

**MEASURING TRUSTWORTHINESS OF WORKERS IN  
THE CROWDSOURCED COLLECTION OF  
SUBJECTIVE JUDGEMENTS**

Gnei Sleemani Nadeera Meedin

198113V

Degree of Doctor of Philosophy

Department of Computer Science & Engineering  
Faculty of Engineering

University of Moratuwa  
Sri Lanka

November 2023

**MEASURING TRUSTWORTHINESS OF WORKERS IN  
THE CROWDSOURCED COLLECTION OF  
SUBJECTIVE JUDGEMENTS**

Gnei Sleemani Nadeera Meedin

198113V

Thesis/Dissertation 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

November 2023

## 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: 22/11/2023

The above candidate has carried out research for the ~~PhD/MPhil/Masters~~ thesis/dissertation 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

Signature of the Supervisor:

Date: 22/11/2023

## **DEDICATION**

Dedicated to my loving mother, husband and brother

## ACKNOWLEDGEMENT

I would like to express sincere appreciation to the people and organizations listed below for being their with me throughout my research journey and helping me.

My supervisor Prof. GIUS Perera for encouraging and guiding me throughout this period and for bearing with my frequent lapses. I appreciate all his contributions of ideas and motivation in multiple circumstances. Moreover, his passion for the research was infectious and served as a source of motivation for me.

My progress review evaluation panel members; Dr. (Ms.) S. Ahangama, Dr Charith Chithranan and Dr Shehan Perera for providing constructive feedback on improving my research. It would not be possible for me to reach the level without their support and guidance.

My research team members, Dr L. Ranathunga, Dr.(Ms) S. Ahangama, Dr C.R.J. Amalraj, Dr Shalinda Adikari, Ms K.S.A. Walawage, Ms N. Rajapaksha, Mr H.M.M. Caldrea and Ms WASN Perera for the constructive criticism and suggestions provided throughout my research journey.

Accelerating Higher Education Expansion and Development (AHEAD) Operation of the Ministry of Higher Education, funded by the World Bank, was an excellent aid for me.

The directors of the Operations Technical Secretariat(OTS), Prof. Sanath Jayasena and Prof. Shantha Fernando, and the OTS staff for the continuous support extended throughout the PhD journey.

The academic staff of the Department of Computer Science and Engineering, I thank all the lecturers who provided feedback on project progress evaluations. The feedback helped me immensely in bringing the research up to this level.

Prof. K. G. H. U. W Uditha Rathnayake from the Open University of Sri Lanka for the continuous encouragement extended to carry out the research and for encouraging me to complete the research successfully.

Finally, yet importantly, I am grateful for my family and friends who were always behind me and supporting me throughout my life.

## ABSTRACT

Social media platforms have become integral parts of our lives, enabling people to connect, share, and express themselves on a global scale. Alongside the benefits, there are also substantial challenges that arise from the unfiltered and unrestricted nature of these platforms. One such challenge is the presence of inappropriate and hateful content on social media. While platforms employ algorithms and human moderators to identify and remove inappropriate content, they often struggle to keep up with the constant flood of new posts. Social media posts are written in a variety of languages and multimedia formats. As a result, social media platforms find it more difficult to filter these before reaching a more diverse audience range, as moderation of these social media platform posts necessitates greater contextual, social, and cultural insights, as well as language skills.

Social media platforms use a variety of techniques to capture these insights, and linguistic expertise to effectively moderate social media posts. These techniques help platforms better understand the degrees of content and ensure that inappropriate or harmful posts are accurately identified and addressed. These techniques include Natural Language Processing (NLP) algorithms, keyword and phrase detection, image and video recognition, contextual analysis, cultural sensitivity training, machine learning, AI improvement etc. Data annotation forms the foundation for training these algorithms and identifying and classifying various types of content accurately. Often crowdsourcing platforms such as Mechanical Turk and Crowd Flower are used to get the datasets annotated in these techniques.

The accuracy of the annotation process is crucial for effective content moderation on social media platforms. Crowdsourcing platforms take several trust measures to maintain the quality of annotations and to minimize errors. In addition to these procedures, determining the trustworthiness of workers on crowdsourcing platforms is critical for ensuring the quality and reliability of the contributions they give. Accuracy metrics, majority voting, completion rate, inter-rater agreement, and reputation scores are a few such measurements used by existing researchers. Even though majority voting is used to ensure consensus, existing research shows that the annotated results do not reflect the actual user perception and hence the trustworthiness of the annotation is less.

