

**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

Figure	Description	Page
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

Table	Description	Page
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

Abbreviation	Description
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

Appendix	Description	Page
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

CHAPTER 1

INTRODUCTION

Social media has substantially influenced user decision-making, sharing information, forming opinions and attitudes etc. [1]. The most used social media network in the United States as of May 2021 [2] is Facebook and during the third quarter of 2022, Facebook reported almost 1.98 billion daily active users (DAU) [2]. In addition, 8.2 million active social media users were in Sri Lanka, while the country's population was around 21 Million as of January 2022 [3]. These statistics illustrate that social media plays a significant role worldwide and in Sri Lanka.

Social media is used by people to interact with others, share their experiences, and create communities. Users of social media are constantly updating and disseminating messages and comments on these platforms as users can express themselves freely across languages, cultures, and countries.

Nevertheless, to ensure that its users feel secure using their services, social media companies work to uphold their community standards. Facebook community standards [4] are one such example. According to their standards, users should refrain from posting anything that falls into categories like "coordinating harm damage and promoting crime," "violence and provocation," "dangerous individuals and organizations," etc. on Facebook to avoid any potential harm. Facebook prohibits users from posting hate speech to prevent the spread of segregation and extortion, both of which would encourage actual violence. Similar to Facebook, Twitter [5] and YouTube [6] each have policies against hate speech.

However, despite the existence of norms and standards and the removal of offensive information by social media sites, users frequently utilize these channels to disseminate hate speech and social media posts might include offensive content and have a negative social impact. Some of these efforts, such as citizen voices, would directly contribute to the emergence of social problems. There must be a method for recognizing the items that would directly affect society by causing problems.

Social media sites have found it difficult to filter the billions of posts that are made every day in more than a hundred different languages. One such example of moderating social media posts is to detect hate speech. The delicate balance between hate speech and free speech has been a topic of ongoing research and conflict in discussions surrounding freedom of expression. While free speech is a fundamental right that fosters open discourse and the exchange of ideas, hate speech raises concerns about the potential harm it can inflict on individuals and communities and makes it difficult to strike a balance between [7].

There is no accepted definition of hate speech [8] and it is an argumentative term[9]. According to the strategy and action plan of the United Nations, "any kind of communication in speech, writing, or behaviour, that attacks or uses pejorative or

discriminatory language concerning a person or a group based on who they are, that is, based on their religion, ethnicity, nationality, race, colour, descent, gender, or other identity factor" [10] is considered hate speech. Therefore, "free speech" and "hate speech" vary by region. It is crucial to take cultural and social factors, information related to the context, etc. into account while interpreting the speech. One of the difficulties social media networks encounter when they deal with hate speech is identifying it based on user context [7].

Social media platforms use a social media content moderation process to prevent this challenge. Using crowdsourcing, content identification, and content classification techniques with machine learning techniques to identify and moderate social media content is possible. Social media has given researchers studying social computing a unique perspective into how people interact with one another; in particular, Twitter is utilized to analyse touchy subjects like discrimination [11]. Crowdsourcing has the potential to reach a wider audience and gather user opinions and behaviours.

As a result, a crowdsourcing approach was suggested in this research [12] to capture the subjective aspects of the users who use Sinhala and Singlish when posting on social media platforms. The Literature Survey Section goes into great length about the procedure for moderating social media content as well as how crowdsourcing is used in identifying hate speech.

Multiple characteristics of crowdsourcing are described in [13], and the researchers have identified four dimensions in selecting, accessing others' contributions, collecting others' responses and payment. Furthermore, this research focuses on identifying the factors that would be effective for each of these characteristics in the Sri Lankan context to incorporate into the crowdsourcing framework.

Similarly, this research focuses on identifying mechanisms to use with direct and indirect crowdsourcing to collect opinions on the posts shared by Sri Lankan citizens on various topics in social media. The platform could identify and classify the positions if they were related to social issues. Furthermore, the crowdsourcing platform was used to verify the already placed posts.

While evaluating the quality of the crowdsourcing data, it was identified that the consensus-based approach of ensuring the trustworthiness of crowdsourcing participants is highly affected by the biases of the crowd and the Hawthorne effect [14]. The Hawthorne effect impacts the implementation framework's accuracy and performance. 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. The accuracy of the models significantly relies on the annotated dataset. If the annotation does not reflect accurate labelling, it will affect the result of the detection mechanism.

To overcome these problems, this research proposes to compare and analyse the annotation quality of the crowdsourcing platform using consensus-based, reputation-based, and gold standard-based approaches. The research focuses on identifying the

best trust metric and modelling the trustworthiness of the crowd workers of the crowdsourcing platform. A reputation score was assigned to individual contributors based on their past performance on the platform, worker category and biases demonstrated as an appropriate trust metric to measure the trustworthiness of the workers on the platform. The crowdsourcing platform was evaluated for quality, accuracy and performance using crowd.

1.1 Problem Statements

This research addresses the following three problems;

Research Problem 1:

There is a lack of a software platform to annotate Sinhala and Sinhala words written using English letters by acquiring contextual and language proficiency along with cultural and social insights identifying mechanisms to use with direct and indirect crowdsourcing to collect opinions on the social media posts shared by Sri Lankans.

Research Problem 2:

The problem of having low accuracy and performance of the annotation process with a consensus-based approach is because of the high impact of crowd biases and the Hawthorne effect.

Research Problem 3:

The problem of evaluating the quality of the annotated datasets using the crowdsourcing approach for contributor trustworthiness

1.2 Research Questions

This thesis tries to investigate solutions to four distinct questions while considering the context of such problems.

Research Question 1: What are the techniques to implement in the crowdsourcing platform to pre-select contributors, contributor reward, contributor reputation management and moderate hate speech content?

Research Question 2: How to derive quantitative measurements for social media content analysis using crowd?

Research Question 3: How to perform the analysis of user responses to obtain meaningful insights?

Research Question 4: How to measure and ensure the trustworthiness of users in their responses?

1.3 Motivation

Crowdsourcing platforms such as Mechanical Turk give researchers instant access to a distinct set of workers. However, limits workers to register from only 43 countries as of 2022 [15]. One possible issue with Mechanical Turk data collecting is that respondents who complete many surveys or experiments may become bored and pay less attention to their duties or answer questions in the way they perceive the requester's desires. As a result, the responses are frequently of poor quality. Furthermore, the samples were frequently taken from easy-to-reach people, implying that they are not representative of the general population.

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. Furthermore, if the annotation does not reflect accurate labelling, it will affect the result of the detection mechanism.

The main research objective is to implement a suitable crowdsourcing platform to identify different categories of hate in the Sri Lankan context, identify social media posts with Sinhala and Singlish hate speech, identify a hate speech corpus and facilitate researchers to reach a much larger audience to get their datasets labelled and annotated from social media contents. In addition, the proposed crowdsourcing platform would allow researchers to get their dataset annotated under text identification in images, sentiment analysis etc., which are not yet implemented but given the capability to extend the functionality.

1.4 General Objectives

This research's main objective is to:

Determine an appropriate crowdsourcing mechanism to capture user inputs, thereby implementing a framework to moderate social media content by providing a solution to measure the trustworthiness of crowd response ensuring the quality of captured user inputs.

1.5 Specific Research Objectives

The following specific objectives were identified to address the objective specified above.

1. Design an analytical framework with identified techniques to moderate social media content using crowdsourcing.
2. Identify appropriate trust metrics to evaluate the reliability of the crowd response.
3. Implement a trust metric to enable trust modelling and reasoning about crowd trust.
4. Implement the crowdsourcing platform to facilitate inappropriate content identification with necessary quality control and analytical features.
5. Evaluate the results for quality, transparency, accuracy, and performance of the platform using the crowd.

1.6 Organization of the Thesis

The thesis is spread out under the research areas of User experience (UX), Human-centred computing, Computer-supported cooperative work, and natural language processing. Mainly the research contributes to filling the research gap of ensuring the quality of crowd response with the collaborative workforce from different contexts, such as with multiple cultural insights, language use etc. The thesis proposes and originates a framework to allow workers to capture the worker experience of the annotation process. The proposed framework can allow natural language processing researchers to embed their annotation tasks and select the annotation technique they are willing to use by specifying the workflows. Initially, the framework allows users to specify the annotation task in six categories; language and inappropriate content identification, image text identification, hate speech identification, hate speech propagator identification, hate corpus generation and sentiment analysis. The results can improve the annotation process with the use of cutting-edge technology which would eventually create a better cyberspace.

The rest of this thesis is divided into the following chapters. Chapter 2 of the thesis consists of three sub-sections that provide the background knowledge and comprehension of the existing research on the social media content moderation process, algorithms used in hate speech detection and used annotation schemas and crowdsourcing. In addition, the theoretical background based on their use in other research, the challenges faced by different researchers when carrying out similar research, and the methodologies used to overcome the challenges are discussed in each sub-section.

Furthermore, chapter 3 details a systematic review measuring workers' bias to ensure trustworthiness in the crowdsourced collection of subjective judgements. This chapter attempts to find the answers to the following research questions.

1. What methods are used to verify the quality of the submitted annotations?
2. What are the different types of bias and methods used to eliminate the bias?
3. What are the methods used to measure the trust of crowd response?

Chapter 4 compromises with the methodology of the research and the conceptual. The methodology is explained in steps to provide an easy guide to the thesis. Furthermore, this chapter explains the objectives, data collection process, and the methodology of the preliminary face-to-face study conducted. The conceptual model of the proposed crowdsourcing framework and the rationale behind selecting the functionalities of the framework are also elaborated.

Chapter 5 outlines the design and implementation of the suggested crowdsourcing platform solution. This chapter investigates the framework's components, explaining the reasoning behind their intention, as well as discussing their functionalities. The chapter also discusses the user experience measuring technique to evaluate the crowdsourcing platform and suggests a new method to annotate social media posts and

thereby identify hate speech corpus. Finally, the chapter explains the proposed trust measurement model for crowdsourcing platforms.

Chapter 6 summarizes the evaluation and analysis of the proposed approach under four sections; results of the preliminary study performed with face-to-face workers to understand the user in a crowdsourcing platform, identifying the worker demographics and worker types, validating of trustworthiness of user inputs, and measuring quality attributes for the model and the proposed techniques in chapter 4.

Chapter 7 discusses the key findings of the research and how these findings align and differ from the existing literature. Furthermore, this chapter discusses the significance of the findings. The limitations of the study are explained at the end.

Chapter 8 summarizes the research work, discusses the limitations of the research work in length, and suggests future works as the conclusion. Finally, references and the appendixes are included at the end of the report.

CHAPTER 2

LITERATURE SURVEY

Introduction

This section aims to provide the necessary theoretical background on the sub-areas of social media, social media content moderation, algorithms used in hate speech detection and used annotation schemas, crowdsourcing, use of crowdsourcing in social media mining, analytical capabilities implemented in crowdsourcing platforms, identification of hate speech, and the similar research approaches of different researchers. In addition, each section discusses the challenges faced by the researchers and the methodologies they have identified to overcome them. Finally, the identified constraints and the critical facts considered in this research are listed at the end of each section.

2.1 Social Media and Social Media Content Moderation

This section gives an overview of the social media content moderation process, the kinds of content that are not allowed in social media platforms, how this content is identified using crowdsourcing and by the social media content moderators in different platforms and finally, the need for a novel and human-friendly mechanism to moderate this content.

Social media users can generate and diffuse content, access information and potentially reach large audiences. *Content moderation* [16] is the structured activity of reviewing such User Generated Content (UGC) posts, profiles, and user accounts on online platforms. Using humans and automating the process are the two types of content moderation [17].

Social Media platforms have specified the content categories that are not allowed in posting on social media and from being screened in their community standards and policy documents [5], [18], [19]. The procedures used in content moderation vary from platform to platform. Social media platforms engage in content modification as a strategic measure to safeguard content from adverse publicity while concurrently governing the user experience. This multifaceted approach is driven by the platforms' objectives to uphold user satisfaction, maintain a harmonious online environment, and align with prevailing community guidelines and standards.

To protect content from negative publicity, these platforms implement algorithms and content filtering mechanisms that detect and mitigate potentially harmful, offensive, or misleading material. By stopping such content before it gains traction, social media platforms try to prevent its wide dissemination, thereby minimizing reputational risks to both the platform and its users. This proactive perspective also serves to create a more conducive and trustworthy digital ecosystem.

Methods under content moderation techniques include activities such as deleting offensive or insulting information, banning users, using text filters to block specific terms or content, and applying a variety of other moderation measures [20]. People moderate information for several reasons, including prestige, status, and humanitarian concerns. Furthermore, some moderators may get compensated in the form of fees or reduced access to online services [21]. Nonetheless, the volunteer-based method of content moderation is still widely used in many online communities and platforms today.

Acquiring human resources with their own linguistic, geographical, and cultural knowledge and skills is essential for making formal judgments on contested UGC moderation topics, including video, text, or image-based content suitability for a specific site [21].

However, not all content is easily moderatable. According to Roberts [20], handling hate speech poses challenges. Social media content moderators may need to engage with potentially lengthy content to discern user intent when dealing with hate speech. Additionally, hate content might constitute only a small fraction of the entire content.

In the context of social media content monitoring, Daniel Faggella in Emerj [22] outlines two components of content moderation: a trained machine learning algorithm that assesses content's appropriateness, and human reviewers who guide the algorithm through manual approval or disapproval to improve its decision-making.

Although machine learning techniques play a role in automated content moderation [23], a substantial portion of the work involves human efforts across the globe in removing such content from websites [24]. Regular users often flag user-generated content (UGC) that disturbs them, initiating the review process [21]. When examining individual cases, applying a binary judgment to a nuanced image, like one captured by renowned photographer Nick Ut[25], might seem unreasonable. Crowd workers may ask to perform annotation jobs without providing the contextual information [26]. Andreas believes that many kinds of submitted content may necessitate different moderation techniques[27], including before and after the moderation process, automation of the process and getting the moderation done by distributing the work[25], encompassing both user and spontaneous/reactive moderation, as well as hybrid moderation [27].

Among the inappropriate content being examined, hate speech stands out as a significant type that should be eliminated before it reaches a broader audience. Out of all the inappropriate content under investigation, hate speech is one such important content that should be removed before reaching a wider audience. The current content moderation process on social media requires a fresh toolkit to integrate contextual understanding, necessitating an approach that evaluates user experience.

2.2 Algorithms used in hate speech detection and used annotation schemas

The proposed research focuses on only identifying and moderating hate speech among the different types of content available on social media. When using crowdsourcing to identify hate speech, it is necessary to identify the legal instruments. For this, the social and behavioural theories, legal instruments used in Sri Lanka and Social media community standards could be used. This section explains the instruments to look at when firing questions to identify hatred in the contents of social media data.

Twitter is used often to identify hate targets, which is defined as a prejudice towards a specific aspect of a group of individuals. This notion goes beyond racism and homophobia to include a broader range of biases in online social platforms[28]. According to this study, the most common hate targets are sexual orientation and race.

In a related study, researchers, while collecting the Tweets, phrased the tasks by specifying if a particular tweet conveys a racist message or tells that some one as a racist is considered a racist [29]. The researchers ensured participants understood that the task pertained to racist Tweets and let them quit the job as they wanted This precaution was taken due to the task's potential exposure to sensitive content. Neural networks and support vector machines are employed in this study to categorize tweets for racism and homophobia.

The HaterNet[30] framework can detect and monitor occurrences of hate speech on social media platforms mainly on Twitter. It possesses the capability to categorize such content and further monitor and analyse patterns of hate-related discussion as well as other adverse sentiments. HaterNet evaluates various methods for categorizing content, focusing on different ways of representing documents and models for text classification. The Long Short-Term Memory (LSTM) neural network and the Multilayer Perceptron (MLP) neural network is extensively utilized for various purposes. A new method was used by the researchers namely a "double deep learning neural approach" by combining LSTM and MLP. This double deep-learning neural approach emerges as the most effective. It considers the embeddings of words, emojis, and expressions within tweets, along with the input from TF-IDF, enhancing its accuracy and performance. These embeddings are acquired using a word2vec methodology based on neural networks.

In their study on identifying hate speech on Twitter, Burnap[31]examined 2,000 tweets to look for race, religion or ethnicity. Human assessors were asked to determine whether the material was offensive or confrontational. This process produced three categories: "yes," "no," and "undecided." The task was carried out using CrowdFlower, involving at least four human annotators. For each annotated piece of content, an agreement score was assigned based on the majority decision from trustworthy workers.

Similar to HaterNet many other researchers have used Neural Networks to detect hate speech. Agrawal and Aweaker [32] utilized four Deep Neural Network (DNN) models

for cyberbullying detection using manually annotated data sets. Convolutional Neural Networks (CNN) and Bidirectional LSTM (BLSTM) were added to the list of HaterNet users.

Fernando and Asier [33] focused on hate speech as the central factor and adopted a customized method to read and categorize each tweet as either neutral or constituting hate speech. Their classification process involved scrutinizing the content for the glorification of physical violence, explicit provocation or imminent danger of physical harm and hate, and content that might offend collective sensitivity, with subjective text interpretation. Additionally, the study employed the Kappa coefficient to assess agreement levels during subjective analysis.

In [34], the study revealed diverse forms of hate speech across categories such as religion, race, behaviour, gender, appearance, social class, sexual orientation, ethnicity, gender, disability, and more. Prominent targets of hate speech included individuals linked to characteristics like being overweight, dishonest, homosexual, Caucasian, Black, impolite, misinformed, racist, elderly, self-centred, or practising a specific religion. These findings were especially significant in the European countries.

Collecting and labelling data to train automated classifiers for hate speech identification is a complex endeavour. Consensus on whether specific text qualifies as hate speech presents challenges due to the lack of a universally accepted definition [35]. Annotated databases are in use [22], but a dearth of annotated datasets for social media content from Sinhala and Singlish Sri Lankan users exists. This study aims to rectify this by employing a crowdsourcing platform involving users proficient in reading Sinhala and Singlish. The following section explores prior crowdsourcing research across diverse domains.

TABLE 2.1: BENCHMARKS OF LABELLED DATASETS USED TO IDENTIFY HATE SPEECH

Researcher	Social Media Platform	Labelled Categories(LCs)
Davidson[24] Hatebase Twitter	Tweet	Content that is provocative but does not qualify as hate speech. Provocative content Neither
Waseem[36]	Tweet	Racism Sexism Neither
Gibert et al. [37]	Stormfront Internet posts	- Hate No hate Skip

Joni and Maximilian[38] applied Feed Forward Neural Network (FFNN), Support Vector Machines (SVM), Logistic Regression, Extreme Gradient Boosted Decision Trees (XGBoost), Naïve Bayes, and feature representations including Bidirectional Encoder Representations, TF-IDF, Bag-of-words (BoW), Word2Vec, from Transformers (BERT), and their combinations.

2.3 Crowdsourcing

Crowdsourcing is widely being used to obtain an accurately annotated dataset. This section introduces crowdsourcing, explains the possibility of using crowdsourcing in human-computer interaction, characteristics of crowdsourcing, motivators for the crowd, challenges of crowdsourcing and the possible techniques to overcome the challenges identified.

2.3.1 Definition of Crowdsourcing

Coined by Jeff Howe [39], the term "crowdsourcing" seamlessly blends "crowd" and "sourcing," signifying the practice of entrusting a task to a crowd or anonymous individuals. In this conception, Jeff Howe first defined Crowdsourcing as “the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call.”

Considering the diverse backgrounds and varying levels of expertise among web workers, the accuracy of their labelling can sometimes fall below expectations. To address this, it's commonly recommended to have multiple workers label each task, enhancing overall accuracy. The redundant labels then serve as cues to resolve the correct labels effectively.

Crowdsourcing can also help to mitigate bias in the labelling process. When multiple workers from diverse backgrounds and with different perspectives label the same task, it is less likely that any one individual's biases will significantly influence the final labels.

However, it is important to note that redundant labelling can also increase the cost and time required for data annotation. Therefore, it is essential to strike a balance between the number of redundant labels and the resources available for annotation.

2.3.2 Use of Crowdsourcing

Crowdsourcing stands as a dynamic and rapidly growing approach through which organizations tap into the collective insights of online communities to mutual advantage. This innovative method empowers both the organization and contributors to glean the finest ideas. Diverse internet-powered collaboration platforms harness crowdsourcing across realms like crowd wisdom, creativity, innovation, and crowd voting [13]. Noteworthy platforms, such as Amazon Mechanical Turk, CrowdFlower,

Freelancer.com, Utest, and crowdSPRING, exemplify this collaborative approach, requiring a substantial number of registered workers to ensure optimal performance.

The landscape of social media, witnessing the registration of billions of users annually, offers a fertile ground for identifying individuals fitting various profiles. Consequently, social media platforms emerge as an ideal resource for locating the ideal crowd to support crowdsourcing initiatives.

There are two kinds of crowdsourcing: direct crowdsourcing and indirect crowdsourcing. If direct crowdsourcing is used, it is feasible to reach out to the community directly through various channels, such as social media, to solicit feedback on an idea or to assist them with a project. In indirect crowdsourcing, it is possible to use some other platform, such as Mechanical Turk [40]. Furthermore, social media is frequently utilized to promote involvement in current crowdsourcing projects, which should, in theory, lead to higher-quality ideas, services, or the anticipated ultimate result.

Human reviewers play a role in manually endorsing or rejecting content on crowdsourcing platforms. This assists in refining algorithms for improved content moderation decisions over time. The motivations behind individuals participating in voluntary moderation tasks are diverse. Some do it for recognition, reputation, or altruistic reasons (like community improvement). In other instances, moderators are rewarded with non-financial perks, such as complimentary or discounted access to online services [41]. It is challenging to detect hate by identifying lengthy content as respective hate-included content may be only a few words out of all the content[16].

This research is to be carried out based on both direct and indirect crowdsourcing, which provides a platform to gain benefits out of the crowdsourcing transaction that was taken in place.

2.3.3 Crowdsourcing for Human-Computer Interaction

Crowds can be used to generate designs for open-ended problems [42]; the field of HCI is no exception. The questions to be answered of a variety of applications when selecting participants are listed in [43]as; *who contributes, why they participate, what they contribute, and the effect of rewards and controls for quality*[44], [45]. When using crowdsourcing as a tool of HCI, the following questions are to be answered.

- Where do the opportunities lie?
- What does it contribute?
- What is the value for users and contributors?
- How do you phrase the assignment or the task?
- How do you motivate participation?
- Do you use existing applications or build your own?
- Which crowds are appropriate and when?
- Do the roles change, such as user, contributor, designer, engineer, and decision-maker?

- How do the roles change?
- How can designers use the outcomes, and does it integrate into current practices?

2.3.4 Crowdsourcing for Social Media Data Mining

Social media data mining involves extracting user-generated content along with their social connections. People on social media platforms regularly make diverse decisions, and this aspect can be utilized to provide recommendations for moderating social media content. Classical recommendation algorithms are *content-based methods*, *memory-based*, *user-based* and *model-based collaborative filtering* (CF), and *extending individual recommendations to groups of individuals*. Aggregation strategies for a group of individuals contain maximizing average satisfaction, least misery and most pleasure. It is possible to improve the recommendations by extending the classical methods with social context. It is possible to evaluate the recommendations evaluating the accuracy, relevancy, and ranking of recommendations. It is possible to analyze the behaviours of the contributors individually and collectively[46].

2.3.5 Characteristics of Crowdsourcing

As stated in the “Introduction”, multiple characteristics of crowdsourcing are described in [13]. The four dimensions as stated in[14] are illustrated in Fig 2.1. Two-Step Cluster Analysis and Schwarz’s Bayesian cluster criterion are used by the researchers to identify the patterns of different processes.

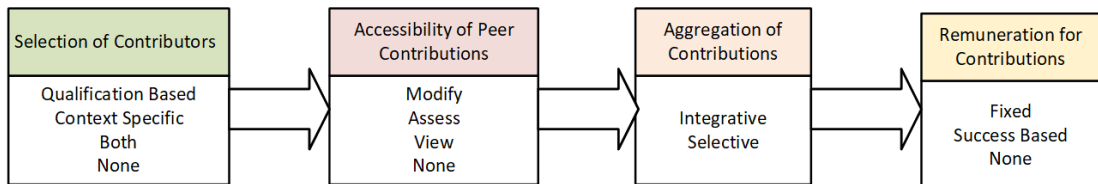


Fig. 2.1: Characteristics of crowdsourcing processes – Adapted from [13]

2.3.6 Motivators for Crowdsourcing Participants

Two types of motivators are discussed in [43] as intrinsic and extrinsic motivators to ensure adequate participation in a crowdsourcing platform. An immediate payoff, an extrinsic motivation, is a standard method used in indirect crowdsourcing platforms, and a list of motivators is shown in Fig. 2.2. One such system is Tختهagle [47], which lets users make a modest income by doing quick activities on their mobile phones for companies that pay them. Some studies discuss three broad approaches economic incentives, social incentives, and intrinsic incentives. For example, a study with Amazon Mechanical Turk[48] found that intrinsic motivators generated more work from crowds than extrinsic motivators did.

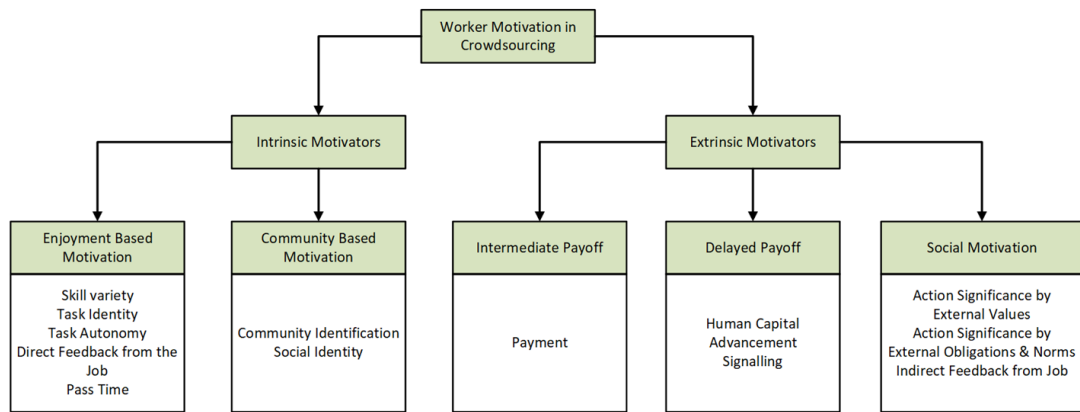


Fig. 2.2: Participant motivators in crowdsourcing – Adapted from [49]

2.3.7 Challenges in Crowdsourcing Platforms

Crowdsourcing, a contemporary approach to collaborative problem-solving, comes with several challenges that need careful examination. These challenges include issues with the credibility of work, the uncertainty of participant engagement, and the management of large groups while maintaining task quality. One significant challenge revolves around the credibility of work contributed by volunteers. Unlike professionals, work completed by volunteers may not always be seen as credible. This calls for ways to ensure the authenticity of their contributions.

Another challenge is the unpredictable nature of participant involvement. Participants may register and then may leave half way through, which can disrupt project continuity. It's important to find strategies to address this potential problem.

Handling a large number of participants introduces complexity. Balancing participant numbers with task quality becomes a complicated task, requiring careful coordination to uphold the quality of the work.

Maintaining transparency on crowdsourcing platforms clashes with data confidentiality. Since these platforms need to be transparent to workers, keeping data confidential is a challenge, and there's a risk that data might become public, raising ethical and security concerns.

The varied motivations of volunteers can lead to differences in the quality of their contributions. With diverse intentions, the reliability of submissions can vary, making quality control necessary. There's a significant challenge in ensuring participant honesty. Some participants might create multiple profiles to gain more rewards, which can distort the purpose of tasks. Preventing such behaviour is crucial.

On the other hand, entrusting tasks to professionals can provide confidence in their expertise. However, the potential for biased self-presentation and fake skill assessments among crowd participants requires careful consideration. In summary, these challenges highlight the complex nature of crowdsourcing. Effective solutions are needed to enhance credibility, manage participant dynamics, and ensure both

quantity and quality. These insights pave the way for refining crowdsourcing practices to make them more effective and reliable[50].

Conflicts in crowdsourcing can arise for various reasons, including unjust reputation systems, sluggish rewards, lack of transparency, discrimination, and socio-spatial disparities[51]. Furthermore, the Hawthorne effect, which refers to changes in a participant's behaviour due to their awareness of being observed in a study, represents one of the drawbacks associated with utilizing crowd intelligence[52].

2.3.8 Selection of Personas from Crowdsourcing Participants

Personas can be chosen according to user experience aspects [53] via the analysis of learning activity patterns that are tracked in real-time. A system attains the status of being context-aware [54] when it leverages context to supply pertinent information and services, with the relevance hinging on the user's specific task. Context, as categorized by Dey and Abowd [55], comprises four core types: Location, Identity, Time, and Activity, all of which help characterize a given entity's situation. For instance, when detecting hate content using source metadata, these context-aware data could be effectively combined with personas.

2.3.9 Crowdsourced Data Management

Any data gathered from the crowd is unreliable and ambiguous. Studies already suggest methods to deal with such problems. One such attempt is to use a worker model based on worker characteristics and adopt various strategies to control quality. A few examples are identifying spammers and removing them. Similarly, the worker answers can be aggregated together and get a quality score calculated[50]. *Worker probability* to use a single parameter to represent the quality of each employee, *confusion matrix* to simulate a worker's performance on a single, optional task, *Bias* and *Variance* to model the ability to perform quantitative jobs, diverse skills across tasks and domains are few of the parameters to model with crowdsourcing workers. *Qualification tests, gold-injected methods, Expectation-Maximization based methods, and graph-based methods* are a few techniques used in the computation of worker model parameters in recent research. Using a sampling-centred method enables the crowd to examine a small portion of the data, which is then used to project their findings to the entire dataset. *Round* and *statistical models* are used in controlling latency.

Here are some examples of metrics used to measure crowdsourcing performance: *the count of messages over a certain timeframe, the number of sources or contributors, the total of collaborating messages, demographics of the most active sources, the progression of solutions over time, the volume of votes received by solutions, the sentiment expressed in comments on solutions, the level of collaboration over time, the total number of participants, the rate of new participants per day, and the duration of time participants spend on the site per visit*[52].

The proposed research will implement a suitable crowdsourcing platform that could be used as the second element stated in[56], and let manual approval. Sinhala and Singlish corpus could not be found for racism as in[44], it is proposed to use the crowdsourcing platform to generate the corpus. In the proposed framework, a gamification model is proposed, which would provide intrinsic and extrinsic motivation to workers, rate users based on their efficiency and accuracy and promote users to their involvement in task completion.

Conclusion

Two primary approaches to moderating social media content are automatic and human-based. However, even to train the automatic algorithms, annotated datasets are required. The current social media content moderation process necessitates new technologies to provide contextual insights, which requires a method that evaluates user experience. Hate speech is one of the main improper contents under examination that needs to be taken down before a more extensive audience sees it. Of the two types of crowdsourcing available, direct and indirect crowdsourcing, direct and indirect crowdsourcing will be used in this study, allowing participants to profit from the crowdsourcing transaction. Indirect crowdsourcing is used to reach a more focused worker community. The proposed crowdsourcing system consists of four major components to select contributors, access the contributions of peers, aggregate contributions and remunerations. These dimensions are detailed in the methodology section, covering the execution of each phase. Additionally, intrinsic and extrinsic motivators are integrated into the framework, and their implementation is further elaborated in the methodology.

The investigated hate categories comprise physical appearance, race, social class behaviour, sexual orientation, disability, ethnicity, gender, religion, and other classifications. Hate targets for the Sri Lankan context have been identified as part of this study.

A considerable portion of workers on crowdsourcing platforms are mainly focused on generating fast, generic responses instead of accurate ones. This approach aims to save time, allowing them to earn more money.

CHAPTER 3

SYSTEMATIC LITERATURE REVIEW

Introduction

This chapter explains the systematic literature review performed on ensuring the trustworthiness of crowdsourcing participants. The consensus-based approach in ensuring the trustworthiness of crowdsourcing participants is highly affected by the bias of the crowd. Research Question No. 04 is “How to measure and ensure the trustworthiness of users in their responses?”. This bias of the workers impacts both the accuracy and the performance of the crowdsourcing framework. 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. Furthermore, if the annotation does not reflect accurate labelling, it will affect the result of the detection mechanism.

It is required to compare and analyse the annotation quality of the crowdsourcing platform using consensus-based, reputation-based and gold standard-based approaches to overcome these problems and ensure crowd responses' trustworthiness. Furthermore, it is proposed to identify the best trust metric and model the crowd workers' trustworthiness of the crowdsourcing platform.

A comprehensive systematic literature review was conducted to achieve the specific research objective No. 02, “Identify appropriate trust metrics to evaluate the reliability of the crowd response”. This research chapter aims to summarise and highlight critical computational efforts to eliminate subjectivity when annotating social media content to ensure trustworthiness.

Before presenting the survey over previous technical endeavours, this chapter first briefly introduces the need for trustworthiness measurement tools embedded in crowdsourcing platforms. Following is the chapter's remaining organization; Section 3.1.1 provides the questions under study in systematic literature review, Section 3.1.2 provides a list of data sources, and Section 3.1.3 outlines the search method for and reviews the literature. Section 3.2 provides definitions for key terms under study. Section 3.3 surveys the efforts that have been taken by researchers to eliminate bias when annotating by answering three research questions. This section briefly examines the worker categorizations of a crowdsourcing platform, identified groups of bias measurement methods, annotation methods and trust metrics. Finally, section 6 concludes the systematic review.

In summary, this chapter provides a systematic overview of methods aimed at mitigating the subjectivity of worker responses and assessing trustworthiness.

3.1 SYSTEMATIC LITERATURE REVIEW METHODOLOGY

This systematic literature review aims to comprehensively and methodically examine existing scholarly literature by answering three research questions, synthesizing and analysing the findings to provide a thorough understanding of the current state of knowledge in that field. The review helped to identify gaps, trends, patterns, and inconsistencies in the existing literature, offering insights for further research and contributing to the development of the field. The preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) model was used to conduct the systematic literature review included in this thesis.

3.1.1 Research Questions

The goal of this systematic review is to compile research findings from various institutions associated with eliminating crowd bias in crowdsourcing platforms and ensuring trust by addressing the subsequent research queries.

1. What approaches and methods are employed to verify the quality of the submitted annotations?
2. What are the different types of bias and methods used to eliminate the bias?
3. What are the methods used to measure the trust of crowd response?

This study aims to explore the research questions mentioned above by examining previous research endeavours. The initial emphasis of this study was in the fields of engineering, computer science, mathematics, and multidisciplinary areas related to crowdsourcing. Subsequently, the search results were analysed to identify pertinent research findings, and this process was iterated to encompass a broad range of relevant discoveries. Fig. 3.1 displays the percentage of research findings found under each subject area.

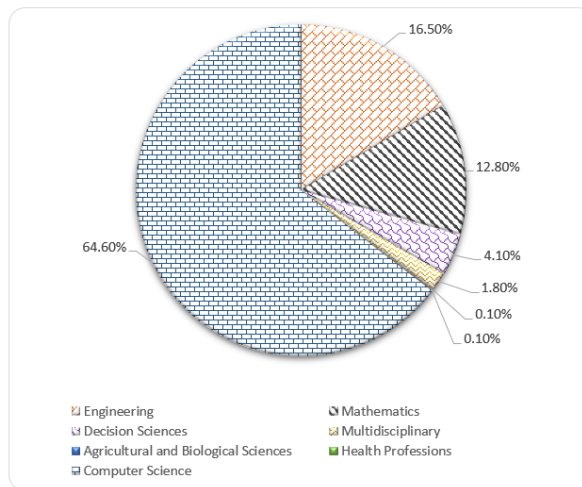


Fig. 3.1: The number of qualified search results in each subject area under review.

3.1.2 Data Source

The initial approach involved querying databases such as IEEE Xplore, ACM Digital Library, and Scopus. Keywords, including but not restricted to measuring trustworthiness, crowdsourcing platforms, and bias, were utilized. Additionally, ongoing research projects, news articles and Google Scholar relevant to the study were included in the study.

