Tackling Hate Speech in Low-resource Languages with Context Experts

DANIEL NKEMELU, Georgia Institute of Technology, USA
HARSHIL SHAH, Georgia Institute of Technology, USA
IRFAN ESSA, Georgia Institute of Technology, USA
MICHAEL L. BEST, Georgia Institute of Technology, USA

Given Myanmar's historical and socio-political context, hate speech spread on social media have escalated into offline unrest and violence. This paper presents findings from our remote study on the automatic detection of hate speech online in Myanmar. We argue that effectively addressing this problem will require community-based approaches that combine the knowledge of context experts with machine learning tools that can analyze the vast amount of data produced. To this end, we develop a systematic process to facilitate this collaboration covering key aspects of data collection, annotation, and model validation strategies. We highlight challenges in this area stemming from small and imbalanced datasets, the need to balance non-glamorous data work and stakeholder priorities, and closed data sharing practices. Stemming from these findings, we discuss avenues for further work in developing and deploying hate speech detection systems for low-resource languages.

CCS Concepts: • **Human-centered computing → Social networking sites**; **Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: hate speech, context experts, digital threats, democracy, low-resource text classification

ACM Reference Format:
Daniel Nkemelu, Harshil Shah, Irfan Essa, and Michael L. Best. 2022. Tackling Hate Speech in Low-resource Languages with Context Experts. In . ACM, New York, NY, USA, 18 pages.

1 INTRODUCTION

The rapid adoption of social media in Myanmar has been accompanied by a surge in the dissemination of problematic content such as hate speech, disinformation, and misinformation [53]. While platforms like Facebook offer opportunities for people to connect online, do business, and participate in online activism [10], they have also seen use as a medium for inciting violence against minority groups from online actors [64, 72]. In 2018, a United Nations independent fact-finding report highlighted the role social media, specifically Facebook, played in spreading hate speech and disinformation that led to a genocide of the Rohingya ethnic minority group in Myanmar. Lee [56] discussed the role of citizen-generated posts and state media-led publication outlets in spreading anti-minority rhetoric that influenced violent narratives about the Rohingya on social media. Violence towards the Rohingya community spurred by the proliferation of hate speech and disinformation on Facebook [34] led to the murder of thousands of civilians, creating almost a million refugees [21, 95]. Myanmar presents a chilling yet increasingly familiar account of the weaponization of Facebook in a nation with a history of armed conflict, authoritarianism, and censorship.

Manuscript submitted to ACM

Outside Myanmar, similar challenges with hate speech have been witnessed in countries like Ethiopia [38] and Sri Lanka [98]. There is a dearth of work investigating effective strategies for real-world hate speech detection in low-resource languages. Current strategies for tackling hate speech in low-resource contexts primarily entail two main steps: exploring the use of sophisticated machine learning tools for detecting hate speech and contracting human content moderators to flag, demote, and ultimately remove problematic content. Both approaches present notable limitations. Machine learning systems require ground truth data and data processing capabilities that are not readily available with low-resource languages. In addition, the vast amount of social media content produced daily makes it infeasible to engage human trackers to detect every instance of hate speech except the most prominent. Even if this was possible for a single context, it is not trivial to scale to new contexts, languages, and countries.

We seek to address this need by exploring a community-driven approach for tackling hate speech in low-resource language settings. This approach involves working with *context experts* on the entire machine learning project pipeline: scoping the project with a local partner focused on issues related to digital threats to democracy, assessing hate speech definitions and guidelines in tandem with legal experts, and working with paid volunteers to generate quality data, train, and validate machine learning models. We use the term *context experts* to highlight their role not merely as language translators but as experts with deep and personal knowledge of the context resulting from their lived experience. We interchangeably use the terms annotator and context experts in some parts of this paper. We only refer to the context experts as annotators when maintaining standard language for data labeling tasks.

