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.

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

Signature of Head of Department

आचार्य 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

List of Tables

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