TABLE 3.1: DOCUMENT TYPES OF RESULTING LITERATURE

Document Type	Number
Conference Paper	108
Journal Articles	19
Conference Review	13
Book Chapter	7
Review	3
Non-technical articles and websites	5
Thesis	3
Editorial	2
Other	29

The generated literature consists of 189 references, outlined in Table 3.1. Two books, Fairness and Machine Learning[57] and Trustworthiness in Crowdsourcing[58], were not surveyed in this work.

3.1.3 Searching Approach

The search is restricted to utilizing online library search engines accessible through university subscriptions. The main searches were carried out utilizing influential digital libraries, and respected conference proceedings were referenced to respond to the research questions, as outlined in the *Data Source* section. No time constraints were imposed on the search, and the publication count per year is depicted in Fig. 3.2. Additionally, it was observed that research activities commenced in the late 2000s.

General Search string – (crowdsourc* AND (trust* OR reliability OR opinion OR (subjective AND (judgement OR component)) OR bias* OR (mitigate AND worker AND bias)))

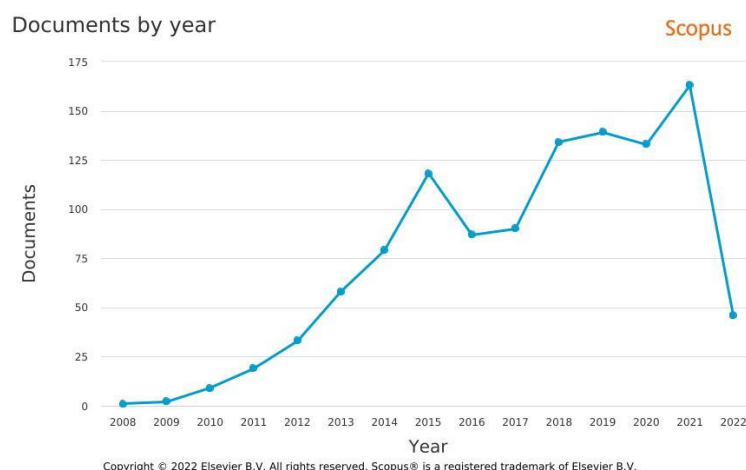


Fig. 3.2: Yearly wise publication count generated by Scopus [59]

3.1.4 Criteria for Inclusion and Exclusion

The criteria for inclusion and exclusion in the PRISMA model dictate the specific parameters used to determine which studies are considered for inclusion in the review and which are excluded. These criteria are essential for maintaining transparency, consistency, and rigour in the study selection process, aligning with the PRISMA guidelines to ensure a systematic and comprehensive approach to the literature review.

Inclusion criteria 1 – Conference papers, journal papers, and other research publications are taken into account as outlined in Table 3.1.

Inclusion criteria 2 – Research studies related to the search query from the selected data sources.

Inclusion criteria 3 – Research studies written in the language English

Inclusion criteria 4 – Research results relevant to the three research questions stated in section 3.1.1

Exclusion criteria 1 – Papers that do not align with the research questions.

Exclusion criteria 2 – Papers with a language that is not English.

Exclusion criteria 3 – Research studies that lack clarity in the domains of social media content moderation, the utilization of crowdsourcing in annotation, and the assessment of trustworthiness in crowdsourcing platforms.

Exclusion criteria 4 – Subject areas chemical engineering, econometrics and finance, business, economics, psychology, earth and planetary sciences, neuroscience, arts and humanities, genetics and molecular biology, energy, material science, medicine, social sciences, environmental science, management and accounting, physics and astronomy and biochemistry.

Exclusion criteria 5 – Keywords network security and blockchain

Since crowdsourcing involves multiple disciplines as a mode of data collection and categorization, research studies have been carried out in almost all the existing subject areas. Therefore, a few subject areas specified in exclusion criteria four were used to limit the research findings under the study area. Furthermore, trust is a significant issue in network security and blockchain applications which are not a focus in this study. Therefore, the search results with the keywords network security and blockchain were eliminated.

3.1.5 Process of Selecting Studies

Microsoft Excel and Zotero were employed as tools for paper analysis and storage. The paper selection process adhered to the PRISMA guidelines [60], depicted in Fig. 3.3 and the filtering was done without bias.

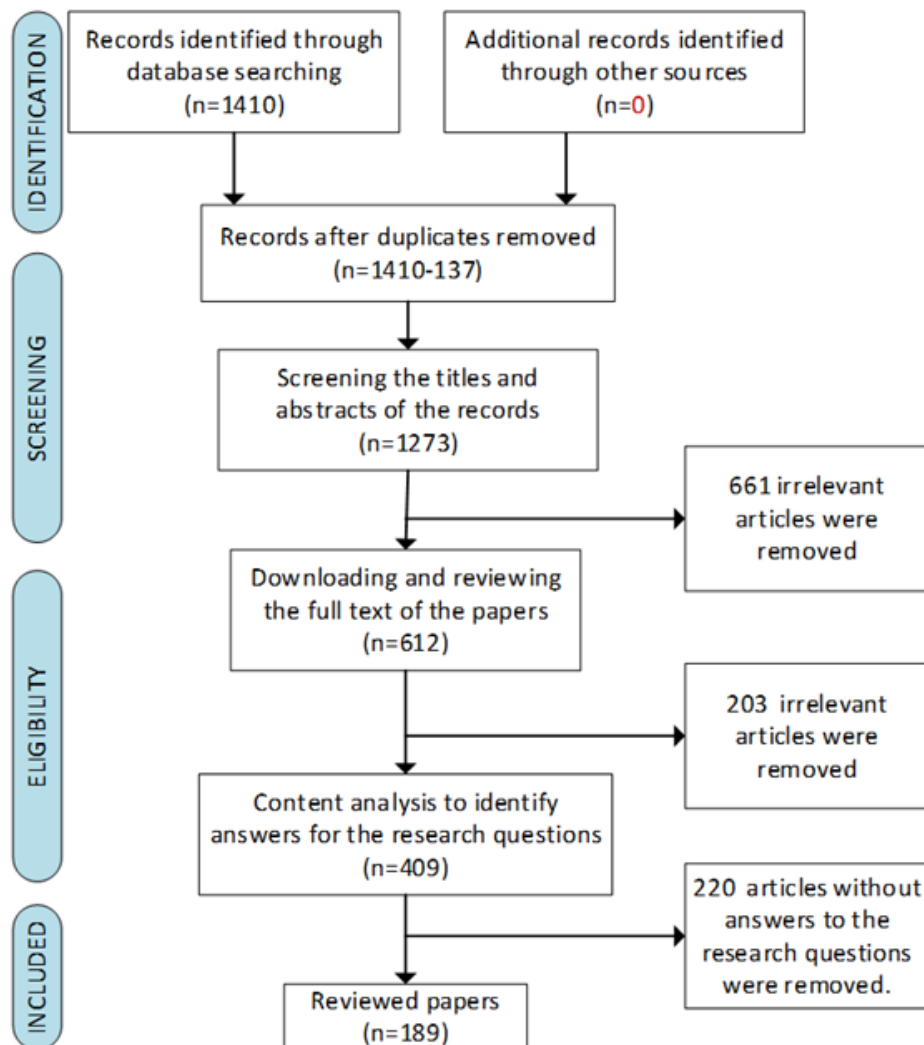


Fig. 3.3 : PRISMA Model

3.2 MEASURING THE TRUSTWORTHINESS OF CROWD PARTICIPANT RESPONSES

To ensure the data's reliability[61], it's vital to assess trustworthiness during the entire process. An effective method for estimating reliability is Krippendorff's alpha coefficient. It's also crucial to vary the number of contributors for different tasks. In a related study, Irene and Annamaria[62] investigated annotator reliability in audio tagging. Multi-annotator competence estimation (MACE) is used to handle multiple labels after calculating a potential ground truth. Both experts and non-experts could be used for this purpose.

For the initial primary classification, employing two contributors is sufficient. However, for tasks demanding more detail, like sentiment strength analysis and identifying hate targets, contributor numbers should increase based on reliability scores. It's important to acknowledge that contributors' beliefs can impact their responses. For this, randomizing worker selection is essential. This ensures various post categories are evenly distributed among workers, thereby reducing bias.

Maintaining crowdsourcing reliability involves evaluating the platform's credibility, the task assigner (crowdsourcer), and the participants' credibility, as shown in Fig. 3.4. This study exclusively focuses on developing techniques to ensure the crowd's trustworthiness, ultimately improving response quality.

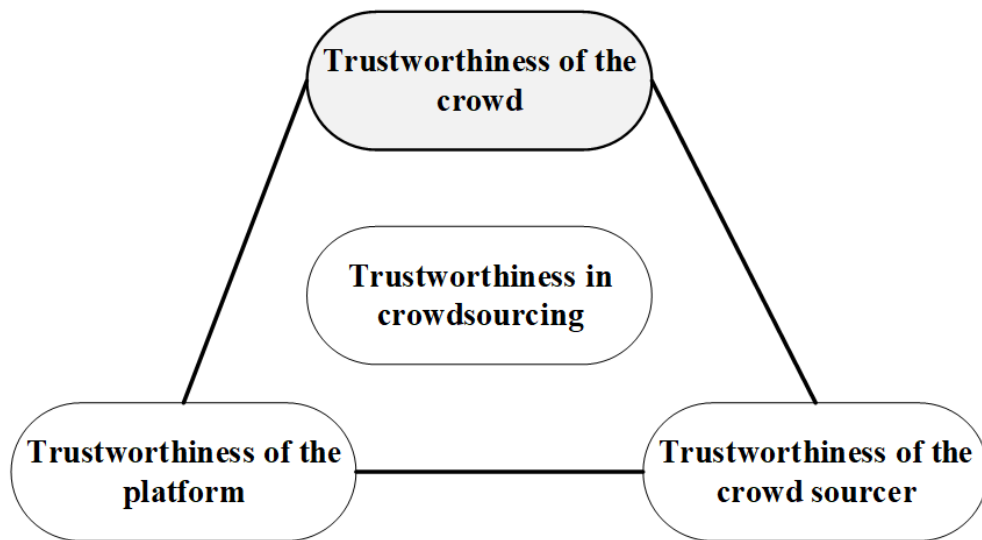


Fig. 3.4 : Aspects in measuring the trustworthiness of crowdsourcing

A central research question when focusing on ensuring trust is, whether it is possible to measure trust. According to Cristiano and Rino [63], trust extends beyond a mere probability calculation and examines a more profound and complex phenomenon, incorporating the theory of mind. The theory of mind[64] refers to the ability to understand that other people have thoughts, beliefs, desires, intentions, and emotions that might be different from one's own. It's essentially the capacity to attribute mental

states to others, allowing us to predict and explain their behaviour based on their mental states.

If trust is measurable, there should be metrics to measure trust. Trust metrics enable trust modelling and reasoning about trust. Let us look at the approaches and methods employed to verify the quality of submitted annotations, the different types of bias and methods used to eliminate the bias and the metrics that the researchers have used to measure trust.

3.2.1 Reputation management in crowdsourcing systems

In their study, Allahbakhsh, Ignjatovic, et al. [65], introduce a novel metric termed the "degree of fairness." This metric is designed to gauge the extent to which evaluators maintain fairness while assessing workers' contributions. The authors employ a majority consensus approach on workers' trustworthiness to measure the proximity of the evaluator's viewpoint to the consensus within the community. Their model encompasses evaluators, workers, evaluations concerning workers' quality, pairwise trust values, and the degree of fairness. Additionally, the researchers present an algorithm using pairwise trust and rank.

3.2.2 Aggregation Techniques in Crowdsourcing

From a collection of responses provided by the crowd workers, hidden ground truth is discovered using aggregation techniques or approaches. Based on their computational approach, the aggregation techniques are broadly divided into two types [66]; Non-iterative and iterative.

A few examples of iterative algorithms are Expectation Maximization (EM)[67], Generative model of Labels, Abilities, and Difficulties (GLAD)[68], Supervised Learning from Multiple Experts (SLME)[69] and Iterative Learning (ITER)[70]. In iterative algorithms first update the aggregated value and then adjust the value based on the crowd response.

A few examples of non-iterative algorithms are Majority voting or majority decision [71], HoneyPot(HP) [72], and Expert Label Injected Crowd Estimation (ELICE).

3.3 MEASURING BIAS OF THE WORKERS

The bias of the crowd has a significant impact on the consensus-based approach to assuring the reliability of crowdsourcing participants. This worker bias affects the crowdsourcing framework's accuracy and effectiveness. Even though majority voting is utilized to verify agreement, existing research demonstrates that the annotated results do not accurately reflect the user perspective, making the annotation less trustworthy. Furthermore, the outcome of the detection technique will be impacted if the annotation does not reflect accurate labelling.

3.3.1 Approaches and methods to verify the quality of the submitted annotations.

Many factors affect the quality of the submitted annotations such as different socio-economic backgrounds, levels of competence and skills, and motivations and aspirations [73] of the crowd workers. Therefore, categorizing worker types and selecting the correct set of workers affects the work quality. Assigning tasks to identified worker types is one way of verifying the quality of submitted annotations. Assigning tasks to different worker types can be achieved by behavioural observation.

After categorizing the worker types, it is essential to measure the consensus among the annotators if majority voting is being used. Therefore, this section provides a brief review of the worker-type classification and the techniques and methods used in measuring consensus among annotators.

Categorizing Worker Types

Kazai et al. [74] divided crowd workers into five types: sloppy, spammer, incompetent, competent, and diligent. These categories were determined by observing worker behaviours, which included parameters such as the number of completed HITs, average task duration, HIT completion time, helpful label proportion, and label accuracy.

Typecasting is often used to match individuals with suitable tasks or assignments within a project or initiative. This process involves assessing the skills, knowledge, and experience of crowd members and then assigning them to tasks that align with their abilities. Type-casting in crowdsourcing can be achieved through various methods, including self-declaration by crowd members, online assessments or tests, evaluation of past work or achievements, and feedback from other participants or project managers. The goal is to ensure that the right individuals are assigned to tasks where they can contribute effectively and maximize the overall quality and efficiency of the crowd-based project.

Other classifications are given in [58] as an Expert worker, Biased worker, Random Spammer, Uniform Spammer, Adversarial Colluded worker, Non-Adversarial Colluded worker, and Non-Adversarial Colluded follower. The researchers [75] classified crowd members into classes such as unenthusiastic, optimistic, pessimistic, etc. Another categorization of workers is sloppy and proper under ethical and random and uniform spammers under unethical workers[76].

In the paper[77], the researchers analyze malicious workers by considering the behaviour patterns using already set criteria. Certain individuals with hidden intentions may intentionally disrupt a task or attempt to rapidly complete it for financial benefits, as noted in reference[67]. In contrast, personality traits also are considered in categorizing users, and the five personality dimensions by [78].

Obtaining consensus of workers is as important as classifying the worker types. Using majority voting to ensure consensus is the most common way of ensuring trust in crowdsourcing platforms. During the annotation process, expert annotation and gold standards are used conventionally[79].

3.3.3 Measuring the quality of annotations through ground truth inference

In statistics and machine learning, the term ground truth is used to refer to the process of validating machine learning results against real-world accuracy. To ensure the quality of submitted annotations, two prevalent strategies are sampling and redundancy [80], [81], [82]. In the redundancy approach, several workers assess the same task (HIT), and a valid HIT is determined through a voting mechanism. For instance, if 4 out of 5 workers label an SMS as positive, the majority voting principle categorizes it as positive, and all four workers receive payment. On the other hand, the sampling approach involves incorporating gold samples into each HIT. Payment is granted to annotators only upon accurately labelling the gold samples. For instance, if a HIT involves tagging 5 SMSes and one gold sample is included for validation, the worker is paid only if the gold sample is accurately labelled.

Majority voting or Majority Decision[MD]

Multiple aggregation methods aggregate independent opinions of unskilled individuals, and majority voting is the simplest method which is a non-iterative method [66]. Majority voting is a decision-making process in which a choice is selected based on the option that receives the most votes from a group. The option with more than half of the total votes is considered the winner, and it's a common method used in various contexts, such as elections, decision-making in committees, and group discussions. It is true for labelling tasks such as if an image consists of a dog or cat, but not in making subjective judgements such as if a post contains hate speech or not.

However, there are many other limitations of using majority voting as the aggregation method such as the homogeneity of workers and the homogeneity of the questions [83]. The homogeneity of workers means that it is assumed that all the workers have the same ability. When majority voting is considered if there is only one true expert and if others have less or no expertise the response favours the majority. Weight is used to solve this problem. The majority vote assumes that questions are of the same complexity is referred to as the homogeneity of the questions or tasks.

When there are two classes, three labellers can always obtain a majority. According to Joni and Ahmed et al. [84], adding more annotators reveals subtle differences between annotators' preferences [85]. From three annotators to nine annotators, they demonstrate how complete agreement among the annotators monotonically declines. The approach of expanding the number of annotators for subjective labelling tasks is supported by the fact that ten annotators have a higher agreement than any other group of annotators. Even with 10 annotators, there is still a minor number of ties (4.11%), so even while it is advantageous, adding more annotators cannot be seen as a fundamental fix for the disagreement in subjective crowdsourcing tasks. To assess the

consistency of agreement among the annotators, Joni and Ahmed et al. recommend having data samples of small sizes with at least five annotators.

The following researchers are using a distinct set of three crowd workers using majority voting for their data validation [86]. The researchers considered a question as valid, they checked three questions which checked if three workers picked the correct answer, and agreed that the two answer options were clear and the context of the pronouns such as “it” was not clear.

Honeypot(HP)[72]

This method of aggregation is non-iterative. In theory, Honeypot (HP) operates similarly to majority decision (MD), with the exception that fraudulent workers are filtered out during the pre-processing stage. During this phase, HP utilizes randomization to merge a collection of questions with corresponding answers (whose true responses are already known). The likelihood that a label might be assigned to each item o_i is then calculated using a majority decision among the remaining workers. This method has certain drawbacks, too, including the fact that it is not always available or that it is frequently created subjectively; for instance, if the trapping questions are too challenging, honest workers can be mistaken for spammers.

Expert Label Injected Crowd Estimation (ELICE)[87]

ELICE is an extension of HP and this method of aggregation is non-iterative. ELICE also employs trapping questions, but it does so to assess every worker's level of competence by comparing the proportion of those answers to the actual answers.

Expectation Maximization (EM) [67]

Expectation Maximization (EM) [88] is a statistical algorithm used for estimating parameters of probabilistic models when dealing with incomplete or missing data. It's an iterative process that alternates between two main steps.

Expectation Step (E-step): In this step, the algorithm calculates the expected values of the missing or unobserved data given the current estimates of the model parameters.

Maximization Step (M-step): In this step, the algorithm updates the model parameters to maximize the likelihood of the observed data, taking into account the expected values computed in the E-step.

The EM algorithm continues to alternate between these two steps until convergence, where the changes in the parameters become very small or negligible. EM is widely used in various fields, such as machine learning, image processing, and natural language processing when dealing with situations where data is incomplete or when there are hidden variables impacting the observed data. Up until all object probabilities remain unaltered, this iteration is repeated. In a nutshell, EM is an iterative method that collects numerous things simultaneously. Running time is a crucial concern because convergence involves many processes. David and Skyne (DS), GLAD [89], RY, and ZenCrowd[90] are the four EM-Based Consensus Algorithms explored in [89]. Another EM-based method is Welinder and Perona.

Dawid-Skene Estimator [91]

An expectation-maximization algorithm was proposed by Dawid when the gold truth was not known.

Generative Model of Labels, Abilities, and Difficulties (GLAD)[68]

This approach is iterative. GLAD is a development of EM. This method considers both the level of skill of the worker and the difficulty of each object's question. It makes two unique examples of its focus. In the first scenario, a worker with great knowledge has a better likelihood of providing an accurate response when several workers are answering the same question. Another instance is when a worker answers numerous questions, the likelihood of answering a question properly is reduced when the question is more challenging. Generally speaking, GLAD and EM-based methods are susceptible to random initializations. The performance guarantees lack a theoretical analysis, thus one must provide one.

Smyth et al. [92]

Researchers are using a Bayesian approach which is an extension of the maximum likelihood framework. Furthermore, researchers state that neglecting subjectivity uncertainty in image labelling can lead to substantial overconfidence in terms of estimation of performance.

RY [84]

In the case of numerous annotators providing labels (potentially noisy), Raykar, Yu, and colleagues offer a probabilistic method for supervised learning in the absence of an unquestionable gold standard. The suggested technique provides an estimation of the hidden labels in addition to evaluating the various experts.

Welinder and Perona[93]

Each annotator is portrayed as a multidimensional entity in their work, including variables for competence, knowledge, and bias. Recognizing the annotators with varying skills and knowledge is allowed by this model.

Ghosh et al [94]

In their work, Ghosh et al assert that all content rating users are reliable and honest. They present a system designed to address the challenge of moderating online content through crowdsourced ratings. This framework tackles the uncertainty surrounding both the content itself and the raters, where the quality of both is uncertain, and some users may exhibit varying levels of unreliability or inaccuracy.

Effectively estimating when one lacks complete information about the raters proves challenging. The authors offer algorithms aimed at precisely identifying misuse, and these algorithms require knowledge of only a single "good" agent's identity—one who consistently evaluates contributions correctly more than half the time. They showcase that their method is capable of deducing the contribution quality with decreasing error as the number of observations increases.

Dalvi et al[95]

This research focuses on analyzing a crowdsourcing system comprising users and binary choice questions. The users' reliability, though fixed and unknown, affects their error rates in answering questions. The main challenge is deducing the questions' true answers solely from user responses. While prior research has explored this, theoretical error bounds have been established only for specific scenarios: complete or random user-question graphs. The paper's novel contribution is addressing a more generalized setup where the user-question graph can be arbitrary. The researchers identify limitations in their algorithm's accuracy and showcase how graph expansion can manage this issue.

LC-ME[96]

LC-ME stands for "linear mixed-effects model for continuous outcomes in longitudinal studies with missing values." This approach uses a linear mixed-effects model to account for the correlation among repeated measurements on the same individual over time, while also accounting for missing data.

Supervised Learning from Multiple Experts (SLME)[84]

EM is also used in supervised learning from multiple experts (SLME), sensitivity and specificity are used in this technique instead of using a confusion matrix.

Iterative Learning (ITER)[97]

The iterative technique known as Iterative Learning (ITER) is based on conventional belief propagation [84]. The worker's skill and the question's difficulty are measured here, but it does so in a somewhat different way. ITER calculates the dependability of each answer separately, in contrast to other methods which use a single value. Additionally, each worker's difficulty score is calculated independently for each question. As a result, the reliability of each worker's replies is added up and weighted according to how challenging the questions that go along with them are. ITER has the benefit of not requiring the initialization of model parameters (such as answer quality or question complexity). Additionally, ITER doesn't presume that workers must respond to every query, unlike other methods.

Gibbs-EM algorithm[98]

Accuracy and Mean square error are considered better by the researchers in[102] when compared with the traditional EM algorithms.

Weighted majority voting (WMV) [99]

Weighted majority voting is a decision-making method that involves combining the opinions of multiple voters or classifiers to arrive at a final decision. In this method, each voter or classifier is assigned a weight or importance value, which reflects their level of expertise or credibility in making the decision.

When using weighted majority voting, each voter or classifier casts a vote in favour of one of the available options, and the votes are then tallied up. The final decision is

made based on the total number of votes each option received, with the weights considered. The weight assigned to each voter or classifier can be determined in various ways, such as by the accuracy of their previous predictions or their level of expertise in the domain. By considering the weights of the voters or classifiers, the method can give more weight to the opinions of those who are more reliable or knowledgeable.

Weighted majority voting can be used in a variety of applications, such as in political elections, where each voter's vote has equal weight, but each voter's opinion may have different levels of credibility, or in machine learning algorithms, where each classifier is given a weight based on its performance on previous tasks.

Adaptive Weighted Majority Voting Algorithm (AWMV) [100]

Ensemble learning is a methodology that amalgamates predictions from multiple models to enhance overall accuracy and resilience. AWMV is an extension of the standard majority voting algorithm used in ensemble learning, where each model's prediction is considered equally important. On the other hand, AWMV gives weights to each model according to how well it did on a validation set. Models that did a good job get higher weights, while models that didn't perform well get lower weights. The weights assigned to each model can also change over time as new data becomes available. This adaptivity feature allows the AWMV algorithm to be more robust to changes in the data distribution over time.

Positive Label frequency Threshold (PLFT) [101]

PLFT is a threshold-based technique used in binary classification problems with imbalanced data. Imbalanced datasets refer to datasets used for machine learning or statistical analysis where the distribution of instances across different classes is highly skewed or disproportionate. Within these datasets, there is often one class (typically termed the "minority class") that contains notably fewer instances in comparison to another class (referred to as the "majority class"). PLFT is used to tackle this bias problem.

Spectral DS (SDS) [102]

Spectral DS (SDS) is a method for measuring the quality of annotations in a dataset, often used in image processing and NLP applications.

The SDS approach is based on "ground truth inference," which entails estimating the accurate labels for a group of data points by contrasting the labels assigned by numerous annotators. The SDS method uses spectral clustering to group the annotations that are most similar to each other and then calculates a measure of agreement between these groups of annotations and the ground truth.

Calculating the Spectral Distance (SD) between the ground truth and the annotations is used to measure annotation quality. The SD measures the discrepancy between the estimated true labels and the annotations assigned by the annotators.

The SDS method can be used to identify unreliable or inconsistent annotators and to estimate the level of agreement among multiple annotators. This method is used in many areas such as image classification, text annotation, and sentiment analysis, among others, where high-quality annotated data is essential for training machine learning models.

Ground Truth Inference using Clustering (GTIC) [103]

GTIC uses different noisy label sets of examples to produce features for a K labelling instance. All of the instances are clustered into K groups thereafter. These are mapped to a distinct class with the K-Means algorithm. A comparable class label is given to examples that belong to the same cluster.

Max-margin formulation.

To enhance the discriminative power of the most widely used majority voting estimator, Tian Tian and Jun Zhu [85] provide a max-margin formulation. They also present a Bayesian generalization that combines the benefits of both generative and discriminative techniques.

The degree of consensus or uniformity among the judges' evaluations is measured using inter-rater reliability, inter-rater agreement, or concordance in statistics which is the degree of agreement among annotators.

Inter-rater agreement metrics

Inter-rater agreement metrics include percent agreement for two annotators, Cohen's Kappa [104], Fleiss Kappa, Scott's pi, Krippendorff's alpha, Limits of agreement, Intra-class correlation coefficient, and correlation coefficients (Interclass correlation, Pearson product-moment correlation coefficient, Spearman's rank correlation coefficient, Kendall rank correlation coefficient, Gwet's AC2 Coefficient), along with Joint probability of agreement.

Ensuring quality when a gold standard is absent

In cases where the true gold standard remains undisclosed, each annotator's performance is assessed concerning sensitivity and specificity relative to this undisclosed benchmark [84]. In situations where multiple annotators provide labels, which may carry inherent noise, without a definitive absolute gold standard, a probabilistic framework for supervised learning is employed.

The research [84] answers the three questions of usage of the supervised learning algorithms and how to change them when many contributors annotate data with subjective responses and if a gold standard is not available., the way to evaluate systems without a gold standard and the way to estimate reliability with many annotators.

Maximum-likelihood estimator

Here the classifier, the actual correct label, and annotator accuracy are learnt at the same time. The final estimation is done by the EM algorithm. Based on the gold standard the performance measure is refined.

Multilevel Bayesian Models of Categorical Data Annotation [105]

Multilevel Bayesian Models of Categorical Data Annotation (MBCDA) is a statistical modelling approach used in the analysis of categorical data annotations, often used in natural language processing and computer vision applications.

A two-stage process is used by MBCDA which is a Bayesian framework that models the annotation process. The first stage models the true, underlying categorical labels for the data, and the second stage models the annotation process itself, which generates the observed annotations. Annotators may have different levels of expertise and biases. It models these differences as random effects and estimates their distributions using Bayesian methods. The MBCDA approach allows for the estimation of the true labels and the identification of unreliable or inconsistent annotators.

CUBAM (Combinatorial Unbiased Bayesian Aggregation Model)[106]

CUBAM (Combinatorial Unbiased Bayesian Aggregation Model) is a machine learning algorithm used in the context of crowdsourcing and collective intelligence. It is designed to combine the judgments of multiple individual annotators or experts to make predictions for a given task.

CUBAM is based on a Bayesian model that accounts for the uncertainty associated with each annotator's responses, as well as the underlying true value being predicted. It allows for the incorporation of multiple sources of evidence, such as the annotator's accuracy and bias, the difficulty of the task, and the correlations between the annotators' responses.

CUBAM can estimate the true label with high accuracy even in cases where individual annotators may have a low level of accuracy. Additionally, CUBAM can handle missing data and can adapt to changes in the annotators' behaviour over time. CUBAM [106] can generalize to multi-class classification but not multi-choice selection.

Irene and Annamaria [62] utilize Krippendorff's alpha and multi-annotator competence estimation (MACE) in a scenario involving multiple labels. They demonstrate how MACE can be applied to estimate a potential ground truth.

Annotator competence estimation

Inter-annotator agreement is used widely for videos[107], images [108] and sentences [109] in crowdsourcing. Crowdsourced data frequently contain discrepancies due to annotator error, overlapping labels, subjectivity, or genuine item uncertainty. As a result, many techniques have been created for learning from data that contains disagreement. One finding from this research is that different approaches seem to perform well based on the dataset's features, such as the amount of noise, for example, while performing subjective tasks like classifying an item as offensive or not. We also

discover that when ambiguity is eliminated through discussion or reasoning, disputes resulting from ambiguity do not perfectly fit into either category [110], [111], [112]

However, the majority of the attention in these earlier publications has been placed on answer correctness, under the presumption that each disagreement can be settled by a single accurate response [110]. Due to numerous types of disagreement, many categorization jobs are unclear in reality. Previous research demonstrates that verbal justifications exchanged can greatly increase answer accuracy when compared to aggregation procedures.

Many classification tasks have ambiguity leading to disagreement [112], [113], [114] due to reasons such as missing context, imprecise questions, contradictory evidence, and different annotator expertise levels. Previous research identified expert disagreement types, including personality-based, judgment-based, and structural disagreement. Disagreement can stem from ambiguous and subjective questions, vague visual evidence, varying expertise, and vocabulary mismatch.

Rank aggregation [115]

Rank aggregation is a technique used in crowdsourcing and collective intelligence applications to combine the rankings or preferences of multiple individuals into a single overall ranking or preference.

In crowdsourcing, rank aggregation is often used to collect preferences or opinions from a large number of people and to use them to inform a decision or recommendation. For example, in the context of movie or product recommendations, rank aggregation can be used to combine the preferences of many people to generate a list of highly rated movies or products.

There are several methods for rank aggregation, including Borda count, Kemeny-Young method, and Markov chain Monte Carlo (MCMC) methods. These methods differ in their assumptions about the underlying distribution of preferences or rankings and their computational complexity.

Rank aggregation can also be used in conjunction with other techniques for aggregating annotations or predictions, such as majority voting or Bayesian models. In these cases, the rank aggregation method is used to aggregate the individual annotations or predictions into a ranking or preference order, which can then be used as input to the other aggregation method.

Overall, rank aggregation is a powerful technique for combining the preferences or opinions of many individuals into a single, consensus ranking or preference. It has found application in diverse fields, including recommendation systems, sports rankings, voting systems, etc.

TABLE 3.2: ALGORITHMS FOR GENERAL GROUND TRUTH INFERENCE
[116]

Algorithm	Function	Type	Methodology
David and Skeme (1979)	Inference	Multi-class	EM-based
Smyth et al. (1995)	Inference	Binary	EM-based
Majority Vote (2008)	Inference	Binary	Statistics-based
GLAD (2009)	Inference	Binary	EM-based
RY (2010)	Inference & Learning	Binary	EM-based
HoneyPot (2010)	Inference	Binary	Statistics-based
Welinder and Perona (2010)	Inference	Multi-class	EM-based
Ghosh et al (2011)	Inference	Multi-class	Algebra-based
Zen Crowd (2012)	Inference	Multi-class	EM-based
Dalvi et al (2013)	Inference	Binary	Algebra-based
KOS (2014)	Inference & Learning	Binary	Algebra-based
PLAT (2015)	Inference	Binary	Statistics-based
LC-ME (2015)	Inference	Binary	Statistics-based
GTIC (2016)	Inference	Binary	Statistics-based
Shan et al (2018)	Inference	Binary	Statistics-based

3.3.2 What are the different types of bias and methods used to eliminate the bias?

One unanswered research challenge is dealing with annotations or answers with bias. These biases are subject to cultural backgrounds and personal preferences [117]. In [33], the present authors define data bias, population bias, behavioural bias, content production bias, linking bias, temporal bias, and redundancy in different cycles from data origins to collection and processing.

Cognitive Biases

Cognitive biases frequently affect human decision-making processes [118]. Among these, anchoring bias[119] compels individuals to heavily rely on initial information, distorting their assessment of new data based on that anchor. This inclination to assess new information from the lens of the anchor rather than objectively can obscure judgment and impede necessary adjustments to plans or projections.

In another study[117], researchers delve into the Hawthorne effect and observer-expectancy effect—specific cognitive biases. The former involves individuals altering their behaviour due to awareness of observation, while the latter stems from a researcher's cognitive bias impacting experiment participants. Newer research suggests the Hawthorne Effect's genuineness is debatable, casting doubt on the initial study's validity.

Addressing cognitive biases, Tim et al. present a checklist [120] featuring various biases, including:

Overconfidence or Optimism Bias: Can crowd workers overestimate their task performance?

- Self-interest Bias: Does the task facilitate motivated errors?
- Affect Heuristic: Can crowd workers be influenced by their affinity for annotated items?
- Groupthink or Bandwagon Effect: Does the task design inadvertently convey others' item evaluations?
- Salience Bias: Could judgments be swayed by the prominence of specific information?
- Confirmation Bias: Could preconceived notions sway crowd workers' judgments?
- Availability Bias: Does the task involve judgments likely to elicit stereotypical associations?
- Anchoring Effect: Is there a chance workers overly fixate on a specific reference point?
- Halo Effect: Are judgments susceptible to irrelevant information?
- Sunk Cost Fallacy: Are the task's time requirements and expectations transparent?
- Disaster Neglect: Are participants informed about task consequences?
- Loss Aversion: Does the design raise concerns about fair payment?

Further categorization [60] encompasses biases relating to data, algorithms, and users. For instance, Measurement Bias, Omitted Variable Bias, and Representation Bias are linked to data to algorithms, while Algorithmic Bias and Popularity Bias are algorithm-to-user biases. Various measures are proposed, including social projection[121], AfLite algorithm [89], and worker bias measurement [121].

A notable complexity emerges when biases aggravate protected groups, necessitating the exploration of social biases in models and strategies to mitigate them algorithmically [122]. Efforts like SoPro[121], AfLite [86], and worker bias measurement [121] strive to mitigate biases. However, existing bias mitigation techniques tailored for classification don't seamlessly extend to the realm of truth discovery [117].

Design choices also influence judgments' fairness [123], particularly in crowd judgment tasks, potentially amplifying biases. Studies [124] indicate that systematic biases in crowdsourced answers might be less prevalent than anticipated, but their impact magnifies with increasing group size.

In sum, cognitive biases play a pivotal role in influencing decisions, and efforts are being made to identify, categorize, and mitigate these biases across various domains of research and application.

Behavioural biases[125]

Behavioural biases in crowdsourcing refer to the systematic deviations from rational decision-making that can occur among individuals participating in a crowdsourcing task. These biases can impact the quality and reliability of the contributions made by crowd members and can affect the overall outcomes of the crowdsourcing process.

Some common behavioural biases that can be observed in crowdsourcing include:

Anchoring Bias: This initial information, or "anchor," influences their subsequent thoughts and choices, often leading to an inadequate adjustment away from that anchor, even if it's irrelevant or arbitrary. Anchoring bias can distort judgment and decision-making processes, potentially leading to inaccurate assessments or choices. This bias highlights the psychological tendency to base decisions on an initial reference point, rather than objectively evaluating all available information.

