

LB/TH/38/2025
TH5967

**THE RANKING OF HATE SPEECH PROPAGATORS IN
SOCIAL MEDIA USING A MULTIMODAL PROFILING
FRAMEWORK**

Weerakoon Achchige Suresha Nilmini Perera

218031V

Degree of Doctor of Philosophy

Department of Computer Science and Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

August 2025

THE RANKING OF HATE SPEECH PROPAGATORS IN SOCIAL MEDIA USING A MULTIMODAL PROFILING FRAMEWORK

Weerakoon Achchige Suresha Nilmini Perera

218031V

Thesis submitted in partial fulfillment of the requirements for the degree

Degree of Doctor of Philosophy

Department of Computer Science and Engineering

Faculty of Engineering

University of Moratuwa

Sri Lanka

August 2025

DECLARATION

I declare that this is my own work and this thesis/dissertation does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other University or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature:

Date: 20th August 2025

The above candidate has carried out research for the PhD under my supervision. I confirm that the declaration made above by the student is true and correct.

Name of Supervisor: Prof. G.I.U.S. Perera

Name of Co-Supervisor: Dr. (Ms) S. Ahangama

Signature of the Supervisor:

Date: 21/08/2025

Signature of the Co-Supervisor:

Date: 21/08/2025

DEDICATION

I dedicate this to my parents, husband, my daughter and son.

ACKNOWLEDGEMENT

I would like to extend my deepest gratitude to those who have supported and guided me throughout this journey.

My heartfelt thanks to my supervisor, Prof. GIUS Perera, for his exceptional guidance and constant encouragement. His patience and understanding have been invaluable, especially during occasional setbacks. I deeply appreciate his insightful contributions and encouragement, which provided direction in many situations. Furthermore, his exceptional knowledge and enthusiasm for research were inspiring and became a driving force for my motivation. I am also immensely grateful to my co-supervisor, Dr.(Ms) S. Ahangama, for her continuous support, invaluable knowledge, and unwavering encouragement. Her outstanding support and guidance throughout this journey helped me to reach this level.

I would like to thank the progress review evaluation team, including Dr. Charith Chitraranjan, Dr. Shehan Perera, and Dr. L. Ranathunga, for their insightful feedback, suggestions for improvement, and their appreciation and encouragement.

I would like to express my deep gratitude for the free education opportunities in Sri Lanka, which enabled me to pursue my studies from primary school to postgraduate level.

I am grateful to the staff of the Department of Computer Science and Engineering, Faculty of Engineering, University of Moratuwa, for sharing their knowledge and supporting my academic pursuits.

Special thanks to my cousin, Dr.(Ms) I. Udayamalee, for encouraging me during challenging times and providing me with the necessary articles throughout the study.

I am thankful to the Higher Education Expansion and Development (AHEAD) of Operation of the Ministry of Higher Education, which is funded by the World Bank, for providing the necessary funds to publish my papers.

I would also like to thank Mr. Duminda De Silva for motivating me to start this journey and for his continued encouragement. I also thank the entire staff of the Department of Computer Science, Faculty of Natural Sciences, The Open University of Sri Lanka.

Finally, I sincerely thank my family and friends for their continuous support and motivation.

ABSTRACT

Social media platforms have revolutionized communication by enabling the sharing of opinions, ideas, and beliefs at an unprecedented scale. However, the rapid growth of these platforms has also facilitated the widespread propagation of hate speech and harmful behaviour, posing both technological and societal challenges. Understanding hate speech propagators, their influence, reach, and impact is crucial for mitigating the spread of harmful content and fostering safer online communities. This research aims to address this challenge by developing a robust hate speaker ranking system to identify, analyze, and rank hate users based on their behaviour and influence. The proposed framework combines profile features, characteristics, and machine learning techniques to evaluate hate speech. The system provides a dynamic, multimodal ranking mechanism to rank hate users into three levels labelled as: HIGH, MEDIUM, and LOW, based on their level of influence. The framework is trained and validated on datasets from Twitter, ensuring scalability and adaptability across diverse environments. Results indicate that the system effectively identifies influential hate users with high accuracy, offering actionable insights for moderation teams. This study also addresses ethical considerations, ensuring fairness, reducing bias, and maintaining freedom of expression. The research highlights the unequal influence of hate speech propagators and emphasizes the importance of personalized strategies to combat their impact. By providing a scalable and dynamic platform for ranking hate users, this study contributes to the ongoing efforts to combat online hate, supporting moderation efforts and promoting a healthier digital environment.

