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

Nadeera Meedin, Maneesha Caldera, Suresha Perera, Indika Perera
Dept. of Computer Science and Engineering, University of Moratuwa, Katubedda, Sri Lanka

***Abstract*—The number of user-generated posts is growing exponentially with social media usage growth. Promoting violence against or having the primary purpose of inciting hatred against individuals or groups based on specific attributes via social media posts is daunting. As the posts are published in multiple languages with different forms of multimedia, social media finds it challenging to moderate before reaching the audience and assessing the posts as hate speech becomes sophisticated due to subjectivity. Social media platforms lack contextual and linguistic expertise and social and cultural insights to identify hate speech accurately. Research is being carried out to detect hate speech on social media content in English using machine learning algorithms, etc., using different crowdsourcing platforms. However, these platforms' workers are unavailable from countries such as Sri Lanka. The lack of a workforce with the necessary skill set and annotation schemes symbolizes further research essentiality in low-resource language annotation. This research proposes a suitable crowdsourcing approach to label and annotates social media content to generate corpora with words and phrases to identify hate speech using machine learning algorithms in Sri Lanka. This paper summarizes the annotated Facebook posts, comments, and replies to comments from public Sri Lankan Facebook user profiles, pages and groups of 52,646 instances, unlabeled tweets based on 996 Twitter search keywords of 45,000 instances of YouTube Videos of 45,000 instances using the proposed annotation scheme. 9%, 21% and 14% of Facebook, Twitter and YouTube posts were identified as containing hate content. In addition, the posts were categorized as offensive and non-offensive, and hate targets and corpus associated with hate targets focusing on an individual or group were identified and presented in this paper. The proposed annotation scheme could be extended to other low-resource languages to identify the hate speech corpora. With the use of a well-implemented crowdsourcing platform with the proposed novel annotation scheme, it will be possible to find more subtle patterns with human judgment and filtering and take preventive measures to create a better cyberspace.**

***Keywords*—Annotation; crowdsourcing; hate speech detection; social media data analytics**

I. Introduction

Social media users keep updating and sharing posts and comments on social media platforms at an exponential rate. Users can express themselves freely across countries and cultures in dozens of languages. People use social media to share their experiences, connect with friends and family, and build communities. However, social media platforms try to maintain their community standards so that their users feel safe using their products. One such example is the Facebook community standard [1]. In their standards, they have specified the content not to be posted by users to prevent possible harms related to content on Facebook under categories such as violence and incitement, dangerous individuals and organization, coordinating damage and publicizing crime etc. For example, Facebook does not allow users to post content, including hate speech on Facebook to avoid creating an environment of intimidation and exclusion and which would promote real-world violence. Similarly, Twitter and YouTube have their hateful policy conduct [2] [3].

However, policies and standards exist, and social media platforms take action to remove posts with inappropriate content; users tend to spread hate speech using social media. The information shared on social media can be offensive and could lead to creating social issues. Social media research includes analyzing social media data on various topics. Some of these topics would have a direct impact on creating social issues. Therefore, it is vital to have a mechanism to identify the contents that directly impact social issues.

It has been a challenge for social media platforms to moderate billions of daily posts in more than a hundred languages. It has been found that it is impossible to maintain a balance between what is considered "hate speech" and "free speech" since social media is global. Hate speech is a broad and contested term [4], and there is no common standard definition for hate speech [5]. Multiple definitions used by different authorities and platforms are explained in the Literature Survey section. United Nations strategy and plan of action on hate speech describe hate speech as "any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language regarding a person or a group based on who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factors"[6]. Hence both "hate speech" and "free speech" are determined by region. When analyzing the speech, it is essential to consider context-specific details, social and cultural factors, etc. Determining hate speech based on user context is one of the challenges faced by social media platforms as they deal with more than a hundred languages [7] and nationalities.

By incorporating a crowdsourcing approach, it can reach a much larger audience and capture user opinions and behaviours. In social computing research, social media has provided a unique window into people's social experience; in particular, Twitter is used for assessing sensitive topics, such as discrimination[8]. Therefore, when capturing user opinions, it is required to ask the appropriate questions in the appropriate order to obtain responses effectively. Similarly, the captured responses should be represented in such a way after ensuring their quality.

This research proposes a novel solution to address this problem: an adaptive questionnaire to identify hate speech, detect hate targets and keywords related to hate and label the social media posts written in Sinhala and Singlish. The main contributions of this paper are the following;

- A crowdsourcing framework for annotating hate speech.
- An annotating scheme to generate a hate speech corpus.
- Annotated dataset and hate speech corpus.