Confirmation Bias: This bias leads people to actively favour information that aligns with their existing viewpoints while disregarding or downplaying evidence that contradicts those beliefs. As a result, confirmation bias can hinder rational decision-making and lead to the reinforcement of existing biases.

Bandwagon Effect: The tendency to adopt or conform to the majority opinion or behaviour. In crowdsourcing, the bandwagon effect can result in a biased aggregation of opinions, as individuals may be influenced by the choices or preferences of others, rather than independently evaluating the information.

Availability Heuristic: The availability heuristic is a cognitive bias that influences people's judgments. Similar to anchoring relying on the immediate examples or information that come to mind easily. This bias leads individuals to overestimate the significance or likelihood of events, situations, or outcomes based on how easily they can recall relevant examples or information from their memory. Essentially, if something is more readily available in one's mind, it's often perceived as being more common or probable, even if this perception isn't necessarily accurate. The availability heuristic can lead to biases in assessing risks, making judgments, and forming opinions.

Overconfidence Bias: Overconfidence bias refers to the tendency of individuals to overestimate their abilities, knowledge, or the accuracy of their predictions. It involves having excessive confidence in one's judgments, often leading to an underestimation of risks and errors.

Cognitive Reflection Test [125]

The Cognitive Reflection Test (CRT) is a psychological assessment tool designed to measure an individual's tendency to override initial intuitive responses and engage in reflective thinking. It consists of several short questions that require individuals to resist impulsive or intuitive answers and instead engage in deliberate and analytical thinking. In the context of crowdsourcing, the CRT can be used as a means of assessing the cognitive abilities or thinking styles of crowd members. By incorporating the CRT as part of the screening process for participants in a crowdsourcing project, organizers can identify individuals who demonstrate higher levels of reflective thinking and analytical reasoning.

Integrating the CRT into crowdsourcing can have several potential benefits. It can help ensure that participants possess the cognitive skills necessary for more complex tasks that require analytical thinking. Additionally, it can enhance the quality of

contributions by selecting individuals who are less prone to certain cognitive biases and more likely to provide thoughtful and well-reasoned responses.

Furthermore, the CRT scores of crowd members can be used as a variable in data analysis or as a means of stratifying participants into different subgroups based on their cognitive reflection abilities. This stratification can aid in understanding how cognitive reflection influences task performance, decision-making, or problem-solving within the context of crowdsourcing.

Raven's Standard Progressive Matrices (RSPM) [126]

RSPM uses diagrammatic patterns in a matrix format to assess the reasoning and fluid intelligence.

Heuristics-and-Biases Test (HBT)[127]

The Heuristics-and-Biases Test (HBT) can be utilized to assess the presence of cognitive biases and reliance on heuristics among participants contributing to a crowdsourcing project. By incorporating the HBT as part of the evaluation or screening process for crowd members, organizers can gain insights into the decision-making tendencies and potential biases that may impact the quality of their contributions.

Syllogistic Reasoning Test (SRT) [128]

The Syllogistic Reasoning Test (SRT) is a cognitive assessment tool that measures an individual's ability to reason logically and draw conclusions based on given premises using syllogisms. A syllogism is a type of logical argument that consists of two premises and a conclusion.

In the context of crowdsourcing, the SRT can be used to evaluate the logical reasoning skills of participants contributing to a crowdsourcing project. By incorporating the SRT as part of the screening or evaluation process, organizers can identify individuals who demonstrate strong deductive reasoning abilities, which can be valuable in tasks that require logical thinking and problem-solving. General biases and issues stated in [33] are populations, behavioural, content, linking, temporal, and redundancy.

Population biases

Population bias refers to systematic discrepancies in demographics or user attributes between a platform population and another focused population. This bias can manifest in several ways:

1. Diverse user demographics are often attracted to distinct social platforms.
2. User demographics tend to interact with platform mechanisms in varying manners.
3. The reliability of proxies for user traits or demographic criteria may vary.

Content biases

Content biases are expressed as different features of languages such as lexical, semantic etc.

Common issues:

- Language usage varies among and within countries and populations.
- Contextual considerations influence how users communicate.
- Popular or "expert" user material contrasts with regular stuff.
- Different populations have differing proclivity to debate various issues.

Aspect identification

As proposed in [129], the unsupervised Attention-based Aspect Extraction (ABAE) technique can identify relevant aspects. This technique has been employed by [130]. The model is trained to generate sentences resembling the original ones and simultaneously acquire the skill to emphasize significant words[67]. The absence of segregation leads to inaccurate evaluation of workers' abilities. It's worth noting that the constructed dataset might exhibit biases specific to the dataset [131]. In [132] researchers validate the presence of "in-batch annotation bias".

3.3.3 What are the methods used to measure the trust of crowd response?

Trust holds the ability to impact the quality of outcomes and expenses. Therefore, choosing workers with higher levels of trust could offer an improved solution for workflow efficiency and cost management.

As highlighted in their research [133], McGeer and Pettit emphasize that when someone places trust in you within a specific domain, it can enhance your ability to show reliability and gain trust. This dynamic can have a positive and empowering effect on your psychological state. Task allocation aims to optimize both the overall quality of completed work and the total cost of the entire workflow.

Two methods can be used in evaluating trustworthiness [58] by considering the credibility of the actions and the credibility based on the members' affiliations, such as their participation in a contract or connection with a university. Commonly used approaches to assess trustworthiness are the reputation-based approach, the gold standard approach[122], and the consensus-based approach. Additionally, researchers use the expectation-maximization (EM) Algorithm and Majority Voting (MV) [134] to evaluate trust. Vijay et al. [131] propose a joint learning model to simultaneously train a classifier and a deferrer in a multiple-experts setting.

In this paper [135], researchers tackle the challenge of combining unreliable reports from a crowd of observers while learning the trustworthiness of individuals. To achieve this, they construct a likelihood model of users' trustworthiness by considering uncertainty and trustworthiness parameters. This model is integrated into a fusion method that combines estimates based on trust parameters. An inference algorithm is provided that computes the fused output and individual trustworthiness using a maximum likelihood framework.

The primary contributions of the work [136] are (1) a unique profiling approach utilizing multi-criteria crowdsourced data to build pairwise trust models and (2) k-NN prediction of user ratings using trust-based neighbour selection. Trusted modelling involves Pearson correlation and user interaction evaluation. Trust values among users are updated with each user event. For instance, if a user frequently selects recommendations from a neighbour k , their trustworthiness with k increases. This approach relies on data streams, such as a sequence of events that add new ratings and enhance existing models to provide more accurate recommendations.

Pairwise comparisons [137], [138], [139]

The major theme of this article[154] is collaborative ranking issues for user preference prediction from crowdsourced pairwise comparisons. To estimate the underlying weight/score matrix and predict the ranking list for each user, a maximum likelihood estimation (MLE) strategy based on the Bradley-Terry-Luce (BTL) model is suggested.

Strong stochastic transitivity (SST) model [140]

The SST model is a mathematical framework employed for the analysis of preference or ranking data. It is specifically designed to analyze and predict individual preferences or choices based on pairwise comparisons.

The SST model assumes that individuals have underlying preferences for a set of items or alternatives and that these preferences can be inferred from the observed pairwise comparisons.

The SST model assigns probabilities to each pairwise comparison outcome, considering the uncertainty or noise associated with individual judgments. It allows for the possibility that preferences may not always follow strict transitivity due to factors like measurement errors, individual differences, or random variations in decision-making.

Estimating the parameters of the SST model typically involves maximizing the likelihood of the observed pairwise comparison data, taking into account the probabilistic nature of the model. Once the model parameters are estimated, they can be used to make predictions about individuals' preferences or rank items based on the observed pairwise comparisons.

Truth Discovery

The majority of current truth-finding techniques rely on iterative updates, optimization, or probabilistic models[141]. Un-supervised methods, MajorityVoting, TruthFinder [141], Average-Log, Investment, PooledInvestment, CRH, CATD, SimpleLCA, GuessLCA and semi-supervised methods, SSTF, ClaimEval methods are being used in the paper [142] to compare their proposed model for truth discovery.

The paper[143] introduces an optimization-based strategy that employs labels provided by workers to determine the optimal set of combined annotations. They use

Majority Voting, CRF-MA, HMM-crowd, BSC-seq and OptSLA as the baseline method to compare the proposed method.

In their work, Yanying, Taipei, and Wendy [117] introduce a novel concept of fairness termed " θ -disparity" within the context of truth discovery. The paper [77] discusses three important research questions, Are distinct strategies employed by untrustworthy workers when accomplishing tasks, along with variations in their exhibited behaviors? Furthermore, is it possible to discern and quantify behavioural trends among malicious workers within the crowd?

The advanced truth discovery method is the Latent Truth Model (LTM) [144]. The Latent Truth Model is a statistical model that infers the true values by jointly estimating the reliability of the sources and the true values themselves. It assumes that each source has an unknown reliability score that indicates its accuracy or trustworthiness. Additionally, the model supposes that the data or information observed is created based on a concealed truth that is shared among all sources.

Xiu et al.[145] assess the credibility of individual workers using measurement. Leveraging this measurement along with existing domain knowledge, scholars determine nuanced worker credibility for specific tasks. To prevent task allocation to workers who merely replicate answers, researchers employ Bayesian analysis for copier detection.

Furthermore, the researchers introduce a crowdsourcing system named SWWC, comprising two key stages: task assignment and truth discovery. In this framework, they utilize an iterative approach to compute estimated truth and worker credibility. Differential privacy (DP)[146] has recently become more prominent in truth finding.

Nasim et al. [143] propose an optimization-based method In addition to the Kazai categorization of worker types Spammers were categorized as random spammers and uniform spammers by considering the [58] categorization.

By considering [75]classification, and trying to identify an unenthusiastic assessor, the assessor is not interested in reading or understanding documents and wants to complete the job. Though identifying optimistic annotators is important it was not practical to identify optimistic annotators and pessimistic annotators. The standard 10-question test [147] was used to classify the crowd workers by measuring the five personality dimensions by Goldberg.

CHAPTER 4

RESEARCH METHODOLOGY AND CONCEPTUAL MODEL

Introduction

This chapter outlines the approach taken to conduct the research, gather data, analyze it and draw conclusions. Moreover, this chapter describes the key concepts, variables, relationships, and assumptions related to the topic of study.

4.1 Research Design

This research contains two major stages; the first stage is the design and implementation of a crowdsourcing framework to facilitate social media content moderation and the second is to build a model to measure the trustworthiness of crowd workers. The listed tasks in Table 4.1 were conducted under each stage.

In this research to best address the research questions and meet the objectives, a mixed-methods research design was adopted. A qualitative research approach was followed for the preliminary face-to-face study to identify how social media users respond towards different types of social media content. An experimental research approach was followed by designing and implementing the crowdsourcing platform to capture the worker responses. The crowdsourcing platform was refined after the analysis of data to increase the quality of crowd responses by measuring the trustworthiness of worker responses.

Three steps out of the five under Stage 1 and four steps in Stage 2 are elaborated as sections of this chapter. Developing research problems, identifying the research gap and forming the research objectives specified under Stage 1 are elaborated in Chapter 1. The findings of the literature survey carried out as step 2 on using crowdsourcing towards NLP applications including hate speech detection and identification, image classification etc. are elaborated briefly under Chapter 2 in this thesis.

Stage 3 involves the preliminary study carried out as a face-to-face session to identify how social media users respond towards different types of social media content. The objectives of this study, and the methodology of data collection and analysis, are explained in section 3.2 of this chapter.

Stage 4 and Stage 5 involve designing and implementing a crowdsourcing platform to moderate social media content. These stages and the system overview of the proposed crowdsourcing platform are discussed in section 3.3. Solution design and implementation are further explained in Chapter 4.

The systematic literature review on measuring bias of the workers to ensure trustworthiness in the crowdsourced collection of subjective judgements given as stage 6 is explained in Chapter 2 under Section 3. In addition to the initial process, literature surveys were carried out during each phase of the research when it was required.

Stage 7, the implementation of trust metric to enable trust modelling and reasoning about crowd trust is explained under Chapter 4.

Stage 8, Data collection using the implemented crowdsourcing platform, data pre-processing and processing, data analysis and interpretation is explained in section 3.3 of this chapter. Results and the verification of the results are elaborated in Chapter 5.

TABLE 4.1 : MAIN STEPS AND STAGES OF RESEARCH DESIGN

	Detail
STAGE 1	
Step 1	Develop research questions and plan research design
Step 2	Perform a literature survey for background study, identifying the research gap, and identifying the analytical features of the proposed crowdsourcing platform.
Step 3	Preliminary face-to-face study to identify how social media users respond towards different types of social media content.
Step 4	Designing an analytical framework with the identified techniques to moderate social media content using crowdsourcing.
Step 5	Implement the crowdsourcing platform to facilitate inappropriate content identification with necessary quality control and analytical features.
STAGE 2	
Step 6	Perform a systematic literature review in measuring the bias of the workers to ensure trustworthiness in the crowdsourced collection of subjective judgements and in identifying appropriate trust metrics to evaluate the reliability of the crowd response.
Step 7	Implement a trust metric to enable trust modelling and reasoning about crowd trust.
Step 8	Data collection using the implemented crowdsourcing platform, data pre-processing and processing, data analysis and interpretation.
Step 9	Refine the platform and verify the results with other existing methods

4.2 Preliminary face-to-face study with social media users.

4.2.1 Objectives

The preliminary study aimed to understand how users on social media react to various types of content and, based on these insights, determine which features should be incorporated into the crowdsourcing platform.

4.2.2 Data collection for preliminary study

Data collection was done by interviewing and studying the facial expressions, gestures, and body language of the participants. Each participant was shown a selected set of

social media posts and the observations were recorded. These social media posts consisted of Facebook and Twitter posts.

Web crawling was used to gather Facebook data and Twitter API to collect Twitter data.

4.2.3 Methodology of the Preliminary Study

The preliminary study tried to explore the user perception of different types of content based on their religion, cultural practices and political viewpoint and the workflow is illustrated in Fig. 4.1.

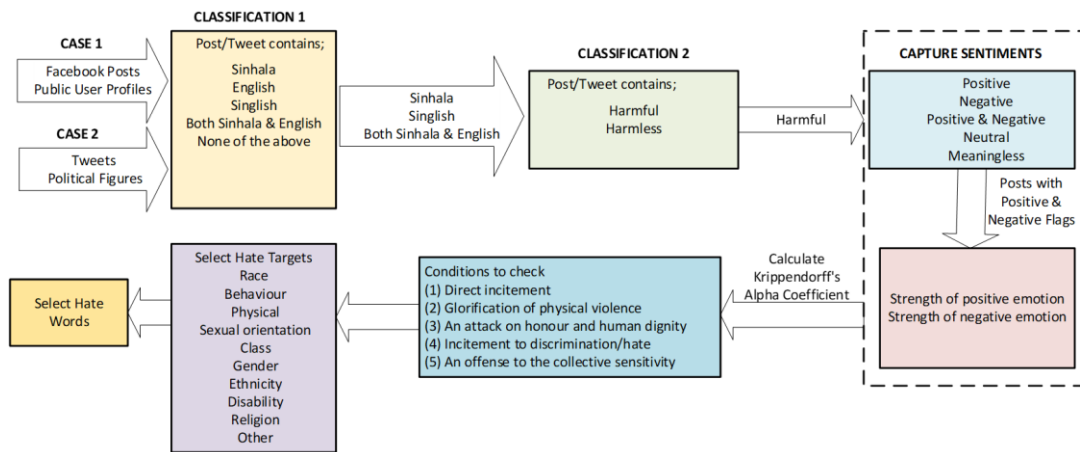


Fig. 4.1 : Workflow diagram of the preliminary study

The initial study was conducted, taking into account two cases. The initial case involved an unlabeled dataset collected from public Facebook user profiles in Sri Lanka. The second case involved an unlabeled dataset sourced from tweets posted by specific political figures, as indicated in Table 4.2.

The Facebook posts were in Sinhala, English, and Singlish with text posts, comments, replies to comments, images, and thumbnails from the video posts, other attributes such as no. of likes for posts and comments, smiles, etc.

The preliminary study was carried out with the help of 100 social media users and only 38 participant responses were considered after removing 52 participants identified as spammers. Monetary or any other reward was not given to the participants of the study. The observations are recorded of the 38 participants.

The task assigned encompassed three distinct classification phases. In the primary classification phase, the objective was to categorize posts based on their written language. The languages were divided into five distinct categories: English, Sinhala, a combination of both English and Sinhala and Singlish and a category denoting none of the above. Among these categories, the investigation focused on posts written using variations of Sinhala language and English. Additional languages were excluded from the research. The number of instances with Sinhala and Singlish contents considered in each case is given in Table 3.2.

TABLE 4.2 : UNLABELLED DATA SETS FOR PRELIMINARY STUDY

Case	Dataset	Tasks
Case 1	Data was collected from public Facebook user profiles originating from Sri Lanka without any labelling.	1000, 10 per user
Case 2	Unlabelled Twitter posts originating from well-known political figures in Sri Lanka(the Year 2018 to Jan 2020).	2000, 20 per user

Harmless or harmful was the second classification involved in shortlisting content to be fired to contributors. Although posts with harmless content were labelled but were not included in the study.

[no positive emotion or energy] 1– 2 – 3 – 4 – 5 [very strong positive emotion]

[no negative emotion] 1– 2 – 3 – 4 – 5 [very strong negative emotion]

The messages as positive, negative, neutral, negative positive and meaningless were the categories considered as options. The sentiment strength of each identified negative and positive post was assessed using a Likert scale, as depicted below.

The subsequent set of questions was formulated to pinpoint the specific hate group being targeted. Every post was presented to a minimum of five users on social media, and the level of agreement was calculated using Krippendorff's alpha coefficient for each instance. Contributors were directed to label a post as hate speech if it met any of the conditions outlined in Table 4.2.

TABLE 4.2 : CONDITIONS USED TO DETERMINE WHETHER OR NOT A MESSAGE CONTAINS HATE SPEECH

Condition in English	Translated to Sinhala
(1) direct incitement/threat of violence	සෘජු උසිගැන්වීම / ප්‍රචණ්ඩත්වය සම්බන්ද තර්ජන
(2) glorification of physical violence	ශාරීරික හිංසනය ප්‍රශංසා කිරීම
(3) an attack on honour and human dignity	ගෞරවයට හා මානව ගරුත්වයට එරෙහි ප්‍රහාරයකි
(4) incitement to discrimination/hate	අසාධාරණ සහ අගතිගාමී ලෙස සැලකීමට හෝ වසිර කිරීමට පෙළඹවීම
(5) an offense to the collective sensitivity	සමූහයකට අපහාස කිරීම හෝ ප්‍රකෝපනය කිරීම

For the content that users indicated as containing hate speech, the specific targets of the hate speech were determined. The classification criteria introduced by Silva[28] were employed. These categories encompass gender, race, sexual orientation, physical traits, social class, ethnicity, disability, religion, conduct, and a miscellaneous

category. If a contributor could not find an option in this list and select “other”, the contributors were asked to enter the hate target that they had identified.

As a final phase, contributors were requested to choose terms meant to arouse to signify the intensity of each word and its association with hatred, which was then put to a hate base. This study's observations and findings are detailed in Chapter 5.

4.3 Conceptual Model of the Crowdsourcing Platform

Based on the observations of the preliminary study the functionalities of the crowdsourcing platform were decided. The observations and the rationale for making decisions are illustrated in Table 4.4.

TABLE 4.4 : OBSERVATIONS FROM THE PRELIMINARY STUDY

Observation	Decision/Functionality
<p>Participants were more interested in classifying Tweets than Facebook posts, mainly due to the context of the Tweets. The posts were taken from the tweets of two well-known candidates during the presidential election period, and the participants knew about this context.</p>	<p>Participants partake in the classification task when they possess a familiarity with the contextual background of the post. Conversely, participants opt to omit the classification task in instances where such familiarity is lacking.</p> <p>It is required to include the context of the post for the crowd workers to decide if a particular post consists of hate content or not. Crowd responses can be aggregated to provide such context.</p>
<p>An inherent challenge was encountered in finding contributors from diverse religious affiliations, particularly given the common Sinhala-speaking and reading demographic. Moreover, many Hindu and Islamic participants exhibited difficulties in comprehending Sinhala-language posts.</p>	<p>Consequently, a preliminary assessment of comprehension levels was considered necessary before participant registration.</p>
<p>The sample composition lacked contributors who held strong and well-defined political viewpoints.</p>	<p>As a result, it was decided not to consider political viewpoints at this stage of the research and to consider future improvements.</p>
<p>It was observed that often contributors consider expressions of anger as posts</p>	<p>It is recommended that training be implemented to effectively perceive posts with hate speech, as contributors</p>

containing hate speech mainly because of the lack of knowledge.	often misattributed expressions of anger as indicative of hate speech. Thereby classifying the hate content based on its severity levels.
A significant number of participants demonstrated a tendency to approach the task with a passing completion mindset, rather than engaging with substantial involvement.	It was required to explain the importance of completing the tasks with due diligence and to incorporate an incentive mechanism along with assessing the trustworthiness of participant responses.
Participants demonstrated behaviour that there is a right response or correct answer for the given classifications rather than answering from their intuition	This behaviour led to decide the necessity of integrating social projection in collecting user responses.
A sample of participants demonstrated a lack of empathy towards identifying sentiments in social media posts	It was decided to opt out of performing the sentiment analysis and to leave it for future work.

The solution design and the implementation of the crowdsourcing platform are explained in Chapter 4. The next section explains the research methodology carried out in collecting data from the crowdsourcing platform and the analysis of the data.

4.3.1 Data Collection

Data collection was done in two stages. Use of pilot dataset for task design. Data was collected from the contributor responses of the implemented crowdsourcing platform.

Pilot Dataset

Three subsets of social media posts extracted from YouTube, Facebook and Twitter were used in this research as the pilot dataset from the year 2022. These subsets were selected based on their significant presence among Sri Lankan users [3]. Singlish denotes the use of Sinhala words in English text. Consequently, our dataset of tweets was procured through the Twitter API, relying on 996 specific search keywords.

The nature of the data is outlined in Table 4.5. The dataset underwent annotation by 20 appointed annotators who successfully cleared a pre-selection assessment and exhibited elevated levels of trustworthiness. Any encoding errors and inconsistencies in the UPF-08 encoding were rectified in the annotated files.

TABLE 4.5 : UNLABELLED DATA SETS

Case	Dataset	Instances
Case 1: Facebook	User profiles, pages, and groups on Facebook within the Sri Lankan context which can be accessed publicly*	52,646
Case 2: Twitter	Tweets**	45,000
Case 3: Youtube	YouTube videos***	45,000

* Comments, posts, responses to comments, images associated with posts and video thumbnails, sourced from.

** A total of 996 targeted search keywords were utilized to gather Twitter content, which included tweet text, corresponding replies, 6317 video thumbnails, and images. On an average basis, each Twitter post is accompanied by approximately 4 comments and replies.

*** YouTube videos were analyzed, covering aspects like video titles, thumbnails, and comments associated with each video. On average, each video contains approximately 6 comments and replies.

JSON Schemas of the Facebook, Twitter and YouTube data used are given below in Fig. 4.3,4.4 and 4.5 respectively. The complete dataset is accessible in the GitHub repository.

```

1  {
2    "postURL": "xxxxxx",
3    "top_level_post_id": 2809687085929207,
4    "PostedBy": "xxxxxxxxxxxxxxxx",
5    "IDofPostedBy": 663973230,
6    "OriginalPostURL": null,
7    "original_content_id": null,
8    "original_content_owner_id": null,
9    "postedDateTime": "2020-08-26 8:31 ",
10   "OriginallyPublishedOn": "xxxxxxxxxxxxxxxx",
11   "OriginallyPublishedOn_id": 2502722529958999,
12   "caption": null,
13   "privacy": "Public",
14   "photo_id": null,
15   "photo_url": null,
16   "MainImageSavedLocation": null,
17   "video_thumbnail": "no",
18   "video_id": null,
19   "VideoThumbnailSavedLocation": null,
20   "sharesCount": 0,
21   "react": {
22     "total": 2,
23     "like": 0,
24     "love": 0,
25     "haha": 2,
26     "wow": 0,
27     "sad": 0,
28     "angry": 0,
29     "care": 0
30   },
31   "commentCount": 0,
32   "comments": [],
33   "group_photo_count": 0,
34   "group_photos": []
35 }

```

Fig. 4.2 : JSON schema for Facebook posts

```

1 {
2   "id": 1229307467638738947,
3   "created_at": "2020-02-17 07:31:54",
4   "favorite_count": 3,
5   "in_reply_to_screen_name": "xxxxxxxxxxxxxx",
6   "in_reply_to_status_id": 1229068050168766464,
7   "in_reply_to_user_id": 887285012453896193,
8   "is_quote_status": "False",
9   "lang": "si",
10  "retweet_count": 0,
11  "text": "xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx",
12  "entities": {
13    "hashtags": [],
14    "urls": [],
15    "user_mentions": [
16      {
17        "screen_name": "Chandhi_03_12",
18        "id": 887285012453896193
19      },
20      {
21        "screen_name": "1KRPA1",
22        "id": 1137191909775073280
23      }
24    ],
25    "medias": []
26  },
27  "original_tweet_id": 1229068050168766464,
28  "user": {
29    "id": 514285275,
30    "name": "xxxxxxxxxxxxxx",
31    "screen_name": "xxxxxxxxxxxxxx",
32    "created_at": "2012-03-04 12:26:54",
33    "favourites_count": 2704,
34    "followers_count": 401,
35    "friends_count": 306,
36    "statuses_count": 1739,
37    "location": "Sri Lanka",
38    "verified": false
39  },
40  "Comments": [

```

Fig. 4.3: A segment of JSON schema for Tweets

```

1 {
2   "VideoURL": "https://m.youtube.com/watch?v=GohYvW5BGAM",
3   "videoId": "GohYvW5BGAM",
4   "Video_title": "Neth Fm Balumgala | කොරෝනාවේ ශ්‍රී ලාංකීය ගැලවුම්කරු|2020-01-31",
5   "tags": [
6     "#Nethfm_Balumgala",
7     "#Subscribe",
8     "#Like"
9   ],
10  "Video_thumbnail_url": "https://i.ytimg.com/vi/GohYvW5BGAM/maxresdefault.jpg",
11  "Video_thumbnail_saved_location": "C:\\Youtube_data\\Video_thumbnails\\GohYvW5BGAM.jpg",
12  "Views": 981,
13  "Likes": 26,
14  "Dislikes": 1,
15  "Channel_Name": "Neth Fm Balumgala",
16  "Subscribers": 95000,
17  "Published_Date": "2020-02-03 00.01",
18  "Description": "#Nethfm_Balumgala\n#Subscribe us on YouTube - https://www.youtube.com/cha",
19  "Comment_Count": 1,
20  "Comments": [
21    {
22      "Comment_id": "Ugxxk6N2-kG1bN5ofFpZ4AaABAg",
23      "Commented_by_URL": "http://www.youtube.com/channel/UCly_byEpNwRj4lkOvx3RPaw",
24      "Commented_by": "glwa j",
25      "Comment_text": "Lankave Production eka patan ganna.",
26      "Comment_Date": "2020-02-03 08:26",
27      "Comment_Like_Count": 0,
28      "Comment_Dislike_Count": 0,
29      "Comment_Reply_Count": 0,
30      "Replies": []
31    }
32  ]
33 }

```

Fig. 4.4 : A segment of JSON schema for YouTube posts

4.3.2 Experimentation

It was evident from the results of the face-to-face preliminary study that it was a complex and challenging task for social media users to identify hate speech content as they were always misled by the subjectivity and the context of the post. Hate speech perpetrators often use coded language, metaphors, sarcasm, or other forms of creative expression to bypass automated filters and human moderators. Hate speech sometimes uses sarcasm or irony to convey offensive messages indirectly, making it difficult for automated systems to distinguish between genuine expressions and disguised hate speech. Hate speech can take on new forms and terms and rapidly changing vocabulary and slang used by hate speech perpetrators. This makes it challenging to detect hate speech accurately. Therefore a training on “Identifying Hate Speech” was conducted on a group of interested participants before getting a sample dataset manually annotated to use as the benchmarking dataset or the golden standard. Majority voting was used to ensure trustworthiness and the number of annotators was varied from three annotators to five annotators by considering the complexity of the task. If an annotator flagged a task as a complex task the number of annotators was increased by one. The details of data stored in a crowdsourcing platform other than the annotated datasets are given in Table 4.6 and the manual social media post annotation process is illustrated in Fig. 4.5.

TABLE 4.6 :DETAILS OF DATA STORED IN THE CROWDSOURCING PLATFORM

Attribute	Details
Worker Details	
Worker ID	Unique identifiers for individual crowd workers, ensuring traceability and quality assessment.
Demographics	Information about workers' demographics, such as age, gender, location, highest educational qualification, religion, and language proficiency.
Performance Metrics	Metrics related to workers' performance, accuracy, speed, and reliability(Listed in Chapter 4).
Responses and Annotations	
Answers	Actual responses provided by crowd workers for the task
Annotations	Additional information or metadata is attached to the responses, such as confidence scores, timestamps, or explanations.
Quality Control and Validation	
Gold Standard Data	High-quality or known-correct data is captured after the manual process of annotation as a reference to assess worker quality.
Agreement and Disagreement	
Inter-Rater Agreement	Krippendorff alpha and kappa coefficient =measure the degree of agreement among different crowd workers for the same task.

	This helps assess the difficulty of the task and the reliability of the collected data.
Task Progress and Completion	
Timestamps	Records of when tasks were assigned, started, and completed by crowd workers.
Completion Status	Indicators of whether a task was completed or if there were any issues or incompletions.
Privacy and Ethics	
Data Anonymization	Measures were taken to ensure the privacy of both crowd workers and any individuals mentioned in the data.
Ethical Considerations	Records of how ethical concerns, such as sensitive content or biased tasks, were addressed during the data collection process.

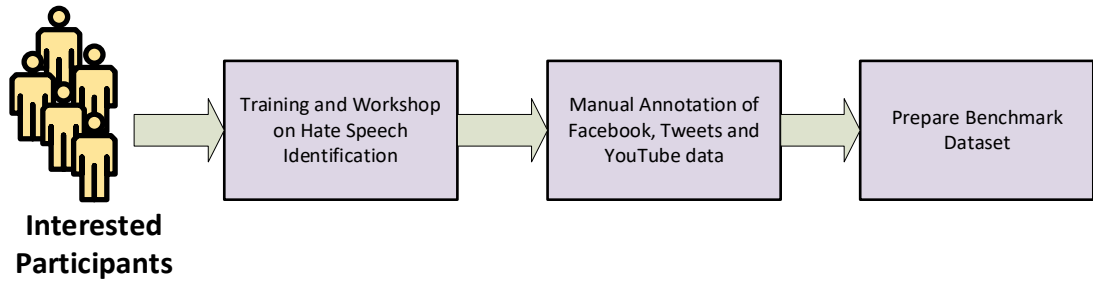


Fig. 4.5 : Workflow of manual data annotation process

After the preparation of the benchmark dataset, the implemented crowdsourcing platform was used to capture user responses. The research examined three distinct scenarios for YouTube, Twitter and Facebook posts as outlined in Table 3.4. These scenarios were utilized for the tasks of identifying hate speech and generating a hate corpus. The crowdsourcing platform was not evaluated for the task designs of hate speech content propagator identification and image text identification. The seven modules of the crowdsourcing platform were evaluated for their accuracy and performance.

Neural Network Architecture to model the crowd worker trustworthiness is illustrated in Fig. 4.6. In this system, seven features were integrated namely Accuracy_R_Gold, Accuracy_R_Consensus, CompletionRate, ReputationScore, BiasnessScore, Weighted Evaluation Metric and Aggregated Trust Score calculated from the crowd responses. Observed feature values are given in Appendix H.

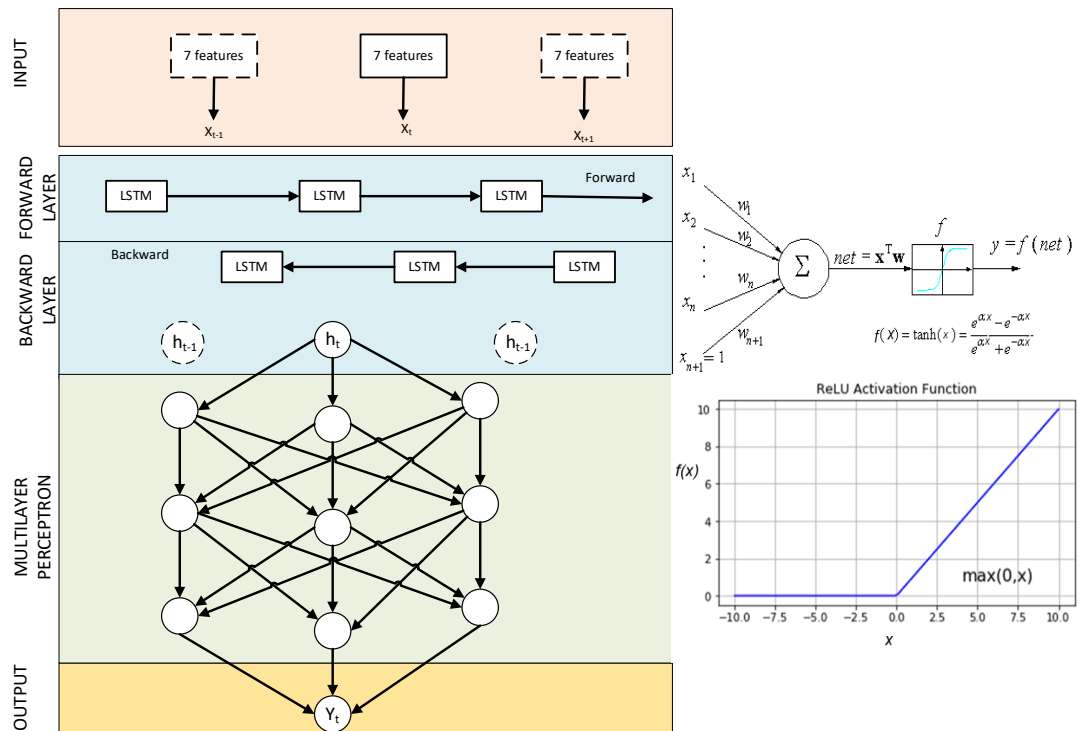


Fig. 4.6 : Neural Network Architecture

Calculations of the feature values are further explained in Chapter 4. To implement the model Bidirectional LSTM along with multilayer perceptron with three hidden layers were used. These elements collaboratively contribute to the determination of the reliability score output. A hyperbolic tangent activation function and the rectified linear unit activation function(ReLU) were adapted.

Conclusion

In summation, this research employed a mixed research methodology encompassing both qualitative and quantitative elements. An initial phase of qualitative investigation was undertaken to comprehend the diverse responses of various social media users to distinct forms of social media content. This phase involved the use of a sample of Facebook posts and Tweets as the basis for task design, with data collection conducted through interviews and observational methods directed at social media users. Subsequently, guided by the insights derived from the preliminary study, the development of a crowdsourcing platform ensued. This platform featured seven distinct modules, each meticulously tested to ensure their quality and efficacy. Central to this endeavour was the establishment of a mechanism advancing the trustworthiness of user responses throughout the content moderation process. In pursuit of this obligation, a comparative analysis was conducted, evaluating the abilities of consensus-based, reputation-based, and gold-standard approaches.

CHAPTER 5

SOLUTION DESIGN AND IMPLEMENTATION

Introduction

This section presents the approach taken to implement the crowdsourcing platform and the model to measure the trust of the crowd response.

5.1 Crowdsourcing Platform to Moderate Social Media Content

The crowdsourcing platform permits interested individuals to sign up by submitting information on the profile of a user. This information includes name, nationality, location of the users, date of birth etc. This incorporates the findings from the initial study. To gather data for essential analysis, tasks related to identifying hate speech and generating a hate corpus were created using the operational crowdsourcing platform. Fig. 5.1 illustrates the architecture of the deployed crowdsourcing platform.

