Methodology for RQ-2 . . . 102
5.3.6 Methodology for RQ-3 . . . 103
5.4 Results and Discussion . . . 103
5.4.1 Results for RQ-1 . . . 104
5.4.2 Results for RQ-2 . . . 108
5.4.3 Results for RQ-3 . . . 110
5.4.4 Error Analysis . . . 112
5.4.5 Discussion . . . 113
5.5 Summary of this Chapter . . . 114

6 Hate Speech Detection in Code-Mixed Social Media Conversations 119
6.1 Problem Statement . . . 120
6.1.1 Tasks Description . . . 121
6.2 Dataset . . . 123
6.3 Methodology . . . 123
6.3.1 Data Pre-processing . . . 123
6.3.2 Methodology for RQ-1 . . . 125
6.3.3 Methodology for RQ-2 . . . 126
6.3.4 Methodology for RQ-3 . . . 127
6.4 Results and Discussion . . . 134
6.4.1 Results for RQ-1 . . . 134
6.4.2 Results for RQ-2 . . . 134
6.4.3 Results for RQ-3 . . . 135
6.4.4 Discussion . . . 135
6.5 Summary of this Chapter . . . 138

7 Code-Mixed Information Retrieval 141
7.1 Problem Statement . . . 141
7.2 Research Questions (RQs) . . . 143
7.3 Building a Text Collection . . . 143
7.3.1 Data Pre-processing . . . 145
7.3.2 Pooling and Judgment . . . 147
7.4 CMIR: A phonetic algorithms-based approaches . . . 150
7.4.1 Experimental Setup . . . 150
7.4.2 Results . . . 156
7.4.3 Discussion . . . 166

7.4.4 Error Analysis 181
7.5 CMIR: A Study on Effects of Stop words Removal 182
7.5.1 Experimental Setup 184
7.5.2 Results 189
7.5.3 Discussion 193
7.6 Summary of this Chapter 195

8 Conclusions and Future Work **197**
8.1 Future Directions 198

References **201**

List of publications **219**