II. Related Work

A. *Use of Crowdsourcing in Annotating Hate Speech*

Daniel Faggella [9] explains how crowdsourcing could be used in social media content moderation and states that moderation at a scale usually involves two elements; a trained machine learning algorithm informed by the user or outside data. In addition, several factors [10] affect the judgement of crowd moderators when deciding on the suitability of text content, such as participants often labelled unsure when they found it challenging to decide on marginal content. Considering these approaches, the factors considered in registering workers and how to direct workers to annotate, limiting the judgements are further explained in the Section III Crowdsourcing Platform.

Amazon Mechanical Turk(MTurk) is the most common crowdsourcing platform for performing labour-intensive tasks, data collection and annotation for hate speech identification [11], especially in NLP. MTurk facilitate quality management [11][12], ranking annotators[13], ensuring trustworthiness. There are no mechanisms in MTurk for detecting unfair evaluations and no metric called reputation. As a result, workers are highly susceptible to misbehaviour [14]. MTurk allows workers from only a limited number of countries to register. Hence the existing workers could not cater for the cultural, linguistic and contextual insights. Furthermore, MTurk is a general-purpose crowdsourcing platform which allows the annotation of many types of tasks and the annotation scheme adaptation for low-resource languages is crucial because of the linguistic challenges in MTurk.

B. *Annotation in Hate Speech Detection*

Both race and sexual orientation are among the top hate speech targets [15]. Our study examined the hate targets in the Sri Lankan context. It was different depending on the cultural, societal and religious differences. Amazon Mechanical Turk (AMT) was used in Relia et al. research for annotation after performing keyword filtering, and Support Vector Machine (SVM) and Neural Network (NN) was used for Twitter classification. As this task involved the exposure of humans to potentially sensitive content, the researchers have indicated the task was about racist Tweets as an individual Human Intelligence Task (HIT), giving workers a chance to discontinue at any point without losing payment if they felt uncomfortable. HaterNet [16] can identify and classify hate speech in Twitter data. The Spanish national office is using it against hate crimes. HaterNet uses a combination of Long Short-Term Memory (LSTM) and Multi-layer Perceptron (MLP) neural networks.

Burnap [17] has annotated text using CrowdFlower, with a minimum of four humans, to assess whether it is likely to be offensive or antagonistic regarding race, ethnicity or religion. Scores are given based on a ternary set of classes, yes, no and undecided. Agrawal and Aweaker have used four DNN-based models for cyberbullying detection; CNN, LSTM, BLSTM, and BLSTM [18]. Here they have used a manually annotated dataset. Fernando and Asier [19] have introduced a new algorithm to detect hate speech messages. They used the Kappa coefficient to measure the degree of agreement when performing the subjective analysis. The researchers have categorized tweets into five categories: direct incitement or threat of violence, an attack on honour and human dignity, incitement to discrimination or hate and an offence to collective sensitivity.

The hate categories observed in [20] are race, behaviour, physical, sexual orientation, class, ethnicity, gender, disability, religion and others. A few of the top hate targets in the US, Canada, and the UK are black, fat, fake, stupid, gay, white, rude, ignorant, racist, old, selfish and religious, which is different in an Asian country Sri Lanka as it was found out in our study.

We followed the approach of Fernando and Asier [19] and used the five conditions given in Table IV to check if a post contains hate. In addition, the hate targets and the hate categories specified in each post were identified using the categories listed in [20]. These categories are listed in Table IV. Finally, we went one step further to identify the words, phrases and sentences which incite hatred in the post.

C. *Benchmark Datasets for Hate Speech Detection*

Table IX at the end of this paper summarizes existing benchmark annotated datasets used for hate speech detection in different languages. The system architecture, design and implementation of the crowdsourcing platform, information architecture of the system, the pre-selection criteria of contributors, and the proposed criterion to identify inappropriate content in social media are explained in the next section.

III. Crowdsourcing Platform

This paper presents a novel crowdsourcing platform that allows any interested participant to register by providing user profile details such as name, age, nationality, date of birth and location of contribution. The system architecture of the implemented crowdsourcing platform is given in Fig. 1.

The intrinsic rewarding process starts with the first digital badge, "Contributor", and after fulfilling the selection criteria specified in Table II, the contributors get the badge "Selected contributor" and the eligibility to earn financial rewards. The worker management process is illustrated in Fig. 2.

