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

LIST OF FIGURES

LIST OF TABLES

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