In this research, a crowdsourcing platform was designed and developed to allow the annotation process by overcoming the limitations of measuring trustworthiness which would facilitate identifying inappropriate social media content using crowd responses. Here the research focus was limited to social media content written in Sinhala and Sinhala words written in English (Singlish) letters as the most popular Mechanical Turk and Crowd Flower do not allow workers from Sri Lanka.

As outcomes of this research, a few novel approaches were proposed, implemented, and evaluated for hate speech annotation, hate speech corpus generation, measuring user experience, identifying worker types and personality traits and hate speech post-identification. In addition, the implemented crowdsourcing platform can extend the task designs to other annotation tasks; language and inappropriate content identification, text identification from images, hate speech propagator ranking and sentiment analysis. When evaluating the quality of the results for accuracy and performance, it was identified that the consensus-based approach of ensuring the trustworthiness of crowdsourcing participants is highly affected by the crowd's biases and the Hawthorne effect. Therefore, a comparison and analysis of the annotation quality of the crowdsourcing platform with consensus, reputation, and gold

standard-based approaches were conducted and a model to measure the trustworthiness of crowd response was developed.

The major outcome of this research is the crowdsourcing platform that can be used for local annotation processes with the assurance of worker reliability. The number of tasks completed by the workers within a given period, the number of tasks attempted by each worker within a given period, the percentage of tasks completed compared to tasks attempted, time taken to complete tasks, the accuracy of responses considering golden rules, time taken to submit responses after each task assignment and the consistency of response time provided were identified as the quantitative measurements to assess the trustworthiness of workers. After this identification, the relationship between reputation score, performance score and bias score was formulated by analysing the worker responses. The worker behaviour model and trust measurement model showed an accuracy of 87% and 91% respectively after comparing with the expert response score which can be further improved by incorporating contextual analysis, worker belief and opinion analysis.

The proposed methodology would accelerate data collection, enhance data quality, and would promote the development of high-quality labelled datasets.

**Keywords:** Annotation, Collaboration, Crowdsourcing, Human-Computer Interaction, Trustworthiness

# TABLE OF CONTENTS

Declaration .....	i
Dedication .....	ii
Acknowledgement.....	iii
Abstract .....	iv
Table of Contents .....	vi
List of Figures .....	x
List of Tables.....	xii
List of Abbreviations.....	xiii
List of Appendices .....	xv
Chapter 1 .....	1
INTRODUCTION .....	1
1.1 Problem Statements .....	3
1.2 Research Questions .....	3
1.3 Motivation .....	4
1.4 General Objectives .....	4
1.5 Specific Research Objectives .....	4
1.6 Organization of the Thesis .....	5
Chapter 2 .....	7
LITERATURE SURVEY .....	7
Introduction .....	7
2.1 Social Media and Social Media Content Moderation .....	7
2.2 Algorithms used in hate speech detection and used annotation schemas.....	9
2.3 Crowdsourcing .....	11
2.3.1 Definition of Crowdsourcing .....	11
2.3.2 Use of Crowdsourcing.....	11
2.3.3 Crowdsourcing for Human-Computer Interaction.....	12
2.3.4 Crowdsourcing for Social Media Data Mining .....	13
2.3.5 Characteristics of Crowdsourcing .....	13
2.3.6 Motivators for Crowdsourcing Participants .....	13



2.3.7	Challenges in Crowdsourcing Platforms.....	14
2.3.8	Selection of Personas from Crowdsourcing Participants.....	15
2.3.9	Crowdsourced Data Management.....	15
	Conclusion.....	16
Chapter 3	.....	17
SYSTEMATIC LITERATURE REVIEW .....		17
	Introduction.....	17
3.1	Systematic Literature Review Methodology.....	18
3.1.1	Research Questions.....	18
3.1.2	Data Source.....	19
3.1.3	Searching Approach.....	19
3.1.4	Criteria for Inclusion and Exclusion.....	20
3.1.5	Process of Selecting Studies.....	21
3.2	Measuring the Trustworthiness of Crowd Participant Responses.....	22
3.2.1	Reputation management in crowdsourcing systems.....	23
3.2.2	Aggregation Techniques in Crowdsourcing.....	23
3.3	Measuring Bias of the Workers.....	23
3.3.1	Approaches and methods to verify the quality of the submitted annotations.....	24
	Categorizing Worker Types.....	24
3.3.3	Measuring the quality of annotations through ground truth inference.....	25
3.3.2	What are the different types of bias and methods used to eliminate the bias?	33
3.3.3	What are the methods used to measure the trust of crowd response?..	37
Chapter 4	.....	40
RESEARCH METHODOLOGY AND CONCEPTUAL MODEL .....		40
	Introduction.....	40
4.1	Research Design.....	40
4.2	Preliminary face-to-face study with social media users.....	41
4.3	Conceptual Model of the Crowdsourcing Platform.....	44
	Conclusion.....	50
Chapter 5	.....	51