1) Pre-selection of contributors: Participants who were literate in Sinhala and Sri Lankan natives were selected as contributors to the evaluation process. Those who did not qualify in the pre-selection process, failed at quality control, and failed at the trustworthiness-ensuring process were eliminated. The list of symbol definitions for pre-selection contributors is given below in Table I.

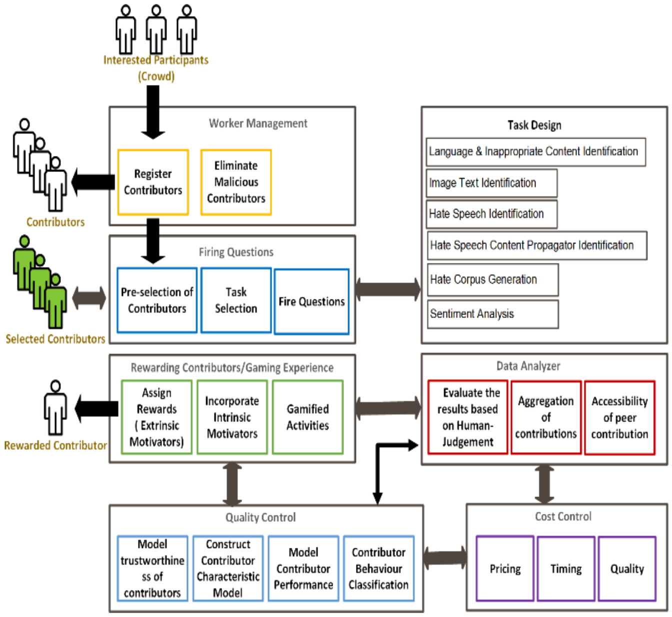
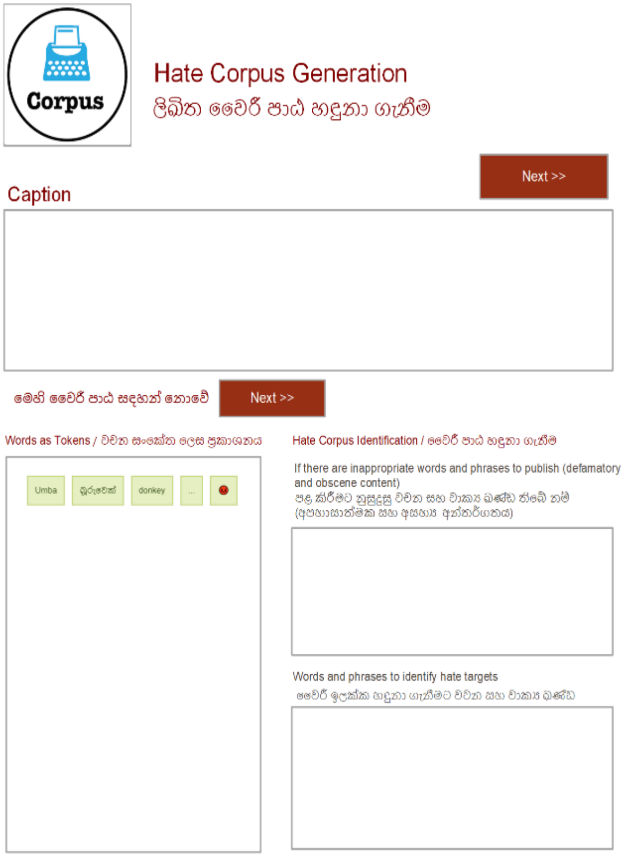

Interested Participants
(Crowd)
Contributors
Selected Contributors
Rewarded Contributor
Worker Management
Register Contributors
Eliminate Malicious Contributors
Firing Questions
Pre-selection of Contributors
Task Selection
Fire Questions
Rewarding Contributors/Gaming Experience
Assign Rewards (Extrinsic Motivators)
Incorporate Intrinsic Motivators
Gamified Activities
Task Design
Language & Inappropriate Content Identification
Image Text Identification
Hate Speech Identification
Hate Speech Content Propagator Identification
Hate Corpus Generation
Sentiment Analysis
Data Analyzer
Evaluate the results based on Human-Judgement
Aggregation of contributions
Accessibility of peer contribution
Quality Control
Model trustworthiness of contributors
Construct Contributor Characteristic Model
Model Contributor Performance
Contributor Behaviour Classification
Cost Control
Pricing
Timing
Quality

Fig. 1. The System Architecture of the Crowdsourcing Platform.