Keywords: Hate Speech, Influence, Social Media Platforms, Social Network Analysis, User Ranking,

TABLE OF CONTENTS

DECLARATION	I
DEDICATION	II
ACKNOWLEDGEMENT	III
ABSTRACT	IV
TABLE OF CONTENTS	V
LIST OF FIGURES.....	VIII
LIST OF TABLES.....	IX
LIST OF ABBREVIATIONS	X
LIST OF APPENDICES	XIII
CHAPTER 1.....	1
INTRODUCTION	1
1.1 BACKGROUND AND MOTIVATION.....	1
1.2 PROBLEM STATEMENT	4
1.3 RESEARCH QUESTIONS.....	4
1.4 GENERAL OBJECTIVE	5
1.5 SPECIFIC OBJECTIVES.....	5
1.6 STRUCTURE OF THE THESIS	5
CHAPTER 2.....	6
LITERATURE REVIEW	6
2.1 FREEDOM OF EXPRESSION	6
2.2 HATE SPEECH.....	8
2.3 SOCIAL NETWORK ANALYSIS.....	9
2.3.1 <i>Social Network Analysis for Identifying Influential Hate Speech Propagators</i>	11
2.4 BEHAVIOURAL ANALYSIS	12
2.4.1 <i>Questionnaire Approach</i>	12
2.4.2 <i>Big Five Personality Traits Model</i>	12
2.5 CHARACTERISTICS OF HATE SPEAKERS.....	16
2.6 STRATEGIES TO REDUCE THE HATE SPEECH FROM SPREADING ON SMPs.....	18
2.7 THE SPEAKER, THE HATE SPEAKER, AND THE HATE SPEECH PROPAGATOR	19
2.8 METHODOLOGY FOR SYSTEMATIC LITERATURE REVIEW	20
2.8.1 <i>Research Questions</i>	21
2.8.2 <i>Search Strategy for Systematic Literature Review</i>	21
2.8.3 <i>Data Collection for the Systematic Literature Review</i>	27
2.9 INCLUDED PAPERS – OVERVIEW	27
2.9.1 <i>Publication Year</i>	27
2.9.2 <i>Paper Publication Venues</i>	27
2.9.3 <i>Frequently Used Terms Analysis with the Abstract</i>	27
2.10 RANK USERS IN SOCIAL MEDIA PLATFORMS	29
2.10.1 <i>Ranking Mechanisms of Social Media Platform Users</i>	29
2.11 CONCLUSION	60
CHAPTER 3.....	62
RESEARCH METHODOLOGY AND CONCEPTUAL FRAMEWORK	62

