

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

To
Baba Vishwanath, Mata Annapurna
and
my beloved family

CERTIFICATE

It is certified that the work contained in the thesis titled **Text Processing on Code-Mixed Social Media Data** by **Supriya Chanda** has been carried out under my supervision and 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 **Ph.D. Degree in Computer Science and Engineering**.

Date of Submission: 07.06.2024



Supervisor

Dr. Sukomal Pal

Associate Professor

Department of Computer Science and Engineering

Indian Institute of Technology (BHU) Varanasi

Varanasi, INDIA, 221005

पर्यवेक्षक/Supervisor
संगणक विज्ञान एवं अभियांत्रिकी विभाग
Department of Computer Sc. & Engg
भारतीय प्रौद्योगिकी संस्थान
Indian Institute of Technology
(काशी हिन्दू विश्वविद्यालय)
(Banaras Hindu University)
वाराणसी/Varanasi-221005

DECLARATION BY THE CANDIDATE

I, **Supriya Chanda**, certify that the work embodied in this thesis is my own bona fide work and carried out by me under the supervision of **Dr. Sukomal Pal** from **July-2018 to June-2024**, at the **Department of Computer Science and Engineering**, 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 research workers 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.

Date: 07.06.2024

Place: Varanasi

Supriya Chanda

Supriya Chanda

CERTIFICATE BY THE SUPERVISOR

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

SPal

Dr. Sukomal Pal

Associate Professor,

Department of Computer Science and Engineering,

Indian Institute of Technology (BHU) Varanasi,

Uttar Pradesh, INDIA, 221005

पर्यवेक्षक/Supervisor

संगणक विज्ञान एवं अभियांत्रिकी विभाग

Department of Computer Sc. & Engg

भारतीय प्रौद्योगिकी संस्थान

Indian Institute of Technology

(काशी हिन्दू विश्वविद्यालय)

(Banaras Hindu University)

वाराणसी/Varanasi-221005

M. Singh
Signature of Head of Department

साचाय व विभागाध्यक्ष

Professor & Head

संगणक विज्ञान एवं अभियांत्रिकी विभाग

Department of Computer Sc. & Engg

भारतीय प्रौद्योगिकी संस्थान

Indian Institute of Technology

(काशी हिन्दू विश्वविद्यालय)

(Banaras Hindu University)

COPYRIGHT TRANSFER CERTIFICATE

Title of the Thesis: Text Processing on Code-Mixed Social Media Data

Name of the Student: Supriya Chanda

Copyright Transfer

The undersigned hereby assigns to the Indian Institute of Technology (Banaras Hindu University) Varanasi all rights under copyright that may exist in and for the above thesis submitted for the award of the *Doctor of Philosophy*.

Date: 07.06.2024

Place: Varanasi

Supriya Chanda

Supriya Chanda

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.

List of Figures

1.1	Market shares of leading SM platforms (in percentages)	3
1.2	Social media infographics	5
1.3	Example of a CM conversation from movie Pink	6
1.4	Example of CM notification push by companies	7
1.5	Example of an original comment and its machine translation version . .	9
1.6	Number of publications over time in *CL and ISCA venues. [1]	10
1.7	Type of NLP task on code-mixed data	11
1.8	Example of language identification task	11
1.9	Overview of the dissertation’s structure	20
2.1	Mapping between language and script	22
2.2	Example of a transliteration	23
2.3	Example of Mono-lingual IR	32
2.4	Example of Cross-lingual IR	32
2.5	Example of multilingual mono script IR	33
2.6	Example of Mixed script IR	33
2.7	Example of Code Mixed IR	34
2.8	Populating the classification outcome in the Confusion Matrix	45
3.1	List of venues (Shared task) with timeline for Language Identification task	51
4.1	Example of a code-mixed sentence along with its language and non- language tag.	68
4.2	Example 2: example of a code-mixed sentence along with its language and non-language tag.	70
4.3	Example 3: example of a code-mixed sentence along with its language and non-language tag.	70
4.4	Graphical comparison of Code-Mixing metrics for the ICON_POS dataset	71
4.5	Graphical comparison of Code-Mixing metrics for the ICON_SAIL dataset	72