Hate Corpus Generation
ලිඛිත වෛරී පාඨ හඳුනා ගැනීම

Caption

මෙහි වෛරී පාඨ සඳහන් නොවේ

Words as Tokens / වචන සංකේත ලෙස ප්‍රකාශනය

Hate Corpus Identification / වෛරී පාඨ හඳුනා ගැනීම

If there are inappropriate words and phrases to publish (defamatory and obscene content)
පළ කිරීමට නුසුදුසු වචන සහ වාක්‍ය ඛණ්ඩ තිබේ නම්
(අපහාසාත්මක සහ අසභ්‍ය අන්තර්ගතය)

Words and phrases to identify hate targets
වෛරී ඉලක්ක හඳුනා ගැනීමට වචන සහ වාක්‍ය ඛණ්ඩ

Fig. 2. Worker Registration and Reward Process.

TABLE I. LIST OF SYMBOL DEFINITIONS FOR PRE-SELECTION OF CONTRIBUTORS

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 II. CRITERIA IN PRE-SELECTION OF 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

The given threshold values were selected for the criteria in assessing prior knowledge in hate speech identification (HST=5), Sinhala language proficiency (LPT=8), ability to read Sinhala words written in English letters (LT=8), comprehension, and analytical skills (CAT=8) and the criteria in pre-selection of contributors is given in Table II.

Task Selection: Selected contributors were asked questions randomly based on a question generation mechanism helping to generate a corpus with hate speech, annotate the posts to detect and classify the intention of the hate speech, identify Sinhala texts from images, identify inappropriate words in the social media content and to sort Sinhala, Singlish and English words in a given set of social media posts. Out of the six types of task designs allowed by the crowdsourcing framework, only the results of hate speech identification and hate corpus generation are explained in the evaluation section of this paper (See Fig. 3).

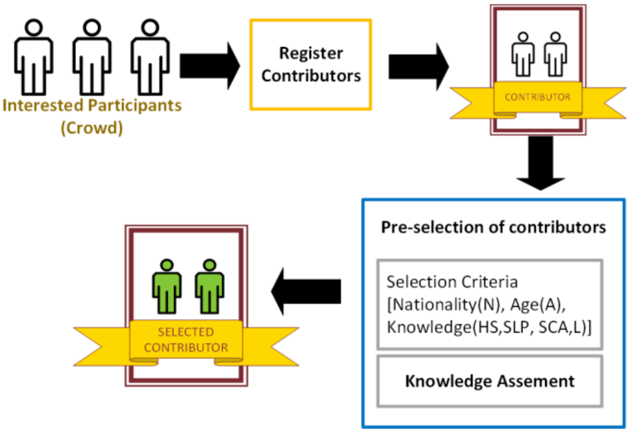


Fig. 3. UI to Capture Inappropriate Contents and Words and Phrases Referring Hate Targets.

2) Fire questions: A rule base was used to assess the pre-selection criteria of contributors with forwarding chaining as the firing mechanism, and a dataset with JSON objects of Twitter, Facebook and Video posts was used during the task assignments.

3) Assign Rewards (Extrinsic and Intrinsic Motivators): Google crowdsource [21] is a global platform which uses the feedback of users to design products to provide a customized experience to its users. For example, they provide a gaming experience as an intrinsic reward to users instead of monetary rewards. After alterations, a similar approach has been taken in our research to motivate the Sri Lankan workers. The proposed reward system was designed after testing a few samples of workers by applying different rewarding methods.

Workers were allowed to earn monetary awards considering the completion levels, accuracy, trustworthiness of the contributor, etc. To admire the effort of the contributors to make cyberspace better monetary rewards were assigned to those who showed higher trustworthiness scores, and based on the human intelligence tasks HITs) completed, digital badges were assigned. In addition, the platform would provide a gaming experience for contributors to retain in the cause. The intrinsic and extrinsic motivators were embedded in the gaming experience, as shown in (See Fig. 4).