Worker management, firing questions, rewarding contributors, quality control, task design and price control are the seven major components of the platform. Each component is briefly explained in the remaining section.

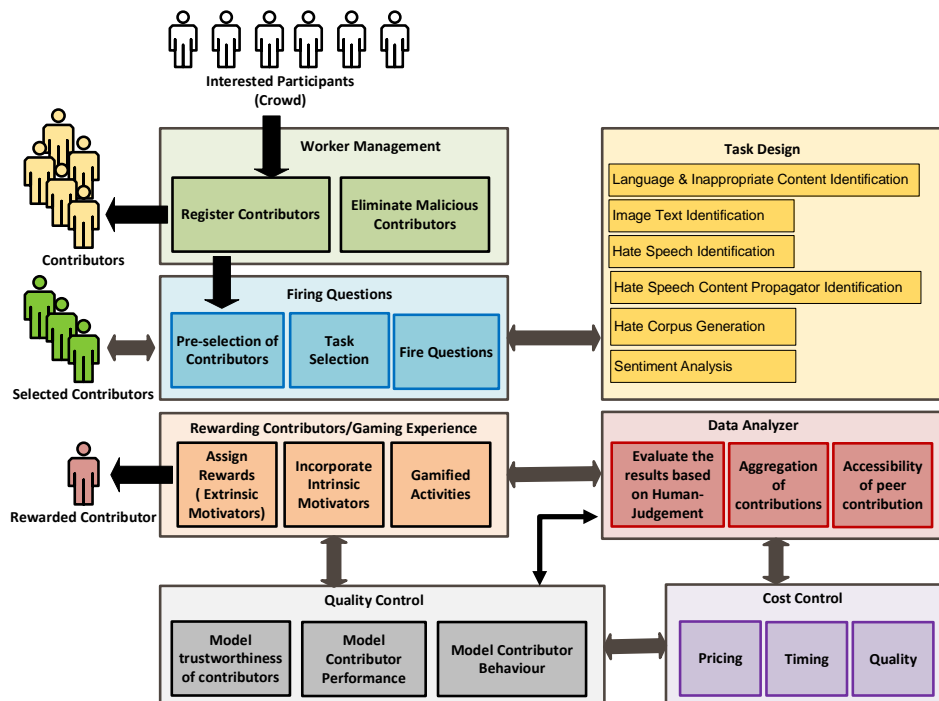


Fig. 5.1 : System architecture of the implemented crowdsourcing platform

5.1.1 Worker Management

Worker management involves contributor registration and a malicious worker elimination process. Anyone interested in serving the cause of creating better cyberspace was allowed to register as a worker through the system by signing up using the “Work with Us” button shown in the user interface given in Fig. 5.2.

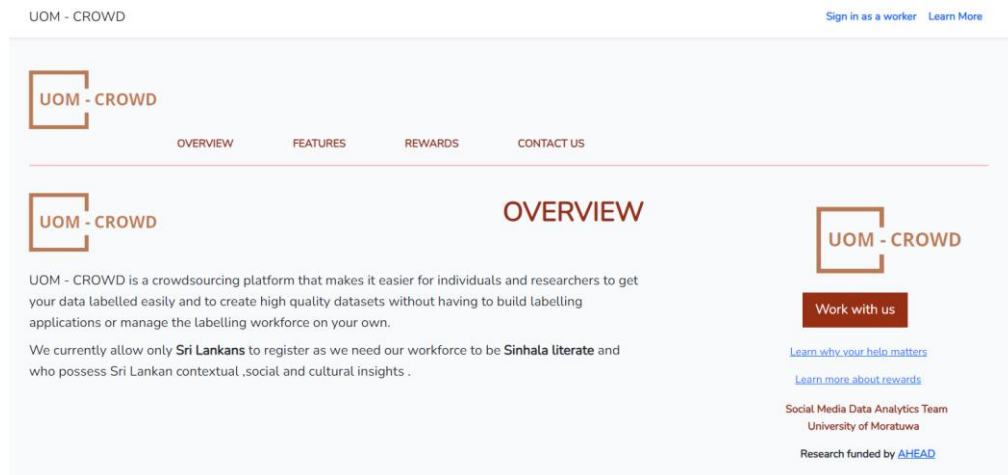


Fig. 5.2 : Crowdsourcing Platform – Home Page

Contributor registration involves three more stages; creating the contributor profile, and checking if the interested user is fit to serve as a contributor and to serve as a contributor.

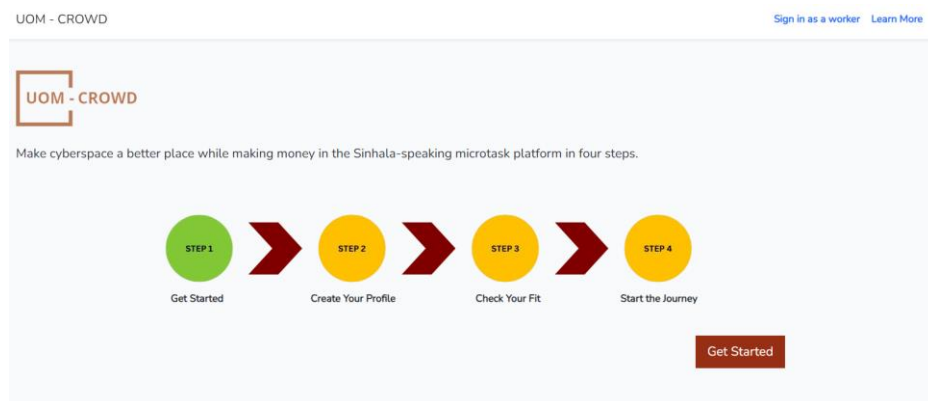


Fig. 5.3 : Worker Registration Page

Assessing fitness to work as a contributor has four assessments with 10 questions in each section to assess Sinhala language proficiency(Appendix A), capability to read mix codes(Appendix B), and familiarity with the process of identifying hate speech. (Appendix C), comprehension, and analytical skills(Appendix D). A sample of this questionnaire for each assessment is given as Appendixes. User consent was taken to the following declaration before the registration process.

The purpose of this research is to design a crowdsourcing platform that allows social media content moderation. For that it is required to examine the social media users beliefs and biasness's towards politics, etc. This platform consists of a registration form which would collect your personal information such as Name, Age, Work place etc. Furthermore the system will direct you to answer few questions about your personal beliefs of the categories *race, behavior, physical, sexual orientation, class, gender, ethnicity, disability, religion, other*.

Your replies will be kept anonymous and will not be associated with you personally. Your involvement is entirely voluntary. Please skip any questions that you do not feel comfortable answering. Thank you for your assistance.

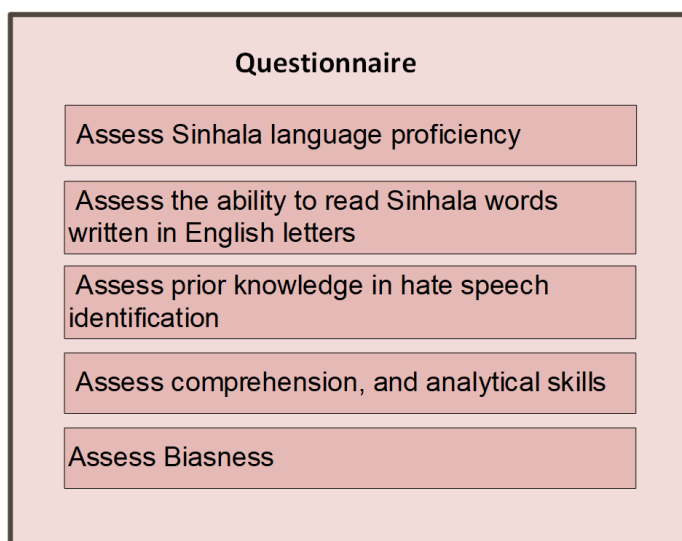
5.1.2 Fire Questions

This module is responsible for firing the respective questions to the contributors. There are two major types of questionnaires as illustrated in Fig. 5.4. Four questionnaires are used to assess whether a contributor meets the pre-selection criteria and to detect any potential biases. The second type of task involves assignments falling within the six categories depicted in the diagram below.

A rule-based system was employed, utilizing forward chaining as the mechanism for triggering actions. (Appendix E).

The following details were stored against each type of task design.

- Task Description: A clear and concise explanation of the task that needs to be completed by the crowd workers.
- Instructions: Specific guidelines or instructions are provided to workers on how to perform the task correctly.
- Examples: Sample inputs and expected outputs to help workers understand the task



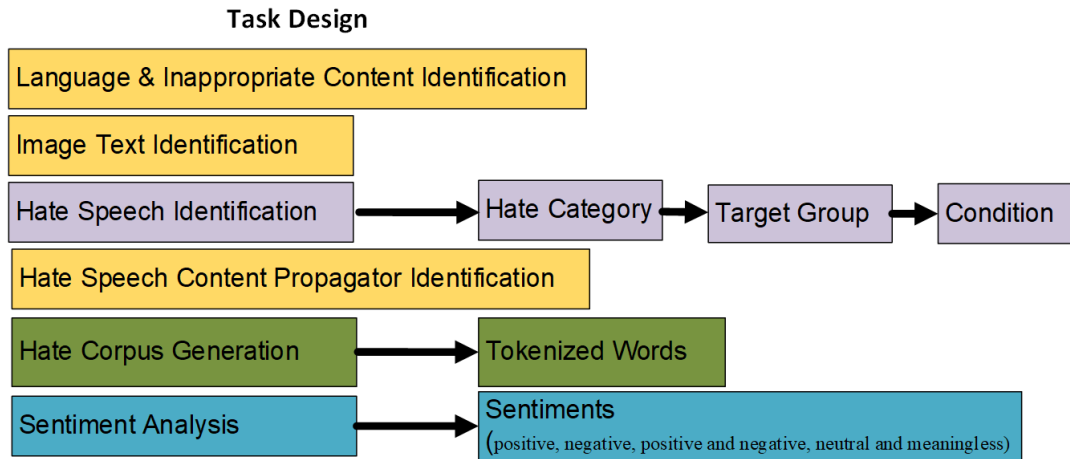


Fig. 5.4 : Question types and task types in the questionnaire

Contributor pre-screening:

Individuals who possessed literacy in Sinhala and were native to Sri Lanka were chosen as contributors to the evaluation process. Participants who didn't meet the pre-selection criteria, didn't pass quality control, or didn't meet the trustworthiness requirements were excluded. The table outlines definitions for the symbols of contributor pre-screening in Table 5.1.

Comprehension and analytical skills (CAT=8), pre-existing familiarity with hate speech detection (HS_T=5), capacity to grasp English letters used to write Sinhala words (LT=8) and Sinhala language competence (LP_T=8) were chosen as assessment criteria. Table 5.1 outlines the criteria for pre-selection of contributions.

TABLE 5.1 : SYMBOL DEFINITIONS LIST FOR CONTRIBUTOR PRE-SELECTION

Symbol	Definition
N	Nationality
A	Age
HS, HS _T	Knowledge level of hate speech, the Threshold value
LP, LP _T	Language proficiency (Sinhala), the Threshold value
CA, SCA _T	Comprehension & Analytical skill assessment (Sinhala), the Threshold value
L, L _T	Ability to read mixed codes (Sinhala words written in English letters), the Threshold value

TABLE 5.2 : PRE-SELECTION CRITERIA FOR CONTRIBUTORS

Pre-selection of contributors
if {(Nationality="Sri Lankan") and (A>=18) and (HS>=HS _T) and (LP>=LP _T) and (CA >= CA _T) and (L >=L _T)}
Badge="Selected Contributor"
Else
Eliminate contributor
Endif

5.1.3 Extrinsic and Intrinsic Rewarding Process

The process of intrinsic rewards begins with the initial digital badge named "Contributor." Once the selection criteria outlined in Table 5.2 are met, contributors are awarded the "Selected Contributor" badge, gaining eligibility for receiving financial rewards. The worker management process is depicted in Fig. 5.5.

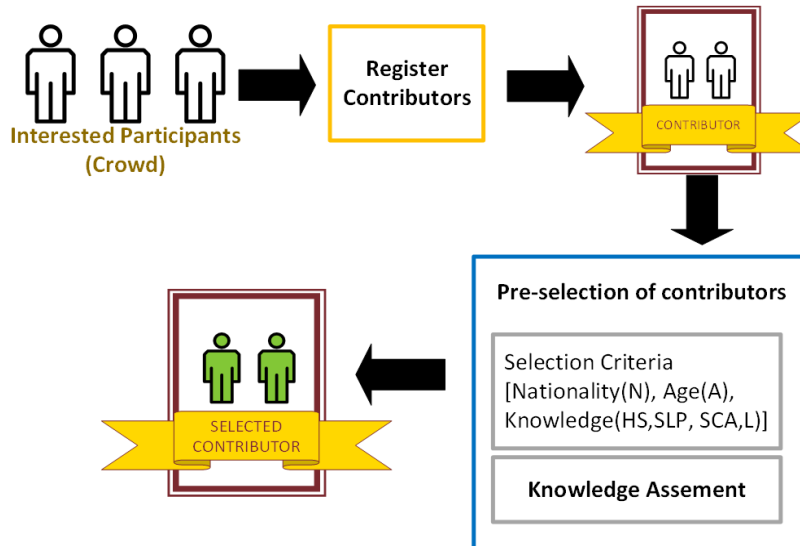


Fig. 5.5 : Worker registration and reward process

Drawing inspiration from the crowdsourcing platform of Google[148], a comparable strategy has been adopted in this study to motivate workers from Sri Lanka. The rewards system was developed after experimenting with various reward techniques using a sample of workers.

Participants or contributors of the platform can earn money as extrinsic rewards. These are assigned by considering how far they have completed different tasks and the level that they are in, the score obtained for trustworthiness and the accuracy after comparing with the Golden standards. This process would provide an incentive for the contribution to create the cyber space a better one.

Furthermore, digital badges were awarded based on the completion of human intelligence tasks (HITs). Additionally, the platform offered a gaming experience to engage contributors and maintain their involvement in the initiative. This gaming experience was enriched with both intrinsic and extrinsic motivators, as illustrated in Fig. 5.6.

ALL LEVELS AND BADGES



























<p> REGISTRATION</p> <p> Unlock your CONTRIBUTOR badge</p>	0 Points	
<p> LEVEL 0</p> <p> Unlock your SELECTED CONTRIBUTOR badge</p>	0 Points	
<p> LEVEL 1</p>	50 Points	
<p> LEVEL 2</p>	75 Points	
<p> LEVEL 3</p> <p> Unlock your first monetary reward</p>	100 Points	
<p> LEVEL 4</p> <p> Unlock your second monetary reward</p>	200 Points	
<p> LEVEL 5</p>	350 Points	
<p> LEVEL 6</p> <p> Unlock your third monetary reward</p>	500 Points	
<p> LEVEL 7</p>	750 Points	
<p> LEVEL 8</p> <p> Unlock your fourth monetary reward</p>	1000 Points	

Fig. 5.6 : Assignment of rewards for chosen contributors

5.1.4 Quality Control

This module comprises a model to measure the trustworthiness of crowd responses, constructing a contributor characteristics model, model contributor performance and contributor behaviour classification. Each of these models is further explained from Section 4.2 onwards.

5.2 Novel Annotation Scheme

Annotation of data was performed using the scheme specified in Table 4.3. and the labelled data can be found in GitHub.

TABLE 5.3 : COMPILATION OF SYMBOL DEFINITIONS FOR LABELLING PURPOSES.

Symbol	Definition
L ₁	Content analysis
L ₂	Hate speech identification
L ₃	Who does a particular post target?
L ₄	Hate categories
C ₁ , C ₂ , C ₃ , C ₄ , C ₅	Direct incitement or threat of violence, an attack on honour and human dignity, Incitement to discrimination or hate, An offence to the collective sensitivity or Other.
L ₄ {1 to 16}	1. Race and Ethnicity 2. Religion 3. Nationality 4. Sexual Orientation 5. Disability 6. Disease 7. Immigration 8. Victims of a major violent event and their kin 9. Veteran Status/Profession 10. Cast 11. Political 12. Regional 13. Gender 14. Economic & Business 15. A particular individual 16. Other social groups

Definition to label L₁ to L₄

The starting set of labels L₁ is:

$$L_1 = \{Offensive\ content,\ no\ offensive\ content,\ cannot\ tell\}$$

The starting set of labels L₂ is:

$$L_2 = \{C_1,\ C_2,\ C_3,\ C_4,\ C_5\}$$

The starting set of labels L₃ is:

$$L_3 = \{Individual,\ Group,\ cannot\ tell\}$$

The starting set of labels L₄ is:

$$L_4 = \{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16\}$$

5.1.5 Data Analyzer

Evaluate the results based on Human Judgement

This module is designed to assess the outcomes or responses provided by human contributors participating in the crowdsourcing tasks. These responses are reviewed and analyzed by human experts after providing training.

Aggregation of contributions

The level of agreement among the contributors' responses when addressing specific tasks or questions within the crowdsourcing platform is referred to as the degree of consensus. To determine the level of agreement the following agreement scores were computed using Krippendorff's alpha statistical, Fleiss' kappa and Cohen's kappa coefficient.

Accessibility of peer contributions

This module is proposed for contributors or participants within a crowdsourcing platform to access and review the contributions made by their fellow participants. When the participants are assigned tasks or questions, and once they submit their responses, these responses are made accessible to other contributors for evaluation or further processing.

5.1.7 Cost Control

The cost control unit was recognized as a viable component in any crowdsourcing platform, although its execution remained pending. Specifically, the aspects encompassing pricing, timing, and quality modules were among the domains where this unit's potential impact was identified. However, despite its recognition, the practical implementation of the cost control unit has yet to be realized within these domains.

5.2 Worker Behaviour Model for Crowdsourcing Platform

Developing worker categorization mechanisms to ensure worker trustworthiness in crowdsourcing platforms is an important problem in the ever-growing collaborative platforms.

This section explains the approach taken to build a model to describe worker behaviour on the crowdsourcing platform. To model the behaviour of workers on a crowdsourcing platform effectively, it is required to capture a variety of worker interactions and actions that reflect worker engagement, performance, and decision-making. The next section explains the type of interactions considered in the study and the overall system is illustrated in the following Fig. 5.7.

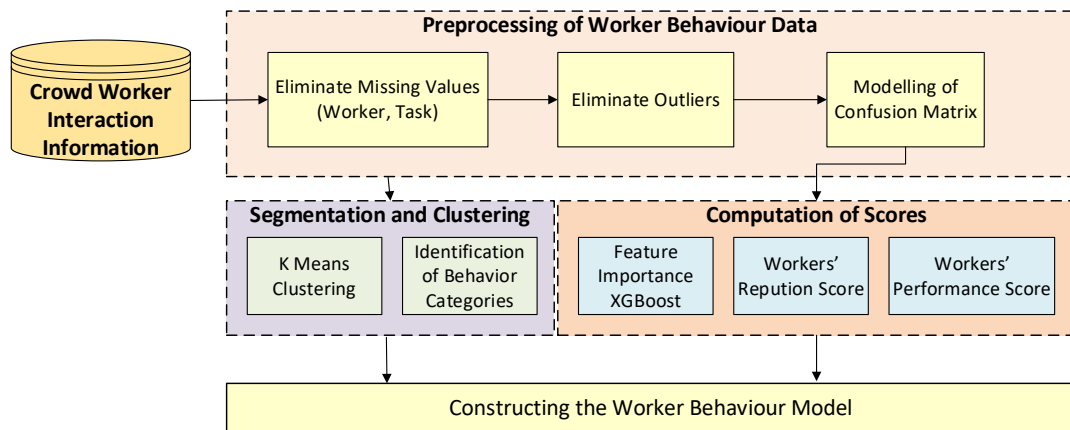


Fig. 5.7 : System Overview of Worker Behaviour Model

Data Collection and Features:

The interactions of crowd participants were recorded starting from issuing the batch “Contributor” of the workers. The recorded data against each Worker consists of the following;

TABLE 5.4 : FEATURES CONSIDERED FROM CONTRIBUTOR RESPONSES

Symbol	Definition
Task Participation and Completion	
T _i	Task ID
Total_Tasks	Number of tasks completed
Total_Completed	Number of tasks completed within a given period
Total_Attempted	Number of tasks attempted by each worker within a given period
CompletionRate	Percentage of tasks completed compared to tasks attempted (Total_Completed/ Total_Attempted)
{ Task ID, Time T Completion }	Time taken to complete tasks
Accuracy and Quality	
Accuracy_R_Gold	Accuracy of responses considering Golden Rules
Accuracy_R_Consensus	Accuracy of Responses Considering Consensus
Response Patterns	
Response_Time	Time taken to submit responses after task assignment
Consistent {0,1-true}	Consistency of response time

Data Preprocessing:

After identifying which features have missing values it was observed that the majority of the workers have opted out of labelling the same set of tasks and it was chosen to drop rows with missing values. Scatter plots were used to identify extreme values that deviate significantly from the majority of the data and chosen to drop the rows.

Segmentation and Clustering:

K means clustering algorithm was used to identify worker groups with different characteristics and to divide workers into clusters based on similar behaviour patterns by considering the feature vector given in Table 5.4 above.

Identification of Behaviour Categories:

Spammers, high-performing, low-performing, inconsistent, high quality and low-quality workers were identified as behaviour categories by considering the worker feature characteristics.

Computation of Feature Importance Scores:

XGBoost was used to compute feature importance scores based on the contribution of each feature to improving the model's performance

Performance Metrics:

Completion rate, accuracy and response time were used as the performance metrics relevant to the crowdsourcing platform to model the performance of each crowd worker.

Behavioural Patterns:

Analysis of temporal patterns in worker behaviour was conducted to check the consistency of response time and thereby the response time trends. The correlation between the behaviour pattern and task types was identified. A predictive model was built to estimate worker performance on future tasks. The dataset was divided into three sets: training, validation, and test sets. The training set was used to train the model, and the validation set was utilized to avoid overfitting by assessing the model's performance. The loss function used was mean squared error (MSE), and the Adam optimizer was employed for gradient descent. Under anomaly detection identification of workers with unusual behaviour was done such as sudden drop in performance. To determine how worker behaviour affects their reputation and trustworthiness the past behaviour score was used.

5.3 Assess the trustworthiness of contributors

The subsequent diagram given in Fig. 5.8 provides a system overview of the model designed to assess the trustworthiness of contributors.

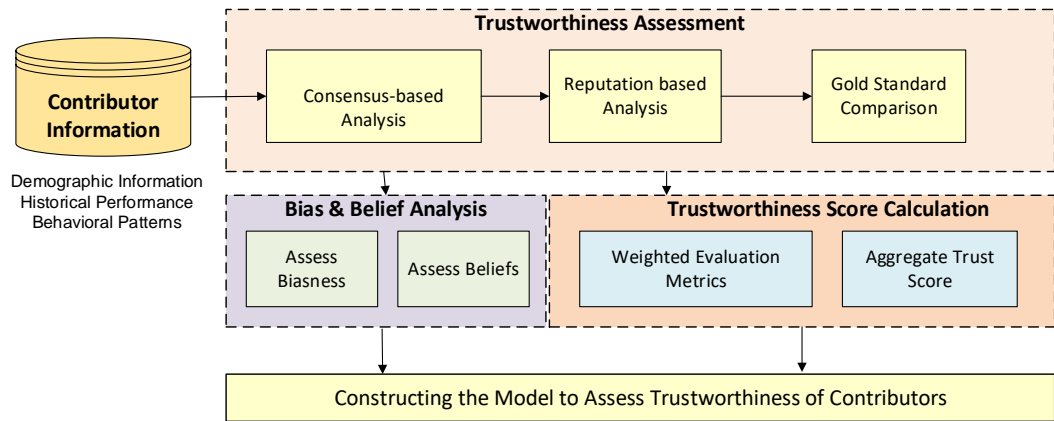


Fig. 5.8 : System overview of the model designed to assess the trustworthiness of contributors

Trustworthiness Assessment

Trustworthiness assessment is performed based on consensus-based and reputation-based analysis and by considering the gold standard. Consensus-based Analysis compares a contributor's responses to those of the crowd, seeking agreement to determine trustworthiness. Reputation scores are assigned based on historical performance and contributions. Contributors with higher reputation scores are considered more trustworthy. Trustworthiness is assessed by comparing contributors' responses to a set of known accurate responses (the "gold standard").

Bias and Belief Analysis involves analyzing contributors' biases and beliefs to assess the impact on trustworthiness. The approach taken to calculate the worker reputation score is given below;

TABLE 5.5 : VARIABLE DEFINITIONS LIST FOR CALCULATING REPUTATION SCORE

Symbol	Definition
IWRS	Initial Workers' Reputation Score
PWPS	Past Workers' Performance Score
GRQ	Quality Score – Golden Rule based
CBQ	Quality Score – Consensus based
PIS	Provided Information Score

$$IWRS = (W1 * PWPS) + (W2 * GRQ) + (W3 * CBQ) + (W4 * PIS) \quad (1)$$

Where:

W1, W2, W3 and W4 are weights assigned to each factor. These weights are summed up to 1 to ensure the final score is within a reasonable range.

Past Workers' Performance Score represents the score calculated considering the contributor's past performance, completion rates, response time, etc.

The provided Information Score accounts for the score obtained at the pre-selection of contributor process based on the answers provided for the four questionnaires.

Quality Score is calculated both by considering the responses for the Golden Rules fired at the beginning and the accuracy score obtained for the consenses-based calculation.

Anomaly Detection: Train machine learning models to detect anomalies in response patterns, helping identify contributors who consistently provide inaccurate or spammy content.

The research encompassed the development of an internally embedded mechanism tailored to ascertain the trustworthiness of individual contributors' responses, subsequently awarding a badge signifying their trustworthiness status. Central to this mechanism, a set of ten established golden rules was employed as a cornerstone for evaluating trustworthiness. The outcomes derived from these rules were juxtaposed with predicted trustworthiness scores, thereby facilitating a comprehensive assessment.

In a deviation from conventional computational trust models, this novel approach distinctly segregates the notion of worker beliefs from their competence within varying contextual spaces. Furthermore, it systematically accommodates subjectivity inherent in the evaluation of a particular trustee by diverse trustors, acknowledging the nuances that arise in this multifaceted evaluation process.

A crowd consensus mechanism is implemented where multiple contributors review and rate each other's responses. Contributions that receive high ratings from multiple trusted contributors contribute positively to the contributor's reputation. The number of contributions was decided by considering the complexity of the task for annotation. Fig. 5.9 illustrates how the ground truth inference is performed.

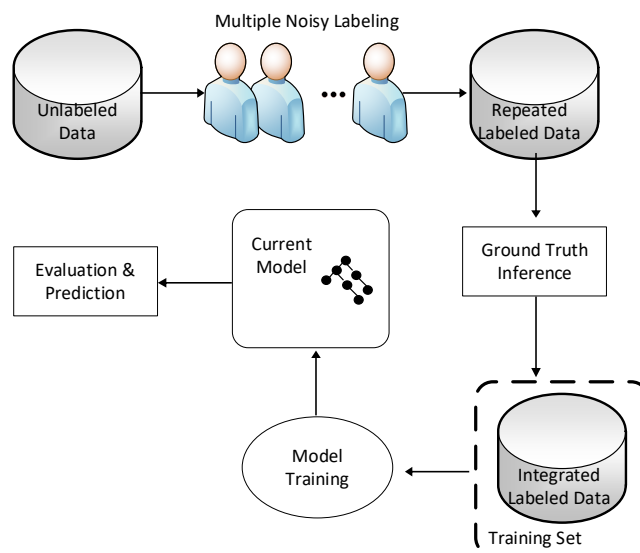


Fig. 5.9 : Ground Truth Inference

The research additionally conducted real-time experimental studies to empirically contrast the efficacy of the proposed integrity belief model with alternative trust models documented in the literature. These experiments spanned diverse user behaviour patterns to comprehensively assess performance. The results underscored the superiority of the proposed model, particularly in predicting the behaviour of users exhibiting instability or volatility in their interactions.

In essence, the research not only introduced a sophisticated mechanism for assessing trustworthiness but also charted new ground by unravelling beliefs and competence in the evaluation of trust, all while achieving sharp predictive performance through empirical validation. This holistic approach holds the potential to profoundly impact the realm of trust modelling and user behaviour prediction.

5.5 Implementation

Software tools used are listed below;

Crowdsourcing Platform Implementation: Laravel Framework with MySQL

Data Analysis and Modelling: Open source neural network libraries: Keras library, running on the TensorFlow

The datasets, source codes for the completed project, hate targets, hate corpus and the hate-related search keywords employed on Twitter can be accessed in the provided GitHub repository:

<https://github.com/gsnadeerameedin/HateSpeechCorpus>

Conclusion

In conclusion, this chapter outlined the implementation strategy of the crowdsourcing platform and the trustworthiness assessment model. The crowdsourcing platform was designed to facilitate user registration and task allocation, emphasizing the importance of pre-selection and task variety. The architecture of the platform depicted the seamless integration of key components like worker management, task distribution, reward allocation, quality control, and more. The implementation followed a multi-stage process, ensuring the fitness of contributors through comprehensive assessments of language proficiency, hate speech knowledge and analytical skills. Meanwhile, the task distribution mechanism was powered by a rule-based approach, providing an effective way to direct questions to contributors. To motivate contributors, a rewarding system was established, drawing inspiration from Google Crowdsourcing's approach. Contributors were encouraged through intrinsic and extrinsic rewards, including digital badges and monetary incentives. These incentives were tailored based on the completion levels, accuracy, and trustworthiness of contributors.

Quality control played a key role in maintaining the reliability of crowd responses. An innovative model was designed to measure contributor trustworthiness, incorporating consensus-based analysis and reputation scores. This approach successfully addressed biases and beliefs in the response assessment process. Additionally, the behavioural

patterns of contributors were identified, categorized, and modelled using machine learning techniques.

Moving forward, the novel annotation scheme facilitated effective data annotation for hate speech identification. The evaluation process involved the application of Krippendorff's alpha, Fleiss' kappa coefficient and Cohen's kappa coefficient, to measure agreement levels and enhance reliability.

Furthermore, this chapter also stressed the accessibility of the annotated datasets, source codes, and hate corpus via the GitHub repository provided. A live demonstration of the implemented project was offered through the provided link, enhancing the transparency and replicability of the research.

In conclusion, this chapter demonstrated a comprehensive implementation of the crowdsourcing platform and the trustworthiness assessment model, contributing significantly to the development of an effective solution for identifying and managing hate speech on social media platforms. The subsequent chapters will delve deeper into the evaluation of the proposed solution, discussing the results and implications in greater detail.

CHAPTER 6

EVALUATION AND ANALYSIS

Introduction

This chapter presents the critical assessment of the methodologies employed, the outcomes obtained, and the implications derived from the research objectives. Through systematic evaluation and analysis, this chapter provides a deeper explanation of the research outcomes, contributing to the validation and refinement of the proposed concepts and frameworks. In the subsequent sections, a detailed exploration of the collected data, examination of patterns and trends, and the interpretation of results will collectively sort out the extent to which the research objectives have been met.

6.1 Preliminary face-to-face study with social media users.

The preliminary study tried to explore the user perception of different types of content based on their religion, cultural practices and political viewpoints and how social media users respond towards social media posts. The study intended to identify the functionalities to include in the crowdsourcing platform. The demographic characteristics of the preliminary study participants are given in Table 6.1.

TABLE 6.1 : DEMOGRAPHIC CHARACTERISTICS OF THE PRELIMINARY STUDY PARTICIPANTS.

Age	M	F	Experience Facebook use	M	F	Social Media Use	M	F	Political Viewpoint	M	F
18-24	3	0	<=6 months	4	8	Twitter Account	32	12	Strongly Agree	12	10
25-34	19	10	7-12 months	4	8	Using YouTube	60	30	Neutral	38	22
35-44	15	17	1-2 years	16	6				Strongly Disagree	13	5
45-54	10	4	2-3 years	15	10						
55-64	10	5	3-10 years	14	3						
65+	6	1	10+ years	10	2						

6.2 Analyzing the Annotation Method for Building a Hate Speech Corpus

Distribution of the annotator profiles' demographics

Fig. 6.1 displays the demographic breakdown of the chosen contributors who participated in the annotating process.

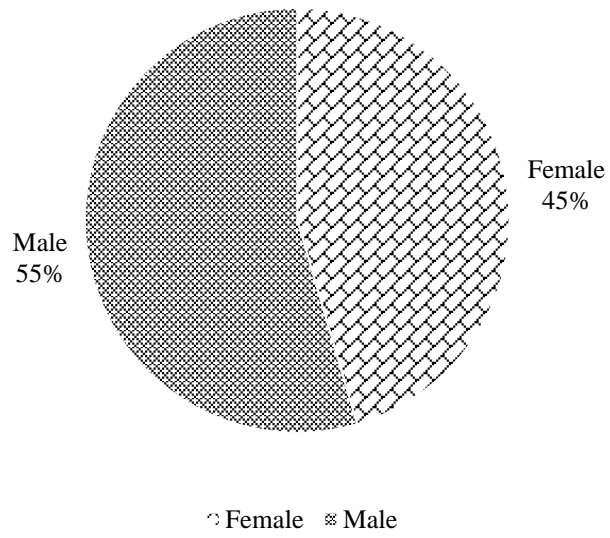


Fig. 6.1 : Gender Distribution of Selected Contributors

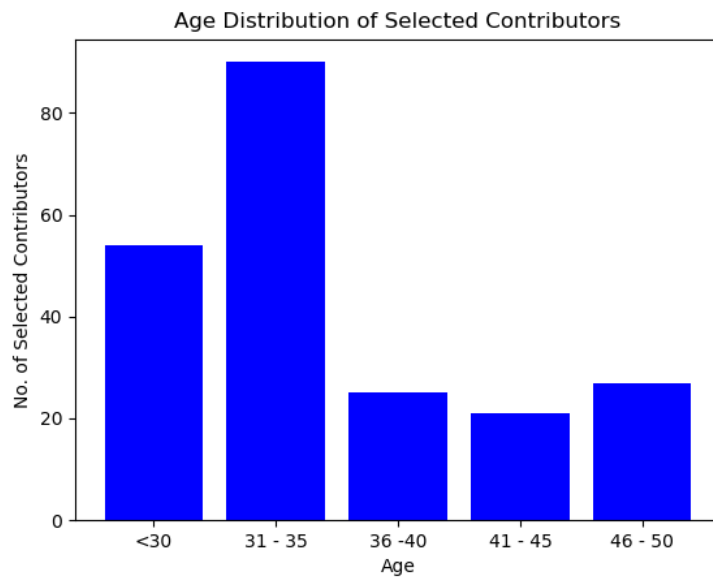


Fig. 6.2 : Contributors' age distribution

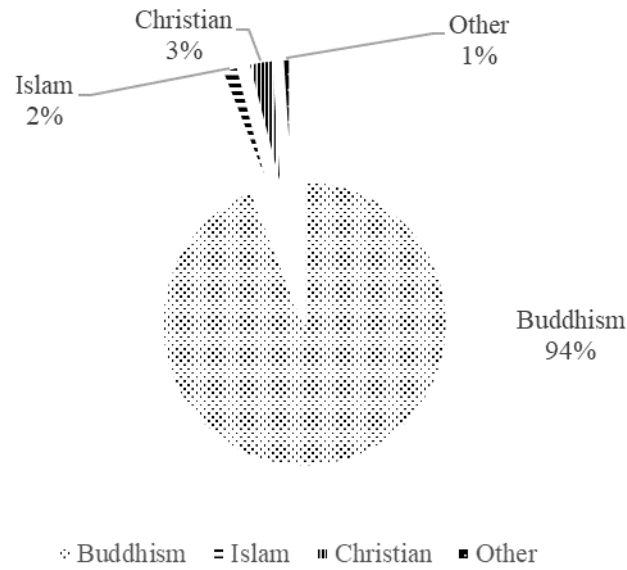


Fig. 6.3 : Contributors' Religious Distribution

6.2.1 Inter-annotator agreement

Table 6.2 displays the obtained agreements, represented as the average per cent agreement (agr_i), Krippendorff's alpha (α), average Cohen's kappa coefficient (avg k) and Fleiss' kappa coefficient (Fleiss). Additionally, the table provides the count of annotated tweets, comments, or posts for every sample.