4.6	Graphical comparison of Code-Mixing metrics for the LinCE dataset	73
4.7	Model Architecture of our proposed system	76
4.8	Architecture of a LSTM block	77
4.9	Graphical comparison of Precision for baseline and best performing models on ICON_POS (BN-EN) dataset	79
4.10	Graphical comparison of Recall for baseline and best performing models on ICON_POS (BN-EN) dataset	80
4.11	Graphical comparison of F_1 score for baseline and best performance model on ICON_POS (BN-EN) dataset	80
4.12	Graphical comparison of Precision for baseline and best performing models on ICON_POS (HI-EN) dataset	80
4.13	Graphical comparison of Recall for baseline and best performing models on ICON_POS (HI-EN) dataset	81
4.14	Graphical comparison of F_1 score for baseline and best performing models on ICON_POS (HI-EN) dataset	81
4.15	Graphical comparison of Precision score for baseline and best performing models on ICON_SAIL (BN-EN) dataset	82
4.16	Graphical comparison of Recall score for baseline and best performing models on ICON_SAIL (BN-EN) dataset	82
4.17	Graphical comparison of F_1 score for baseline and best performing models on ICON_SAIL (BN-EN) dataset	82
4.18	Graphical comparison of Precision score for baseline and best performing models on ICON_SAIL (HI-EN) dataset	83
4.19	Graphical comparison of Recall score for baseline and best performing models on ICON_SAIL (HI-EN) dataset	84
4.20	Graphical comparison of F_1 score for baseline and best performing models on ICON_SAIL (HI-EN) dataset	84
4.21	Graphical comparison of Precision score for baseline and best performing models on LinCE (HI-EN) dataset	84
4.22	Graphical comparison of Recall score for baseline and best performing models on LinCE (HI-EN) dataset	85
4.23	Graphical comparison of F_1 score for baseline and best performing models on LinCE (HI-EN) dataset	85
4.24	Graphical comparison of Precision score for baseline and best performing models on LinCE (ES-EN) dataset	86

4.25	Graphical comparison of Recall score for baseline and best performing models on LinCE (ES-EN) dataset	86
4.26	Graphical comparison of F_1 score for baseline and best performing models on LinCE (ES-EN) dataset	86
4.27	Confusion Matrices for the baseline and proposed models on the ICON_POS (BN-EN) Corpus Test Set	88
5.1	Model Architecture for identifying monolingual and code-mixed data	101
5.2	Model architecture for multi-class classification with ruled-based language tag	103
5.3	Model architecture for hierarchical approach with mBERT	104
6.1	The structure of Conversational code-mixed Tweet	120
6.2	Example of conversational code-mixed data from Twitter	121
6.3	Creation of datasets using anchor-positive and anchor-negative pair	128
6.4	Architecture diagram for fine-tuning mBERT	129
6.5	Architecture of proposed methodology for Task 2A	131
6.6	Architecture of proposed methodology for Task 2B	133
6.7	Confusion matrix on ICHCL 2021 test data for English-Hindi code mixed language (Submission 1 for Subtask 2)	137
7.1	Screenshot of a social media post showing a query and the corresponding documents	145
7.2	An example of code-mixed query (Topic Number 18)	146
7.3	An example of code-mixed query (Topic Number 1)	147
7.4	An example of code-mixed query (Topic Number 7)	148
7.5	An example of code-mixed document	148
7.6	Length distribution of queries	149
7.7	Length distribution of documents	149
7.8	Process flow for retrieval	152
7.9	MAP on different retrieval model	163
7.10	Query wise performance changes (Run 0, Run 9 and Run 16) by BM25 model	163
7.11	Performance changes (Run 0, Run 1, Run 8 and Run 15) across queries by BM25 model	165
7.12	Query wise performance changes (Run 0, Run 13 and Run 20) by BM25 model	165