3.1	DESIGN STEPS OF THE RESEARCH.....	62
3.2	THE PRELIMINARY STUDY OF THE IDENTIFICATION OF HATE SPEAKERS ON THE SOCIAL MEDIA PLATFORM	69
3.2.1	<i>The Selection of the Social Media Platform.....</i>	69
3.2.2	<i>Data Collection.....</i>	70
3.2.3	<i>Labelling of Tweets as Hate and Non-Hate.....</i>	75
3.3	CONCLUSION	79
CHAPTER 4.....		80
DESIGN AND IMPLEMENTATION OF THE PROPOSED SOLUTION		80
4.1	HATE SPEAKER CHARACTERISTICS	80
4.1.1	<i>Sentiment Analysis.....</i>	85
4.1.2	<i>Emotion Analysis</i>	86
4.1.3	<i>Big Five Personality Traits Model.....</i>	86
4.1.4	<i>Severity Level of a Hate Text by Hate Users</i>	88
4.2	RANKING HATE USERS IN TWITTER SMP.....	89
4.2.1	<i>Feature Engineering</i>	90
4.2.2	<i>Final Dataset for Hate Speaker Ranking</i>	100
4.2.3	<i>Ranking the User</i>	100
4.2.4	<i>Hyperparameter Tuning for Machine Learning Algorithms.....</i>	105
4.2.5	<i>Evaluation.....</i>	107
4.3	THE IMPLEMENTATION OF THE MODEL	110
4.4	CONCLUSION	110
CHAPTER 5.....		111
RESULTS, EVALUATION AND ANALYSIS		111
5.1	CHARACTERISTICS OF THE HATE SPEAKERS.....	111
5.2	FEATURES AFFECT THE INFLUENCE OF HATE USERS.....	123
5.3	FEATURES AFFECTING THE RANKING OF THE INFLUENCE OF HATE USERS	124
5.4	RANK HATE USERS	124
5.5	EVALUATION AND VALIDATION	125
5.6	ANALYSIS OF PERFORMANCE COMPARISON	126
5.7	CONCLUSION	127
CHAPTER 6.....		128
DISCUSSION.....		128
6.1	CONTRIBUTION TO THE SCIENTIFIC COMMUNITY	128
6.1.1	<i>Research Publications</i>	129
6.1.2	<i>Research Paper Contribution</i>	129
6.2	INTERPRETATION OF RESULTS AND CONTRIBUTION TO EXISTING LITERATURE	139
6.2.1	<i>Research Findings Aligned with Objectives</i>	141
6.3	LIMITATIONS OF THE RESEARCH STUDY	143
6.4	FUTURE WORK.....	146
6.5	CONCLUSION	146
CHAPTER 7.....		147
CONCLUSION		147
7.1	SIGNIFICANCE OF FINDINGS AND IMPLICATIONS FOR FUTURE RESEARCH AND PRACTICE.....	147
7.2	IMPLEMENTED FRAMEWORK FOR THE RANKING OF THE HATE SPEAKER INFLUENCE	148
7.2.1	<i>Solutions for Research Questions</i>	149
7.3	RECOMMENDATIONS AND SPECIAL CONCERNS WHEN RANKING THE HATE USERS IN TWITTER SMP	150
REFERENCES		153

