Acc No #825

AUTOMATIC CYBERBULLYING DETECTION THROUGH MACHINE LEARNING TECHNIQUES

FARWA KHALID
03-243191-003

A thesis submitted in the fulfilment of the
requirements for the award of the degree of
Master of Science (Computer Science)

Department of Computer Sciences

BAHRIA UNIVERSITY LAHORE

NOVEMBER 2021

Approval For Examination

Scholar's Name: Farwa Khalid Registration No. 60588

Program of Study: Master of Science in Computer Science

Thesis Title: **Automatic Cyberbullying Detection through Machine Learning Techniques**

It is to certify that the above scholar's thesis has been completed to my satisfaction and, to my belief, its standard is appropriate for submission for examination. I have also conducted plagiarism test of this thesis using HEC prescribe software and found similarity index 11% that is within the permissible limit set by the HEC for the MS degree thesis. I have also found the thesis in a format recognized by the BU for the MS thesis.

Principal Supervisor' Signature:

Mustafa

Date: 02/12/2021

Name: **Dr. Ghulam Mustafa**

Author's Declaration

I, Farwa Khalid here by state that my MS thesis titled **Automatic Cyberbullying Detection Through Machine Learning Techniques** Is my own work and has not been submitted previously by me for taking any degree from this university Bahria University Lahore Campus or anywhere else in the country/world.

At any time if my statement is found to be incorrect even after my graduation, the university has the right to withdraw/cancel my MS Degree.

Name of scholar: Farwa Khalid

Date: 01-12-2021

Plagiarism Undertaking

I, solemnly declare that research work presented in the thesis titled **"Automatic Cyberbullying Detection through Machine Learning Techniques"** is solely my research work with no significant contribution from any other person. Small contribution / help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero-tolerance policy of the HEC and Bahria University towards plagiarism. Therefore, I as an Author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred / cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS degree, the university reserves the right to withdraw /revoke my MS degree and that HEC and the University has the right to publish my name on the HEC / University website on which names of scholars are placed who submitted plagiarized thesis.

Scholar / Author's Sign: 

Name of the Scholar: Farwa Khalid

ACKNOWLEDGEMENT

It is only by the blessings of Almighty Allah, that I was able to complete this tedious task of thesis writing.

I wish to express my sincerest regards to my supervisor, Dr. Ghulam Mustafa, Assistant professor, Bahria University Lahore campus. His interest in teaching, experience in the field of research, depth of technical guidance and most of all, his moral support have been an invaluable asset to me. I express my gratitude to my co-supervisor, Dr. Zia ul Rahman, Assistant Professor, Bahria University Lahore campus for his invaluable guidance and help in making this seemingly impossible task, possible. I would be unfair on my part if I do not remember my patients, whose contribution made this research work possible. May Allah bless them for their contribution in this Sadqa-e-Jariya. I am really thankful to Colleagues for their co-operation and providing me the facilities to carry out the research project. I am extremely grateful to my sister, brother, or family members, who helped me in the compilation of this manuscript. I have no words to express my thanks to my father, who has been my ideal and a constant source of inspiration in everything I do and without whose motivation and prayers this project could not be completed

Last but not the least, I dedicate this thesis to the memory of my dearest mother, for without her prayers I would not be where I am today. Though she left me in the beginning of studies, her prayers remained around me throughout this work and they are still surrounding me.