7.13	Performance changes (Run 0 and Run 20) across queries by BM25 model	167
7.14	Performance changes (Run 0 and Run 15) across queries by BM25 model	168
7.15	Performance changes (Run 0 and Run 8) across queries by BM25 model	170
7.16	Performance changes (Run 0 and Run 1) across queries by BM25 model	171
7.17	Performance changes (Run 0 and Run 16) across queries by BM25 model	175
7.18	Percentage of word tags in each query	176
7.19	Performance changes (from Run 21 wrt Run 0) across queries using BM25 model	179
7.20	Changes in performance on human-annotated and machine-annotated data for the Run 20 experiment across queries by the In BM25 model .	180
7.21	Performance gain and loss on both annotation data for different experi- ment (Run 8, Run 20 and Run 21)	181
7.22	Process flow for retrieval	187
7.23	Process flow for retrieval	189
7.24	Find the best threshold value for corpus-based language-dependent stop words list	192
7.25	Performance gain from Run 3 to Run 12 wrt Run 0	193
7.26	Performance changes (from Run 12 wrt Run 0) across queries by InL2 model	194
7.27	Performance changes (from Run 12 wrt Run 2) across queries using InL2 model	194

List of Tables

2.1	Soundex phonetic codes	29
2.2	Phonix phonetic codes	30
2.3	Hindex codes	31
2.4	Hindex Codes for vowels pair	31
3.1	System papers of MSIR 2013 task: Ad hoc retrieval for Hindi song lyrics	62
3.2	System papers of MSIR 2014 task: Ad hoc retrieval for Hindi song lyrics	63
3.3	System papers of MSIR 2015 task: Ad hoc retrieval for Hindi song lyrics	64
3.4	Approaches on ‘IR on code-mixed Hindi–English tweets’, FIRE 2016 .	65
4.1	Corpus Statistics	69
4.2	Class distributions of ICON_POS (BN-EN) test data	79
4.3	Class distributions of ICON_POS (HI-EN) test data	81
4.4	Class distributions of ICON_SAIL (BN-EN) test data	83
4.5	Class distributions of ICON_SAIL (HI-EN) test data	83
4.6	Class distributions of LinCE (HI-EN) development data	85
4.7	Class distributions of LinCE (ES-EN) development data	87
4.8	Comparison of baseline with best proposed model on ICON_POS (BN-EN) Data	90
4.9	Comparison of baseline with best proposed model on ICON_POS (HI-EN) Data	90
4.10	Comparison of baseline with best proposed model on ICON_SAIL (HI-EN) Dataset	90
4.11	Comparison of baseline with best proposed model on ICON_SAIL (BN-EN) Dataset	91
4.12	Comparison of baseline with best proposed model on LinCE (HI-EN) development data	91

4.13	Comparison of baseline with best proposed model on LinCE (ES-EN) development data	91
4.14	Effect of Bi-LSTM layer	93
5.1	Data Distribution for sentiment detection of code-mixed text in Dravidian languages	97
5.2	Example of code-mixed text in Dravidian languages from three language pairs for all classes	98
5.3	The statistics of monolingual and code-mixed data involved in training, development, and test datasets of all three language pairs.	105
5.4	Level of code-mixing (CMI values) involved in training, development, and test datasets of all three language pairs.	105
5.5	Precision, recall, F_1 -scores, and support for all experiments on Tamil-English test data	106
5.6	Precision, recall, F_1 -scores, and support for all experiments on Kannada-English test data	106
5.7	Precision, recall, F_1 -scores, and support for all experiments on Malayalam-English test data	107
5.8	Weighted average F_1 -scores and support for three language pairs where model is trained only on CM data but tested on all, Monolingual and CM data	108
5.9	Comparison of Precision, recall, F_1 -scores, and support with our proposed model for RQ-2 on Tamil-English test data	109
5.10	Comparison of Precision, recall, F_1 -scores, and support with our proposed model for RQ-2 on Kannada-English test data	109
5.11	Precision, recall, F_1 -scores, and support with our proposed model for RQ-2 on Malayalam-English test data	109
5.12	Comparison of Precision, recall, F_1 -scores, and support with our proposed model for RQ-3 on Tamil-English test data	111
5.13	Comparison of Precision, recall, F_1 -scores, and support with our proposed model for RQ-3 on Kannada-English test data	111
5.14	Comparison of Precision, recall, F_1 -scores, and support with our proposed model for RQ-3 on Malayalam-English test data	111
5.15	Errors in gold standard	112
5.16	Errors made by the LID Model	113
5.17	Rank list for Task A: Tamil track published by organiser	115
5.18	Rank list for Task A: Malayalam track published by organiser	115