APPENDIX	169
-----------------------	------------

LIST OF FIGURES

FIG. 2.4.1: A COMMON FRAMEWORK FOR PERSONALITY PREDICTION	15
FIG. 2.8.1: PRISMA MODEL	23
FIG. 2.9.1: NUMBER OF PAPERS ACCORDING TO THE PUBLICATION YEAR	28
FIG. 2.10.1: DIFFERENT FEATURES USED BY DIFFERENT SCHOLARS.....	54
FIG. 2.10.2: NUMBER OF PAPERS RELEVANT TO EACH CATEGORY	55
FIG. 2.10.3: THE DISTRIBUTION OF SOCIAL NETWORKS IN RANKING USER INFLUENCE.....	56
FIG. 3.1.1: THE CONCEPTUAL FRAMEWORK FOR RESEARCH METHODOLOGY	67
FIG. 3.2.1: NUMBER OF TWEETS ACCORDING TO THEIR LANGUAGE DISTRIBUTIONS.	71
FIG. 3.2.2: THE RELATIONSHIP AMONG ALL THE DATASETS.....	72
FIG. 3.2.3: THE DISTRIBUTION OF ALL THE TWEETS BY 530 USERS ACROSS ALL LANGUAGES	72
FIG. 3.2.4: NUMBER OF HATE AND NON-HATE POSTS ANNOTATED USING THE CROWDSOURCING PLATFORM.....	76
FIG. 3.2.5: HATE AND NON-HATE TWEETS IN THE DATASET02 AS ANNOTATED USING THE CROWDSOURCING PLATFORM ON SINHALA AND ENGLISH TWEETS.....	76
FIG. 3.2.6: ANNOTATION PROCESS OF SINHALA AND ENGLISH TWEETS IN THE DATASET03 AS HATE/NON-HATE	77
FIG. 4.1.1: ARCHITECTURAL DIAGRAM FOR TWITTER USER CHARACTERISTICS IDENTIFICATION.....	81
FIG. 4.1.2: INTERACTIONS BY TWITTER USERS WITH THEIR SOCIAL NETWORK	82
FIG. 4.1.3: FRAMEWORK FOR MAPPING HATE USERS TO PERSONALITY TRAITS	87
FIG. 4.1.4: THE ARCHITECTURAL DIAGRAM OF THE XLM-ROBERTA CLASSIFICATION MODEL FOR IDENTIFYING HATE TEXT SEVERITY LEVELS.....	89
FIG. 4.2.1: DISTRIBUTION OF FEATURES FOR ENGAGEMENT, REACH, TRUSTWORTHINESS, AND ACTIVITY METRICS ..	94
FIG. 4.2.2: VIF ANALYSIS FOR FEATURE MULTICOLLINEARITY	96
FIG. 4.2.3: VIF ANALYSIS AFTER REMOVING NON_HATE_TWEETS_REPLY_COUNT.....	97
FIG. 4.2.4: VIF ANALYSIS AFTER REMOVING NON_HATE_TWEETS_REPLY_COUNT AND HATE_TWEETS_REPLY_COUNT	97
FIG. 4.2.5: THE FRAMEWORK FOR RANKING HATE USERS IN THE TWITTER SMP	102
FIG. 4.2.6: CONCEPTUAL DIAGRAM OF THE PROPOSED MULTIMODAL FRAMEWORK	104
FIG. 4.2.7: THE WEIGHTS ASSIGNED FOR THE TWITTER USERS' INFLUENCE VARIABLES. ADOPTED FROM [108].....	108
FIG. 5.1.1: THE CDFS AND FOLLOWER-FOLLOWING COUNTS OF HATE AND NON-HATE USERS.....	112
FIG. 5.1.2: THE DISTRIBUTION OF HATE/NON-HATE USERS ACCORDING TO THEIR FOLLOWER AND FOLLOWING COUNTS	113
FIG. 5.1.3: THE COMPARISON OF THE ACCOUNT AGE OF HATE AND NON-HATE USERS IN THE TWITTER SMP	114
FIG. 5.1.4: THE TIME INTERVAL FROM ACCOUNT CREATION TO THE FIRST TWEET POSTING BY USERS.	114
FIG. 5.1.5: THE LANGUAGE DISTRIBUTION OF TWITTER HATE USERS.....	115
FIG. 5.1.6: THE DIFFERENCE BETWEEN THE MOST RECENT AND THE FURTHEST TWEET OF HATE AND NON-HATE USERS	116
FIG. 5.1.7: AVERAGE AND MEDIAN ENGAGEMENT METRICS OF HATE AND NON-HATE USERS.....	117
FIG. 5.1.8: SENTIMENT ANALYSIS OF SINHALA AND ENGLISH TWEETS BY HATE AND NON-HATE USERS.....	119
FIG. 5.1.9: EMOTION ANALYSIS OF ENGLISH TWEETS BY HATE AND NON-HATE USERS.....	119
FIG. 5.1.10: HIGH SCORE COMPARISON OF BFPT BETWEEN HATE AND NON-HATE USERS.....	120
FIG. 5.1.11: LOW SCORE COMPARISON OF BFPT BETWEEN HATE AND NON-HATE USERS	120
FIG. 6.1.1: PERCENTAGES OF REACTIONS, SHARES, AND COMMENTS FOR HATE AND NON-HATE POSTS.	132
FIG. 6.1.2: THE VALUES FOR EACH SOCIAL MEDIA BEHAVIOUR FACTOR IN THIS STUDY	132
FIG. 6.1.3: VISUALIZATION OF THE TWITTER MENTION NETWORK BEFORE THE CRISIS.	136
FIG. 6.1.4: VISUALIZATION OF THE TWITTER MENTION NETWORK DURING THE CRISIS.....	136