In the case where the collection of items is $\{i | i \in I\}$, and the cardinality is i .

For the values agr_i , if the observed agreement is given by A_0 for all items $i \in I$

$$A_o = \frac{1}{i} \sum_{i \in I} agr_i \quad (2)$$

$$agr_i = \begin{cases} 1, & \text{if the two coders assign } i \text{ to the same category} \\ 0, & \text{if the two coders assign } i \text{ to different categories} \end{cases}$$

If expected agreement is given as A_e , Cohen Kappa coefficient(k) can be calculated as;

$$k = \frac{A_0 - A_e}{1 - A_e} \quad (3)$$

The results of the calculations of these coefficients are given in Table 6.2.

TABLE 6.2 : INTER ANNOTATOR AGREEMENT FOR L1

Case	Instances	agr_i	k	Fleiss	α
Case 1: Facebook	52,646	90.4	0.623	0.615	0.800
Case 2: Twitter	45,000	89.4	0.624	0.613	0.740
Case 3: YouTube	45,000	88.7	0.598	0.600	0.757

6.2.2 Datasets after Annotation

L2 and L3 were used to analyze hate posts, comments, and tweets. If a minimum of one option was chosen and the agreement between raters exceeded 0.6, the content was categorized as hate posts.

TABLE 6.3 : ANNOTATION RESULTS FOR FACEBOOK, TWITTER AND YOUTUBE DATA – HATE AND NO HATE

	Facebook	Twitter	YouTube
Hate	9%	21%	14%
No Hate	88%	70%	83%
Skip	3%	9%	3%

TABLE 6.4 : ANNOTATION RESULTS FOR FACEBOOK -OFFENSIVE AND NONE OFFENSIVE

	Offensive	No offensive content
Facebook		
Caption	1.68%	>98%
Video thumbnail	0.18%	>99%
Main image	0.015%	>99%
Comment text	0.023%	>99%
Comment image	0.002%	>99%
Reply text	0.78%	>99%
Group photo	0.04%	>99%
Twitter		
Tweet text	1.54%	>98%
Comment text	2.27%	>97%
Reply text	1.03%	>98%
YouTube		
Video title	0.81%	>99%
Video thumbnail	0.57%	>99%
Comment text	2.27%	>97%
Reply text	1.01	>98%

Upon comparing Table 5.3 and Table 5.4, annotation of content, whether offensive or non-offensive, demonstrates reduced percentages in contrast to the hate and non-hate classification.

6.2.3 Lexical Distribution

Table 5.5 lists the ten most used words to reference hate targets, along with the highest incidence rate for each category. Recognized hate targets, search keywords in Twitter,

and corpora reflect hate targets are included in the lexical distribution. This is available in the repository of GitHub. To obtain the annotated datasets from YouTube, Twitter and Facebook, please reach out to the authors of this thesis via email.

TABLE 6.5: THE MOST FREQUENT TERMS FOUND IN HATE SPEECH RELATED TO HATE TARGETS.

Word	An explanation of the term in English	Distribution
දෙමළා	Employed as a derogatory and disdainful term to refer to a specific targeted community.	0.65%
හමිලා	Employed as a derogatory and disdainful term to refer to a specific targeted community.	0.67%
තමිබියෙක්		0.65%
තමිබියා		1.9%
තමිබි	Employed as a derogatory and disdainful term to refer to a specific targeted community.	1.01%
තමිබියෝ		1.83%
තමිබිලා		0.38%
ගණයා	Employed as a derogatory and disrespectful term to address a priest belonging to a specific targeted religion.	0.25%
අන්තවාදී	Employed to label an individual as an extremist.	0.35%
සිංහලේ	Utilized to uplift a particular ethnicity while belittling all other ethnicities.	0.65%

68% of the phrases that mentioned hate targets included at least one term from the provided list.

Discussion

Upon completing this research, several limitations have been identified:

1. Workers found it challenging to determine the harmfulness or harmlessness of a comment solely from its content.
2. The required number of annotators varied based on the task, leading to fluctuations in annotator needs.
3. Efforts were made to mitigate annotator bias in responses.

4. The research found that viewing the original post, replies, and associated images was crucial to making informed decisions, suggesting the need to address entire comment threads instead of isolated comments.
5. Images and the applicable data should be given with the task to enhance the accuracy of the question.
6. Although subjective responses were collected during labelling, the annotation scheme should be designed to ensure unbiased and conscious judgment by annotators.
7. Pre-selection of contributors should utilize Krippendorff's alpha coefficient and vary the number of contributors for different tasks based on reliability.
8. Initial classifications can be performed by two contributors, while more complex tasks expect examining the strength of sentiments may require higher contributor numbers due to reliability concerns.
9. Representing equal percentages for each religion is challenging due to the predominantly Buddhist population in Sri Lanka. Randomized worker selection is necessary to prevent bias and ensure diverse post-categorization.

Future research should consider identifying multiple comments related to Facebook posts for preventive measures against hate spread, incorporating relevant images and context-specific data in crowdsourcing tasks, ensuring conscious judgment by annotators, measuring worker biases and beliefs, and exploring clustered worker types for different question categories.

For future research, the focus should be on the following areas:

1. Moving beyond a single comment and analyze at least a few comments. This approach could be more effective in preempting the spread of hate.
2. Revise the design of the tasks in crowdsourcing to encompass required images if any and data that is specific to the context. This approach enhances question accuracy and contextual understanding.
3. Explore the formation of how clusters of workers vary and tailor questions based on the cluster to enhance task efficiency and response accuracy.
4. Develop a mechanism that ensures annotators consciously evaluate comments based on provided criteria, eliminating intuitive responses.
5. Biases and beliefs of the workers should be identified and assessed to guarantee the crowd responses to ensure reliability.

6.3 Worker Behaviour Model

Either underfitting or overfitting can fail to capture the inherent structure and patterns within the data. This can lead to outcomes that are either overly generalized or excessively specific, driven by noise. In the latter case, clusters might lack meaningful

significance. Therefore Fig. 6.4 illustrates the methods used in finding an optimal number of clusters which was 4 in this scenario.

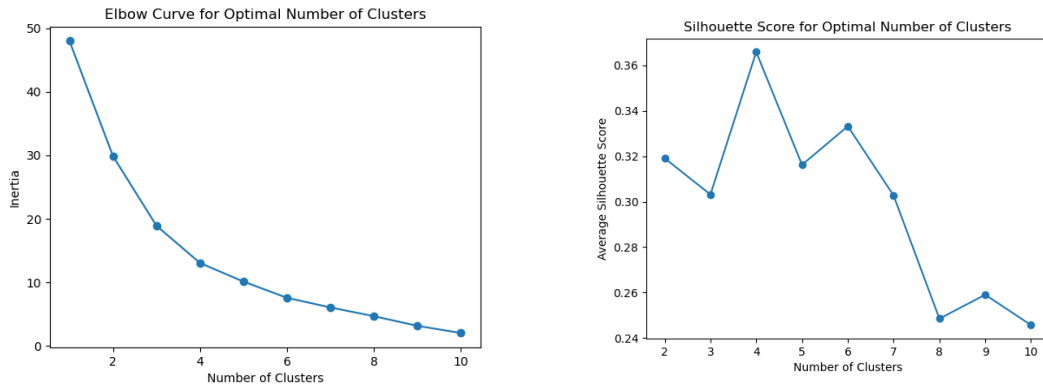


Fig. 6.4 : Find the optimal number of clusters

6.4 Model to measure trustworthiness

A comparison of the Accuracy and Precision of the consensus-based, reputation-based and gold-standard approach are given in this section.

The accuracy of the three measurements consensus-based, reputation and gold standard approaches for 20 annotators for a sample of 100 Facebook caption posts annotations are given below and illustrated in Fig. 6.5. The formula given as (4) was used to calculate the accuracy of the annotation process.

Symbol	Explanation	Example
True Positives (TP)	Instances where the measurement correctly identifies a positive case.	N number of annotators correctly identify 5 posts as hate posts when the post has hate content.
False Positives (FP)	Instances where the measurement incorrectly identifies a positive case.	N number of annotators incorrectly identify 95 posts as hate posts when the post has hate content.
False Negatives (FN)	Instances where the measurement incorrectly identifies a negative case.	N number of annotators incorrectly identify 5 posts as hate posts when the post has no hate content.
True Negatives (TN)	Instances where the measurement correctly identifies a negative case.	N number of annotators incorrectly identify 95 posts as a none hate posts when the post has no hate content.

(4)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

(5)

$$Precision = \frac{TP}{TP + FP}$$

Accuracy of Consensus-based approach: 0.79

Accuracy of Reputation-based approach: 0.82

Accuracy of Gold Standard approach: 0.89

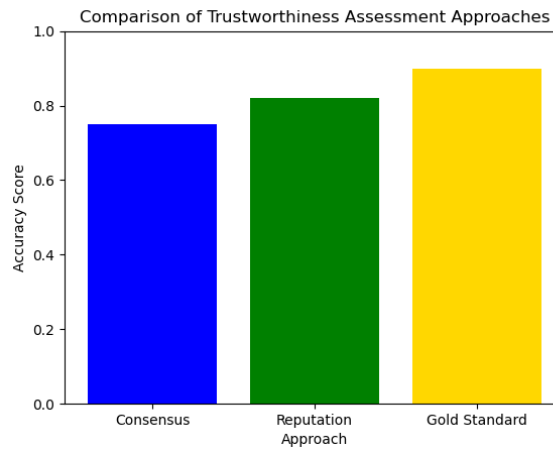


Fig. 6.5. : Comparison of the Accuracy of Consensus-based, reputation based and Gold standard approach

Results of accuracy and precision after varying the number of workers when annotating 100 posts are given in Table 6.6 below.

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.

N	Approach	True Positives (TP)	False Positives (FP)	False Negatives (FN)	True Negatives (TN)	Precision	Accuracy
1	CB	61	39	11	89	0.61	0.75
	RB	89	11	54	46	0.89	0.68
	GS	77	23	75	25	0.77	0.51
2	CB	51	49	15	85	0.51	0.68
	RB	80	20	4	96	0.80	0.88

	GS	89	11	24	76	0.89	0.83
3	CB	68	32	7	93	0.68	0.81
	RB	80	20	8	92	0.80	0.86
	GS	88	12	70	30	0.88	0.59
4	CB	52	48	10	90	0.52	0.71
	RB	78	22	37	63	0.78	0.71
	GS	79	21	28	72	0.79	0.76
5	CB	65	35	13	87	0.65	0.76
	RB	65	35	6	94	0.65	0.80
	GS	91	9	24	76	0.91	0.84
6	CB	53	47	28	72	0.53	0.63
	RB	81	19	12	88	0.81	0.85
	GS	80	20	2	98	0.80	0.89
7	CB	67	33	48	52	0.67	0.60
	RB	65	35	64	36	0.65	0.51
	GS	89	11	70	30	0.89	0.60
8	CB	81	19	2	98	0.81	0.90
	RB	66	34	27	73	0.66	0.70
	GS	79	21	11	89	0.79	0.84
9	CB	85	15	17	83	0.85	0.84
	RB	67	33	3	97	0.67	0.82
	GS	96	4	68	32	0.96	0.64
10	CB	78	22	49	51	0.78	0.65
	RB	91	9	74	26	0.91	0.59
	GS	70	30	71	29	0.70	0.50
11	CB	80	20	4	96	0.80	0.88
	RB	92	8	68	32	0.92	0.62
	GS	72	28	64	36	0.72	0.54
12	CB	67	33	60	40	0.67	0.54
	RB	91	9	77	23	0.91	0.57
	GS	73	27	13	87	0.73	0.80
13	CB	86	14	65	35	0.86	0.61
	RB	92	8	70	30	0.92	0.61
	GS	74	26	4	96	0.74	0.85
14	CB	72	28	11	89	0.72	0.81
	RB	61	39	6	94	0.61	0.78

	GS	76	24	54	46	0.76	0.61
15	CB	89	11	51	49	0.89	0.69
	RB	65	35	54	46	0.65	0.56
	GS	70	30	30	70	0.70	0.70
16	CB	77	23	76	24	0.77	0.51
	RB	70	30	6	94	0.70	0.82
	GS	70	30	66	34	0.70	0.52
17	CB	80	20	79	21	0.80	0.51
	RB	73	27	39	61	0.73	0.67
	GS	69	31	48	52	0.69	0.61
18	CB	80	20	20	80	0.80	0.80
	RB	86	14	10	90	0.86	0.88
	GS	99	1	13	87	0.99	0.93
19	CB	75	25	19	81	0.75	0.78
	RB	79	21	25	75	0.79	0.77
	GS	78	22	42	58	0.78	0.68
20	CB	82	18	24	76	0.82	0.79
	RB	75	25	12	88	0.75	0.82
	GS	79	21	1	100	0.79	0.89

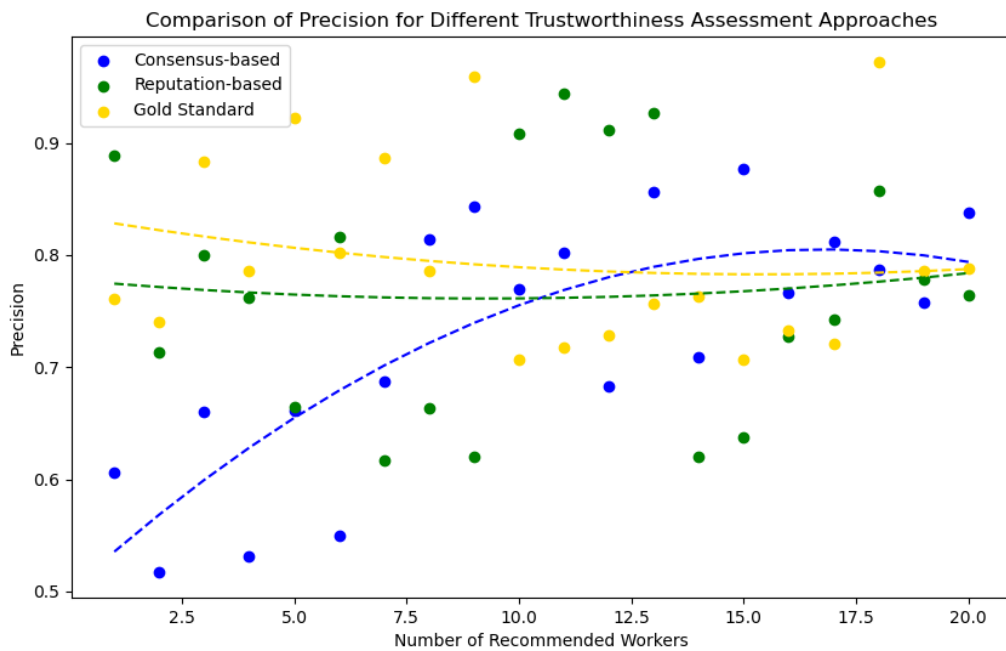


Fig. 6.6 :Comparison of precision for Consensus-based, reputation-based and Gold standard approaches for annotating 100 Facebook captions.

The scatter plot displayed above illustrates the comparison of precision scores for three distinct trustworthiness assessment approaches – consensus-based, reputation-based, and gold standard. The horizontal axis represents the number of recommended workers, while the vertical axis represents the precision scores achieved by each approach.

Across the range of recommended workers, trends were observed in the precision scores for each approach. Notably, the dashed lines depict polynomial trend lines that capture the general direction of the precision scores as the number of recommended workers increases.

For the consensus-based approach (depicted in blue colour), the precision scores exhibit a moderately increasing trend. This suggests that as the number of recommended workers grows, the reliability of the trustworthiness assessment tends to improve, resulting in higher precision.

The reputation-based approach (illustrated in green colour) portrays a steeper upward trend in precision scores. This implies that increasing the number of recommended workers more significantly contributes to elevated precision, underscoring the effectiveness of reputation-based assessments in enhancing the reliability of contributor evaluations.

The gold standard approach (shown in gold colour) reveals a consistent and gradual rise in precision scores. This suggests that, as anticipated, utilizing a gold standard for comparison serves as a reliable baseline for trustworthiness assessment, leading to incremental improvements in precision with the augmentation of recommended workers.

In summary, the scatter plot with trend lines provides valuable insights into the behaviour of different trustworthiness assessment approaches concerning precision. The observed trends can guide decision-making when determining the number of recommended workers for each approach, ensuring accurate and dependable contributor evaluations across diverse scenarios.

Based on the observation the reputation score(y) increases with performance score(x) and (y) decreases with bias and belief score(z). Therefore a relationship can be formulated as $y=f(x,z)$.

Where $y=f(x,z)$ is a function that captures the above behaviour. This behavior can be represented as a function of x and z and a simple example formula that demonstrates this behavior is $y=ax-bz$.

In this formula a and b are constants. When x increases, the positive ax term causes y to increase. When z increases, the negative bz term causes y to decrease.

This is illustrated in Fig. 6.7 for a sample of 100. The x-axis represents the performance score (x) ranging from 0.5 to 1, which is a measure of a contributor's performance or ability. The z-axis represents the bias and belief score (z) ranging from 0 to 0.5,

indicating the extent of bias or belief in the contributor's responses. The y-axis represents the reputation score (y) influenced by the performance score and bias and belief score, reflecting the contributor's reputation or level of trustworthiness. Each blue circle on the plot represents an individual contributor's data point.

Relationship between Reputation, Performance, and Bias/Belief

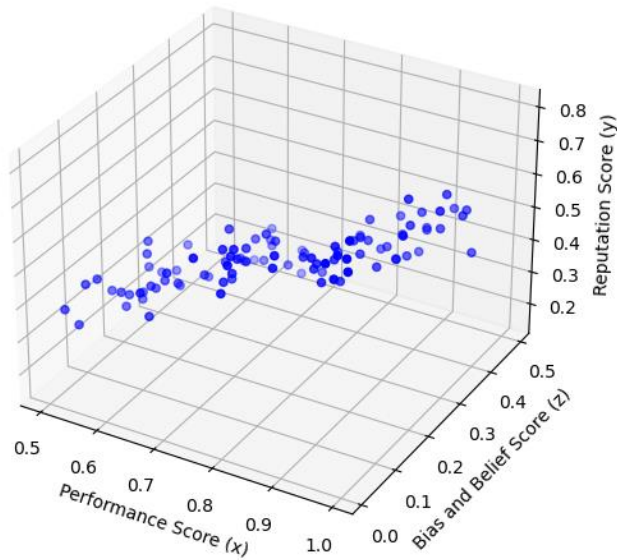


Fig. 6.7 : Relationship between reputation score, performance score and bias score

It is possible to notice an upward trend of reputation score as performance score (x) increases. This suggests that contributors with higher performance scores tend to have higher reputation scores, indicating trustworthiness or credibility. As bias and belief score (z) decreases, reputation score (y) tends to increase. This implies that contributors with lower bias and belief scores (indicating less bias and stronger belief) tend to have higher reputation scores.

Fig. 6.8 provides a visual representation of the clusters identified by the K-means algorithm. K=3 provided the most meaningful and interpretable segmentation. It is evident that the reputation score of the workers is concentrated in the (0.20, 0.45) interval, and the performance score is concentrated in the (0.5, 1.0) interval.

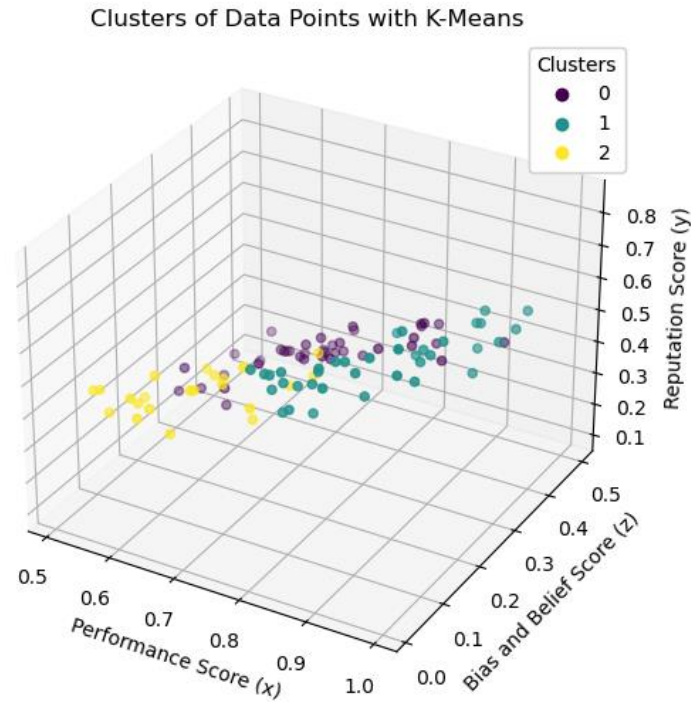


Fig. 6.8 : Clusters of Data points with K-Means Algorithm

Cluster 1 consists of moderate to high-performing contributors with low bias and high reputation scores. Cluster 2 consists of low-performing contributors with moderate bias. Cluster 0 involves contributors with moderate performance, varying bias, and reputation scores.

6.5 Usability Assessment of the Crowdsourcing Platform

The crowdsourcing platform was evaluated for usability to ensure its user-friendliness for platform contributors. The subsequent sections of this thesis offer insights into the evaluation of the crowdsourcing platform's usability. This assessment aimed to guarantee that the application is user-friendly, particularly for participants actively contributing to the platform. After developing the crowdsourcing platform, a summative usability assessment was conducted to ascertain its ease of use for participants in a standard usability study.

Participants

Participants were 30 volunteers who contributed to the annotation process previously through the crowdsourcing platform as a part of their job submission and scored greater than 0.5 as the trustworthiness core, and 5 research assistants contributed to the research team and who used the crowdsourcing platform to get the datasets annotated. The thirty volunteers were selected from those who responded between the 13th of November 2023 and to 18th of November 2023. None of the participants were excluded from the data analysis. Out of the 35 participants, sixteen were female and nineteen

were male and the age ranged from 22 to 45. The questions used in measuring 10 aspects of usability using the System Usability Scale(SUS) are given in Appendix G. The 10 questions include positive and negative statements.

- Positive Statements: Questions 1, 3, 5, 7, and 9 are positively worded.
- Negative Statements (Reverse Coded): Questions 2, 4, 6, 8, and 10 are negatively worded.

The combination of positive and negative statements helps to reduce response bias and provides a more comprehensive assessment of the user's perception of usability.

The sum of the corresponding score calculated for the 10 questions is given in Table 6.6 sorted in ascending order. Appendix G lists the ten questions pertaining to the evaluation of usability along with their corresponding calculations.

The result of this study is given below.

The mean time took to complete the SUS: was 36 seconds

Standard deviation:10 seconds

Range: 20 to 52 seconds.

The overall SUS score calculated based on the sample of contributors is 65.2

Interpret the SUS scores based on standard benchmarks:

- Scores above 68 are considered above average.
- Scores above 80 are considered excellent.
- Scores below 68 may indicate usability issues that need attention.

The SUS score can be interpreted as a moderate level of perceived usability. The SUS score is normalized to fall within a range of 0 to 100, where higher scores indicate better usability. This suggests that users find the system usable to some extent but may have encountered challenges or aspects that could be improved. By considering this as the starting point for identifying areas of improvement in the system's user experience Table 6.7 was examined to identify the aspects to be improved.

TABLE 6.7 : SUS SCORE – USABILITY ASSESSMENT OF THE OVERALL SYSTEM

Question	Aspect Assessed	Total Corresponeded Score
Positive Statements		
Q5	Features are well integrated.	98
Q3	Easiness to use	96
Q7	Easiness to learn	92
Q9	How confident are the users to access the service?	91

Q1	Likeness to use this system frequently.	86
Negative Statements		
Q4	Need the help of a technical person to use this service	83
Q6	Inconsistency in this service.	88
Q2	Complexity	89
Q8	Cumbersome to use	94
Q10	Need prior knowledge to use	96

After analysing the above table it is evident that users have scored less marks for the likeness to use this system frequently. To address this, it is recommended to embed more intrinsic and extrinsic motivators to encourage users to visit the platform frequently.

Based on the task design of the annotation process the learning of the subject knowledge assessed by the contributors varied. Therefore, training was provided to the participants on how to use the platform. It is recommended to introduce an onboarding process that guides users through the system's features. Only the tutorials on how to identify hate content are included in the platform. Instead suggested offering tutorials, tooltips, or interactive guides on how to use the system for new users.

CHAPTER 7

DISCUSSION

Introduction

Unlike automated processes, crowdsourcing involves human decision-making. Humans bring subjectivity to their decisions, influencing the final responses or outcomes. Crowdsourcing platforms often involve a diverse group of contributors with varied backgrounds, experiences, and expertise. Due to this diversity, responses to tasks or assignments can be subjective, reflecting the individual characteristics of the contributors. The output or contributions generated by participants on a crowdsourcing platform are influenced by individual perspectives, opinions, or interpretations rather than being objectively determined. Variances in the skills and knowledge levels of contributors contribute to subjectivity. Subjectivity may arise when contributors inject their preferences or biases into the completion of tasks.

Subjective judgments are inherently more challenging to measure compared to objective metrics. Measuring trustworthiness becomes complex when dealing with a heterogeneous group of contributors who approach subjective judgments differently. Unlike objective tasks where a ground truth can be established, subjective judgments often lack a definitive correct answer. The absence of a clear ground truth complicates the assessment of the accuracy and trustworthiness of worker judgments. Workers in crowdsourcing platforms may have varying levels of expertise or knowledge on the subject matter. Determining trustworthiness requires accounting for differences in expertise and ensuring that workers are appropriately qualified for the tasks.

This thesis proposed a novel crowdsourcing platform, because of the unavailability of crowdsourcing platforms for annotating social media posts composed in Sinhala, as well as Sinhala words written using English characters with the participants of required contextual and linguistic proficiency, along with social and cultural insights in the most common crowdsourcing platforms such as Mechanical Turk and Crowd Flower.

There are no benchmarks to compare the trustworthiness of contributors with the same task designs with contributors with similar contributor profiles and social and cultural contexts because of this unavailability. Therefore to evaluate the crowdsourcing platform we used the accuracy of the annotation results by considering the following cases.

7.1 Contribution of the Research Papers

This section outlines how the following research papers contribute to the attainment of the thesis objectives.

A Novel Annotation Scheme to Generate Hate Speech Corpus through Crowdsourcing and Active Learning[149]

This research describes a unique annotation approach for crowdsourcing to generate a corpora including hate speech from social media data. The drawbacks identified after performing this research are listed below.

(1) Challenges arise in determining the potential harm and to annotate of a post as harmful or not solely by examining its content.

(2) Needs to identify any businesses with the contributor responses to eliminate the biases.

(3) How many annotators we can use to perform a task should be varied by considering the complexity and nature of the post and needs to have a mechanism to calculate the complexity of the task.

It is vital to adapt the tasks to incorporate pertinent visuals and any context-specific data with the post. This is because the users must read the initial post and the replies if any before responding. As a result, advocating preventive steps would be ineffective if only one comment was removed; instead, a set of comments would need to be erased.

It was observed that the selected contributors doing the job intuitively rather than analysing the tasks by considering the given criteria. Therefore a mechanism should be identified to eliminate this practice.

The approach employed for assessing trustworthiness

The trustworthiness of users was scored using two different techniques when selecting participants. The reliability estimation method employed was Krippendorff's alpha coefficient, and the number of contributors assigned to each task category was varied. Primary classifications were conducted by two contributors, and further in-depth analyses, including analysing the strength of the sentiment and identifying hate targets, should be expanded by assessing the reliability score.

To maintain unbiased it is required to select a sample of participants representing different religions in an equal percentage. Since the majority of the Sri Lankans are following Buddhism it is extremely challenging.

To mitigate religious bias, future research should consider the outcomes of this study and focus on the following aspects:

1. Preventive measures against hate speech:

Rather than focusing on individual comments or captions, future research should aim to identify multiple comments on Facebook posts for removal as a preventive measure against the dissemination of hate.

2. Measuring Beliefs and Bias:

Measuring bias and beliefs is crucial as it provides insights into the subjective perspectives and potential predispositions of individuals. This information is essential

for understanding the factors influencing decision-making, behaviour, and responses, allowing for a more comprehensive analysis of data and outcomes. Additionally, it aids in addressing and mitigating potential biases, promoting fairness, objectivity, and transparency in various processes, including research, decision-making, and the development of systems and models.

3. Ensuring conscious judgement:

Ensuring conscious judgment involves implementing measures to guarantee that individuals assessing or making judgments are doing so thoughtfully, deliberately, and with full awareness. This is crucial in various contexts, such as decision-making, evaluation, or analysis, to minimize the risk of impulsive or unconscious biases that might influence the outcomes. By fostering a conscious approach, it enhances the quality and fairness of judgments, contributing to more reliable and unbiased results.

4. Tailoring Questions to Worker Types:

Tailoring questions to worker types involves customizing queries or tasks based on the characteristics, skills, or expertise of different groups of workers. This strategy recognizes that individuals may have varied strengths, backgrounds, or proficiencies, and tailoring questions ensures that tasks are well-suited to the capabilities of specific worker types. This approach enhances the efficiency and effectiveness of crowdsourcing initiatives by optimizing task distribution and matching workers with tasks that align with their skill sets, ultimately improving the overall quality of contributions and results.

This research contributed toward deciding considering bias and belief-related scores to ensure the trustworthiness of crowd response other than considering Krippendorff alpha and Cohen kappa coefficient.

A comparative study of the characteristics of hate speech propagators and their behaviours over the Twitter social media platform[150]

This research focuses on studying the interaction between the users who spread hate and non-hate content. A corpus of 102,882 posts from 530 Twitter user profiles are categorized as hate and non-hate. This research explores the distinctive attributes of those who propagate hate speech and non-hate users, after analysing emotions, sentiments and the social network for linguistics.

It is observed that the hate users have higher counts of followers and following and exhibit restrained expression, limited geotagging, and infrequent account verification. Interaction with Twitter users engaging in hate speech is notably scarce, evidenced by fewer likes, retweets, and replies. This restrained engagement underscores the potential significance of audiences in actively countering and mitigating the impact of hate speech.

An analysis of sentiment across languages highlights a polarization of negative tweets toward Sinhala. English language users, in a synergistic effect, utilize a positive tone to disseminate harmful content.

The research contributes towards ranking and categorizing the hate users in the social media platforms and policy reforms. Moreover, the speech propagator characteristic and data annotation process are compared in this study. Annotation of the dataset was performed using 19 crowd participants with the tweets without performing any preprocessing.

The platform introduced in this thesis was employed to annotate the dataset, involving three contributors and employing a majority voting system. For each case, the Krippendorff alpha coefficient was calculated, allowing users with a score of at least 0.5 to contribute to ten posts.

Cohen's Kappa coefficient for each occurrence was checked and further assessed. To classify a post as reliable the Kappa value was considered if greater than 0.6. This methodology ensured a comprehensive and reliable annotation process for the raw tweets and comments.

This approach facilitates a comprehensive examination of the dataset by allowing annotators to provide varied perspectives and interpretations. Rigorous quality control procedures were implemented throughout the labelling process, enhancing the robustness of the research findings.

A Peer Recommendation Model to Avoid Hate Speech Engagements in Multiplex Social Networks [151]

Presented in this paper is a groundbreaking algorithm designed to establish connections for a new social peer in MSN with the objective of neutralizing hate discussions. The algorithm's key attribute lies in suggesting a peer possessing a sufficiently influential factor capable of countering hate speech effectively. The introduction of a new cluster that actively contributes to the elimination of hate speech is deemed highly significant, representing a noteworthy social innovation for any social media platform.

The crowdsourcing platform proposed in this thesis was used to identify the features to train the peer recommendation model by identifying a peer who does have a good enough influence factor to block the hate speech.

A User Experience Measuring Technique to Moderate Social Media Content Through Crowdsourcing [152]

This paper presents a method for soliciting user perspectives on the suitability of social media content using crowdsourcing. This innovative crowdsourcing model contributes to the process of moderating social media content, aiming to enhance the overall user experience. The paper discusses the designing instruments to capture user opinions.

The results of the research affirm the effectiveness of employing criteria to assess the appropriateness of media content shared on social platforms. UX designers can use this approach to perform different research with users.

The crowdsourcing methodology that is proposed is positioned to benefit designers seeking to gather responses specifically from their focused users, thereby enriching the user-centric design process.

Time Series-Based Trend Analysis for Hate Speech on Twitter During COVID-19 Pandemic[153]

During the COVID-19 outbreak in Sri Lanka, social media platforms played a significant role as a medium for information propagation. The impact of social media engagement trends was greatly influenced by the context of the posts. This research is primarily focused on understanding the propagation of hate speech on the Twitter platform during this critical period.

The paper presents the findings of a study that explores how content in the Sinhala language, including Sinhala words written using English text, was disseminated on Twitter. Additionally, the trend analysis techniques employed in identifying hate speech propagation trends over the observed period are discussed.

Notably, the study observed a rapid interaction with social media platforms during the initial posts identified, with a gradual decrease in interactions in the latter part. Finally, this abstract outlines a trend line depicting the identified hate speech posts on Twitter data, providing insights into the dynamics of hate speech propagation during the COVID-19 outbreak.

Conducted within a concise timeframe, this research aimed to discern hate speech trends in a limited span of Twitter posts and predict commenting behaviour through time series analysis. Data collection spanned from April 1, 2020, to May 15, 2020, focusing exclusively on Twitter posts in the Sinhala language.

The dataset, comprising posts containing both English language and Sinhala words written in English text, underwent preprocessing. This involved the removal of personal identifiers and outliers. The subsequent categorization classified the dataset into two main categories: posts or content by a user and replies.

To ascertain the nature of the posts, a crowdsourcing approach was employed, with manual comment analysis conducted by distributing the comments among 50 university students. Each post was further distributed among five students, facilitating a comprehensive labelling and classification process into hate, not hate, or neutral categories. This multi-step methodology contributed to a nuanced understanding of hate speech trends and user commenting behaviour on Twitter within the specified period.

Trust measurement method

The number of annotators was changed based on the difficulty of the manual annotation and the optimum number of annotators was decided for majority voting as 5.

Classification of Trending Videos on YouTube [154]

Building upon our prior research that introduced an innovative method for capturing user perspectives on social media content appropriateness through crowdsourcing, this current study adopts the same technique. However, in this instance, the approach is employed as a complementary method to pinpoint the crucial factors of trend analysis, aided by paid workers.

Long-Term Trend Analysis for Social Media Content Published During COVID-19 Pandemic [155]

A primary constraint in this research is the incapability of text representations to conduct polarity sentiment analysis, necessitating conversion into English. Unfortunately, there is minimal software support for polarity analysis in the Sinhala language, despite the existence of a few Singlish-to-Sinhala converters. Compounded by the fact that end users adopt varied text combinations, the indispensable role of the crowdsourcing platform becomes evident in the overall research framework.

Significance of the study

Based on the results of the trust evaluation, the proposed worker selection model can pick out reliable and capable workers.

7.2 Recommendations on maintaining trustworthiness

By considering the research findings the following list of recommendations are proposed to maintain the trustworthiness of workers.

1. To assess the consistency of agreement among the annotators, majority voting is used by researchers and Joni and Ahmed et al. recommend having data samples of small sizes with at least five annotators. However, according to our findings, it's evident that the number of participants should be varied by considering the trustworthiness score of the participants.
2. We recommend finding the balance between the number of redundant labels and the resources available for annotation to reduce the cost and time required for data annotation. The number of participants varies from 3 to 5 based on the contributor's expertise.
3. While benchmarks are valuable tools in assessing performance in objective tasks, their applicability diminishes when dealing with subjective outputs. Instead, the proposed methodology can be used as a robust method for quality control, training, and validation specific to the subjective nature of the tasks at hand.
4. Some workers may provide intentionally biased or dishonest subjective responses. A benchmark may not effectively detect or account for such intentional deviations. Therefore during the pre-processing stage, it is required to identify and remove the intentionally biased or dishonest subjective responses before calculating the trustworthiness score.
5. In an ideal scenario, we need to find participants who are not biased and we need to calculate the correlation between belief systems and bias, beliefs vs accuracy

etc. Since we cannot calculate the trustworthiness score we considered three scores.