5.19	Rank list for Task A: Kannada track published by organiser	116
6.1	Statistical overview of the datasets from ICHCL 2021, 2022, and 2023 . .	124
6.2	Example tweets from ICHCL 2021, ICHCL 2022 and ICHCL 2023 dataset for all classes (HI for Hindi, EN for English and Lang for Language) . .	125
6.3	Rule-based approach for Task 2B	133
6.4	Evaluation results on ICHCL 2021 test data (Submission number in bracket)	134
6.5	Evaluation results on ICHCL 2022 test data (Submission number in bracket)	135
6.6	Evaluation results for Task 2A and 2B on Hindi-English test data (ICHCL 2023)	136
6.7	ICHCL Task 2A results published by organiser	136
6.8	ICHCL Task 2B results published by organiser	136
7.1	Collection statistics	144
7.2	An example of query expansion	155
7.3	All Experiment setup	155
7.4	The results of retrieval effectiveness measured by MAP, R-prec, and P@10 when LID task is done by Machine. ↑ sign denotes that the model for this setup performs better than the baseline (Run 0). ↓ sign denotes that the model for this setup performs less than the baseline.	159
7.5	The results of retrieval effectiveness measured by MAP, R-prec, and P@10 when LID task is done by Human. ↑ sign denotes that the model for this setup performs better than the baseline (Run 0). ↓ sign denotes that the model for this setup performs less than the baseline.	161
7.6	An example of query term expansion (Query Number 51)	173
7.7	An example of query term expansion (Query Number 56)	173
7.8	An example of query term expansion (Query Number 15)	174
7.9	All experiment setup	189
7.10	The results of retrieval effectiveness measured by MAP, Recip-rank, nDCG, P@5, and P@10 for all the experimental setups.	191

List of Abbreviations

Abbreviation	Description
ACL	Association for Computational Linguistics
AP	Average Precision
BERT	Bidirectional Encoder Representations from Transformers
Bi-LSTM	Bidirectional Long Short-Term Memory
BN	Bengali
CALCS	Computational Approaches to Linguistic CodeSwitching
CM	Code Mixing
CMI	Code Mixing Index
CHOF	Contextual Hate and Offensive
CLIR	Cross-lingual Information Retrieval
CMIR	Code Mixed Information Retrieval
CMTET	Code-Mixed Telugu-English Text
CNNs	Convolutional Neural Networks
CRF	Conditional Random Field
CS	Code Switching
EMNLP	Empirical Methods in Natural Language Processing
EN	English
ES	Spanish
FN	False Negative
FP	False Positive
FIRE	Forum for Information Retrieval Evaluation
GloVe	Global Vectors for Word Representation
HASOC	Hate Speech and Offensive Content Identification
HI	Hindi
HOF	Hate and Offensive
HSL	Hindi song lyrics
ICHCL	Identification of Conversational Hate-Speech in Code-Mixed Languages

Abbreviation	Description
ICON	International Conference on Natural Language Processing
IR	Information Retrieval
ISCA	International Symposium on Computer Architecture
KA	Kannada
LID	Language Identification
LinCE	Linguistics Code-switching Evaluation
LOR	Log Odds Ratio
LSTM	Long Short-Term Memory
MA	Malayalam
MAP	Mean Average Precision
MLP	Multilayer Perceptron
MR	Movie Reviews
MSIR	Mixed-Script Information Retrieval
MSRI	Microsoft Research India
MT	Machine Translation
NE	Named Entity
NER	Named Entity Recognition
NLP	Natural Language Processing
NOT	Non-offensive
OOV	out-of-vocabulary
POS	Parts of Speech
ReLU	Rectified Linear Unit
RNNs	Recurrent Neural Networks
RO	Research Objective
RQ	Research Question
SA	Sentiment Analysis
SAIL	Sentiment Analysis of Indian Language
SGD	Stochastic Gradient Descent
SVM	Support Vector Machine
SW	SeedWords
TA	Tamil
TN	True Negative
TP	True Positive
SHOF	Standalone Hate and Offensive
SM	Social Media

Abbreviation	Description
SPC	Simon Personal Communicator
TF-IDF	Term Frequency - Inverse Document Frequency
UGC	User Generated content
CNN	Convolutional Neural Network
DNN	Deep Neural Networks