ABSTRACT

Cyberbullying is defined in many ways but there is vast consensus in the current literature which have cruel, intentional, and repetitive behavior between peers, through social media networks that are online social applications. The incredible fame of social media platform like Facebook, Twitter has been providing more freedom for online users to upload their content on social media sites without any constraints. Many researchers proposed many solutions for the detection of cyberbullying in English, Arabic, Dutch, Hindi, and Marathi. However, little work is done in the detection of cyberbullying in the Roman Urdu language. Many techniques are used in the detection of cyberbullying, mostly Machine Learning and Natural Language Processing. Our proposed method implemented for cyberbullying detection. For this purpose we crawl data from twitter normalized it with some preprocessing techniques. Six different ML classifiers that is logistic regression, random forest, SVM, decision tree, Naïve Bayes and KNN are trained on our benchmark balanced binary class (bullied=1, non-bullied=0) dataset of 3786. Feature are extracted by TF-IDF feature extraction technique and n-gram (unigram and bigram) method is also applied. The benchmark dataset is train and six different ML classifiers are applied and performance is evaluated by accuracy, precision, recall, F1 score and confusion matrix are plotted separately for uni-gram and bi-gram. Also cross validation with K-fold 10 is applied against each ML classifiers. It is noticed that if the N-gram range increases, the performance of supervised ML classifier change. Naïve Bayes performs well among all other classifiers having accuracy 76% and recall score is 78% with unigram. On the other hand SVM have 74.3% accuracy with uni gram and increase in recall score when bi-gram is apply. Recall score with Bi-gram for SVM with k-fold is 90%. It is noticed that SVM recall score is greater than baseline studies with our proposed methodology.

TABLE OF CONTENTS

CHAPTER	TITLE	PAGE
	APPROVAL FOR EXAMINATION	**ii**
	AUTHOR'S DECLARATION	**iii**
	PLAGIARISM UNDERTAKING	**iv**
	DEDICATION	**v**
	ACKNOWLEDGMENT	**vi**
	ABSTRACT	**vii**
	TABLE OF CONTENTS	**viii**
	LIST OF TABLES	**xiii**
	LIST OF FIGURES	**xiv**
	LIST OF ABBREVIATIONS	**xvi**
	LIST OF APPENDICES	**xvii**
	LIST Of SYMBOLS	**xviii**
1	**INTRODUCTION**	**1**
	1.1 Background	1
	1.1.1 Bullying	1
	1.1.2 Cyberbullying	1
	1.2 Prevention of Electronic Crimes Law in Pakistan	2
	1.3 Types of cyberbullying	2
	1.3.1 Flooding	3
	1.3.2 Masquerade	3
	1.3.3 Flaming or bashing	3
	1.3.4 Harassment	3
	1.3.5 Denigration	4
	1.3.6 Outing	4

1.3.7 Trickery 4
1.3.8 Exclusion 4
1.3.9 Cyber stalking 5
1.3.10 Cyber threats 5
1.3.11 Grooming 5
1.3.12 Racism 5
1.3.13 Pull a P** 6
1.4 Significance of the Research 6
1.5 Problem Statement 7
1.6 Key Research Questions 7
1.7 Aims and Objectives 7
1.8 Main contribution 8
1.9 Thesis organization 8
1.10 Summary of Chapter 1 9

LITERATURE REVIEW 10

2.1 Detection of cyber bullying 10
2.2 Effects of Cyberbullying 12
2.2.1 Social media platform involve in cyberbullying 13
2.3 Identifying definition criteria for cyberbullying 14
2.3.1 Cyberbullying –a social media problem 14
2.3.2 Twitter feature in cyberbullying 15
2.3.3 Offensive Language 16
2.3.4 Hate Speech 17
2.3.5 Sentiment analysis on cyberbullied dataset 17
2.4 Machine Learning 17
2.4.1 Types of Machine Learning 18
2.4.1.1 Supervised Machine Learning 18
2.4.1.2 Unsupervised Machine Learning 19
2.5 Cyberbullying in Pakistan 20
2.6 Datasets for cyberbullying detection 21
2.7 Related Work of Cyberbullying Detection 22
2.8 Bilingual Cyberbullying Detection 24

2.9 Related work of Cyberbullying and lexicon Normalization 29
2.10 Sentiment detection for Roman Urdu Language 31
2.11 Draw back for detection of cyberbullying 34
2.11.1 UCD consists of two main components 36
2.11.1.1 Hierarchical Network to model (HAN) 36
2.11.1.2 UCD benefits from the following design mechanisms 37
2.11.1.3 Complementary temporal analysis 37
2.11.1.4 Joint optimization 37
2.12 Classification of severity of cyberbullying 37
2.13 Comparison of datasets, parameters, and preprocessing techniques 38
2.13.1 Summary of datasets 40
2.14 Comparison of various cyberbullying detection techniques 41
2.15 Summary of Chapter 2 44