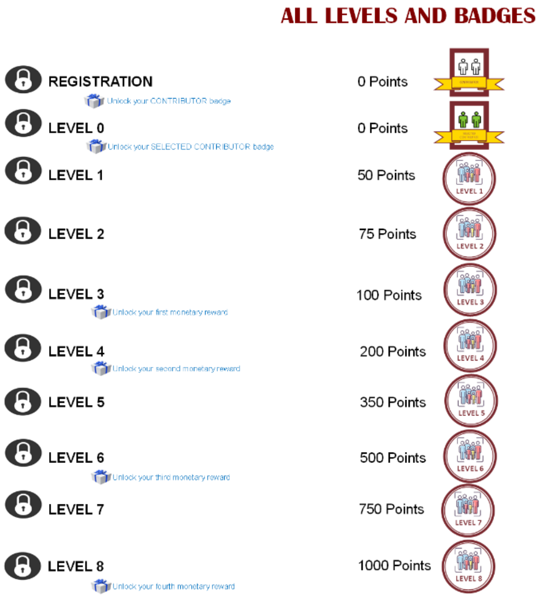


Fig. 4. Intrinsic and Extrinsic Rewards Assignment for Selected Contributors.

4) Model the trustworthiness of a contributor: An inbuilt mechanism was built to check the trustworthiness of a particular contributor's responses and assign a badge for trustworthiness. Ten golden rules were used as the primary method of measuring trustworthiness and were compared with the predicted trustworthiness score. In addition, a higher weightage of validity to the response from a trustworthiness badge-owned contributor was considered in assessing the quality of the response.

5) Aggregation of contributions: Evaluates the results based on human judgment for each post, store the results, ignored, hate speech or not, category of the hate nature, etc., to aggregate the contributions and generate advanced questions later. The agreement scores were calculated using Cohen's kappa coefficient, Fleiss' kappa coefficient and Krippendorff's alpha.

IV. Experimentation

The research was carried out considering three cases, the first using Facebook posts, the second using Tweets and the third using YouTube posts, as specified below in the dataset section. The initial classification step involved classifying the posts in the language in which the post was written. Here we categorized the language into five categories Sinhala, English, Singlish, Both English and Sinha-la, and None of the above. Out of the five categories, the posts written in Sinhala, Singlish and both Sinhala and English were considered under the study. Posts written in English and other languages were not considered under the study. The second step involved the task design for hate post identification, and the third in hate speech corpus generation.

V. Pilot Dataset

There is no large publicly available data set of Facebook, YouTube and Twitter written in Sinhala, Singlish and mixed code. Therefore, in this research, we selected three subsets from the three social media platforms as our cases under study, with the most significant number of Sri Lankan users as of 2022[22]. Singlish is Sinhala words written using English. Therefore, our dataset tweets were collected using Twitter API based on 996 Twitter search keywords. Web crawling was used to collect Facebook posts from public Sri Lankan Facebook user profiles, pages and groups, and YouTube posts and comments were captured from popular Sinhala YouTube channels as shown in Table III. The dataset was annotated by 20 selected annotators who passed the pre-selection test and showed a higher trustworthiness score. UPF-08 encoding errors and inconsistencies were corrected in the labelled files.

TABLE III. Unlabelled Data Sets for Hate Speech Identification

Case	**Dataset**	**Instances**
Case 1: Facebook	Unlabeled posts, comments, and replies to comments from public Sri Lankan Facebook user profiles, pages and groups(Post image, Comments, video thumbnail)	52,646
Case 2: Twitter	Unlabeled tweets based on 996 Twitter search keywords (Tweet text, replies for each tweet, 6317 video thumbnails and pictures). The average number of comments and replies per Twitter post is 4	45,000
Case 3: Youtube	Youtube Videos(Video title, thumbnail, Comments for each video). The average number of comments and replies per video is 6	45,000

VI. Annotation Scheme to Generate a Hate Speech Corpus

The annotation scheme specified below in Table IV was used to get the data annotated, and the annotated datasets were published in the GitHub repository.

TABLE IV. LIST OF SYMBOL DEFINITIONS FOR LABELLING

Symbol	Definition
L_1	Content analysis
L_2	Hate speech identification
L_3	Who does a particular post target?
L_4	Hate categories
C1, C2, C3, C4, C5	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 Orientation5. Disability 6. Disease 7. Immigration 8. Victims of a major violent event and their kin 9. Veteran Status/Profession10. Caste11. Political12. Regional 13. Gender 14. Economic & Business 15. A particular individual 16. Other social groups

Definition to label L_1 to L_4

Starting set of labels L_1 is:

L_1=\{*Offensive content, No offensive content, Cannot tell*\}

Starting set of labels L_2 is:

L_2=\{*C1, C2 , C3, C4, C5*\}

Starting set of labels L_3 is:

L_3=\{*Individual, Group, Cannot tell*\}

Starting set of labels L_4 is:

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

VII. RESULTS AND DISCUSSION

A. Demographic Distribution of the Annotator Profiles

The demographic distribution of the selected contributors involved with the annotation process is given in Fig. 5 to 7.