6. The relationship between these scores is complex and context-dependent, and understanding the dynamics between reputation, performance, and bias is crucial for evaluation.

CHAPTER 8

CONCLUSION & FUTURE DEVELOPMENT

Introduction

A software platform is currently absent for annotating social media posts composed in Sinhala, as well as Sinhala words written using English characters. This is mainly because of the absence of workers with contextual and linguistic proficiency, along with social and cultural insights in the most common crowdsourcing platforms such as Mechanical Turk and Crowd Flower. This is a problem for the researchers who perform Natural Language Processing(NLP) research. Therefore this research seeks to identify mechanisms for both direct and indirect crowdsourcing to gather opinions on posts shared by Sri Lankan citizens across various topics on social media.

Having a quality control mechanism is crucial when implementing a crowdsourcing platform to serve this purpose. The method of achieving trustworthiness among crowdsourcing participants through a consensus-based approach is significantly influenced by the biases present within the crowd and the Hawthorne effect. These biases and the Hawthorne effect have a direct influence on the accuracy and performance of the implementation framework. The challenge lies in assessing the calibre of annotated datasets using the crowdsourcing method, particularly in gauging the trustworthiness of contributors.

8.1 Research contribution

The primary contribution of this study lies in the introduction of a novel approach aimed at ensuring the trustworthiness of crowd responses' annotation. This research advocates for the implementation of an algorithm designed to pre-select contributors (refer to Table 5.2). This selection process takes into account various criteria, including language proficiency, the ability to comprehend Sinhala words written in English, analytical skills, and domain-specific knowledge, particularly in the context of hate speech identification. Subsequently, a trustworthiness score (outlined in Section 5.3) is computed for each worker. This score is derived by evaluating past worker performance, a quality score determined by comparing the quality of work against established golden rules, a quality score based on consensus, and an information score provided by the workers.

Various established metrics, such as Cohen's Kappa, Krippendorff's Alpha, Fleiss' Kappa, and Gwe'ts AC2 Coefficient, have traditionally been employed by researchers to calculate the inter-rater reliability of annotation tasks. These methods assess reliability based on consensus. Notably, Krippendorff's Alpha is acknowledged as the most robust measure of inter-rater reliability, despite its computational complexity.

Considering this, the present research assesses the proposed methodology by adopting Krippendorff's Alpha as the preferred inter-rater agreement methodology. This evaluation is specifically applied to a consensus-based approach for workers whose scores exceed 0.6, providing a comprehensive examination of the reliability and trustworthiness of the annotation process.

The crowdsourcing platform proposed in this study has been utilized by researchers for annotating their datasets, as documented in references [149], [150], [153], [154] and [155]. The same approach can be further extended by the researchers to perform their NLP research.

Moreover, the annotation scheme put forth for hate speech identification and the corresponding hate speech corpora can serve as valuable resources for researchers in their ongoing NLP investigations. In addition to the annotated datasets mentioned above, researchers can also leverage annotated datasets of Sinhala and Singlish social media posts for their studies.

The next section provides a summary of the research performed throughout and the findings under each research question. This research tried to address the following three problems.

Research Problem 1 :

There is a lack of a software platform to annotate Sinhala and Sinhala words written using English letters by acquiring contextual and language proficiency along with cultural and social insights identifying mechanisms to use with direct and indirect crowdsourcing to collect opinions on the social media posts shared by Sri Lankans.

Research Problem 2 :

The consensus-based approach of ensuring the trustworthiness of crowdsourcing participants is highly affected by the crowd's biases and the Hawthorne effect. These biases and the Hawthorne effect impact the implementation framework's accuracy and performance.

Research Problem 3 :

The problem of evaluating the quality of the annotated datasets using the crowdsourcing approach for contributor trustworthiness.

This thesis made an effort to investigate solutions to the three problems by finding answers to the four distinct questions stated below.

Research Question 1 :

What are the techniques to implement in the crowdsourcing platform to pre-select contributors, contributor reward, contributor reputation management and moderate hate speech content?

Research Question 2 :

How to derive quantitative measurements for social media content analysis using crowd?

Research Question 3 :

How to perform the analysis of user responses to obtain meaningful insights?

Research Question 4 :

How to measure and ensure the trustworthiness of users in their responses?

This research's main objective was to determine an appropriate crowdsourcing mechanism to capture user inputs, thereby implementing a framework to moderate social media content by providing a solution to measure the trustworthiness of crowd response to ensure the quality of captured user inputs.

Specific Research Objectives of this research are restated below;

Design an analytical framework with identified techniques to moderate social media content using crowdsourcing.

1. Identify appropriate trust metrics to evaluate the reliability of the crowd response.
2. Implement a trust metric to enable trust modelling and reasoning about crowd trust.
3. Implement the crowdsourcing platform to facilitate inappropriate content identification with necessary quality control and analytical features.
4. Evaluate the results for quality, transparency, accuracy, and performance of the platform using the crowd.

The first key finding of this research is the software that allows the moderation process of social media content using crowdsourcing for Sinhala and Singlish. The second key finding is the annotated datasets of YouTube, Twitter and Facebook posts if hate speech content is present or not. Third the hate speech corpus generated by the crowd workers.

In addition to this software after analyzing the data collected several key findings emerged. The study could find that there is a positive correlation between worker reliability measures and their overall job performance. This would suggest that reliable workers tend to perform better in their roles. Response time and accuracy were comparatively high and the completion time was comparatively low for reliable workers. Comparative analysis of reliability measurement mechanisms revealed that embedding reputation-based trustworthiness measurement scores extends the quality of the worker responses along with the consensus-based trustworthiness and gold standard. The most significant outcome of this research is the models for worker behaviour and worker trustworthiness.

This research proposes three novel methodologies; first the crowdsourcing approach for social media content moderation, the second, a novel annotation scheme to identify hate speech and the third approach to measure the trustworthiness of the crowd responses. The original contribution of this research to the field is the crowdsourcing framework suggested for social media content moderation. The practical application of the implementation of the proposed crowdsourcing framework. A novel annotation

scheme to annotate social media posts. New insights would be necessary to consider reputation scores other than the consensus-based approach and gold standard approach.

The following section discusses how this research answered the research questions formulated at the outset of this study.

Research Question 1 :

What are the techniques to implement in the crowdsourcing platform to pre-select contributors, contributor reward, contributor reputation management and moderate hate speech content?

A criterion for the pre-selection of contributors was proposed in this research and implemented in the crowdsourcing platform and the details are given in Chapter 4, the questionnaires used in assessing different skills are given from Annexure A to Appendix D. A methodology for setting contributor rewards was proposed with the gamification and further explained in Chapter 4 under “Assignment of intrinsic and extrinsic rewards for chosen contributors”. Managing contributor reputation was done with the use of the proposed model for worker behaviour modelling and the details are given in Chapter 4. How the proposed crowdsourcing platform was the speech identification and annotation is explained under “Task Design” in Chapter 4. The findings section reveals the best practices and effective strategies for each of these aspects in crowdsourcing platforms.

Research Question 2 :

How to derive quantitative measurements for social media content analysis using crowd?

This research led to the development of quantitative metrics and measurements for analyzing social media content through crowd contributions. Findings include the quantitative measurements identified; *Number of tasks completed within a given period, Number of tasks attempted by each worker within a given period, Percentage of tasks completed compared to tasks attempted, Time taken to complete tasks, Accuracy of responses considering Golden Rules, Time taken to submit responses after task assignment and Consistency of response time provide* and these were measured during the data collection process and the analysis provided valuable insights into content trends, sentiment analysis, and user engagement.

Research Question 3 :

How to perform the analysis of user responses to obtain meaningful insights?

How the analysis of user responses was performed is explained in Chapter 3. During the preliminary study user interviews and observations were conducted to identify the user reactions to the social media posts. Based on the findings the functionalities were identified for the proposed crowdsourcing framework. The implemented crowdsourcing framework was used in collecting the data and the analysis was performed. A Likert scale was used in collecting the sentiments of the contributors for

the social media posts and users were given ternary choices to identify if a particular post contains hate content, or not or to ignore. The contributors were requested to choose the rules that justified their selections. Furthermore, the contributors were shown the posts with hate speech labels and asked to identify the words in which the hatred was exhibited and a corpus was generated. The worker behavioural model and the model to measure trustworthiness were evaluated after dividing the dataset into two parts: a training set and a testing (or validation) set. Accuracy, precision, recall, F1-score, mean squared error and root mean squared error were used as evaluation metrics.

Research Question 4 :

How to measure and ensure the trustworthiness of users in their responses?

It proposed criteria to calculate a reputation score accepted with the past workers' performance score, quality score based on a golden rule base and quality score based on consensus-based and the provided information score at the registration. This will potentially lead to better quality control in crowdsourcing platforms.

The thesis spans research domains encompassing user experience, computer-supported cooperative work, social networks, and text processing. Primarily, this research addresses the existing gap in assuring the quality of crowd responses within a collaborative workforce originating from diverse contexts, encompassing cultural insights and language usage.

The thesis introduces a novel framework that facilitates workers in capturing their experiences throughout the annotation process. Through this proposed framework, researchers in natural language processing gain the ability to integrate their annotation tasks and select preferred annotation techniques by specifying workflows. Initially, this framework empowers users to define annotation tasks within six distinct categories: identification of language and inappropriate content, image text recognition, detection of hate speech, identification of hate speech propagators, generation of hate corpora, and sentiment analysis.

The outcomes of this endeavour have the potential to enhance the annotation procedure, thereby extending its applications to areas such as machine learning, data mining, and natural language processing. Ultimately, this contributes to the creation of an improved cyberspace environment.

The positive practical implications of implementing a crowdsourcing platform in a country like Sri Lanka enhance the efficiency in distributing the moderation workload across a large number of contributors, allowing for faster review and response times to flagged content. Furthermore, crowdsourcing enables platforms to scale their content moderation efforts to handle the vast amount of user-generated content that is uploaded daily. By utilizing a distributed workforce, platforms can potentially reduce costs associated with hiring and maintaining a full-time moderation team. Contributors from different backgrounds and cultures can provide more diverse perspectives on what constitutes inappropriate content, leading to a more inclusive moderation

process. Crowdsourcing allows platforms to adapt quickly to changing trends and user behaviours by adjusting moderation guidelines and tasks. Platforms can implement quality control mechanisms to ensure consistency and accuracy in moderation decisions.

Research Publications

A list of published abstracts, extended abstracts, conference papers and journal papers published by the research team members and I are given below;

- Two Scopus-indexed journal papers with the research titles, “A Comparative Study of the Characteristics of Hate Speech Propagators and Their Behaviours over Twitter Social Media Platform” and “A Novel Annotation Scheme to Generate Hate Speech Corpus through Crowdsourcing and Active Learning” in Scopus-indexed Q1(Heliyon) & Q3(IJACSA) respectively.
- Four indexed international conference papers with the titles; “A Peer Recommendation Model to Avoid Hate Speech Engagements in Multiplex Social Networks”, “A User Experience Measuring Technique to Moderate Social Media Content Through Crowdsourcing”, “Classification of Trending Videos on YouTube”, “Long-Term Trend Analysis for Social Media Content Published During COVID-19 Pandemic” and “Time Series Based Trend Analysis for Hate Speech in Sri Lankan Social Media Platforms During COVID-19 Pandemic”.
- One abstract for a poster presentation that could obtain the best poster award for the research “Hate Speech Corpus Generation using Crowd”.

8.2 Research Limitations

This section explains the limitations of this research. One limitation of the study was not allowing the contributors to select the reason for opting out of completing the tasks. It could easily identify the reason for contributors to do that as after identifying which features have missing values it was observed that the majority of the workers have opted out of labelling the same set of tasks and it was chosen to drop rows with missing values. The following were not considered under the study when modelling the behaviour of workers.

- Interactions with other workers through comments, discussions, or messages.
- Types of tasks preferred by the worker.
- Patterns in task selection based on difficulty, payout, or topic.
- Types of tasks aligned with the contributor skills
- Frequency of logins and platform engagement.
- Ratings and feedback provided by the worker to other contributors or requesters.
- Patterns in the type of feedback given.

The initial research plan aimed to furnish all annotators with training through the facilitation of a workshop. This approach was intended to ensure that contributors were adequately informed about the identification of hate content. However, the unforeseen

lockdown measures necessitated by the COVID-19 pandemic rendered the in-person workshop ineffective. As a result, the decision was made to transition to an online format. Regrettably, the online workshop experienced a reduced level of participation.

8.3 Future Work

Future investigations should endeavour to identify multiple comments associated with Facebook posts rather than an isolated comment or caption. This approach seeks to proactively counteract the dissemination of hate speech. When designing annotation tasks more details about the post should be given. This extension ensures a comprehensive understanding when seeking responses from the crowd. Furthermore, to ensure the thorough assessment of comments in alignment with stipulated criteria, the implementation of a mechanism that prompts annotators to consistently evaluate content, avoiding intuitive responses, is vital. A comprehensive bias and belief analysis should be done. Subsequent investigations should incorporate a rigorous evaluation of biases and belief systems inherent in the contributing workforce. This scrutiny is pivotal in ensuring the credibility and reliability of crowd responses. The identification of distinct clusters of worker profiles permits consideration. In subsequent research, it is advisable to tailor various types of questions from different categories to each type of worker, thereby optimizing the specificity and relevance of inquiries.

Trustworthy crowd responses are pivotal for improving the quality and accuracy of data annotations, content moderation, and user-generated content analysis. This, in turn, enhances the reliability of insights drawn from these processes. Trustworthy responses contribute to sound decision-making processes by ensuring that the data used for analysis and decision-making are dependable and free from biases or misinformation. In contexts where data privacy regulations apply, trustworthy responses are essential to ensure compliance with laws and regulations. Implementing mechanisms to measure trustworthiness strengthens the reputation of the crowdsourcing platform, attracting more participants and stakeholders. Policymakers, businesses, and researchers can make informed decisions based on reliable data insights derived from trustworthy crowd responses.

In conclusion, implementing a crowdsourcing platform for content moderation can provide practical benefits in terms of efficiency, scalability, and diverse perspectives. However, platforms must also carefully address challenges related to quality control, bias, sensitive content exposure, and legal and ethical considerations to ensure a successful and responsible moderation process.

REFERENCES

- [1] 'Voramontri and Klieb - 2019 - Impact of Social Media on Consumer Behaviour.pdf'. Accessed: Feb. 12, 2020. [Online]. Available: https://www.researchgate.net/profile/Leslie_Klieb/publication/326098250_Impact_of_Social_Media_on_Consumer_Behaviour/links/5b38ccb8aca2720785feb863/Impact-of-Social-Media-on-Consumer-Behaviour.pdf
- [2] 'Facebook: global daily active users 2022', Statista. Accessed: Nov. 02, 2022. [Online]. Available: <https://www.statista.com/statistics/346167/facebook-global-dau/>
- [3] 'Digital 2022: Sri Lanka', DataReportal – Global Digital Insights. Accessed: Nov. 02, 2022. [Online]. Available: <https://datareportal.com/reports/digital-2022-sri-lanka>
- [4] 'Community Standards'. Accessed: Feb. 02, 2020. [Online]. Available: https://www.facebook.com/communitystandards/content_related_requests
- [5] 'Twitter's policy on hateful conduct | Twitter Help'. Accessed: Aug. 07, 2022. [Online]. Available: <https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy>
- [6] 'Hate speech policy - YouTube Help'. Accessed: Feb. 02, 2020. [Online]. Available: <https://support.google.com/youtube/answer/2801939?hl=en>
- [7] C. Curtis, 'Facebook's global content moderation fails to account for regional sensibilities', *The Next Web*. Accessed: Feb. 02, 2020. [Online]. Available: <https://thenextweb.com/socialmedia/2019/02/26/facebooks-global-content-moderation-fails-to-account-for-regional-sensibilities/>
- [8] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder, 'Hate speech detection: Challenges and solutions', *PloS One*, vol. 14, no. 8, p. e0221152, 2019.
- [9] I. Gagliardone, D. Gal, T. Alves, G. Martínez, and Unesco, *Countering online hate speech*. 2015.
- [10] A. Guterres, 'What is hate speech?', UNITED NATIONS STRATEGY AND PLAN OF ACTION ON HATE SPEECH, May 2019. Accessed: Feb. 04, 2020. [Online]. Available: https://www.un.org/en/genocideprevention/documents/advising-and-mobilizing/Action_plan_on_hate_speech_EN.pdf
- [11] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze, 'Annotating named entities in Twitter data with crowdsourcing', p. 9.
- [12] G. S. N. Meedin, H. M. M. Caldera, and I. Perera, 'A User Experience Measuring Technique to Moderate Social Media Content Through Crowdsourcing', in *2020 From Innovation to Impact (FITI)*, Colombo, Sri Lanka: IEEE, Dec. 2020, pp. 1–6. doi: 10.1109/FITI52050.2020.9424911.
- [13] D. Geiger, S. Seedorf, T. Schulze, R. C. Nickerson, and M. Schader, 'Managing the Crowd: Towards a Taxonomy of Crowdsourcing Processes', p. 12, 2011.

- [14] K. Hansson and T. Ludwig, ‘Crowd Dynamics: Conflicts, Contradictions, and Community in Crowdsourcing’, *Comput. Support. Coop. Work CSCW*, vol. 28, no. 5, pp. 791–794, Sep. 2019, doi: 10.1007/s10606-018-9343-z.
- [15] ‘Amazon Mechanical Turk’. Accessed: Nov. 12, 2022. [Online]. Available: <https://www.mturk.com/>
- [16] S. T. Roberts, ‘Commercial Content Moderation: Digital Laborers’ Dirty Work’, *Dirty Work*, p. 12.
- [17] Y. Gerrard and H. Thornham, ‘Content moderation: Social media’s sexist assemblages’, *New Media Soc.*, vol. 22, no. 7, pp. 1266–1286, Jul. 2020, doi: 10.1177/1461444820912540.
- [18] ‘Community Standards’. Accessed: Oct. 16, 2020. [Online]. Available: <https://www.facebook.com/communitystandards/>
- [19] ‘Hate speech policy - YouTube Help’. Accessed: Aug. 07, 2022. [Online]. Available: <https://support.google.com/youtube/answer/2801939?hl=en>
- [20] S. T. Roberts, ‘Content Moderation’, in *Encyclopedia of Big Data*, L. A. Schintler and C. L. McNeely, Eds., Cham: Springer International Publishing, 2017, pp. 1–4. doi: 10.1007/978-3-319-32001-4_44-1.
- [21] S. T. Roberts, *Content moderation*. 2017. Accessed: Oct. 16, 2020. [Online]. Available: <https://escholarship.org/uc/item/7371c1hf>
- [22] D. Faggella, ‘Crowdsourced Content Moderation – How it Works and What’s Possible’, *Emerj*. Accessed: Feb. 04, 2020. [Online]. Available: <https://emerj.com/partner-content/crowdsourced-content-moderation-how-it-works-and-whats-possible/>
- [23] P. Burnap and M. L. Williams, ‘Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making: Machine Classification of Cyber Hate Speech’, *Policy Internet*, vol. 7, no. 2, pp. 223–242, Jun. 2015, doi: 10.1002/poi3.85.
- [24] T. Davidson, D. Warmesley, M. Macy, and I. Weber, ‘Automated Hate Speech Detection and the Problem of Offensive Language’, *ArXiv170304009 Cs*, Mar. 2017, Accessed: May 31, 2021. [Online]. Available: <http://arxiv.org/abs/1703.04009>
- [25] Y. Ibrahim, ‘Facebook and the Napalm Girl: Reframing the Iconic as Pornographic’, *Soc. Media Soc.*, vol. 3, p. 205630511774314, Oct. 2017, doi: 10.1177/2056305117743140.
- [26] G. Koushik, K. Rajeswari, and S. K. Muthusamy, ‘Automated Hate Speech Detection on Twitter’, in *2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, Pune, India: IEEE, Sep. 2019, pp. 1–4. doi: 10.1109/ICCUBEA47591.2019.9128428.
- [27] A. Veglis, ‘Moderation Techniques for Social Media Content’, in *Social Computing and Social Media*, vol. 8531, G. Meiselwitz, Ed., in Lecture Notes in Computer Science, vol. 8531. , Cham: Springer International Publishing, 2014, pp. 137–148. doi: 10.1007/978-3-319-07632-4_13.

- [28] L. Silva, M. Mondal, D. Correa, F. Benevenuto, and I. Weber, ‘Analyzing the Targets of Hate in Online Social Media’, *ArXiv160307709 Cs*, Mar. 2016, Accessed: Feb. 04, 2020. [Online]. Available: <http://arxiv.org/abs/1603.07709>
- [29] K. Relia, M. Akbari, D. Duncan, and R. Chunara, ‘Socio-spatial Self-organizing Maps: Using Social Media to Assess Relevant Geographies for Exposure to Social Processes’, *Proc. ACM Hum.-Comput. Interact.*, vol. 2, no. CSCW, pp. 1–23, Nov. 2018, doi: 10.1145/3274414.
- [30] J. C. Pereira-Kohatsu, L. Quijano-Sánchez, F. Liberatore, and M. Camacho-Collados, ‘Detecting and Monitoring Hate Speech in Twitter’, *Sensors*, vol. 19, no. 21, p. 4654, Oct. 2019, doi: 10.3390/s19214654.
- [31] P. Burnap and M. L. Williams, ‘Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making: Machine Classification of Cyber Hate Speech’, *Policy Internet*, vol. 7, no. 2, pp. 223–242, Jun. 2015, doi: 10.1002/poi3.85.
- [32] S. Agrawal and A. Awekar, ‘Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms’, in *Advances in Information Retrieval*, vol. 10772, G. Pasi, B. Piwowarski, L. Azzopardi, and A. Hanbury, Eds., in Lecture Notes in Computer Science, vol. 10772. , Cham: Springer International Publishing, 2018, pp. 141–153. doi: 10.1007/978-3-319-76941-7_11.
- [33] A. Olteanu, C. Castillo, F. Diaz, and E. Kıcıman, ‘Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries’, *Front. Big Data*, vol. 2, 2019, Accessed: May 31, 2022. [Online]. Available: <https://www.frontiersin.org/article/10.3389/fdata.2019.00013>
- [34] G. Li, J. Wang, Y. Zheng, and M. J. Franklin, ‘Crowdsourced Data Management: A Survey’, p. 23.
- [35] R. Gupta, ‘Using Social Media for Global Security’, p. 458.
- [36] Z. Waseem and D. Hovy, ‘Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter’, in *Proceedings of the NAACL Student Research Workshop*, San Diego, California: Association for Computational Linguistics, 2016, pp. 88–93. doi: 10.18653/v1/N16-2013.
- [37] O. de Gibert, N. Perez, A. García-Pablos, and M. Cuadros, ‘Hate Speech Dataset from a White Supremacy Forum’, in *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 11–20. doi: 10.18653/v1/W18-5102.
- [38] J. Salminen, M. Hopf, S. A. Chowdhury, S. Jung, H. Almerexhi, and B. J. Jansen, ‘Developing an online hate classifier for multiple social media platforms’, *Hum.-Centric Comput. Inf. Sci.*, vol. 10, no. 1, p. 1, Dec. 2020, doi: 10.1186/s13673-019-0205-6.
- [39] J. Howe, *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*, 0 edition. New York: Currency, 2009.
- [40] M. Dontcheva, ‘Crowdsourcing and Creativity’, p. 4.
- [41] S. T. Roberts, ‘Content Moderation’, in *Encyclopedia of Big Data*, L. A. Schintler and C. L. McNeely, Eds., Cham: Springer International Publishing, 2017, pp. 1–4. doi: 10.1007/978-3-319-32001-4_44-1.

- [42] B. E. Tidball and P. J. Stappers, ‘Crowdsourcing Contextual User Insights for UCD’, p. 4.
- [43] J. Rogstadius, V. Kostakos, A. Kittur, B. Smus, J. Laredo, and M. Vukovic, ‘An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets’, p. 8.
- [44] K. Relia, M. Akbari, D. Duncan, and R. Chunara, ‘Socio-spatial Self-organizing Maps: Using Social Media to Assess Relevant Geographies for Exposure to Social Processes’, *Proc. ACM Hum.-Comput. Interact.*, vol. 2, no. CSCW, pp. 1–23, Nov. 2018, doi: 10.1145/3274414.
- [45] M. Dontcheva, ‘Crowdsourcing and Creativity’, p. 4.
- [46] P. Gundecha and H. Liu, ‘Mining Social Media: A Brief Introduction’, in *2012 TutORials in Operations Research*, INFORMS, 2012, pp. 1–17. doi: 10.1287/educ.1120.0105.
- [47] N. Eagle, ‘txteagle: Mobile Crowdsourcing’, in *Internationalization, Design and Global Development*, vol. 5623, N. Aykin, Ed., in *Lecture Notes in Computer Science*, vol. 5623. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 447–456. doi: 10.1007/978-3-642-02767-3_50.
- [48] N. Kaufmann, T. Schulze, and D. Veit, ‘More than fun and money. Worker Motivation in Crowdsourcing – A Study on Mechanical Turk’, 2011.
- [49] G. Barbier, R. Zafarani, H. Gao, G. Fung, and H. Liu, ‘Maximizing benefits from crowdsourced data’, *Comput. Math. Organ. Theory*, vol. 18, no. 3, pp. 257–279, Sep. 2012, doi: 10.1007/s10588-012-9121-2.
- [50] G. Li, J. Wang, Y. Zheng, and M. J. Franklin, ‘Crowdsourced Data Management: A Survey’, p. 23.
- [51] K. Hansson and T. Ludwig, ‘Crowd Dynamics: Conflicts, Contradictions, and Community in Crowdsourcing’, *Comput. Support. Coop. Work CSCW*, vol. 28, no. 5, pp. 791–794, Sep. 2019, doi: 10.1007/s10606-018-9343-z.
- [52] R. Gupta, ‘Using Social Media for Global Security’, p. 458.
- [53] ‘All UX evaluation methods « All About UX’. Accessed: Feb. 12, 2020. [Online]. Available: <https://www.allaboutux.org/all-methods>
- [54] B. Das, A. M. Seelye, B. L. Thomas, D. J. Cook, L. B. Holder, and M. Schmitter-Edgecombe, ‘Using smart phones for context-aware prompting in smart environments’, in *2012 IEEE Consumer Communications and Networking Conference (CCNC)*, Las Vegas, NV, USA: IEEE, Jan. 2012, pp. 399–403. doi: 10.1109/CCNC.2012.6181023.
- [55] G. D. Abowd, A. K. Dey, P. J. Brown, N. Davies, M. Smith, and P. Steggle, ‘Towards a Better Understanding of Context and Context-Awareness’, in *Handheld and Ubiquitous Computing*, vol. 1707, H.-W. Gellersen, Ed., in *Lecture Notes in Computer Science*, vol. 1707. , Berlin, Heidelberg: Springer Berlin Heidelberg, 1999, pp. 304–307. doi: 10.1007/3-540-48157-5_29.
- [56] D. Faggella, ‘Crowdsourced Content Moderation – How it Works and What’s Possible’, *Emerj*. Accessed: Feb. 02, 2020. [Online]. Available: <https://emerj.com/partner-content/crowdsourced-content-moderation-how-it-works-and-whats-possible/>

- [57] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, ‘A Survey on Bias and Fairness in Machine Learning’, *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–35, Jul. 2021, doi: 10.1145/3457607.
- [58] K. Abhinav, ‘Trustworthiness in Crowdsourcing’, p. 61.
- [59] ‘About Scopus - Abstract and citation database | Elsevier’. Accessed: May 09, 2023. [Online]. Available: <https://www.elsevier.com/solutions/scopus>
- [60] M. J. Page *et al.*, ‘The PRISMA 2020 statement: an updated guideline for reporting systematic reviews’, *BMJ*, p. n71, Mar. 2021, doi: 10.1136/bmj.n71.
- [61] A. Abusafia and A. Bouguettaya, ‘Reliability Model for Incentive-Driven IoT Energy Services’. arXiv, Jan. 19, 2021. Accessed: Jun. 11, 2022. [Online]. Available: <http://arxiv.org/abs/2011.06159>
- [62] I. Martin-Morato and A. Mesaros, ‘What is the ground truth? Reliability of multi-annotator data for audio tagging’. arXiv, Apr. 09, 2021. Accessed: Jun. 11, 2022. [Online]. Available: <http://arxiv.org/abs/2104.04214>
- [63] C. Castelfranchi and R. Falcone, ‘Trust Is Much More than Subjective Probability: Mental Components and Sources of Trust’, *Rd Hawaii Int. Conf. Syst. Sci.*, p. 10, 2000.
- [64] H. D. Schlinger, ‘Theory of Mind: An Overview and Behavioral Perspective’, *Psychol. Rec.*, vol. 59, no. 3, pp. 435–448, Jul. 2009, doi: 10.1007/BF03395673.
- [65] M. Allahbakhsh, A. Ignjatovic, B. Benatallah, S.-M.-R. Beheshti, E. Bertino, and N. Foo, ‘Reputation management in crowdsourcing systems’, in *8th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom)*, Oct. 2012, pp. 664–671. doi: 10.4108/icst.collaboratecom.2012.250499.
- [66] N. Quoc Viet Hung, N. T. Tam, L. N. Tran, and K. Aberer, ‘An Evaluation of Aggregation Techniques in Crowdsourcing’, in *Web Information Systems Engineering – WISE 2013*, vol. 8181, X. Lin, Y. Manolopoulos, D. Srivastava, and G. Huang, Eds., in *Lecture Notes in Computer Science*, vol. 8181, Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 1–15. doi: 10.1007/978-3-642-41154-0_1.
- [67] P. G. Ipeirotis, F. Provost, and J. Wang, ‘Quality management on Amazon Mechanical Turk’, in *Proceedings of the ACM SIGKDD Workshop on Human Computation*, Washington DC: ACM, Jul. 2010, pp. 64–67. doi: 10.1145/1837885.1837906.
- [68] J. Whitehill, T. Wu, J. Bergsma, J. R. Movellan, and P. L. Ruvolo, ‘Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise’.
- [69] V. C. Raykar *et al.*, ‘Supervised Learning from Multiple Experts: Whom to trust when everyone lies a bit’.
- [70] ‘Bi et al. - Iterative Learning for Reliable Crowdsourcing Syst.pdf’.
- [71] L. I. Kuncheva, C. J. Whitaker, C. A. Shipp, and R. P. W. Duin, ‘Limits on the majority vote accuracy in classifier fusion’, *Pattern Anal. Appl.*, vol. 6, no. 1, pp. 22–31, Apr. 2003, doi: 10.1007/s10044-002-0173-7.

- [72] K. Lee, J. Caverlee, and S. Webb, ‘The social honeypot project: protecting online communities from spammers’, in *Proceedings of the 19th international conference on World wide web*, Raleigh North Carolina USA: ACM, Apr. 2010, pp. 1139–1140. doi: 10.1145/1772690.1772843.
- [73] J. Ross, L. Irani, M. S. Silberman, A. Zaldivar, and B. Tomlinson, ‘Who are the crowdworkers?: shifting demographics in mechanical turk’, in *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, Atlanta Georgia USA: ACM, Apr. 2010, pp. 2863–2872. doi: 10.1145/1753846.1753873.
- [74] G. Kazai, J. Kamps, and N. Milic-Frayling, ‘Worker types and personality traits in crowdsourcing relevance labels’, in *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11*, Glasgow, Scotland, UK: ACM Press, 2011, p. 1941. doi: 10.1145/2063576.2063860.
- [75] B. Carterette and I. Soboroff, ‘The effect of assessor error on IR system evaluation’, in *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '10*, Geneva, Switzerland: ACM Press, 2010, p. 539. doi: 10.1145/1835449.1835540.
- [76] J. Vuurens, A. P. de Vries, and C. Eickhoff, ‘How Much Spam Can You Take? An Analysis of Crowdsourcing Results to Increase Accuracy’, p. 8.
- [77] U. Gadiraju, R. Kawase, S. Dietze, and G. Demartini, ‘Understanding Malicious Behavior in Crowdsourcing Platforms: The Case of Online Surveys’, in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, Seoul Republic of Korea: ACM, Apr. 2015, pp. 1631–1640. doi: 10.1145/2702123.2702443.
- [78] D. R. Smith and W. E. Snell, ‘Goldberg’s bipolar measure of the Big-Five personality dimensions: reliability and validity’, *Eur. J. Personal.*, vol. 10, no. 4, pp. 283–299, Nov. 1996, doi: 10.1002/(SICI)1099-0984(199611)10:4<283::AID-PER264>3.0.CO;2-W.
- [79] A. Cocos, T. Qian, C. Callison-Burch, and A. J. Masino, ‘Crowd control: Effectively utilizing unscreened crowd workers for biomedical data annotation’, *J. Biomed. Inform.*, vol. 69, pp. 86–92, May 2017, doi: 10.1016/j.jbi.2017.04.003.
- [80] Y. Nehme, P. L. Callet, F. Dupont, J.-P. Farrugia, and G. Lavoue, ‘Exploring Crowdsourcing for Subjective Quality Assessment of 3D Graphics’, in *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*, Tampere, Finland: IEEE, Oct. 2021, pp. 1–6. doi: 10.1109/MMSP53017.2021.9733634.
- [81] J. Wang, T. Kraska, M. J. Franklin, and J. Feng, ‘CrowdER: crowdsourcing entity resolution’, *Proc. VLDB Endow.*, vol. 5, no. 11, pp. 1483–1494, Jul. 2012, doi: 10.14778/2350229.2350263.
- [82] J. Fan, M. Lu, B. C. Ooi, W.-C. Tan, and M. Zhang, ‘A hybrid machine-crowdsourcing system for matching web tables’, in *2014 IEEE 30th International Conference on Data Engineering*, Chicago, IL, USA: IEEE, Mar. 2014, pp. 976–987. doi: 10.1109/ICDE.2014.6816716.
- [83] ‘Aggregation Methods’, Aggregation Methods. Accessed: May 30, 2022. [Online]. Available: <https://toloka.ai/knowledgebase/aggregation/>

- [84] G. Hermosillovaladez, C. Florin, L. Bogoni, L. Moy, and N. Org, ‘VIKAS.RAYKAR@SIEMENS.COM SHIPENG.YU@SIEMENS.COM’.
- [85] T. Tian, J. Zhu, and Y. Qiaoben, ‘Max-Margin Majority Voting for Learning from Crowds’, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2480–2494, Oct. 2019, doi: 10.1109/TPAMI.2018.2860987.
- [86] K. Sakaguchi, R. L. Bras, C. Bhagavatula, and Y. Choi, ‘WinoGrande: an adversarial winograd schema challenge at scale’, *Commun. ACM*, vol. 64, no. 9, pp. 99–106, Sep. 2021, doi: 10.1145/3474381.
- [87] F. K. Khattak and A. Salleb-Aouissi, ‘Quality Control of Crowd Labeling through Expert Evaluation’.
- [88] E. Estellés-Arolas and F. González-Ladrón-de-Guevara, ‘Towards an integrated crowdsourcing definition’, *J. Inf. Sci.*, vol. 38, no. 2, pp. 189–200, Apr. 2012, doi: 10.1177/0165551512437638.
- [89] J. Zhang, V. S. Sheng, Q. Li, J. Wu, and X. Wu, ‘Consensus algorithms for biased labeling in crowdsourcing’, *Inf. Sci.*, vol. 382–383, pp. 254–273, Mar. 2017, doi: 10.1016/j.ins.2016.12.026.
- [90] G. Demartini, D. E. Difallah, and P. Cudré-Mauroux, ‘ZenCrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking’, in *Proceedings of the 21st international conference on World Wide Web*, Lyon France: ACM, Apr. 2012, pp. 469–478. doi: 10.1145/2187836.2187900.
- [91] A. P. Dawid and A. M. Skene, ‘Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm’, *Appl. Stat.*, vol. 28, no. 1, p. 20, 1979, doi: 10.2307/2346806.
- [92] P. Smyth, U. M. Fayyad, M. C. Burl, P. Perona, and P. Baldi, ‘Inferring Ground Truth from Subjective Labelling of Venus Images’, p. 8.
- [93] P. Welinder, S. Branson, P. Perona, and S. J. Belongie, ‘The Multidimensional Wisdom of Crowds’, p. 9.
- [94] A. Ghosh, S. Kale, and P. McAfee, ‘Who moderates the moderators?: crowdsourcing abuse detection in user-generated content’, in *Proceedings of the 12th ACM conference on Electronic commerce*, San Jose California USA: ACM, Jun. 2011, pp. 167–176. doi: 10.1145/1993574.1993599.
- [95] N. Dalvi, A. Dasgupta, R. Kumar, and V. Rastogi, ‘Aggregating crowdsourced binary ratings’, in *Proceedings of the 22nd international conference on World Wide Web*, Rio de Janeiro Brazil: ACM, May 2013, pp. 285–294. doi: 10.1145/2488388.2488414.
- [96] I. Arora, J. Guo, S. I. Levitan, S. McGregor, and J. Hirschberg, ‘A Novel Methodology for Developing Automatic Harassment Classifiers for Twitter’, in *Proceedings of the Fourth Workshop on Online Abuse and Harms*, Online: Association for Computational Linguistics, 2020, pp. 7–15. doi: 10.18653/v1/2020.alw-1.2.
- [97] D. R. Karger, S. Oh, and D. Shah, ‘Iterative Learning for Reliable Crowdsourcing Systems’.