METHODOLOGY **45**
3.1 Research Methodology 45
3.1.1 Methodology Step 45
3.2 Literature Review and Find Research Gap 46
3.3 Dataset Generation 47
3.4 Data Authentication 48
3.5 Data Pre-processing 49
3.5.1 Stage 1 50
3.5.1.1 Labeling of tweets 51
3.5.3 Stage 2 52
3.5.3.1 Tokenization 52
3.5.3.2 Stop words Removal 53
3.5.3.3 Normalization 53
3.5.4 Stage 3 54
3.6 Feature Extraction 54
3.7 TF-IDF 55
3.7.1 Term Frequency 56
3.7.2 Inverse Document Frequency 56
3.7.3 TF-IDF 56
3.7.4 Vectorizer 57

3.7.5 N-gram 57
3.8 Splitting dataset into Training and Testing 58
3.8.1 Train_Test_Split Function 58
3.8.2 Training and Testing 59
3.9 Summary of Chapter 3 59

RESULTS AND EVALUATION 60

4.1 Word Cloud Representation 60
4.1.1 Bilingual Corpus 60
4.1.2 Roman Urdu Corpus 61
4.1.3 English Corpus 61
4.2 Statistics of data 62
4.2.1 Bullied and non-bullied dataset representation 64
4.2.2 Language based dataset Representation 63
4.3 Bullied and Non Bullied Dataset Representation Against Language 64
4.4 Experimental Setup 65
4.5 Evaluation Measures 66
4.5.1 Accuracy 66
4.5.2 K-fold Cross Validation 66
4.5.3 Confusion Matrix 67
4.5.4 Precision 67
4.5.5 Recall 67
4.5.6 F1-Score 68
4.5.7 Roc Curve 68
4.6 Machine learning algorithms 68
4.6.1 Machine learning classifiers with TF-IDF + Uni-gram 68
4.6.2 Machine learning classifiers with TF-IDF + Bi-gram 69
4.7 Logistic Regression 70
4.7.1 Confusion matrix of Logistic Regression (Uni-gram) 71
4.7.2 Logistic regression confusion matrix (Bi-gram) 72
4.8 Random Forest 72
4.8.1 Confusion matrix of Random Forest (Uni-gram) 73
4.8.2 Confusion Matrix of Random Forest (Bi-Gram) 73

4.9 Decision Tree Classifier **74**
4.9.1 Confusion matrix of Decision Tree (Uni-gram) 76
4.9.2 Confusion matrix of Decision Tree (Bi-gram) 76
4.10 SVM 77
4.10.1 Confusion Matrix of SVM (Uni-gram) 78
4.10.2 Confusion Matrix of SVM (Bi-gram) 78
4.11 Multinomial NB 79
4.11.1 Confusion Matrix of Multinomial NB (Uni-gram) 80
4.11.2 Confusion Matrix of Multinomial NB (Bi gram) 81
4.12 K-Nearest Neighbours (KNN) 81
4.12.1 Confusion Matrix of KNN (Uni-gram) 83
4.12.2 Confusion Matrix of KNN with K fold (Uni-gram) 83
4.12.3 Confusion Matrix of KNN (Bi gram) 84
4.12.4 Confusion Matrix of KNN with K Fold (Bi gram) 84
4.13 Discussion 85
4.14 Summary of Chapter 4 87