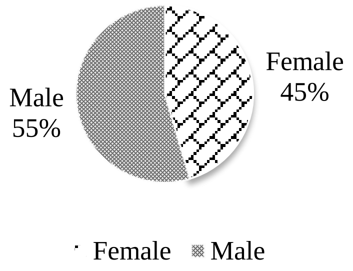


Fig. 5. Gender Distribution of Selected Contributors.

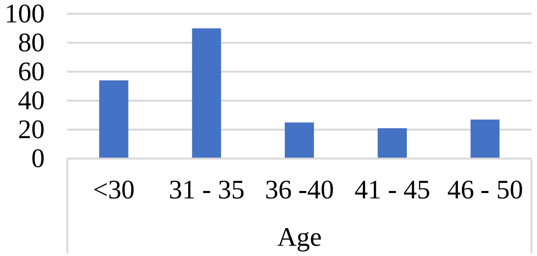


Fig. 6. Age Distribution of Selected Contributors.

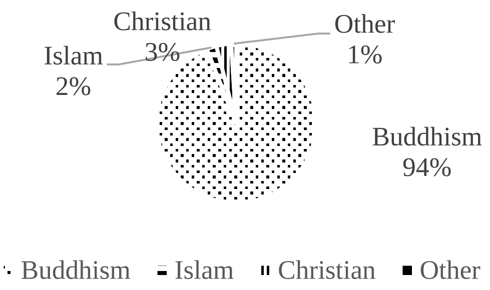


Fig. 7. Religion-wise Distribution of Selected Contributors.

B. Inter Annotator Agreement

Table V shows the agreements obtained in terms of the average percent agreement (agr_i), average Cohen's kappa coefficient (avg k), Fleiss' kappa coefficient (Fleiss) and Krippendorff's alpha (α). The number of annotated posts/comments/tweets is also given for each batch.

If the set of items is $\{i|i \epsilon I\}$, and is of cardinality *i*.

Observed agreement(A_0) over the values agr_i for all items $i \in I$.

$$A_o \frac{1}{i}\sum_{i\in I} agr_i \tag{1}$$

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

Cohen Kappa coefficient(k) where expected agreement is A_e

$$K = \frac{A_0 - A_e}{1 - A_e} \tag{2}$$

TABLE V. INTER ANNOTATOR AGREEMENT ON CASE1, CASE 2 AND CASE 3 FOR L_1

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

C. Annotated Datasets

Hate posts/comments and tweets were identified after analyzing L_2 and L_3. If at least one choice was selected and the inter-rater agreement was more than 0.6, the posts were identified as hate posts.

If we compare Table VI and Table VII, it is evident that intuitive annotation of offensive and non-offensive content shows lesser percentages when compared with hate and non-hate classification.

TABLE VI. DISTRIBUTION OF POSTS/COMMENTS/TWEETS ANNOTATIONS

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

D. Lexical Distribution

Table VIII lists the ten most frequently occurring words to refer to the hate targets with the most frequent occurrence for each class. Twitter search keywords, identified hate targets and the corpus associated with hate targets focusing on an individual or group are given in the GitHub repository. The annotated Facebook, Twitter, and YouTube datasets can be requested by emailing the authors of this paper.

TABLE VII. DISTRIBUTION OF POSTS/COMMENTS/TWEETS OVER CATEGORIES IN THE CLEAN DATASET-FACEBOOK

	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%

TABLE VIII. DISTRIBUTION OF THE TEN MOST FREQUENTLY OCCURRING TERMS IN HATE SPEECH REFERRING HATE TARGETS

Word	**Meaning of the word in English**	**Distribution**
දෙමළා	Used as an insulting and contemptuous term for a person from a target community	0.65%
හම්බා	Used as an insulting and contemptuous term for a person from a target community	0.67%
තම්බියෙක්		0.65%
තම්බියා		1.9%
තම්බි	Used as an insulting and contemptuous term to refer target community	1.01%
තම්බියෝ		1.83%
තම්බිලා		0.38%
ගණයා	Used as an insulting and contemptuous term to refer to a priest from a target religion	0.25%
අන්තවඞී	Used to refer to a person as an extremist	0.35%
සිංහලේ	Used to elevate an ethnicity and demean all the other ethnicities	0.65%