This paper offers a report on our remote study of hate speech in Myanmar. In the months leading up to the 2020 Myanmar national elections, we worked remotely, due to the COVID-19 pandemic, with context experts to curate a dataset of 226 Burmese hate speech posts from Facebook through its CrowdTangle API service [86]. Our work contributes to research in machine learning for development (ML4D) seeking to understand ways to tackle hate speech on social media. We develop a process for coordinating machine learning work within low-resource language settings and show that working with context experts offer a key solution to the problem of hate speech in these settings. We also provide an early look into downstream classification tasks for the Burmese language using classical and neural-network-based machine learning models.

2 BACKGROUND

This section provides some background on Myanmar's political history and the evolution of hate speech targeted explicitly against the Rohingya community. We broadly discuss hate speech detection algorithms and datasets in resource-rich and low-resource languages and highlight Burmese natural language processing work.

2.1 Social Media in Myanmar and Hate Speech

After gaining independence in January 1948, Myanmar's over 14 years-long democratic government was interrupted by a military coup and subsequent dictatorship in 1962 that lasted almost 50 years. The military at the time pushed against calls for autonomy by non-Burman ethnic groups, which they labeled as anti-nationalist and anti-unity. For most of its years post-independence, Myanmar has faced many ethnic and religious conflicts and wars. When Myanmar re-transitioned to democratic rule in 2011, these conflicts remained persistent [101]. The government was dominated by the majority Bamar Buddhists who had exclusive control of military and civilian institutions, despite the country's cultural and linguistic diversity. This unequal power share led to the marginalization of ethnic minority groups resulting in both armed conflict and non-violent political actions [84, 87].

An estimated 1 million Rohingyas living in the Rakhine state have historically faced discriminatory practices from the military government, including but not limited to restricted access to education, employment, and citizenship identity cards [34, 101]. These practices appeared to have worsened since the transition to democracy. Lee [55] pointed to the liberalization of media and political freedom stemming from the transition as a motivating factor that amplified political polarization, which fuelled the agendas of ultra-nationalist anti-Rohingya Buddhist groups.

McLaughlin [64] discusses how violence against the Rohingya was preceded by the viral spread of hateful rhetoric and disinformation on Facebook primarily targeted against Muslims. Facebook had experienced massive growth in adoption between 2016 and 2017 through its "free basics" program, which allowed users to sign up for a free, limited Facebook version without a mobile internet plan [9, 64]. According to Fink [34], Facebook ignored several warnings by local rights and technology civil society organizations to act on dangerous speech posted on the platform calling on Buddhists to pick up arms in preparation for Muslim attacks and vice versa in 2017. The United Nation's independent fact-finding mission on Myanmar confirmed the significant negative role of hate speech and disinformation spread on Facebook played in heating the polity [21]. The platform has struggled to effectively moderate content in Myanmar's context due to the non-locally resident nature of its moderation system. The escalation of conflict in August 2017 lasted more than two months and led to over 750,000 fleeing as refugees [21].

As Myanmar geared for its national election in late 2020, observers were concerned about the possibility of online actors exploiting existing distrust of both the government and media in the country to foster violent responses to election results. This worry prompted opportunities for local and international civil society organizations to explore localized hate speech tracking and mitigation projects. It is within this socio-political context that this work is situated.

2.2 Automated Hate Speech Detection

Earlier works in hate speech detection have mostly leveraged a keyword-based approach that relied on the presence of a derogatory term typically used in hate speech to make decisions about whether a post is hate speech or not, e.g. [51]. However, keyword-based approaches have been shown to offer little performance value [58, 78]. More sophisticated computational approaches for tackling online hate speech have gained attention in recent years, and machine learning techniques have since been applied to the task of hate speech detection [78, 80, 96]. Prior work have explored bag-of-words, word-, and character-level n-grams features [65, 73, 97] and TF/IDF weighted embedding methods [22], with algorithms such as support vector machines [60], balanced random forests[13], and logistic regression models [22]. Recent works have adopted neural network-type approaches such as convolutional neural networks (CNNs) [39], CNNs combined with a Gated Recurrent Unit (GRU) network [99], and Transformer-based models such as the Bidirectional Encoder Representations from Transformers (BERT) [69, 70].