LIST OF TABLES

TABLE 2.4.1: BFPT PREDICTION USING DIFFERENT FEATURES AND ALGORITHMS IN VARIOUS SMPS	14
TABLE 2.5.1: METADATA: PROFILES, ACTIVITIES, AND USER INTERACTIONS	18
TABLE 2.7.1: THE SPEAKER PARAMETER OF THE RABAT FRAMEWORK - THE UNS AND PLAN OF ACTION ON HATE SPEECH. ADAPTED FROM: [95]	20
TABLE 2.8.1: SEARCH QUERIES USED IN EACH DATABASE AND THE NUMBER OF RESEARCH ARTICLES	24
TABLE 2.8.2: FILTERED STUDY COUNT BY SEARCH STRING ACROSS DATABASES	24
TABLE 2.8.3: NUMBER OF ACCEPTED AND REJECTED PAPERS ACCORDING TO THE TITLE, ABSTRACT OR FULL PAPER READ	26
TABLE 2.8.4: NUMBER OF PAPERS ACCEPTED AFTER THE TITLE, ABSTRACT, AND FULL PAPER READ	26
TABLE 2.9.1: PUBLICATION VENUE OF THE PAPERS CONSIDERED IN SLR	28
TABLE 2.9.2: FREQUENCY OF THE MOST USED TERMS IN THE ABSTRACT SECTION	28
TABLE 2.10.1: RANKING METHODS USED TO RANK USERS IN VARIOUS SMPS	34
TABLE 3.1.1: THE MAIN STAGES AND THEIR PHASES IN THE RESEARCH DESIGN	64
TABLE 3.2.1: THE DESCRIPTION OF THE DATASET01	70
TABLE 3.2.2: FEATURES EXTRACTED FROM TWEETS BY A PARTICULAR USER	73
TABLE 3.2.3: FEATURES EXTRACTED FROM A PARTICULAR TWEETER USER	74
TABLE 3.2.4: FEATURES EXTRACTED FROM TWITTER SMP ARE CATEGORIZED ACCORDING TO THE FOUR CATEGORIES AS OF CONCEPTUAL FRAMEWORK	74
TABLE 3.2.5: THE USE OF SINHALA AND ENGLISH LANGUAGE IN THE FORM OF MONOLINGUAL, CODE-MIX AND CODE-SWITCH FORMATS	78
TABLE 4.1.1: FEATURES COLLECTED AND ANALYSED TO IDENTIFY THE CHARACTERISTICS OF THE HATE USERS	83
TABLE 4.1.2: ATTRIBUTES FOR PREDICTING PERSONALITY TRAITS	87
TABLE 4.2.1: KEY FEATURES FOR IDENTIFYING THE INFLUENCE OF HATE USERS ON TWITTER SMP	91
TABLE 4.2.2: THE SELECTED FEATURES FOR THE FEATURE ANALYSIS AFTER THE EXPERTS OPINION AND THE LITERATURE	98
TABLE 4.2.3: HYPERPARAMETER USED IN RF, LR AND SVM	106
TABLE 4.2.4: THE OPTIMAL HYPERPARAMETERS OBTAINED BY THE GS AND RS FOR EVERY ML ALGORITHM	106
TABLE 5.1.1: FOLLOWINGS AND FOLLOWERS CENTRAL TENDENCY MEASUREMENTS OF HATE AND NON-HATE USERS	113
TABLE 5.1.2: PERCENTAGE DISTRIBUTION OF INTERACTIONS WITH HATE TWEETS	116
TABLE 5.1.3: PERCENTAGE DISTRIBUTION OF INTERACTIONS WITH NON-HATE TWEETS	117
TABLE 5.1.4: PERFORMANCE COMPARISON OF SUPERVISED ML ALGORITHMS FOR BFPT PREDICTION	122
TABLE 5.4.1: PERFORMANCE METRICS OF DIFFERENT MACHINE LEARNING MODELS - EXPERT-BASED ANNOTATION	125
TABLE 5.4.2: PERFORMANCE METRICS OF DIFFERENT MACHINE LEARNING MODELS - METRICS-BASED VALIDATION	125
TABLE 5.5.1: COMPARISON OF PERFORMANCE METRICS OF PROPOSED ML APPROACH WITH STATE-OF-ART METHODS	126
TABLE 6.1.1: RAW DATASET FOR HATE SPEECH DETECTION	133
TABLE 6.1.2: ANNOTATION SUMMARY OF POSTS, COMMENTS, AND TWEETS	133
TABLE 6.1.3: TOP 10 HIGH-IMPACT TWITTER ACCOUNTS PROMOTING HARMFUL CONTENT RANKED BY EC AND PR METRICS	137
TABLE 6.1.4: SNA ANALYSIS OF THE MENTION NETWORK FOR HATE USERS: PRE-CRISIS VS. DURING CRISIS	137