TABLE IX. SUMMARY OF EXISTING BENCHMARK ANNOTATED DATASETS USED FOR HATE SPEECH DETECTION FOR DIFFERENT LANGUAGES

Dataset	Social Media Platform	Language	Criteria/Based on	Annotation question/s or Categories	Size	Intercoder-agreement score	Output
Davidson [23] Hatebase Twitter	Tweet	English	The words appearing in a given tweet and the context in which they were used.	Labelled categories: hate speech, offensive but not hate speech, or neither offensive nor hate speech	24,802 tweets	Majority Voting 92%	5% - hate speech 76% - Offensive 17% - Neither
Waseem[24]	Tweet	English	Criteria based on Critical Race Theory(CRT) : 1. uses a sexist or racial slur. 2. attacks a minority. 3. seeks to silence a minority. 4. criticizes a minority (without a well-founded argument). 5. promotes, but does not directly use, hate speech or violent crime. 6. criticizes a minority and uses a straw man argument. 7. blatantly misrepresents truth or seeks to distort views on a minority with unfounded claims. 8. shows support for problematic hashtags. E.g. "#BanIslam", "#whoriental", and "#whitegenocide" 9. negatively stereotype a minority. 10. defends xenophobia or sexism. 11. contains an offensive screen name, as per the previous criteria, the tweet is ambiguous (at best), and the tweet is on a topic that satisfies any of the above criteria	Labelled categories: Sexism, Racism, Neither	16,914 tweets	Cohen's kappa coefficient K=0.84	Racism – 12% Sexism – 20% Neither - 68
Gibert et al.[25]	Stormfront - Internet posts	English	a) deliberate attack, b) directed towards a specific group of people, c) motivated by aspects of the group's identity.	Hate, No hate, Skip	10,568 sentences	Average percent agreement 91.03 Cohen's kappa coefficient 0.614 Fleiss' kappa coefficient 0.607	Hate 1196 Skip 72 No hate 9674

Most of the phrases referring to the hate targets consisted of at least one term from the given list.

VIII. CONCLUSION

After performing this research, the following drawbacks were identified. (1) The workers found it challenging to identify if a particular comment is harmful or harmless by looking only at it. (2) based on the task, the number of annotator requirements should be identified and vary (3) eliminate the bias of the annotator response.

To make the decisions, they needed to see the original post, the replies against it, and the images associated with the comments, if any. Therefore, suggesting preventive measures would not work if we only remove a single comment. Instead, it would be required to remove a set of comments. Therefore when designing tasks for crowdsourcing platforms, it is required to redesign the tasks to include the relevant images and any context-specific data along with the post when asking questions from the crowd, as proposed in this research.

Though subjective responses were captured in labelling, it is mandatory to design the annotation scheme such that the annotator would consciously judge the comment based on the given criteria without intuitively completing the crowdsourcing tasks.

Two different techniques should be used to score the trustworthiness of users during the pre-selection of contributors to the crowdsourcing platform. Krippendorff's alpha coefficient could be used as the reliability estimation method, and the number of contributors for each type of task would be very, as shown in the results. Two contributors can be used to perform primary classifications, while in-depth analyses such as sentiment strength analysis and hate target identification should be increased by checking the reliability score.

It is essential to use equal percentages to represent each religion which is nearly impossible in this context as the majority of the Sri Lankan population is Buddhists, to avoid bias toward religion.

The beliefs of the contributor would affect their response. Therefore, it is required to randomize the worker selection to eliminate the bias and to fire different categories of posts for each worker type.

Despite the outcomes of this research, future research should aim to (1) identify multiple comments against the Facebook posts instead of a single comment or caption to remove as a preventive measure in spreading hate, (2) redesign crowdsourcing tasks to include the relevant images and any context-specific data along with the post when asking questions from the crowd (3) having a mechanism to ensure annotator would consciously judge the comment based on the given criteria without intuitively completing the crowdsourcing tasks (4) measure the biases and beliefs of workers to ensure the trustworthiness of crowd response (5) identifying clusters of worker types and fire different categories of questions to each type.

The outcomes of this research consist of a crowdsourcing framework for annotating hate speech for Sinhala, Sinhala written in English and English social media posts, an annotation scheme to generate a hate speech corpus and an annotated dataset. These outcomes can be used by NLP researchers in performing linguistic research and getting annotation done for local languages, and policymakers to take preventive measures in identifying inappropriate content, hate targets and hate categories.