Data is a critical resource for determining performance, robustness, and scalability in machine learning systems [43]. To this end, several hate speech datasets have been released by researchers working in this area [14, 22, 37, 96, 97]. In their survey of the hate speech detection literature, Fortuna and Nunes [36] found the majority of the datasets to be in English, with few exceptions in Dutch [88], German [77], and Italian [25]. These datasets range in size from as small as 36 tweets [8], to as large as 150,000 multimodel (image-text) tweets [41] sourced from publicly available internet data.

A large proportion of the works mentioned earlier (e.g., [15, 37]) leverage crowdsourcing platforms like Figure-Eight (formerly Crowdflower) [1], or use publicly sourced comments online as ground truth (e.g., Warner and Hirschberg [94] used several thousand comments from Yahoo!). In some cases, the research team labels the data themselves. We note that several of these resources may not be available in many low-resource contexts. First, crowdsourcing platforms do not have universal coverage across languages and geographic regions. Even if they did, finding the right annotators for

a hate speech task that requires knowledge of social contexts can be challenging. There is also not a broad diversity of web platforms in low-resource language settings to scrape potential ground truth data. Most times, people who understand the language and the socio-political context may not be the researchers themselves, thus requiring the kind of collaboration we explore in this paper.

2.2.1 Low-Resource Hate Speech Detection Data. We refer to low-resource strictly within the context of limited availability of technical resources such as labeled training data; linguistic tools for tasks such as semantic analysis, named-entity recognition, and parts of speech tagging; or digitized texts that can serve as supervised/unsupervised training data for language models. Researchers have also explored the task of hate speech detection in these contexts. Mubarak et al. [71] studied the use of abusive language in the Arabic language on Twitter. Ishmam and Sharmin [45] explored machine learning approaches for classifying public Facebook posts in the Bengali language. Similar studies have been conducted in Amharic [68], Indonesian [4], Hindi [85], and Vietnamese [91]. Our investigation into these works shows that none point to the relevance of working with context experts as central to scaling hate speech detection in low-resource contexts. The authors also offered limited visibility into the data curation process, making it difficult to replicate it in new environments.

In their critical analysis of existing hate speech detection datasets, Madukwe et al. [59] highlight that several works do not make their data publicly available, making it difficult to benchmark. When provided on-request, the data may suffer from data degradation—a case where a dataset re-generated on demand by the researcher no longer produces the same amount or quality of data as at the time of publication. While data collection and sharing are vital for scientific progress, we acknowledge that authors cannot often do so for several reasons, including privacy concerns, platform restrictions, and the potential dissemination of harmful content. Anane-Sarpong et al. [5] discuss how a host of structural, organizational, cultural, and ethical complexities influence a researcher's decision to share their data. For instance, according to CrowdTangle's terms of use, we do not have Facebook's permission to share our dataset from Myanmar. However, we detail the steps we undertook for the hate speech dataset curation and provide materials for future use in new contexts and possible replication studies.

2.3 Burmese Language Processing

Burmese is the official language of Myanmar and the native language of the Bamar people. It is a largely monosyllabic and analytic language with a subject-object-verb word ordering and belongs to the Sino-Tibetan family. Like Chinese, Burmese morphemes can be combined freely with no changes [46]. Ding et al. [28] describes a challenge that arises for Burmese language processing because the boundaries of what implies a "word" are not clearly defined. This challenge emerges because Burmese has no specific rule or convention on how spaces separate words. Traditional Burmese does not use white spaces to mark word boundaries. However, modern variants of Burmese do use white spaces between phrases to improve readability.

Another significant challenge is the lack of a consistent, standardized font encoding. The most widely-used Burmese font for reading and writing on modern computers and smartphones is the Zawgyi font. Zawgyi is not defined as a standard character encoding and is not part of the standard Unicode character set. As a result, it is not typically built into major operating systems. Nonetheless, Facebook supports Zawgyi as an optional encoding on their app and website, and this option is widely used in Myanmar.