Solution Design and Implementation.....	51
Introduction .....	51
5.1 Crowdsourcing Platform to Moderate Social Media Content .....	51
5.2 Novel Annotation Scheme.....	57
5.2 Worker Behaviour Model for Crowdsourcing Platform .....	58
5.3 Assess the trustworthiness of contributors .....	60
Chapter 6 .....	65
Evaluation and Analysis.....	65
Introduction .....	65
6.1 Preliminary face-to-face study with social media users.....	65
6.2 Analyzing the Annotation Method for Building a Hate Speech Corpus .....	65
6.3 Worker Behaviour Model.....	70
6.4 Model to measure trustworthiness.....	71
6.5 Usability Assessment of the Crowdsourcing Platform.....	77
Chapter 7 .....	80
Discussion .....	80
Introduction .....	80
7.1 Contribution of the Research Papers .....	80
7.2 Recommendations on maintaining trustworthiness.....	85
Chapter 8 .....	87
Conclusion & future development .....	87
Introduction .....	87
8.1 Research contribution.....	87
8.2 Research Limitations .....	92
8.3 Future Work .....	93
References .....	94
APPENDIX A .....	106
QUESTIONNAIRE TO ASSESS THE KNOWLEDGE ON SINHALA LANGUAGE .....	106
APPENDIX B .....	109
QUESTIONNAIRE TO ASSESS THE KNOWLEDGE ON SINGLISH READING .....	109

APPENDIX C .....	111
Sample questionnaire to assess comprehension & analytical skills (Sinhala).....	111
APPENDIX D .....	116
Sample questionnaire to assess the ability to read Singlish .....	116
APPENDIX E .....	122
Implementation of the pre-selection of contributors .....	122
APPENDIX F.....	124
MODEL TO MEASURE TRUSTWORTHINESS OF CROWD RESPONSES USING LOGISTIC REGRESSION .....	124
Appendix G .....	125
System Usability Scale (SUS) Assessment.....	125
APPENDIX H.....	129
FEATURE VALUES TO ASSESS TRUSTWORTHINESS .....	129

## LIST OF FIGURES

<b>Figure</b>	<b>Description</b>	<b>Page</b>
Figure 2.1	Characteristics of crowdsourcing processes	13
Figure 2.2	Participant motivators in crowdsourcing	14
Figure 3.1	The number of qualified search results in each subject area under review.	18
Figure 3.2	Yearly-wise publication count generated by Scopus	20
Figure 3.3	PRISMA Model	21
Figure 3.4	Aspects in measuring the trustworthiness of crowdsourcing	22
Figure 4.1	Workflow diagram of the preliminary study	42
Figure 4.2	JSON schema for Facebook posts	47
Figure 4.3	A segment of JSON schema for Tweets	47
Figure 4.4	A segment of JSON schema for YouTube posts	48
Figure 4.5	Workflow of manual data annotation process	50
Figure 4.6	Neural Network Architecture	50
Figure 5.1	System architecture of the implemented crowdsourcing platform	52
Figure 5.2	Crowdsourcing Platform – Home Page	53
Figure 5.3	Worker Registration Page	53
Figure 5.4	Question types and task types in the questionnaire	55
Figure 5.5	Worker registration and reward process	57
Figure 5.6	Assignment of rewards for chosen contributors	58
Figure 5.7	System Overview of Worker Behaviour Model	60
Figure 5.8	System overview of the model designed to assess the trustworthiness of contributors	62
Figure 5.9	Ground Truth Inference	63
Figure 6.1	Gender Distribution of Selected Contributors	67
Figure 6.2	Contributors' age distribution	67
Figure 6.3	Contributors' Religious Distribution	68
Figure 6.4	Find the optimal number of clusters	71

Figure 6.5	Comparison of the Accuracy of Consensus-based, reputation based and Gold standard approach	72
Figure 6.6	Comparison of precision for Consensus-based, reputation based and Gold standard approach	73
Figure 6.7	Relationship between reputation score, performance score and bias score	74
Figure 6.8	Clusters of Data points with K-Means Algorithm	75