LIST OF ABBREVIATIONS

Abbreviation Description

ACRA	Aggregation Consensus Rank Algorithm
AHP	Analytic Hierarchy Process
API	Application Programming Interface
AS	Audience Sentiment
BC	Betweenness Centrality
BFPTM	Big Five Personality Traits Model
CC	Closeness Centrality
CDFs	Cumulative Distribution Functions
CLC	Local Centrality with a Coefficient
DC	Degree of Centrality
DEV	Daily Engagement Velocity
EDA	Exploratory Data Analysis
EC	Eigenvector Centrality
EC-SLR	Exclusion Criteria Systematic Literature Review
EO	Engagements & Outreach
FB	FaceBook
FF	Forest Fire Model
flerCnt	Followers Count
flingCnt	Followings Count
GoSL	Government of Sri Lanka
hashCnt	Hashtags Count

HEV	Hourly Engagement Velocity
HITS	Hyperlink-Induced Topic Search
HMCDM	Hierarchical Multiple-Criteria Decision-Making
IB	Influencing Behavior
IC	Independent Cascade Model
IC-SLR	Inclusion Criteria Systematic Literature Review
ICCPR	International Covenant on Civil and Political Rights
IDC	In-Degree Centrality
imgUrl	Availability of the Profile Image URL
lang	Language
len	Length of the Tweet
LIWC	Linguistic Inquiry and Word Count
listCnt	Listed Count
loc	Location
LSTM	Long-Short Term Memory
menCnt	Mentions Count
MPR	Motif-based PageRank
MRC	Medical Research Council
mimgUrl	Availability of the User's Mini Profile Image URL
NB	Naive Bayes
NLTK	Natural Language Toolkit
NLP	Natural Language Processing
OF	Overall Footprint

OSN	Online Social Network
PR	PageRank
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
retweetCnt	Retweets Count
RNN	Recurrent Neural Network
SDS-PBRR	Statistic Dice Similarity-Based Probabilistic Binomial Regression and Ranking
sim	Similarity of the Tweets
SIR	Susceptible-Infected-Recovered Model
SMP	Social Media Platform
SMPs	Social Media Platforms
SMN	Social Media Network
SMUs	Social Media Users
SNA	Social Network Analysis
SPLICE	Structured Programming for Linguistic Cue Extraction
SLR	Systematic Literature Review
SVM	Support Vector Machine
tweetCnt	Tweets Count
tweetFavCnt	Tweet Favorites Count
UDHR	United Nations Universal Declaration of Human Rights
urlCnt	URLs Count
URL	Uniform Resource Locator
userFavCnt	The Number of Favorites the User Has Marked on Others' Tweets
VPNs	Virtual Private Networks

LIST OF APPENDICES

APPENDIX A: INSTITUTIONAL PERSPECTIVES ON HATE SPEECH	169
--	------------