CONCLUSION **88**

5.1 Research Objective Discussion 89
5.2 Main contribution of our study 89
5.3 Limitations 90
5.4 Future work 90

REFERENCES **91**

Appendix A **99**

LIST OF TABLES

TABLE NO.	TITLE	PAGE
2.1	Comparison of datasets, parameters, and preprocessing techniques	39
2.2	Comparison of various cyberbullying detection techniques	41
3.1	Sample of Dataset	48
3.2	Data view after 1st stage of pre-processing	50
3.3	Data view after 2nd stage of pre-processing	52
3.4	Data view after 3rd stage of pre-processing	52
3.5	List of Roman Urdu Stopwords	53
3.6	Data view after complete 3rd stage of pre-processing	54
3.7	Data view after 4th stage labeling	54
3.8	TF-IDF of doc1 and doc2	56
3.9	Dataset splitting	58
4.1	Statistics of data	62
4.2	Class * Language Cross tabulation	64
4.3	Measured values of Logistic Regression	70
4.4	Measured values of Random Forest	73
4.5	Measure values of Decision Tree	75
4.6	Measure values of SVM	77
4.7	Measured values of Multinomial Naïve Bayes	80
4.8	Measured values of KNN	82
4.9	Comparison with baseline of Naïve Bayes	86
4.10	Comparison with baseline of SVM	87

LIST OF FIGURES

FIGURE NO.	TITLE	PAGE
1.1	Types of cyberbullying	6
1.2	Organization of Thesis	9
2.1	Supervised Machine Learning	18
2.2	Unsupervised Machine Learning	19
3.1	Proposed Methodology	46
3.2	Flow diagram for pre-processing	49
3.3	Flow diagram performance evaluation	55
3.4	TF-IDF Vector	57
3.5	Word Feature	57
4.1	Cloud Representation of Bilingual Corpus	61
4.2	Cloud Representation of Roman Urdu Corpus	61
4.3	Cloud Representation of English Corpus	62
4.4	Bullied and non-bullied dataset	63
4.5	Language Based graph	63
4.6	Language and Class Graph	65
4.7	K-Fold	66
4.8	ML classifiers performance with Uni-gram	69
4.9	ML classifiers performance with Bi-gram	69
4.10	Confusion matrix of Logistic Regression (Uni-gram)	71

4.11	Confusion matrix of Logistic Regression (Bi-gram)	72
4.12	Confusion matrix of Random Forest (Uni-gram)	74
4.13	Confusion matrix of Random Forest (Bi-gram)	74
4.14	Confusion Matrix of Decision tree (Uni-gram)	76
4.15	Confusion Matrix of Decision tree (Bi-gram)	76
4.16	Confusion Matrix of SVM (Uni-gram)	78
4.17	Confusion Matrix of SVM (Bi-gram)	79
4.18	Confusion Matrix of NB (Uni-gram)	81
4.19	Confusion Matrix of NB (Bi-gram)	81
4.20	Confusion Matrix of KNN (Uni-gram)	83
4.21	Confusion Matrix of KNN (Uni-gram with k-fold 10)	83
4.22	Confusion Matrix of KNN (Bi-gram)	84
4.23	Confusion Matrix of KNN (Bi-gram with k-fold 10)	84
5.1	Main contribution of our study	89

LIST OF ABBREVIATIONS

ICT	-	Information and Communication Technologies
Ml	-	Machine Learning
NLP	-	Nature language processing
BOW	-	Bag of words
CSMT	-	Code Mixed social media text
IBK	-	Instance Based Learner
NB	-	Naïve Bayes
SVM	-	Support vector machine
DT	-	Decision Tree
RF	-	Random Forest
LR	-	Logistic regression
ADB	-	AdaBoost
SNAc	-	Social Network Analysis
MLR	-	Multinomial Logistic Regression
ANN	-	Artificial neural Network
KNN	-	k-nearest neighbor's algorithm
SMN	-	Social Media Network
LSA	-	Latent Semantic Analysis
LDA	-	Latent Dirchilet Allocation
EBoW	-	Embeddings-enhanced Bag-of-Words Model
TERUN	-	Transliteration based Encoding for Roman Hindi/Urdu text normalization
OAG	-	Overtly Aggressive
CAG	-	Covertly Aggressive
LSTM	-	Using Long Short-Term Memory

LIST OF APPENDICES

APPENDIX	TITLE	PAGE
A	Code	99

LIST OF SYMBOLS

SYMBOLS	TITLE	PAGE
σ	Non-linear Activation Function	25