- [98] Y. Du, Y.-E. Sun, H. Huang, L. Huang, H. Xu, and X. Wu, ‘Quality-aware online task assignment mechanisms using latent topic model’, *Theor. Comput. Sci.*, vol. 803, pp. 130–143, Jan. 2020, doi: 10.1016/j.tcs.2019.07.033.
- [99] D. Tao, J. Cheng, Z. Yu, K. Yue, and L. Wang, ‘Domain-Weighted Majority Voting for Crowdsourcing’, *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 1, pp. 163–174, Jan. 2019, doi: 10.1109/TNNLS.2018.2836969.
- [100] J. Zhang and X. Wu, ‘Multi-Label Truth Inference for Crowdsourcing Using Mixture Models’, *IEEE Trans. Knowl. Data Eng.*, pp. 1–1, 2019, doi: 10.1109/TKDE.2019.2951668.
- [101] X. S. Lu, M. Zhou, H. Liu, and L. Qi, ‘A Comparative Study on Two Ground Truth Inference Algorithms based on Manually Labeled Social Media Data’, in *2019 IEEE 16th International Conference on Networking, Sensing and Control (ICNSC)*, Banff, AB, Canada: IEEE, May 2019, pp. 436–441. doi: 10.1109/ICNSC.2019.8743287.
- [102] G. Dawson and R. Polikar, ‘OpinionRank: Extracting Ground Truth Labels from Unreliable Expert Opinions with Graph-Based Spectral Ranking’, arXiv, arXiv:2102.05884, May 2021. Accessed: May 30, 2022. [Online]. Available: <http://arxiv.org/abs/2102.05884>
- [103] J. Zhang, V. S. Sheng, J. Wu, and X. Wu, ‘Multi-Class Ground Truth Inference in Crowdsourcing with Clustering’, *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 4, pp. 1080–1085, Apr. 2016, doi: 10.1109/TKDE.2015.2504974.
- [104] J. Carletta, ‘Squibs and Discussions: Assessing Agreement on Classification Tasks: The Kappa Statistic’, *Comput. Linguist.*, vol. 22, no. 2, p. 6.
- [105] B. Carpenter, ‘Multilevel Bayesian Models of Categorical Data Annotation’, p. 52.
- [106] A. Sheshadri and M. Lease, ‘SQUARE: A Benchmark for Research on Computing Crowd Consensus’, p. 9.
- [107] A. Uma, D. Almanea, and M. Poesio, ‘Scaling and Disagreements: Bias, Noise, and Ambiguity’, *Front. Artif. Intell.*, vol. 5, p. 818451, Apr. 2022, doi: 10.3389/frai.2022.818451.
- [108] J. Nassar, V. Pavon-Harr, M. Bosch, and I. McCulloh, ‘Assessing Data Quality of Annotations with Krippendorff Alpha For Applications in Computer Vision’. arXiv, Dec. 20, 2019. Accessed: Aug. 03, 2022. [Online]. Available: <http://arxiv.org/abs/1912.10107>
- [109] T. Spinde, D. Krieger, M. Plank, and B. Gipp, ‘Towards A Reliable Ground-Truth For Biased Language Detection’, arXiv, arXiv:2112.07421, Dec. 2021. Accessed: May 30, 2022. [Online]. Available: <http://arxiv.org/abs/2112.07421>
- [110] M. Schaekermann, J. Goh, K. Larson, and E. Law, ‘Resolvable vs. Irresolvable Disagreement: A Study on Worker Deliberation in Crowd Work’, *Proc. ACM Hum.-Comput. Interact.*, vol. 2, no. CSCW, pp. 1–19, Nov. 2018, doi: 10.1145/3274423.
- [111] R. Drapeau, L. B. Chilton, J. Bragg, and D. S. Weld, ‘MicroTalk: Using Argumentation to Improve Crowdsourcing Accuracy’, p. 10.

- [112] D. Gurari and K. Grauman, ‘CrowdVerge: Predicting If People Will Agree on the Answer to a Visual Question’, in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, Denver Colorado USA: ACM, May 2017, pp. 3511–3522. doi: 10.1145/3025453.3025781.
- [113] J. L. Mumpower and T. R. Stewart, ‘Expert Judgement and Expert Disagreement’, *Think. Reason.*, vol. 2, no. 2–3, pp. 191–212, Jul. 1996, doi: 10.1080/135467896394500.
- [114] J. C. Chang, S. Amershi, and E. Kamar, ‘Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets’, in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, Denver Colorado USA: ACM, May 2017, pp. 2334–2346. doi: 10.1145/3025453.3026044.
- [115] Z. Yan, G. Li, and J. Liu, ‘Private rank aggregation under local differential privacy’, *Int. J. Intell. Syst.*, vol. 35, no. 10, pp. 1492–1519, Oct. 2020, doi: 10.1002/int.22261.
- [116] J. Lu, W. Li, Q. Wang, and Y. Zhang, ‘Research on Data Quality Control of Crowdsourcing Annotation: A Survey’, p. 8.
- [117] Y. Li, H. Sun, and W. H. Wang, ‘Towards Fair Truth Discovery from Biased Crowdsourced Answers’, in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Virtual Event CA USA: ACM, Aug. 2020, pp. 599–607. doi: 10.1145/3394486.3403102.
- [118] D. Kahneman, ‘A perspective on judgment and choice: Mapping bounded rationality.’, *Am. Psychol.*, vol. 58, no. 9, pp. 697–720, 2003, doi: 10.1037/0003-066X.58.9.697.
- [119] I. Cho, R. Wesslen, A. Karduni, S. Santhanam, S. Shaikh, and W. Dou, ‘The Anchoring Effect in Decision-Making with Visual Analytics’, in *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, Phoenix, AZ: IEEE, Oct. 2017, pp. 116–126. doi: 10.1109/VAST.2017.8585665.
- [120] T. Draws, A. Rieger, O. Inel, U. Gadiraju, and N. Tintarev, ‘A Checklist to Combat Cognitive Biases in Crowdsourcing’, p. 12.
- [121] C. Hube, B. Fetahu, and U. Gadiraju, ‘Understanding and Mitigating Worker Biases in the Crowdsourced Collection of Subjective Judgments’, in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow Scotland UK: ACM, May 2019, pp. 1–12. doi: 10.1145/3290605.3300637.
- [122] A. Balayn, C. Lofi, and G.-J. Houben, ‘Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems’, *VLDB J.*, vol. 30, no. 5, pp. 739–768, Sep. 2021, doi: 10.1007/s00778-021-00671-8.
- [123] X. Duan, C.-J. Ho, and M. Yin, ‘The Influences of Task Design on Crowdsourced Judgement: A Case Study of Recidivism Risk Evaluation’, in *Proceedings of the ACM Web Conference 2022*, Virtual Event, Lyon France: ACM, Apr. 2022, pp. 1685–1696. doi: 10.1145/3485447.3512239.
- [124] A. Vercammen, A. Marcoci, and M. Burgman, ‘Pre-screening workers to overcome bias amplification in online labour markets’, *PLOS ONE*, vol. 16, no. 3, p. e0249051, Mar. 2021, doi: 10.1371/journal.pone.0249051.

- [125] E. I. Hoppe and D. J. Kusterer, ‘Behavioral biases and cognitive reflection’, *Econ. Lett.*, vol. 110, no. 2, pp. 97–100, Feb. 2011, doi: 10.1016/j.econlet.2010.11.015.
- [126] W. B. Bilker, J. A. Hansen, C. M. Brensinger, J. Richard, R. E. Gur, and R. C. Gur, ‘Development of Abbreviated Nine-Item Forms of the Raven’s Standard Progressive Matrices Test’, *Assessment*, vol. 19, no. 3, pp. 354–369, Sep. 2012, doi: 10.1177/1073191112446655.
- [127] M. E. Toplak, R. F. West, and K. E. Stanovich, ‘The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks’, *Mem. Cognit.*, vol. 39, no. 7, pp. 1275–1289, Oct. 2011, doi: 10.3758/s13421-011-0104-1.
- [128] H. Markovits and G. Nantel, ‘The belief-bias effect in the production and evaluation of logical conclusions’, *Mem. Cognit.*, vol. 17, no. 1, pp. 11–17, Jan. 1989, doi: 10.3758/BF03199552.
- [129] S. S. Tekiroğlu, Y.-L. Chung, and M. Guerini, ‘Generating Counter Narratives against Online Hate Speech: Data and Strategies’, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online: Association for Computational Linguistics, 2020, pp. 1177–1190. doi: 10.18653/v1/2020.acl-main.110.
- [130] R. Mukherjee, H. C. Peruri, U. Vishnu, P. Goyal, S. Bhattacharya, and N. Ganguly, ‘Read what you need: Controllable Aspect-based Opinion Summarization of Tourist Reviews’, in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Virtual Event China: ACM, Jul. 2020, pp. 1825–1828. doi: 10.1145/3397271.3401269.
- [131] V. Keswani, M. Lease, and K. Kenthapadi, ‘Towards Unbiased and Accurate Deferral to Multiple Experts’, p. 12, 2021.
- [132] H. Zhuang and J. Young, ‘Leveraging In-Batch Annotation Bias for Crowdsourced Active Learning’, in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, Shanghai China: ACM, Feb. 2015, pp. 243–252. doi: 10.1145/2684822.2685301.
- [133] P. Pettit, ‘The Cunning of Trust’, *Philos. Htmlemt Glyphamp Asciiamp Public Aff.*, vol. 24, no. 3, pp. 202–225, Jul. 1995, doi: 10.1111/j.1088-4963.1995.tb00029.x.
- [134] C. Huang, H. Yu, J. Huang, and R. A. Berry, ‘Strategic Information Revelation in Crowdsourcing Systems Without Verification’, arXiv, arXiv:2104.03487, Apr. 2021. Accessed: May 30, 2022. [Online]. Available: <http://arxiv.org/abs/2104.03487>
- [135] M. Venanzi, A. Rogers, and N. R. Jennings, ‘Trust-Based Fusion of Untrustworthy Information in Crowdsourcing Applications’, p. 8.
- [136] F. Leal, B. Malheiro, H. González-Vélez, and J. C. Burguillo, ‘Trust-based Modelling of Multi-criteria Crowdsourced Data’, *Data Sci. Eng.*, vol. 2, no. 3, pp. 199–209, Sep. 2017, doi: 10.1007/s41019-017-0045-1.
- [137] J. Dong, K. Yang, and Y. Shi, ‘Ranking from Crowdsourced Pairwise Comparisons via Smoothed Riemannian Optimization’, *ACM Trans. Knowl. Discov. Data*, vol. 14, no. 2, pp. 1–26, Mar. 2020, doi: 10.1145/3372407.

- [138] R. Heckel, M. Simchowitz, K. Ramchandran, and M. J. Wainwright, ‘Approximate Ranking from Pairwise Comparisons’. arXiv, Jan. 04, 2018. Accessed: Jun. 27, 2022. [Online]. Available: <http://arxiv.org/abs/1801.01253>
- [139] J. C. S. J. Junior, A. Lapedriza, C. Palmero, X. Baro, and S. Escalera, ‘Person Perception Biases Exposed: Revisiting the First Impressions Dataset’, in *2021 IEEE Winter Conference on Applications of Computer Vision Workshops (WACVW)*, Waikola, HI, USA: IEEE, Jan. 2021, pp. 13–21. doi: 10.1109/WACVW52041.2021.00006.
- [140] N. B. Shah, S. Balakrishnan, A. Guntuboyina, and M. J. Wainwright, ‘Stochastically Transitive Models for Pairwise Comparisons: Statistical and Computational Issues’. arXiv, Sep. 27, 2016. Accessed: Jun. 27, 2022. [Online]. Available: <http://arxiv.org/abs/1510.05610>
- [141] Xiaoxin Yin, Jiawei Han, and P. S. Yu, ‘Truth Discovery with Multiple Conflicting Information Providers on the Web’, *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 6, pp. 796–808, Jun. 2008, doi: 10.1109/TKDE.2007.190745.
- [142] S. Lyu, W. Ouyang, Y. Wang, H. Shen, and X. Cheng, ‘Truth Discovery by Claim and Source Embedding’, *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 3, pp. 1264–1275, Mar. 2021, doi: 10.1109/TKDE.2019.2936189.
- [143] N. Sabetpour, A. Kulkarni, S. Xie, and Q. Li, ‘Truth Discovery in Sequence Labels from Crowds’, in *2021 IEEE International Conference on Data Mining (ICDM)*, Auckland, New Zealand: IEEE, Dec. 2021, pp. 539–548. doi: 10.1109/ICDM51629.2021.00065.
- [144] B. Zhao, B. I. P. Rubinstein, J. Gemmell, and J. Han, ‘A Bayesian approach to discovering truth from conflicting sources for data integration’, *Proc. VLDB Endow.*, vol. 5, no. 6, pp. 550–561, Feb. 2012, doi: 10.14778/2168651.2168656.
- [145] X. Fang *et al.*, ‘Selecting Workers Wisely for Crowdsourcing When Copiers and Domain Experts Co-exist’, *Future Internet*, vol. 14, no. 2, p. 37, Jan. 2022, doi: 10.3390/fi14020037.
- [146] D. Wang, J. Ren, Z. Wang, X. Pang, Y. Zhang, and X. S. Shen, ‘Privacy-preserving Streaming Truth Discovery in Crowdsourcing with Differential Privacy’, *IEEE Trans. Mob. Comput.*, pp. 1–1, 2021, doi: 10.1109/TMC.2021.3062775.
- [147] A. Alamsyah, N. Dudija, and S. Widiyanesti, ‘New Approach of Measuring Human Personality Traits Using Ontology-Based Model from Social Media Data’, *Information*, vol. 12, no. 10, p. 413, Oct. 2021, doi: 10.3390/info12100413.
- [148] ‘Crowdsource by Google’. Accessed: Nov. 21, 2022. [Online]. Available: <https://crowdsource.google.com/>
- [149] N. Meedin, M. Caldera, S. Perera, and I. Perera, ‘A Novel Annotation Scheme to Generate Hate Speech Corpus through Crowdsourcing and Active Learning’, *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 11, 2022, doi: 10.14569/IJACSA.2022.0131146.
- [150] S. Perera, N. Meedin, M. Caldera, I. Perera, and S. Ahangama, ‘A comparative study of the characteristics of hate speech propagators and their behaviours over

Twitter social media platform’, *Heliyon*, vol. 9, no. 8, p. e19097, Aug. 2023, doi: 10.1016/j.heliyon.2023.e19097.

- [151] H. M. M. Caldera, G. S. N. Meedin, and I. Perera, ‘A Peer Recommendation Model to Avoid Hate Speech Engagements in Multiplex Social Networks’, in *2020 From Innovation to Impact (FITI)*, Colombo, Sri Lanka: IEEE, Dec. 2020, pp. 1–6. doi: 10.1109/FITI52050.2020.9424905.
- [152] G. S. N. Meedin, H. M. M. Caldera, and I. Perera, ‘A User Experience Measuring Technique to Moderate Social Media Contents through Crowdsourcing’.
- [153] H. M. M. Caldera, G. S. N. Meedin, and I. Perera, ‘Time Series Based Trend Analysis for Hate Speech in Twitter During COVID 19 Pandemic’, in *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*, Colombo, Sri Lanka: IEEE, Nov. 2020, pp. 1–2. doi: 10.1109/ICTer51097.2020.9325491.
- [154] H. M. M. Caldera, S. Perera, G. S. N. Meedin, and I. Perera, ‘Classification of Trending Videos in YouTube’, in *2021 From Innovation To Impact (FITI)*, Colombo, Sri Lanka: IEEE, Dec. 2021, pp. 1–6. doi: 10.1109/FITI54902.2021.9833039.
- [155] H. M. M. Caldera, N. Meedin, S. Perera, and I. Perera, ‘Long-Term Trend Analysis for Social Media Content Published During COVID-19 Pandemic’, in *2022 2nd International Conference on Advanced Research in Computing (ICARC)*, Belihuloya, Sri Lanka: IEEE, Feb. 2022, pp. 108–113. doi: 10.1109/ICARC54489.2022.9753776.

APPENDIX A

QUESTIONNAIRE TO ASSESS THE KNOWLEDGE ON SINHALA LANGUAGE

[1]. ඔවුන් දෙදෙනා වැඩ කළේ දිවෙන් දිව ගා ගෙනය. වාක්‍යයෙහි තද කළු අකුරෙන් මුද්‍රිත කොටසේ අර්ථයට වඩාත් සමීප අර්ථය දෙන උත්තරය තෝරන්න.

- a) ඥාතීන් මෙන් ය
- b) හැම විටම එක තැන සිටිය
- c) ඉතා මිත්‍රශීලී ව ය.
- d) දිවි එකිනෙක ස්පර්ශ කරමින් ය.

[2]. පාලකයන් සමග කටයුතු කළ යුත්තේ දැලි පිහියෙන් කිරි කන්නා සේ ය. වාක්‍යයෙහි තද කළු අකුරෙන් මුද්‍රිත කොටසේ අර්ථයට වඩාත් සමීප අර්ථය දෙන උත්තරය තෝරන්න.

- a) ඉතා පරිස්සමිනි.
- b) කපෙනවාට බියෙනි.
- c) දුරින් සිටිමිනි.
- d) ඇඟලුම්කම් පාමිනි.

[3]. අක්කාගේ විවාහය ප්‍රමාද වූයේ අම්මාගේ වංසේ කබලේ ගැ නිසාය. වාක්‍යයෙහි තද කළු අකුරෙන් මුද්‍රිත කොටසේ අර්ථයට වඩාත් සමීප අර්ථය දෙන උත්තරය තෝරන්න.

- a) තම පරම්පරාව ගැන කියූ නිසා
- b) තමා ගැන පුරසාරම් කියූ නිසා
- c) මනාල පාර්ශවයේ අඩුපාඩු කියූ නිසා
- d) තම පරම්පරාව ගැන පුරසාරම් කියූ නිසා

[4]. වර්තමාන ශිෂ්‍ය පරම්පරාව වෙනිත් කෙමෙන් කෙමෙන් ආත්ම ඉවත් වනු පෙනෙයි. හිස්තැනට වඩාත් සුදුසු පදය තෝරන්න.

- a) අධීක්ෂණය
- b) ශික්ෂණය
- c) සමීක්ෂණය
- d) හිංසනය

[5]. විවාද කණ්ඩායමේ නායකයා තම කරුණු ඔප්පු කළේය.හිස්තැනට වඩාත් සුදුසු පදය තෝරන්න.

- a) තර්කානුකූලව
- b) නිත්‍යානුකූලව
- c) ධර්මානුකූලව
- d) ආගමානුකූලව

[6]. අන්ධ විශ්වාසවල එල්ල සිටින්නෝ තම පාවා දෙති.හිස්තැනට වඩාත් සුදුසු පදය තෝරන්න.

- a) ආත්මය
- b) දෛවය
- c) බුද්ධිය
- d) ධනය

[7]. ප්‍රත්‍යූපකාර කරන්නෝ අප සමාජයේ විරල ය. වාක්‍යයෙහි තද කළ අකුරෙන් මුද්‍රිත කොටසේ අර්ථය නිවැරදිව දැක්වෙන වරණය තෝරන්න.

- a) බෙහෙවින් උපකාර
- b) පරෝපකාර
- c) නැවත උපකාර
- d) නොකඩවා උපකාර

[8]. සමහරු වැඩිහිටියන් කෙරෙහි සානුකම්පික වෙති. වාක්‍යයෙහි තද කළ අකුරෙන් මුද්‍රිත කොටසේ අර්ථය නිවැරදිව දැක්වෙන වරණය තෝරන්න.

- a) අනුකම්පා සහගත
- b) අනුකම්පා විරහිත
- c) අනුකම්පා හරිත
- d) අනුකම්පා පූර්වක

[9]. අපි රට ගොඩනැගීමට එකාබද්ධව කටයුතු කරමු. වාක්‍යයෙහි තද කළ අකුරෙන් මුද්‍රිත කොටසේ අර්ථය නිවැරදිව දැක්වෙන වරණය තෝරන්න.

- a) එකා මෙන් එකට බැඳී
- b) එකට ඇලී
- c) එකට බැඳී
- d) එකාවන් ව නැගී සිට

[10]. දී ඇති අදහසට වඩාත් ම ගැලපෙන ප්‍රස්තාව පිරුළ අන්තර්ගත වරණය තෝරන්න.

"නිෂ්ඵල ක්‍රියාවක නිරත වීම"

- a) අබරා සිල් ගත්තා වාගේ
- b) පරංගියා කෝට්ටේ ගියා වගේ
- c) අටුවා කඩා පුටුව හදන්න වාගේ
- d) අන්දරේ සීනි කෑවා වාගේ

[11]. දී ඇති අදහසට වඩාත් ම ගැලපෙන ප්‍රස්තාව පිරුළ අන්තර්ගත වරණය තෝරන්න.

"වඩාත් උඩඟු වී ක්‍රියා කිරීම"

- a) පාළු ගෙයි වළන් බිඳිනවා වගේ
- b) අශ්වයාට අං ලැබුණා වාගේ
- c) අහේ ඉඳන් කන කනවා වාගේ
- d) නැගපු ඉණිමගට පයින් ගහනවා වාගේ

APPENDIX B

QUESTIONNAIRE TO ASSESS THE KNOWLEDGE ON SINGLISH READING

Select the correct Sinhala word/phrase written in English letters.

ඉංග්‍රීසි අකුරින් ලියා ඇති නිවැරදි සිංහල වචනය/වාක්‍ය බණ්ඩය තෝරන්න.

[1]. Tawa language ekak danagattata paadu wenne naha

- a) තව language එකක් දැනගත්තට පාඩු වෙන්නෙ නැහැ .
- b) තව එකක් දැන පාඩු නැහැ
- c) ඔයාගේ දුරකතනයෙන් පාඩු වෙන්නේ නැහැ

[2]. karunakarala mata sanasillay inna denna

- a) කරුණාකරලා මට සැනසිල්ලේ ඉන්න දෙන්න
- b) කරුණාකරලා සැනසිල්ලේ දෙන්න
- c) කරුණාවෙන් සවන් දෙන්න

[3]. oya math ekka poddak natanna kamathida?

- a) ඔයා මාත් එක්ක පොඩ්ඩක් නටන්න කැමතීද
- b) ඔයා නටන්න කැමතීද
- c) ඔයා මාත් එක්ක එනවද

[4]. oba sinhala kathā karanavadha?

- a) ඔබ සිංහල කතා කරනවද?
- b) ඔබ කතා කරනවා
- c) ඔයා සිංහල කතා කරනවද?

[5]. mata obava therum ganna baha

- a) මට ඔබව තේරුම් ගන්න බැහැ
- b) මට තේරුම් ගනු බෑ
- c) මට තේරුම් ගන්න බහ ඔබව

[6]. sadarayen piligannawa

- a) සාදරයෙන් පිළිගන්නවා
- b) සාදර පළිගන්නවා
- c) ඔබව මම සාදරයෙන් පිළිගන්නවා

[7]. Anna araya huu kiyanawa

- a) අන්න අරයා හු කියනවා
- b) අන්න හු කියන
- c) අරයට හු කියන්න

[8]. Rata hadana sawbhaagyaye dakma

- a) රට හදන සවිභාගයේ දැක්ම
- b) රට සවිභාගය දැක්ම
- c) රට හදන්න අත්‍යවශ්‍යයි

[9]. Uthsawa sabhawa amatamin janadhipathi apekshaka uthsawa sabhawa amathiya

- a) උත්සව සභාව අමතමින් ජනාධිපති අපේක්ෂක උත්සව සභාව අමතයි
- b) උත්සවය අමතා ජනාධිපති අමතයි
- c) උත්සව සභාව අමටන්න අවශ්‍යයි

[10]. Lankawe anka eke wyapaarikaya

- a) ලංකාවේ අංක එකේ ව්‍යාපාරිකයා
- b) ලංක ව්‍යාපාරික
- c) අංක එකේ ව්‍යාපාරික ප්‍රජාව

APPENDIX C

SAMPLE QUESTIONNAIRE TO ASSESCOMPREHENSION & ANALYTICAL SKILLS (SINHALA)

Assess the knowledge on Hate Speech(වෛරී කථනය පිළිබඳ දැනුම තක්සේරු කරන්න)

These questions are also mentioned in the English language for the convenience of the service providers (Contributors) who have proficiency in the English language.

Some of the excerpts from the social media posts may be emotional and provocative.

ඉංග්‍රීසි භාෂාවෙහි නිපුණත්වයක් ඇති සේවා සපයන්නන් (Contributors)ගේ පහසුව සඳහා ඉංග්‍රීසි භාෂාවෙන්ද මෙම ප්‍රශ්න සඳහන් කර ඇත.

සමාජ මාධ්‍ය දැන්වීම් වලින් උපුටා ගන්නා ලද කරුණු සමහරක් ආවේගශීලී හා ප්‍රකෝපකාරී විය හැකිය.

Read to learn / කියවා ඉගෙන ගන්න

<https://docs.google.com/document/d/1NL5ztVoH6-5SvkRWCGglFFiewOM-If4iUXwAB9MGOzE/edit?usp=sharing>

How familiar are you with hate speech and related issues? To find out, answer our quiz.

වෛරී ප්‍රකාශය සහ ඒ ආශ්‍රිත ගැටලු පිළිබඳව ඔබ කෙතරම් හුරුපුරුදුද? සොයා ගැනීමට, අපගේ ප්‍රශ්නාවලිය ට පිළිතුරු ලබා දෙන්න.

[1]. What is true of hate speech, as it is commonly understood?

සාමාන්‍යයෙන් තේරුම් ගත හැකි පරිදි වෛරී ප්‍රකාශ සම්බන්ධ සත්‍ය ප්‍රකාශය කුමක්ද?

- It is discriminatory and/or pejorative towards an individual, or a group.
- It refers to real, purported or imputed identity factors, such as religion, ethnicity, nationality, race, colour, descent, and gender.
- It is communicated through various means, encompassing expressions such as symbols, art objects, memes, cartoons, images, gestures etc.
- All of the above

- රූප, කාටූන්, Memes , කලා වස්තු, අභිනයන් සහ සංකේත - සහ/හෝ මාධ්‍ය ඇතුළුව ඕනෑම ආකාරයක ප්‍රකාශනයක් හරහා එය ප්‍රකාශ කෙරේ.
- එය පුද්ගලයෙකුට හෝ කණ්ඩායමකට වෙනස් කොට සැලකීමක් සහ/හෝ අපහාසාත්මක වේ.
- එය ආගම, වාර්ගිකත්වය, ජාතිකත්වය, ජාතිය, වර්ණය, සම්භවය, ස්ත්‍රී පුරුෂ භාවය වැනි සැබෑ, අරමුණු කරගත් හෝ ආරෝපණය කරන ලද අන්‍යන්‍ය සාධක වෙත යොමු කරයි.
- ඉහත සියල්ලම

[2]. According to the United Nations definition, what “identity factor(s)” must hate speech refer to in a discriminatory and/or pejorative way?

එක්සත් ජාතීන්ගේ නිර්වචනයට අනුව, වෛරී කථනය වෙනස් කොට සැලකීමේ සහ/හෝ නින්දිත ආකාරයෙන් සඳහන් කළ යුතු “අන්‍යායනා සාධක(ය)” මොනවාද?

- a) Solely pertaining to nationality, religion, race, ethnicity, descent, color, and/or gender.
- b) At least 2 of the following: ethnicity, colour, religion, , nationality, race, gender, and/or descent
- c) Any characteristics conveying identity in a broad sense, such as religion, ethnicity, nationality, race, colour, descent, gender, but also language, economic or social origin, disability, health status, or sexual orientation and any other identity factors
- d) None - speech need not reference any identity factor to be considered hateful.

- a) ආගම, වාර්ගිකත්වය, ජාතිකත්වය, ජාතිය, වර්ණය, සම්භවය, සහ/හෝ ස්ත්‍රී පුරුෂ භාවය පමණි.
- b) පහත සඳහන් අවම වශයෙන් 2: ආගම, වාර්ගිකත්වය, ජාතිකත්වය, ජාතිය, වර්ණය, සම්භවය, සහ/හෝ ස්ත්‍රී පුරුෂ භාවය
- c) ආගම, වාර්ගිකත්වය, ජාතිකත්වය, ජාතිය, වර්ණය, සම්භවය, ස්ත්‍රී පුරුෂ භාවය, නමුත් භාෂාව, ආර්ථික හෝ සමාජ සම්භවය, ආබාධිත තත්වය, සෞඛ්‍ය තත්වය හෝ ලිංගික දිශානතිය සහ වෙනත් අන්‍යායනා සාධක වැනි පුළුල් අර්ථයකින් අන්‍යායනාව ප්‍රකාශ කරන ඕනෑම ලක්ෂණයක්
- d) කිසිවක් නැත - කථනය ද්වේශ සහගත ලෙස සැලකීමට කිසිදු අන්‍යායනා සාධකයක් සඳහන් කිරීම අවශ්‍ය නොවේ.

[3]. According to the United Nations, who can be targeted by hate speech?

එක්සත් ජාතීන්ගේ සංවිධානයට අනුව, වෛරී ප්‍රකාශ මගින් ඉලක්ක කළ හැක්කේ කාටද?

- a) States and their offices and symbols, public officials, religious leaders, or tenets of faith.
- b) Individuals or groups of individuals based on who they are.
- c) Minority groups only.
- d) All of the above.

- a) රාජ්‍යයන් සහ ඒවායේ කාර්යාල සහ සංකේත, රාජ්‍ය නිලධාරීන්, ආගමික නායකයන්, හෝ පුද්ගලයෙකු හෝ, බොහෝ විට, පුද්ගල කණ්ඩායමක් විසින් ගරු කරනු ලබන මූලධර්මයක් හෝ විශ්වාසයක්.
- b) ඔවුන් කවුරුන්ද යන්න මත පදනම්ව පුද්ගලයන් හෝ කණ්ඩායම්.
- c) සුළු ජාතික කණ්ඩායම් පමණයි.
- d) ඉහත සියල්ලම.

[4]. Online hate speech can contribute to causing real harm.

පරිගණකයකින් පාලනය වන හෝ සම්බන්ධ කර සිදු කරන වෛරී කථනය සැබෑ හානියක් කිරීමට දායක විය හැක.

- a) True
- b) False
- a) නිවැරදි
- b) වැරදි

[5]. What are some of the ways to tackle hate speech recommended by the Strategy and Plan of Action on Hate Speech?

වෛරී කථනය පිළිබඳ උපාය මාර්ගය සහ ක්‍රියාකාරී සැලැස්ම මගින් නිර්දේශ කර ඇති වෛරී ප්‍රකාශය මැඩලීමේ ක්‍රම මොනවාද?

- a) Engage and support the victims.
- b) Tackle the underlying reasons, catalysts, and contributors to hate speech.
- c) Track and assess hate speech.
- d) Use education as a preventive tool.
- e) Engage with all relevant actors and the media.
- f) All of the above.
- a) වෛරී කථනය නිරීක්ෂණය කිරීම සහ විශ්ලේෂණය කිරීම.
- b) වෛරී ප්‍රකාශයේ මූල හේතු, ගෙන යන්නන් සහ ක්‍රියාකාරීන් අමතන්න.
- c) වින්දිතයින් හෝ ගොදුරු වූ පුද්ගලයන් සමඟ සම්බන්ධ වී සහය වන්න.
- d) වැළැක්වීමේ මෙවලමක් ලෙස අධ්‍යාපනය භාවිතා කරන්න.
- e) අදාළ සියලුම ක්‍රියා කරන්නන් සහ මාධ්‍ය සමඟ සම්බන්ධ වන්න.
- f) ඉහත සියල්ලම.

[6]. Fighting hate speech is the responsibility of:

වෛරී කථනයට එරෙහිව සටන් කිරීම වගකීම වන්නේ:

- a) Government
- b) Targets of hate speech
- c) The United Nations
- d) Social media platforms
- e) All of us
- a) ආණ්ඩුව
- b) වෛරී ප්‍රකාශයේ ඉලක්ක
- c) එක්සත් ජාතීන්ගේ සංවිධානය
- d) සමාජ මාධ්‍ය වේදිකා
- e) අපි හැමෝම

[7]. Abusive or indecent comments on social media are hate speech even if they are not targeted at a person or organization.

- a) Yes
- b) No

c) May be

අපවාදාත්මක හෝ අයික්ෂිත යෙදීම් සඳහන් සමාජ මාධ්‍යයෙහි පලවෙන ප්‍රකාශ පුද්ගලයකු හෝ ආයතනයක් ඉලක්ක කොට නොගත්තද වෛරී ප්‍රකාශයක් වේ.

- a) ඔව්
- b) නැත
- c) විය හැක

[8].Hate groups describe "the other" in ways that emphasize difference, making them seem alien and even subhuman. This is often done through caricatures or name-tape associations or ideologies. In some cases, hate groups will claim that others are not human.

Indicate whether you agree with this statement.

වෛරී කණ්ඩායම් "අනෙකා" විස්තර කරන්නේ ඔවින් ආගන්තුකයන් මෙන්ම අමනුෂ්‍යයන් ලෙස පවා පෙනෙන්නට සලස්වමින් වෙනස්කම් අවධාරණය වන ආකාරවලට බවයි. මෙය බොහෝ විට කෙරෙන්නේ විකට චිත්‍ර හෝ නම් පට බැඳීම් හෝ මතවාද මගිනි. සමහර අවස්තාවලදී වෛරී කණ්ඩායම් අනෙකුත් සැබවින්ම මිනිසුන් නොවේ යැයි ප්‍රකාශ කරනු ඇත.

ඔබ මෙම ප්‍රකාශයට එකඟ වේදැයි සඳහන් කරන්න.