The datasets, implemented project, hate targets, hate-related search key terms used with Twitter, and hate corpus could be found in the;

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

ACKNOWLEDGEMENT

This research is aided by the Accelerating Higher Education Expansion and Development (AHEAD) Operation of the Ministry of Higher Education with World Bank funds.

REFERENCES

[1] ‘Facebook Community Standards | Transparency Centre’. https://transparency.fb.com/en-gb/policies/community-standards/ (accessed Aug. 07, 2022).

[2] ‘Twitter’s policy on hateful conduct | Twitter Help’. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy (accessed Aug. 07, 2022).

[3] ‘Hate speech policy - YouTube Help’. https://support.google.com/youtube/answer/2801939?hl=en (accessed Feb. 02, 2020).

[4] I. Gagliardone, D. Gal, T. Alves, G. Martinez, and Unesco, Countering online hate speech. Paris: United Nations Educational, Scientific and Cultural Organization, 2015.

[5] 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.

[6] 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.

[7] C. Curtis, ‘Facebook’s global content moderation fails to account for regional sensibilities’, The Next Web, Feb. 26, 2019. https://thenextweb.com/socialmedia/2019/02/26/facebooks-global-content-moderation-fails-to-account-for-regional-sensibilities/ (accessed Feb. 02, 2020).

[8] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze, ‘Annotating named entities in Twitter data with crowdsourcing’, p. 9.

[9] S. Juumta, ‘Crowdsourced Content Moderation - How it Works and What’s Possible’, Emerj Artificial Intelligence Research. https://emerj.com/partner-content/crowdsourced-content-moderation-how-it-works-and-whats-possible/ (accessed Aug. 07, 2022).

[10] D. Hettiachchi and J. Goncalves, ‘Towards Effective Crowd-Powered Online Content Moderation’, in Proceedings of the 31st Australian Conference on Human-Computer-Interaction, Fremantle WA Australia, Dec. 2019, pp. 342–346. doi: 10.1145/3369457.3369491.

[11] P. Ipeirotis, ‘Crowdsourcing using mechanical turk: quality management and scalability’, in Proceedings of the 8th International Workshop on Information Integration on the Web: in conjunction with WWW 2011, 2011, p. 1.

[12] P. G. Ipeirotis, F. Provost, and J. Wang, ‘Quality management on Amazon Mechanical Turk’, in Proceedings of the ACM SIGKDD Workshop on Human Computation - HCOMP ’10, Washington DC, 2010, p. 64. doi: 10.1145/1837885.1837906.

[13] V. C. Raykar and S. Yu, ‘Ranking annotators for crowdsourced labeling tasks’, in In Advances in neural information processing systems, 2011, pp. 1809–1817.

[14] M. Allahbakhsh, A. Ignjatovic, B. Benatallah, S.-M.-R. Beheshti, E. Bertino, and N. Foo, ‘Reputation Management in Crowdsourcing Systems’, in Proceedings of the 8th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing, Pittsburgh, United States, 2012. doi: 10.4108/icst.collaboratecom.2012.250499.

[15] 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.

[16] 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.

[17] P. Burnap and M. L. Williams, ‘Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making’, Policy Internet, vol. 7, no. 2, pp. 223–242, 2015.

[18] 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. Cham: Springer International Publishing, 2018, pp. 141–153. doi: 10.1007/978-3-319-76941-7_11.

[19] F. Miró-Llinares, A. Moneva, and M. Esteve, ‘Hate is in the air! But where? Introducing an algorithm to detect hate speech in digital microenvironments’, Crime Sci., vol. 7, no. 1, p. 15, Dec. 2018, doi: 10.1186/s40163-018-0089-1.

[20] M. Mondal, L. A. Silva, and F. Benevenuto, ‘A Measurement Study of Hate Speech in Social Media’, in Proceedings of the 28th ACM Conference on Hypertext and Social Media - HT ’17, Prague, Czech Republic, 2017, pp. 85–94. doi: 10.1145/3078714.3078723.

[21] ‘Crowdsource by Google’. https://crowdsource.google.com/ (accessed Nov. 21, 2022).

[22] ‘Digital 2022: Sri Lanka’, DataReportal – Global Digital Insights. https://datareportal.com/reports/digital-2022-sri-lanka (accessed Nov. 02, 2022).

[23] T. Davidson, D. Warmsley, 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.

[24] 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, 2016, pp. 88–93.

[25] 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, Oct. 2018, pp. 11–20. doi: 10.18653/v1/W18-5102.