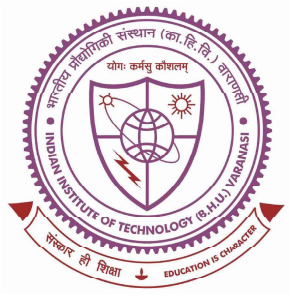
The Impact of Social Media Text: A study on e-Governance, Election and Disaster Management



Thesis submitted in partial fulfilment
for the Award of
DOCTOR OF PHILOSHPHY (PHD)
in
COMPUTER SCIENCE AND ENGINEERING

by
Anita Saroj

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY
BANARAS HINDU UNIVERSITY
VARANASI - 221 005

ROLL NUMBER
16071006

YEAR OF SUBMISSION
2022

I would like to dedicate this thesis to my loving parents …

Certificate

It is certified that the work contained in this thesis entitled **"The Impact of Social Media Text: A study on e-Governance, Election and Disaster Management"** by **"Anita Saroj"**, Roll Number **16071006** has been carried out under my supervision and that 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**.

Supervisor

Dr. Sukomal Pal

Department of Computer Science and Engineering

Indian Institute of Technology (Banaras Hindu University)

Varanasi-221005

Declaration

I, **Anita Saroj**, 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 2016 to July 2022 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 whenever and wherever their works have been cited in my work in this thesis. I further declare that I have not wilfully copied any others' 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 14 July 2022 Signature Anita

Place Varanasi Anita Saroj

Certificate by the Supervisor

It is certified that the above statement made by the student is correct to the best of my knowledge.

Supervisor

Dr. Sukomal Pal

Department of Computer Science and Engineering

IIT (BHU)

Varanasi

Signature of the Head of the Department

(Computer Science and Engineering)

Professor & Head
Department of Computer Sc. & Engg
Indian Institute of Technology
(Banaras Hindu University)
Varanasi-221005

Copyright Transfer Certificate

Title of the Thesis :**The Impact of Social Media Text: A study on e-Governance, Election and Disaster Management**

Name of the Student : **Anita Saroj**

Copyright Transfer

The undersigned hereby assigns to the 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"** in **Computer Science and Engineering**.

Date 14 July 2022 Signature anita

Place Varanasi (Anita Saroj)

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.

Acknowledgements

This thesis is the end of my journey in obtaining my Ph.D. I have not traveled in a vacuum in this journey. This thesis has been kept on track and been seen through to completion with the support and encouragement of numerous people including my well wishers, my friends, colleagues and various institutions. At the end of my thesis, I would like to thank all those people who made this thesis possible and an unforgettable experience for me. At the end of my thesis, it is a pleasant task to express my thanks to all those who contributed in many ways to the success of this study and made it an unforgettable experience for me.

Firstly, I would like to express my sincere gratitude to my advisor Dr.Sukomal Pal for the continuous support of my Ph.D study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.

I specially thank the members of my Research Progress Evaluation Committee, Dr. A. K. Singh, Dr. Sandip Ghosh , DPGC Convener Dr. Bhaskar Biswas, Dr. Pratik Chattopadhyay for their invaluable suggestions regarding the thesis, their insightful comments and encouragement, but also for the hard question which incented me to widen my research from various perspectives.

Most importantly, my deepest gratitude is for my family for their constant support, inspiration, guidance, and sacrifices. My parents were constant source of motivation and inspiration. Their affection and guidance was instrumental in me choosing Engineering and eventually continuing on to my Ph.D.

Very special gratitude goes out to my colleagues and friends with special mention to Mr. Rajesh Kumar Munodtiya, Mr. Siba Shankar Shau, and Mr. Supriya Chanda for being great friends and the best advisors I could be ever have. Their advice, encouragement, and critics were always sources of inspiration. This thesis would not have been possible without their invaluable remarks and persistent help. We really enjoyed our discussions, which spanned on every topic under the sun other than academics.