The Burmese language is referred to as part of a class of low-resource languages due to the limited availability of tools, datasets, and studies focused on the language. A series of recent works have been published, mostly in the past five

years, laying the groundwork for Burmese natural language processing. These include morphological analyses such as syllable-based tokenization [27], part-of-speech tagging [29], word segmentation [28], named-entity transliteration [67], and the development of a Burmese treebank [30] as part of the Asian Language Treebank Project [76]. Burmese has also featured as a constituent language in monolingual [47] and multilingual [19] learning tasks for language models. This work provides an early look into downstream classification tasks for automatic hate speech detection in the Burmese language using classical and neural network-based machine learning models.

3 AUTOMATIC HATE SPEECH DETECTION WITH CONTEXT EXPERTS

Most automatic hate speech detection systems rely on machine learning algorithms trained on existing ground truth data. However, the resources needed to facilitate this task are limited to a small set of languages and organizations. Similarly, publicly-available pretrained large language models only work on a handful of languages. This limitation implies that progress in automatic hate speech detection tends to follow a top-down approach where only privileged languages gain engineering attention. To balance this disproportion, we aim to design a collaborative process with context experts in low-resource language settings to scope the problem of hate speech detection and develop machine learning models to address them.

To concretize this collaboration, we focus on the task of developing machine learning models in the Burmese language to automatically detect hate speech posted on social media within the context of the Myanmar general election. A process approach helps us temporally order the crucial aspects of this work to support replication, reuse, and revalidation. By developing this process, we hope to provide a recipe for practitioners and researchers interested in collaborative work that addresses hate speech in low-resource language settings. Our approach consists of the following four key steps:

(1) establish partnerships to co-design project scope, identify and recruit paid volunteers.
(2) contextualize hate speech definitions and annotation guidelines for local relevance with legal experts.
(3) generate quality data and train machine learning models.
(4) validate trained machine learning models, and iterate step (3).

To facilitate this process, context experts take on two roles. First, they serve as technologists, actively contributing to shaping the overall direction of our research, and subsequently as model validators, identifying opportunities for improving the machine learning system.

3.1 Context Experts as Technologists

Attygale [6] discusses the idea of engaging with context experts as an intentional process of co-creating solutions in partnership with people who know the opportunities for and barriers to impact through their own experiences. According to Attygale [6], context experts offer perspectives that add depth and breadth to the technical expertise of "content experts" (in our case, machine learning researchers). In this work, we adopt context experts to refer to Burmese natives working in civil society and with an expressed interest in their nation's political system. **We argue that sustainably addressing hate speech in low-resource language settings involves empowering context experts with the technology-driven tools needed to tackle the problem locally**. The goal is to center the voices of the context experts from the onset of the project and ensure they have a sense of ownership. This type of collaborative model is often broadly described as community consultation, deliberation, engagement, or participation [75]. It differs from simply relying on them for feedback or buy-in, as in a contextual inquiry.

3.1.1 Recruitment and Training. Our partners in Myanmar advertised recruitment calls on public social and political Facebook pages. These pages were selected due to the number of accounts following the page and its focus on social and political issues. The call sought people who were active on Facebook and interested in the Myanmar's political system. The earliest 12 respondents were interviewed on a roll-in basis. The interviews were conducted by our resident project manager and focused on the interviewees' knowledge of relevant political issues and their ability to use a computer. After reviews and interviews, eight people were recruited. Five of them identified as female and three identified as male. All the context experts were native Burmese speakers and resident in Myanmar. The context experts attended an initial virtual session to connect and discuss the motivation for the project concerning the forthcoming Myanmar national elections. A subsequent training session to prepare the context experts for the data curation process necessary for the task. We provided labeling guidelines in the English and Burmese languages. The training was facilitated by an experienced project manager, an expert in Myanmar politics. The facilitator conducted the training session in the Burmese language through a video conference call that lasted for 90 minutes. We then created a shared workspace to facilitate communication between researchers and context experts. The context experts were monetarily compensated for their work throughout the project.