## LIST OF TABLES

<b>Table</b>	<b>Description</b>	<b>Page</b>
Table 2.1	Benchmarks of labelled datasets used to identify hate speech	10
Table 3.1	Document types of resulting literature	19
Table 3.2	Algorithms for general ground truth inference.	33
Table 4.1	Main steps and stages of research design	41
Table 4.2	Unlabelled data sets for preliminary study	43
Table 4.3	Conditions used to determine whether or not a message contains hate speech	43
Table 4.4	Observations from the Preliminary Study	44
Table 4.5	Unlabelled data sets	46
Table 4.6	Details of data stored in the crowdsourcing platform	50
Table 5.1	Symbol definitions list for contributor pre-selection	57
Table 5.2	Pre-selection criteria for contributors	57
Table 5.3	List of symbol definitions for labelling	60
Table 5.4	Features considered from contributor responses	62
Table 5.5	Variable definitions list for calculating reputation score	64
Table 6.1	Demographic characteristics of the preliminary study participants	67
Table 6.2	Inter annotator agreement for L1	70
Table 6.3	Annotation results for Facebook, Twitter and Youtube data-Hate and No Hate	70
Table 6.4:	Annotation results for Facebook -Offensive and None Offensive	70
Table 6.5	The most frequent terms found in hate speech related to hate targets	71
Table 6.6 :	Observed TP, FP, FN and TN values along with calculated precision and accuracy for Consensus-Based (CB), Reputation-Based(RB) and Golden Standard(GS) approaches.	77
Table 6.6	The most frequent terms found in hate speech related to hate targets	78

## LIST OF ABBREVIATIONS

<b>Abbreviation</b>	<b>Description</b>
ABAE	Attention-based Aspect Extraction
ACF	Adversarial Colluded Followers
ACL	Adversarial Colluded Leader
AF	Adversarial Filtering
AggSLC	Aggregation method for Sequential Labels from Crowds
AHEAD	Accelerating Higher Education Expansion and Development
AMT	Amazon Mechanical Turk
API	Application Programming Interface
AWMV	Adaptive Weighted Majority Voting Algorithm
BERT	Bidirectional Encoder Representations from Transformers
BoW	Bag-of-words
BLSTM	Bidirectional LSTM
BTL	Bradley–Terry–Luce model
DAU	daily active users
CF	collaborative filtering
CNN	Convolutional Neural Network
CRT	Critical Race Theory
DS	David and Skyne
DNN	Deep Neural Network
DP	Differential privacy
ELICE	Expert Label Injected Crowd Estimation
EM	Expectation Maximization
XGBoost	Extreme Gradient boosted Decision Trees
FD	Fast Deceivers
FFNN	Feed Forward NN
GLAD	Generative model of Labels, Abilities, and Difficulties
GSP	Gold Standard Preys

GTIC	Ground Truth Inference using Clustering
HTMS	Hierarchical Trust Management System
HBT	Heuristics-and-Biases Test
HP	Honeypot
HCI	Human-Computer Interaction
HIT	Human Intelligence Task
IE	Ineligible Workers
ITER	Iterative Learning
IJACSA	International Journal of Advanced Computer Science and Applications
LCs	Labelled Categories
LSTM	Long Short-Term Memory
MLE	Maximum Likelihood Estimation
MD	Major Decision
MV	Majority Voting
MACE	Multi-Annotator Competence Estimation
MLP	Multilayer Perceptron
NN	Neural Networks
NACL	Non-Adversarial Colluded
NACF	Non-Adversarial Colluded Followers
OTS	Operations Technical Secretariat
PMI	Pointwise Mutual Information
PLAT	Positive LAbel frequency Threshold
RSPM	Raven's Standard Progressive Matrices
RB	Rule Breakers
SD	Smart Deceivers
SDS	Spectral DS
SVM	Support Vector Machine
SLME	Supervised Learning from Multiple Experts
SRT	Syllogistic Reasoning Test
SST	Strong stochastic transitivity model
UGC	User-Generated Content
WMV	Weighted majority voting



## LIST OF APPENDICES

<b>Appendix</b>	<b>Description</b>	<b>Page</b>
APPENDIX - A	Sample questionnaire to assess the knowledge level of hate speech	108
APPENDIX - B	Sample questionnaire to assess language proficiency (Sinhala)	111
APPENDIX - C	Sample questionnaire to assess comprehension & analytical skills (Sinhala)	113
APPENDIX - D	Sample questionnaire to assess the ability to read Singlish	118
APPENDIX - E	Implementation of the pre-selection of contributors	124
APPENDIX - F	Model to measure the trustworthiness of crowd response.	126
APPENDIX - G	System Usability Scale (SUS) assessment	127
APPENDIX – H	Feature values to assess trustworthiness	132