I extend special thanks to the non-teaching staff in the department, particularly, Mr. Ravi Kumar Bharti, Mr. Ritesh Singh, Mr. Shubham Pandey, Mr. Prakhar Kumar, and Mr. Manoj Kumar Rai, Mr. Viplav Biswas, and Mr. Akhilesh Kumar Pal.

I take this opportunity to sincerely acknowledge the Ministry of Human Resource Development, Government of India for providing Ph.D. Fellowship for financial assistance.

It is said that it is no coincidence who you meet on the journey. I have had the fortunate of meeting so many wonderful people on the journey; capable and ready to guide, help, and advice me. I sincerely thank all of them who contributed in for helping me to see the light at the end of every scary tunnel during my Ph.D.

-Anita Saroj

Preface

Social media (SM) is part of our online existence today. We express ourselves on social media every day on different issues like politics, government policies, disasters, commercial products, celebrity statements, controversies and so on. Numerous kinds of social media platforms are available that host blogs, wikis, or multimedia content like YouTube, Instagram and Flickr, or enable social networking like Facebook and LinkedIn, micro blogging services like Twitter, Weibo etc. With Web 2.0, web-users are increasingly creating content than just consuming it. How this trend is entering into and thereby affecting our life and society is an interesting area of research. We chose, in particular, the social and political life and governance. An analysis and review of the adoption of social media by as well as its effects on the public and government sector is of due significance - examining the role and implications of these new technologies with a focus on different emotion detection and analysis in the chosen field from an Indian perspective is attempted in the thesis.

Many in the West believe that ICT-enabled public services positively impact economic growth, inclusion, and quality of life. With the ease-of-use, social media platforms are convenient vehicles for all forms of information exchange and continuously opening up newer avenues of governance, education, healthcare, entertainment and business as well as ushering in social changes. Even though information content in the media can be multi-modal involving image, audio and video, text is the predominant data-type. Social media text carries a lot of information. User-generated content provides diverse and unique information in the form of comments, posts, and tags. Features like friends, followers, connections provide a lot of contextual, structural and social network-related information about the users. The valuable information hidden in the text resources of social media provides opportunities for researchers of different disciplines to mine interesting patterns and metadata that might not be obvious. We conduct three studies to see the nature and

effect of social media communication through text in Indian context, keeping in mind two important stakeholders of the society: Government and general public. SM has been used in the government sector of the West for quite some time. In India, however, use of SM in governance is slowly but steadily increasing. We study the pattern of use, topics of discussion, interaction among different ministries and public participation in the governance in the first task. We look at another event of public-private interaction through social media. Elections are events when governments, political parties and public engage intensely with each other, especially in India. People's interaction with the candidates in social media posts reflects many social trends in a charged atmosphere. People's likes and dislikes of political parties and their leaders, their manifesto, promises and public comments often trigger hate and offensive posts. Sometimes irony is used while expressing opinions. We look at these varied sentiments in social media posts during a parliamentary election of India. Thirdly, we study the public reaction and sentiment from a different perspective like when a crisis or disaster happens. Crawling the SM channels during COVID-19 pandemic to collect data for the study and conducting different sentiment analysis related experiments is another task reported in the thesis.

First, we study the adoption and penetration of social media (SM) in government sector. The study explores the adoption of social media (SM) by different ministries of the Government of India(GoI) (actually 46 ministries). We analyze how government ministries use SM, particularly Twitter and Facebook, to disseminate information, engage with different stakeholders and take feedback on government initiatives. The study based on three years of SM data captures activities, discussion topics, inter-connectedness, public engagement, and popularity of GoI ministries through SM. Out of 8 different SM channels that GoI ministries use, almost all are active in Twitter and Facebook. Primary topics of discussion are meetings and projects like Prime Minister's (PM) projects, Railways projects, GST, and different development schemes. The Ministry of Railways and the

Ministry of Agriculture post maximum in SM, while PM's account is followed the most. The Ministry of Information and Broadcasting serves as a good coordinator. The study involves crawling the data from SM channels, pre-processing raw data, content analysis, time series analysis, citation analysis among different users.