- a) I agree
- b) I don't agree
- a) එකඟ වේ
- b) එකඟ නොවේ

[9].Following are some of the social media posts. This is not hate speech;

පහත සඳහන් වන්නේ සමාජ මාධ්‍ය ජාලාවෙහි පල කිරීම් කිහිපයකි. මින් වෛරී ප්‍රකාශයක් නොවන්නේ;

- a) ඩලස් නැමති කුණු අපුල්ලන්නාට, නැගෙනහිර හම්බයොන්ගේ දරුවන්ට අන්තවාදය හැදෑරීම විශ්ව විද්‍යාල හැදීමට මුදල්වලට කැදර වී මුලිකවූ තමුන්නාන්සේ සිංහල දරුවෙකු මරා දැමූ විට ඒක සාධාරණය කරන අපුරුව. ලැප්පයි මහත්තයෝ තමුන්නාන්සේ ගැන. කටට පණුවෝ ගහනවා ඕයි මේ කියන කතාවලට.
- b) රජය මේ ඉස්ලාම් පිළිලය වැඩි වීමට පෙර නීති සකස් කල යුතුය. මේ රටේ මහ ජාතියට වින කරන කක්කුසියක් පවා ටික දිනකින් පල්ලි කර ගන්නා ඉස්ලාම් අන්තවාදීන් මර්දනය කළ යුතුය. නැතහොත් අනාගතයේ මීට වඩා දරුණු ගැටුම් ඇති විය හැකිය.
- c) බංගලිදේශ හම්බයෙක් හාමුදුරුවන්ට ගහන හැටි ඡේනවද? අපේ සමහර පොන්නයෝ කියනවා හැම ආගමකින්ම කියන්නේ හොඳක් කියලා.
- d) රටේ පවතින විවිධ ජාතීන්ට ප්‍රමුඛ තාවය දීමේ වැඩ පිළිවලට විරුද්ධ තාවය දැක්වීම සඳහා නව නීති වල අවශ්‍යතාව පෙන්වා දීම සඳහා රටෙහි සමස්ත ජනතාව පෙළ ගැසිය යුතුය.

[10]. Which of the following factors affect the detection of hate speech?

වෛරී ප්‍රකාශ හඳුනා ගැනීම සඳහා පහත සඳහන් කුමන කරුණු බලපායිද?

- a) The social class in which they live

- b) His own beliefs
- c) Their core principles
- d) own bias
- e) All of the above
- a) තමන් ජීවත්වන සමාජ ස්ථරය
- b) තමාගේ විශ්වාසයන්
- c) තමන් ගේ හර පද්දතීන්
- d) තමාගේ පක්ෂග්‍රාහීත්වය
- e) ඉහත සියල්ල

APPENDIX D

SAMPLE QUESTIONNAIRE TO ASSESS THE ABILITY TO READ SINGLISH

මෙම ප්‍රශ්නය වාචික සහ නිගාමී තර්ක(deductive reasoning) කුසලතා පරීක්ෂාවකි:

[1]. උපාධි අපේක්ෂකයින් හත් දෙනෙකුගෙන් යුත් කණ්ඩායමකින් (A, B, C, D, E, F, සහ G) හතර දෙනෙකු ශිෂ්‍ය සංගමයට තෝරා ගනු ලැබේ. ඒ සඳහා පහත සඳහන් කොන්දේසි සපුරාලිය යුතුය:

A හෝ B තෝරාගත යුතු නමුත් A සහ B දෙදෙනාම තෝරාගත නොහැක.

E හෝ F යන දෙදෙනාගෙන් කෙනෙකු තෝරාගත යුතුය, නමුත් E සහ F දෙදෙනාම තෝරාගත නොහැක.

C තෝරන්නේ නැත්නම් E තෝරාගත නොහැක.

B තෝරන්නේ නැත්නම් G තෝරාගත නොහැක.

ශිෂ්‍ය සංගමයට F තෝරාගෙන නැති බව අපි දන්නේ නම්, ඉහත නිර්ණායක අනුගමනය කරමින් හතර දෙනෙකුගෙන් යුත් විවිධ කණ්ඩායම් කීයක් සෑදිය හැකිද?

- a) එකක්
- b) දෙකක්
- c) තුනක්
- d) හතරක්
- e) පහක්

This inquiry assesses verbal and deductive reasoning abilities.

Out of a cohort of seven undergraduate students (A, B, C, D, E, F, and G), four individuals will be chosen to present to the students' union. The selection must adhere to the following conditions:

1. Either E or F should be chosen, but not both E and F simultaneously.
2. Either A or B should be chosen, but not both A and B simultaneously.

- 3. G cannot be selected unless B is selected.
- 4. E cannot be selected unless C is selected.

If it's established that F is not chosen for the presentation, how many distinct sets of four individuals can be formed while adhering to the given criteria?

- a) Five
- b) Four
- c) Three
- d) Two
- e) One

[2]. This question assesses numerical (or non-verbal) reasoning by evaluating the ability to identify pattern rules and predict the next element in the sequence

මෙම ප්‍රශ්නය රටා රීති(pattern rules) හඳුනා ගැනීමට සහ ඊළඟට එන දේ පුරෝකථනය කිරීමට සංඛ්‍යාත්මක (හෝ වාචික නොවන) තර්කනයේ පරීක්ෂණයකි.

Examine the images in the top row. Determine the next box in the sequence.

- a) E
- b) D
- c) C
- d) B
- e) A

පින්තූරවල ඉහළ පේළිය දෙස බලන්න. අනුපිළිවෙලින් ඊළඟට එන්නේ කුමන කොටුවද?

- a) A
- b) B
- c) C
- d) D
- e) E

[3]. This question is a test of Non-verbal reasoning question.

මෙම ප්‍රශ්නය වාචික නොවන තර්කනයේ (Non-verbal reasoning) පරීක්ෂණයකි.

ඊළඟට එන අංකය කුමක්ද?

What number comes next?

9, 15, 13, 19, 17, 23...

- a) 18
- b) 29
- c) 19
- d) 20
- e) 21

[4]. What's the next figure in the pattern?

රටාවේ ඊළඟ රූපය කුමක්ද?

- a) A
- b) B
- c) C
- d) D
- e) E

[5]. George, Emmeli, Diyani & and Angela sit in this order in a row from left to right. Janet changes places with Eric, and then Eric changes places with Martin. Who is at the right end of the row?

- a) George
- b) Emmeli
- c) Diyani
- d) Angela

ජෝර්ජ්, එමලී, දියනි සහ ඇන්ජලා වමේ සිට දකුණට අනුපිළිවෙලෙනි ජේලියට වාඩි වී සිටිති. ජැනට් එරික් සමඟ ස්ථාන වෙනස් කරයි, පසුව එරික් මාටින් සමඟ ස්ථාන වෙනස් කරයි. ජේලියේ දකුණු කෙළවරේ සිටින්නේ කවුද?

- අ) ජෝර්ජ්
- ආ) එමලී
- ඇ) දියනි
- ඈ) ඇන්ජලා

[6]. A group of friends lives in a house divided into one flat per floor. Janith is in the flat below Anuradhi & and Samanalee is in the flat above Sara. Sara is in the flat below Janith, and Anuradhi lives with Roger. Peter lives on the top floor. Who is in the bottom flat?

- a) Janith
- b) Anuradhi
- c) Samanalee
- d) Sara
- e) Roger

යහළුවන් පිරිසක් මහල් නිවසක විවිධ තට්ටු වල ජීවත් වෙයි. ජනිත් අනුරාධිට පහල තට්ටුවේද, සමනලී සාරාට උඩ තට්ටුවේද ජීවත් වෙයි.. සාරා ඉන්නේ ජනිත්ට පහල තට්ටුවේය. අනුරාධි ජීවත් වෙන්නේ රොජර් එක්කය. පීටර් ඉහළම මහලේ ජීවත් වේ. පහළම තට්ටුවේ ජීවත් වන්නේ කවුද?

- අ) ජනිත්
- ආ) අනුරාධි
- ඇ) සමනලී
- ඈ) සාරා
- ඉ) රොජර්

[7]. Five cars have a race. The Honda beat the Mitsubishi but couldn't overtake the Nissan. The Mini Cooper failed to overtake the Audi but beat the Nissan. Which car came last?

- a. Honda

- b. Mitsubishi
- c. Nissan
- d. Mini Cooper
- e. Audi

මෝටර් රථ පහක් තරඟයකට ඉදිරිපත් වෙයි. Honda රථය Mitsubishi එක පසු කෙරුවද Nissan රථය පසු කිරීමට නොහැකි විය. Mini Cooper රථය Audi රථය අහිඛවා යාමට අසමත් වූ නමුත් Nissan රථය පරදවයි. අවසානයට ආපු රථය කුමක්ද ?

Honda

Mitsubishi

Nissan

Mini Cooper

Audi

[8]. පහත කවිය කියවා පිළිතුරු සපයන්න
 පොත් කියවන සෑම දෙනා
 නව දැනුමෙන් පිබිදෙනා
 අලුත් අලුත් දේ තනා
 ලොව බැබළෙයි තරු මෙනා

මෙම කවියේ සමස්ත අදහසින් පැවසෙන්නේ;

- a. පොත් කියවීමේ අගය පිළිබඳවය
- b. තරු ලොවේ බැබළීම පිළිබඳවය
- c. අලුත් අලුත් දේ නිර්මාණය පිළිබඳවය

[9]. පහත කවිය කියවා පිළිතුරු සපයන්න

අඳුරෙහි නැති - එළියෙහි ඇති

මටම හිතැති - මගේම සකි

මෙයින් කියවෙන්නේ;

- a. හිතමිතුරා ගැනය
- b. සෙවනැල්ල ගැනය
- c. හිරු ගැනය

[10]. දින පහම පාසල පැවැත්වූ සතියක එක ළඟ දින තුනක්ම සමන්ට පාසල් පැමිණීමට නොහැකි විය. මෙම ප්‍රකාශය අනුව සමන් නිසැකවම පාසල් නොපැමිණි දවස වනුයේ;

බදාදාය

බ්‍රහස්පතින්දාය

සිකුරාදාය

APPENDIX E

IMPLEMENTATION OF THE PRE-SELECTION OF CONTRIBUTORS

```
class RuleBasedSystem:
    def __init__(self):
        self.rules = []

    def add_rule(self, condition, action):
        self.rules.append((condition, action))

    def evaluate(self, facts):
        for condition, action in self.rules:
            if condition(facts):
                return action
        return "Eliminate contributor"

# Rule conditions
def condition_sri_lankan(facts):
    return facts["Nationality"] == "Sri Lankan"

def condition_age(facts):
    return facts["A"] >= 18

def condition_hs(facts):
    return facts["HS"] >= facts["HS T"]

def condition_lp(facts):
    return facts["LP"] >= facts["LP T"]

def condition_ca(facts):
    return facts["CA"] >= facts["CAT"]

def condition_l(facts):
    return facts["L"] >= facts["L T"]

# Rule actions
def action_select_contributor():
    return "Selected Contributor"

# Create a rule-based system
rule_system = RuleBasedSystem()

# Add rule
rule_system.add_rule(
    lambda facts: all([
        condition_sri_lankan(facts),
        condition_age(facts),
        condition_hs(facts),
        condition_lp(facts),
        condition_ca(facts),
        condition_l(facts),
    ]),
    action_select_contributor
```

```
)  
  
# facts  
facts = {  
    "Nationality": "Sri Lankan",  
    "A": 20,  
    "HS": 90,  
    "HS T": 80,  
    "LP": 85,  
    "LP T": 75,  
    "CA": 95,  
    "CAT": 90,  
    "L": 70,  
    "L T": 60,  
}  
# Evaluate the facts  
result = rule_system.evaluate(facts)  
print(result)
```

APPENDIX F

MODEL TO MEASURE TRUSTWORTHINESS OF CROWD RESPONSES USING LOGISTIC REGRESSION

```
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score

np.random.seed(42)
num_responses = 100
trustworthiness_scores = np.random.randint(0, 11, num_responses)
features = np.random.rand(num_responses, 3)

# Define a trustworthiness threshold
trustworthiness_threshold = 6

# Convert trustworthiness scores into binary labels (trustworthy or not)
labels = np.where(trustworthiness_scores >=
trustworthiness_threshold, 1, 0)

# Split data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(features,
labels, test_size=0.2, random_state=42)

# Train a logistic regression model
model = LogisticRegression()
model.fit(X_train, y_train)

# Make predictions on the test set
predictions = model.predict(X_test)

# Calculate accuracy
accuracy = accuracy_score(y_test, predictions)
print(f"Accuracy: {accuracy:.2f}")
```

APPENDIX G

SYSTEM USABILITY SCALE (SUS) ASSESSMENT

Rate your agreement with the following statements: පහත ප්‍රකාශයන් සමඟ ඔබේ එකඟත්වය සඳහන් කරන්න

Sinhala translation for the options are given below.

සිංහලට පරිවර්තනය කරන ලද ප්‍රතිචාර :

Strongly Disagree(දැඩි ලෙස එකඟ නොවේ)

Disagree(එකඟ නොවේ)

Neutral(මධ්‍යස්ථ)

Agree(එකඟයි)

Strongly Agree(දැඩි ලෙස එකඟ වේ)

1. I think that I would like to use this system frequently.

මම මෙම සේවාව නිතර භාවිතා කරනු ඇතැයි මම සිතමි.

Strongly Disagree Disagree Neutral Agree Strongly Agree

2. I think the service is unnecessarily complex.

මම හිතන්නේ මෙම සේවාව අනවශ්‍ය ලෙස සංකීර්ණයි.

Strongly Disagree Disagree Neutral Agree Strongly Agree

4. I think the service is easy to use.

මම හිතන්නේ සේවාව භාවිතා කිරීමට පහසුයි.

Strongly Disagree Disagree Neutral Agree Strongly Agree

5. I think I would need the support of a technical person to use this service.

මෙම සේවාව භාවිතා කිරීමට මට තාක්ෂණික පුද්ගලයෙකුගේ සහාය අවශ්‍ය වනු ඇතැයි මම සිතමි.

Strongly Disagree Disagree Neutral Agree Strongly Agree

6. I think that the service's features are well integrated.

සේවාවේ විශේෂාංග/features හොඳින් ඒකාබද්ධ වී ඇතැයි මම සිතමි.

Strongly Disagree Disagree Neutral Agree Strongly Agree

7. I think there is too much inconsistency in this service.

මම හිතන්නේ මේ සේවාවේ නොගැලපීම වැඩියි.

Strongly Disagree Disagree Neutral Agree Strongly Agree

8. I think that most people would learn to use this service very quickly.

මම හිතන්නේ බොහෝ අය ඉතා ඉක්මනින් මෙම සේවාව භාවිතා කිරීමට ඉගෙන ගනු ඇත.

Strongly Disagree Disagree Neutral Agree Strongly Agree

9. I found the service very cumbersome to use.

සේවාව භාවිතා කිරීම ඉතා අපහසු බව මට පෙනී ගියේය.

Strongly Disagree Disagree Neutral Agree Strongly Agree

9. I feel very confident using the service.

මෙම සේවාව විශ්වාසීව භාවිතා කල හැක.

Strongly Disagree Disagree Neutral Agree Strongly Agree

10. I needed to learn a lot of things before I could use the service.

සේවාව භාවිතා කිරීමට පෙර මට බොහෝ දේ ඉගෙන ගැනීමට අවශ්‍ය විය.

Strongly Disagree Disagree Neutral Agree Strongly AgreeAdditional

Comments:

Calculation of the SUS to assess Usability of the system.

		Q1		Q2		Q3		Q4		Q5		Q6		Q7		Q8		Q9		Q10			
Date and Timestamp	User ID	RS	QS	RS	QS	RS	QS	RS	QS	RS	QS	RS	QS	RS	QS	RS	QS	RS	QS	RS	QS	TOTAL	SUS Score
11/13/2023 22:30:40	user_0073	4	3	1	4	5	4	3	2	3	2	1	4	4	3	2	3	4	3	1	4	32	80
11/15/2023 16:16:49	user_0024	5	4	2	3	4	3	1	4	5	4	1	4	4	3	3	2	5	4	2	3	34	85
11/16/2023 1:16:59	user_0067	5	4	3	2	4	3	1	4	3	2	3	2	4	3	1	4	5	4	1	4	32	80
11/16/2023 2:47:55	user_0041	4	3	3	2	4	3	3	2	5	4	3	2	3	2	1	4	3	2	3	2	26	65
11/16/2023 10:26:50	user_0052	5	4	2	3	4	3	3	2	4	3	2	3	5	4	2	3	3	2	1	4	31	77.5
11/16/2023 14:20:01	user_0077	4	3	3	2	4	3	3	2	5	4	1	4	3	2	3	2	4	3	2	3	28	70
11/16/2023 14:24:30	user_0032	5	4	1	4	4	3	1	4	5	4	1	4	4	3	1	4	4	3	1	4	37	92.5
11/16/2023 14:27:23	user_0018	4	3	2	3	5	4	2	3	5	4	1	4	5	4	3	2	3	2	1	4	33	82.5
11/16/2023 15:27:19	user_0080	3	2	1	4	3	2	3	2	3	2	2	3	3	2	1	4	5	4	1	4	29	72.5
11/16/2023 19:04:46	user_0063	4	3	1	4	5	4	3	2	5	4	1	4	5	4	1	4	4	3	2	3	35	87.5
11/16/2023 19:06:09	user_0051	4	3	2	3	5	4	3	2	4	3	2	3	3	2	1	4	5	4	3	2	30	75
11/16/2023 19:08:51	user_0049	4	3	2	3	5	4	2	3	4	3	2	3	4	3	3	2	4	3	1	4	31	77.5
11/16/2023 21:29:40	user_0027	5	4	2	3	4	3	1	4	5	4	3	2	4	3	2	3	3	2	1	4	32	80
11/16/2023 21:39:13	user_0062	3	2	1	4	3	2	1	4	4	3	3	2	3	2	1	4	4	3	3	2	28	70
11/16/2023 21:50:06	user_0055	4	3	3	2	5	4	1	4	3	2	3	2	3	2	2	3	3	2	1	4	28	70
11/16/2023 23:13:08	user_0025	4	3	3	2	4	3	3	2	4	3	3	2	3	2	2	3	3	2	1	4	26	65
11/16/2023 23:27:49	user_0084	3	2	1	4	5	4	2	3	3	2	2	3	4	3	1	4	4	3	1	4	32	80

11/16/2023 23:28:42	user_0070	3	2	2	3	4	3	3	2	4	3	1	4	3	2	2	3	5	4	2	3	29	72.5
11/16/2023 23:39:43	user_0091	3	2	1	4	3	2	2	3	5	4	1	4	4	3	2	3	4	3	1	4	32	80
11/16/2023 23:58:33	user_0092	4	3	3	2	4	3	3	2	5	4	3	2	5	4	3	2	5	4	1	4	30	75
11/16/2023 23:59:44	user_0010	4	3	3	2	5	4	1	4	4	3	2	3	5	4	1	4	5	4	3	2	33	82.5
11/17/2023 0:01:34	user_0016	4	3	1	4	4	3	3	2	4	3	1	4	4	3	1	4	3	2	1	4	32	80
11/17/2023 0:16:56	user_0042	3	2	1	4	4	3	3	2	5	4	2	3	5	4	1	4	4	3	3	2	31	77.5
11/17/2023 21:43:41	user_0057	3	2	3	2	5	4	3	2	5	4	3	2	5	4	3	2	4	3	3	2	27	67.5
11/18/2023 0:08:41	user_0061	3	2	3	2	4	3	1	4	5	4	3	2	5	4	2	3	5	4	1	4	32	80
11/18/2023 21:48:55	user_0031	3	2	3	2	4	3	3	2	4	3	3	2	5	4	3	2	4	3	2	3	26	65
11/19/2023 1:04:48	user_0093	5	4	1	4	3	2	2	3	3	2	2	3	5	4	3	2	4	3	3	2	29	72.5
11/19/2023 1:05:47	user_0047	4	3	3	2	5	4	3	2	5	4	3	2	3	2	1	4	3	2	3	2	27	67.5
11/19/2023 6:23:43	user_0026	4	3	2	3	4	3	1	4	5	4	1	4	4	3	1	4	5	4	2	3	35	87.5
11/19/2023 14:45:15	user_0019	3	2	2	3	4	3	3	2	4	3	3	2	5	4	3	2	4	3	3	2	26	65

APPENDIX H

FEATURE VALUES TO ASSESS TRUSTWORTHINESS

User ID	Response_time	Accuracy_R_Consensus	Accuracy_R_Gold	Total_Tasks	completion_rate	ResponseTime(s)	Consistency	Total_Completed	Total_Attempted	Performance_score
user_0073	0.61	0.089079	0.053995	410	0.208333	0.74	0	5	24	0.66
user_0074	0.76	0.854341	0.489563	268	0.833333	0.33	0	10	12	0.09
user_0075	0.96	0.444457	0.510082	242	0.108108	0.22	0	4	37	0.86
user_0076	0.75	0.550448	0.007168	426	1	0.46	0	32	32	0.98
user_0077	0.74	0.485561	0.082062	59	0.2	0.24	1	1	5	0.16
user_0079	0.06	0.056744	0.763773	6	1	0.21	1	2	2	0.48
user_0080	0.41	0.895517	0.469237	182	0.454545	0.23	0	5	11	0.36
user_0081	0.76	0.256044	0.583338	58	0.244444	0.96	1	11	45	0.34
user_0082	0.36	0.383495	0.59844	383	0.736842	0.57	0	14	19	0.6
user_0084	0.15	0.488757	0.894656	292	0.185185	0.04	1	5	27	0.83
user_0085	0.14	0.276151	0.453913	416	0.25	0.69	1	6	24	0.28
user_0086	0.71	0.827352	0.647266	31	0.52	0.28	0	26	50	0.53
user_0087	0.69	0.032765	0.950585	410	0.3	0.72	1	6	20	0.4
user_0089	0.77	0.578558	0.708976	97	0.342857	0.33	1	12	35	0.53
user_0090	0.09	0.50211	0.148704	176	1	0.87	0	1	1	0.44
user_0092	0.17	0.469236	0.286837	264	0.25	0.64	1	11	44	0.43
user_0093	0.94	0.664678	0.443729	222	0.625	0.03	1	15	24	0.79
user_0095	0.10	0.60308	0.801813	385	0.0625	0.39	1	1	16	0.16
user_0096	0.16	0.360914	0.09468	394	0.678571	0.17	0	19	28	0.78
user_0097	0.92	0.966413	0.287886	109	1	0.92	0	4	4	0.91

user_0099	0.10	0.534744	0.173349	406	0.684211	0.47	1	13	19	0.86
user_0100	0.22	0.586194	0.206466	438	0.357143	0.06	1	5	14	0.16
user_0101	0.48	0.136157	0.439095	468	0.790698	0.82	0	34	43	0.52
user_0102	0.10	0.328038	0.935239	242	0.454545	0.86	1	20	44	0.83
user_0103	0.47	0.501775	0.121458	342	0.641026	0.52	0	25	39	0.77
user_0104	0.39	0.480893	0.081147	89	1	0.29	1	40	40	1
user_0106	0.09	0.727143	0.959752	469	0.953488	0.16	1	41	43	0.96
user_0107	0.82	0.291637	0.220021	443	0.333333	0.49	0	2	6	0.9
user_0108	0.32	0.673051	0.592726	5	0.857143	0.59	0	24	28	0.25
user_0110	0.61	0.223185	0.241761	117	0.238095	0.26	1	5	21	0.19
user_0111	0.87	0.272424	0.787986	425	0.857143	0.95	0	6	7	0.69
user_0112	0.91	0.464719	0.878474	206	0.071429	0.08	1	1	14	0.87
user_0114	0.80	0.953002	0.990758	303	1	0.44	1	31	31	0.16
user_0115	0.32	0.722724	0.324345	209	0.37037	0.88	0	10	27	0.53
user_0117	0.23	0.894293	0.583241	64	0.785714	0.11	1	11	14	0.73
user_0118	0.21	0.458436	0.3385	420	0.333333	0.43	0	14	42	0.69
user_0119	0.64	0.65434	0.89945	302	0.854167	0.8	0	41	48	0.02
user_0121	0.20	0.145883	0.700748	352	0.222222	0.18	1	2	9	0.56
user_0122	0.52	0.253166	0.366699	241	1	0.07	0	2	2	0.09
user_0124	0.53	0.71644	0.441874	138	0.857143	0.81	1	6	7	0.61
user_0125	0.89	0.300525	0.658546	295	1	0.55	0	31	31	0.17
user_0126	0.14	0.756693	0.838851	196	0.347826	0.16	1	8	23	0.07
user_0128	0.48	0.954897	0.987863	296	0.459459	0.92	1	17	37	0.02
user_0130	0.07	0.650479	0.664895	286	0.6875	0.06	0	22	32	0.66
user_0131	0.41	0.220739	0.621459	109	1	0.1	1	6	6	0.66

user_0132	0.69	0.049609	0.537681	205	0.354839	0.81	1	11	31	0.69
user_0133	0.85	0.732931	0.81756	60	0.375	0.38	0	3	8	0.07
user_0134	0.76	0.002938	0.594955	186	0.391304	0.87	1	18	46	0.06
user_0135	0.19	0.522597	0.986835	423	0.7	0.36	0	14	20	0.29
user_0136	0.70	0.253215	0.488106	472	0.5	0.39	1	8	16	0.7
user_0137	0.60	0.316682	0.081023	462	0.923077	0.99	1	12	13	0.6
user_0138	0.20	0.404254	0.002112	377	1	0.61	0	3	3	0.14
user_0139	0.42	0.75326	0.135129	41	1	0.28	1	7	7	0.22
user_0140	0.21	0.269735	0.151654	385	0.666667	0.59	1	2	3	0.92
user_0142	0.30	0.995529	0.681904	152	0.384615	0.89	0	5	13	0.98
user_0143	0.01	0.737569	0.130269	365	0.85	0.33	0	17	20	0.44
user_0144	0.61	0.018946	0.746948	360	0.083333	0.64	1	1	12	0.69
user_0146	0.84	0.595401	0.414119	393	0.588235	0.05	1	20	34	0.36
user_0148	0.19	0.548272	0.250741	399	0.846154	0.81	0	11	13	0.23
user_0151	0.74	0.94013	0.143608	167	0.333333	0.89	0	2	6	0.65
user_0152	0.20	0.913082	0.593638	135	0.444444	0.13	1	8	18	0.62
user_0154	0.24	0.612129	0.658932	219	0.068182	0.23	1	3	44	0.34
user_0155	0.87	0.092837	0.930913	185	0.655172	0.79	0	19	29	0.87
user_0156	0.54	0.601621	0.675513	115	1	0.36	1	2	2	0.11
user_0158	0.43	0.532833	0.453337	161	0.884615	0.45	1	23	26	0.13
user_0160	0.14	0.233854	0.024622	396	0.64	0.37	1	32	50	1
user_0161	0.20	0.447332	0.194933	325	0.5	0.06	1	11	22	0.75
user_0163	0.47	0.813202	0.351641	342	0.386364	0.55	1	17	44	0.98
user_0164	0.65	0.52386	0.855828	388	0.1875	0.27	0	6	32	0.04
user_0165	0.44	0.730839	0.840613	206	1	0.57	1	10	10	0.72

user_0168	0.24	0.079993	0.954937	253	0.904762	0.82	1	19	21	0.04
user_0169	0.15	0.421261	0.111857	250	0.323529	0.9	0	11	34	0.85
user_0171	0.82	0.980248	0.295278	92	1	0.52	0	5	5	0.85
user_0172	0.35	0.850558	0.491993	169	0.5	0.44	0	1	2	0.83
user_0173	0.89	0.500498	0.894066	87	1	0.46	1	1	1	0.37
user_0174	0.42	0.527472	0.081443	70	0.717391	0.32	0	33	46	0.9
user_0176	0.85	0.417732	0.080704	395	0.590909	0.04	1	26	44	0.21
user_0177	0.77	0.534285	0.676205	366	0.964286	0.5	1	27	28	0.9
user_0178	0.50	0.69126	0.407404	93	0.62069	0.62	1	18	29	0.14
user_0179	0.09	0.594999	0.918836	52	0.954545	0.83	1	21	22	0.11
user_0181	0.13	0.006453	0.660324	493	0.714286	0.22	1	5	7	0
user_0182	0.62	0.058018	0.280269	7	1	0.78	0	2	2	0.12
user_0183	0.85	0.719805	0.73413	1	0.333333	0.91	1	2	6	0.53
user_0185	0.75	0.219044	0.978263	441	0.5	0.94	1	3	6	0.93
user_0186	0.21	0.588929	0.072266	204	0.722222	0.25	1	13	18	0.57
user_0188	0.07	0.363264	0.450238	439	0.304348	0.2	1	14	46	0.47
user_0189	0.63	0.205424	0.928117	177	0.485714	0.94	0	17	35	0.88
user_0191	0.77	0.435139	0.872535	380	0.444444	0.38	1	4	9	0.02
user_0192	0.54	0.252755	0.226821	402	0.458333	0.6	0	11	24	0.91
user_0193	0.71	0.594571	0.997321	499	0.692308	0.7	1	9	13	0.38
user_0195	0.90	0.531191	0.945957	278	0.382979	0.7	1	18	47	0.77
user_0196	0.09	0.487872	0.792095	215	0.181818	0.32	1	2	11	0.25
user_0197	0.74	0.594139	0.156474	46	1	0.89	0	49	49	0.68
user_0199	0.60	0.924548	0.099047	224	0.277778	0.72	1	5	18	0.05
user_0201	0.27	0.274154	0.203423	283	0.866667	0.89	1	26	30	0.42

user_0202	0.37	0.866946	0.768719	383	0.029412	0.49	1	1	34	0.01
user_0203	0.25	0.077552	0.386669	488	0.392857	0.03	1	11	28	0.71
user_0204	0.78	0.237954	0.783244	178	0.714286	0.34	1	5	7	0.55
user_0206	0.23	0.985498	0.167674	59	0.888889	0.72	0	8	9	0.45
user_0209	0.43	0.801399	0.904861	327	0.777778	0.33	0	21	27	0.04
user_0210	0.71	0.486972	0.098643	20	0.894737	0.87	1	17	19	0.74
user_0213	0.93	0.767576	0.41026	53	0.918367	0.72	1	45	49	0.01
user_0214	0.31	0.809323	0.120697	439	0.555556	0.06	0	10	18	0.74
user_0215	0.44	0.202446	0.643912	135	0.083333	0.49	0	1	12	0.22
user_0216	0.71	0.247507	0.635419	194	0.576923	0.21	0	15	26	0.12
user_0217	0.98	0.264599	0.548594	58	1	0.68	0	1	1	0.26
user_0219	0.91	0.821554	0.769262	28	0.625	0.9	1	5	8	0.91
user_0220	0.49	0.566579	0.553054	498	0.4	0.36	0	2	5	0.86
user_0221	0.45	0.358388	0.826049	462	0.235294	0.23	1	4	17	0.53
user_0223	0.30	0.980377	0.412324	490	0.066667	0.58	1	1	15	0.86
user_0224	0.09	0.712656	0.59265	49	0.36	0.01	0	18	50	0.9
user_0225	0.08	0.881864	0.710081	456	0.272727	0.6	0	3	11	0.7
user_0226	0.59	0.942261	0.450888	488	0.62963	0.82	0	17	27	0.24
user_0227	0.97	0.564026	0.507565	360	0.263158	0.34	0	5	19	0.55
user_0230	0.93	0.797096	0.921325	50	0.517241	0.74	1	15	29	0.88
user_0231	0.76	0.293269	0.391602	71	0.4375	0.8	0	21	48	0.06
user_0233	0.23	0.619552	0.627588	377	0.770833	0.66	0	37	48	0.66
user_0234	0.70	0.517477	0.111875	179	0.095238	0.77	0	2	21	0.78
user_0235	0.45	0.987807	0.853535	138	1	0.08	1	3	3	0.06
user_0236	0.89	0.32785	0.934401	299	0.418605	0.82	1	18	43	0

user_0238	0.69	0.541516	0.367337	13	1	0.16	1	19	19	0.28
user_0240	0.30	0.227605	0.221834	254	0.789474	0.69	0	30	38	0.65
user_0241	0.02	0.578366	0.677628	201	0.909091	0.9	0	10	11	0.15
user_0244	0.33	0.028904	0.781686	150	0.461538	0.52	1	18	39	0.86
user_0245	0.59	0.142214	0.563851	128	0.682927	0.68	0	28	41	0.98
user_0247	0.28	0.331903	0.160093	136	0.741935	0.54	0	23	31	0.94
user_0249	0.88	0.086757	0.416928	33	0.259259	0.4	1	7	27	0.57
user_0251	0.76	0.559637	0.698494	263	0.75	0.38	0	15	20	0.74
user_0252	0.41	0.422934	0.468625	492	0.804348	0.41	0	37	46	0.79
user_0253	0.59	0.706666	0.863247	421	0.363636	0.5	0	8	22	0.16
user_0254	0.02	0.924926	0.466315	380	0.4375	0.33	0	14	32	0.81
user_0255	0.80	0.787125	0.410336	423	0.355556	0.79	1	16	45	0.64
user_0258	0.41	0.270118	0.624716	31	0.894737	0.37	0	34	38	0.84
user_0260	0.77	0.447225	0.838465	455	0.676471	0.91	1	23	34	0.01
user_0261	0.33	0.826434	0.082639	449	0.533333	0.86	0	8	15	0.93
user_0262	0.43	0.495346	0.81447	344	0.5	0.18	0	1	2	0.81
user_0265	0.02	0.932872	0.635893	231	1	0.05	1	5	5	0.89
user_0266	0.61	0.701619	0.350469	428	0.285714	0.36	0	4	14	0.79
user_0267	0.96	0.684163	0.559346	78	0.933333	0.72	0	42	45	0.48
user_0268	0.07	0.709644	0.505205	312	0.26087	0.82	0	12	46	0.68
user_0272	0.99	0.027943	0.66536	61	0.315789	0.92	0	6	19	0.48
user_0274	0.51	0.965335	0.026211	29	0.583333	0.88	0	28	48	0.95
user_0276	0.01	0.361058	0.566669	249	0.4	0.85	0	2	5	0.81
user_0280	0.88	0.670226	0.497478	264	0.888889	0.17	1	32	36	0.57
user_0284	0.66	0.185558	0.007236	201	0.888889	0.34	0	32	36	0.65

user_0288	0.56	0.759769	0.857024	496	0.8	0.15	0	4	5	0.53
user_0290	0.62	0.336113	0.880114	55	0.179487	0.78	1	7	39	0.68
user_0294	0.68	0.234825	0.773323	373	1	1	0	14	14	0.83
user_0296	0.95	0.621482	0.826895	346	0.48	0.94	0	12	25	0.37
user_0300	0.40	0.842287	0.9882	328	0.6	0.12	1	3	5	0.42
user_0302	0.33	0.213038	0.765938	235	0.307692	0.46	0	8	26	0.74
user_0304	0.02	0.675953	0.010974	52	0.333333	0.83	0	1	3	0.54
user_0308	0.18	0.942615	0.784376	148	0.153846	0.29	0	2	13	0.05
user_0310	0.71	0.203689	0.201161	162	0.4375	0.58	1	21	48	0.37
user_0312	0.83	0.527956	0.532766	65	0.222222	0.03	0	8	36	0.04
user_0316	0.62	0.648191	0.80324	460	1	0.09	1	2	2	0.24
user_0318	0.95	0.002362	0.248594	44	0.045455	0.66	1	1	22	0.89
user_0320	0.12	0.091643	0.624895	102	1	0.1	1	4	4	0.25
user_0324	0.48	0.703895	0.53413	158	0.971429	0.88	0	34	35	0.85
user_0328	0.87	0.448474	0.561777	7	0.916667	0.04	0	11	12	0.37
user_0330	0.42	0.769882	0.904171	472	0.852941	0.46	1	29	34	0.52
user_0334	0.04	0.397205	0.952785	5	0.333333	0.81	1	1	3	0.01
user_0336	0.21	0.141636	0.57447	263	0.307692	0.16	1	4	13	0.28
user_0344	0.26	0.67999	0.516762	15	0.408163	0.23	1	20	49	0.58
user_0348	0.51	0.167951	0.404419	281	0.06383	0.92	1	3	47	0.51
user_0352	0.71	0.437125	0.95362	318	0.34	0.38	0	17	50	0.2
user_0354	0.49	0.396367	0.612158	387	0.285714	0.57	0	10	35	0.4
user_0356	0.12	0.041589	0.336163	110	0.2	0.22	0	3	15	0.13
user_0358	0.09	0.921278	0.659534	275	0.117647	0.83	1	2	17	0.08