Elections are events that causes boom in the use of SM. In India, governmental agencies, political parties and public engage intensely with each other during elections that sees waves of emotive posts in SM. People often use accusations, counter-accusations, hate speeches, ironies, sarcasms. Verbal irony, an utterance that conveys a spirit completely opposed to the surface meaning expressed, is usually understood by body language and the context of the conversation. However, it is challenging to automatically detect irony in a limited amount of text like in SM posts. To study the issue, we crawl the data from social media posts during the 2019 general election in India and make a standard collection with annotations (Indian General Election 2019 or IGE 2019 dataset). We then focus on automatic irony detection using various machine learning and deep learning (Bidirectional Encoder Representations from Transformers (BERT) and Embeddings from Language Models (ELMo)) models to classify them into irony and non-irony. We propose an ensemble model of machine learning and deep learning approaches. The classifiers are trained using a combined word embedding representation obtained from both BERT and ELMo. A series of experiments on irony detection then are performed including a domain adaptation with SemEval-2018 Task-3 (Sub-task A) dataset (SE-2018 T3 data). Our experiments on IGE 2019 and SemEval-2018 Task-3 data show results comparable to the state-of-the-art performance for irony detection. We also created a dataset on hate speech and offensive content identification during parliamentary election 2019 of India (PEI data-2019). It is a hierarchical classification task where we explore multi-task learning (MTL), a machine learning technique that accomplishes many tasks in a single model by exploiting commonalities and differences across different tasks. On the created dataset

we conduct hate speech and offensive content (HOF) identification and classification task, using MTL with the convolution network (MTL-CN) method. The MTL-CN extracts the shared and private latent features from the text. Our experiments on three different text classification tasks, demonstrate benefits of our approach. We also show that the knowledge learned by the proposed model here can be shared with new tasks. Finally, results on different standard datasets like LREC-2020 (PEI-2019), FIRE-2019, FIRE-2020, and SemEval-2019 datasets are shown to establish that multi-task learning yields better results for different classification tasks, irrespective of the languages.

When a disaster (man-made ones like shootings, bomb-blasts, war or natural calamities like floods, cyclones, tsunami, earthquakes) strikes, people suffer and/or get panicked and react emotionally. SM see a sudden increase with flight of emotions. During the COVID-19 pandemic, the whole world saw SM full with emotive posts. Social distancing caused lack of social interactions. The physical void led to increased online interaction among users on social media platforms. Sentiment analysis of such interactions can help us analyze the general public psychology during the pandemic. However, the lack of data in non-English and low-resource languages like 'Hindi' makes it difficult to study it among native and non-English speaking masses. We create a collection of 'Hindi' tweets on COVID-19 during the pandemic containing 10,011 tweets for sentiment analysis (named sentiment analysis for Hindi or SAFH dataset). We describe the process of collecting, creating, annotating the corpus, and performing sentiment classification task.

Table of contents

List of figures

List of tables

Nomenclature

Acronyms / Abbreviations

BERT Bidirectional Encoder Representations from Transformers

CNN Convolutional Neural Network

DA Domain Adaptation

DL Deep Learning

ELMo Embeddings from Language Model

HOF Hate Speech and Offensive

IGE-2019 Indian general election 2019

KNN K-Nearest Neighbour

LR Logistic Regression

LSTM Long short-term memory

ML Machine Learning

MNB Multinomial Naive Bayes

MTL-CN Multi-task Learning With the Convolution Network

MTL Multi-task Learning

PEI-2019 Parliamentary Election of India 2019

RF Random forest

SGD Stochastic Gradient Descent

SM Social Media

SVM Support Vector Machine

TL Transfer Learning