3.1.2 Hate Speech Definition. There is no universally agreed-on definition for hate speech, and what counts as hate speech in one context may not be considered hate speech in another. This presents a challenge because accurate data annotation requires a standardized framework for consistency and reliability. We aimed to adopt a definition broad enough to potentially cover all instances of hate speech relevant to Myanmar, but was specific enough to avoid ambiguous interpretations by the annotators. To do this, we explored definitions provided by the UN, social media platform companies, and those widely used in hate speech research.

According to the United Nations Strategy and Plan of Action on Hate Speech [42], hate speech is *"any communication in speech, writing, or behavior, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender or other identity factor"*. Researchers have developed definitions such as *"language used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group."* [22] or *"a deliberate attack directed towards a specific group of people motivated by aspects of the group's identity"* [23].

Platforms often provide specific categories within which a person or group may be a target of hateful speech. For example, Facebook defines hate speech as *"a direct attack against people — rather than concepts or institutions— on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease"* [32] and Twitter defines it as *"promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease."* [90]. Since Myanmar had not adopted a national definition for hate speech, we adapted a description for our annotation guideline composed of the UN's hate speech definition and included protected characteristics mentioned by Facebook. .

3.1.3 Annotation Guidelines. Together with the context experts, we developed an annotation procedure designed to help standardize the annotation process. The annotation guideline for the task was made available in the English and Burmese languages. It contained our adapted definition of hate speech. We emphasized that hate speech must constitute dehumanizing or demeaning sentiment and be expressed because of who the target is based on some protected characteristics. We further added that hate speech may be directed towards an individual or a group [31] and could be explicit or implicit—the implicit case requiring an understanding of the context. We added a link to the original

posts from the annotation files to enable this context inquiry. We highlighted instances when a speech is not regarded as hate speech, such as defamatory speech that does not invoke a protected characteristic or benign attacks against government policy. For example, a post saying *"The Burmese military is corrupt!"* might be an attack on the military's integrity but does not constitute hate speech since the military is not a protected group. Annotators were provided with examples of posts that were hate speech and not hate speech, according to the guidelines. To validate our annotation guideline, we shared a copy with a team of legal experts based in Yangon, Myanmar, for feedback and edits. The experts provided three main feedback to improve the guideline.

i The guideline should use an exhaustive list of protected characteristics instead of sampling of few examples that can leave too much room for annotator subjectivity.
ii The guideline should clarify what action to take on hate speech regarding political holders. While political holders are protected from aforementioned attacks to protected characteristics, speech that attacks political decisions or ideology of the office holder is not hate speech.
iii The annotator training should acknowledge the inherent challenges with defining an "attack" or a "demeaning post", especially for cases where hate speech is implicit. Since a post can be considered an attack or demeaning based on the perception of the recipient, there could be potential uncertainty in the labelled data.

We edited our guidelines to address the concerns raised in (i) and (ii). For (iii), we relied on the competence of the context experts to reduce uncertainty. A final copy of our annotation guideline is provided in our Github repository.[1]

3.1.4 Data Collection. Using Hatebase [44], an online repository of multilingual hate speech terms, we retrieved 128 crowdsourced Burmese hate terms. Three context experts manually inspected and removed 56 of the keywords considered overly context-sensitive and could potentially result in many benign posts. This inspection reduced the number of Burmese hate terms from Hatebase to 72. During the data collection phase, Phandeeyar [2], a technology organization based in Myanmar, released a lexicon of 88 hate terms resulting from their work tracking COVID-19 related hate speech on social media. We checked for possible appearances of the same words within the Hatebase and Phandeeyar lexicon sets. We found only two occurrences of an exact match. Another five were cases where a hate term in one set is the subset of another term in the other set. We combined both lexicons for a total of 158 hate terms for the next step. The list is also provided in our Github repository.

Next, we use CrowdTangle [86], a public insights service provided by Facebook which enables access to public groups and pages, to retrieve posts containing any of these select hate terms. Together with the context experts, we curated a dashboard of social and political pages in Myanmar using a combination of direct Facebook searches and local knowledge of popular social and political groups and pages. To identify new groups and pages, we snowball from already known pages to new pages that interact with them via shares or mentions. We downloaded historical Facebook posts from the CrowdTangle dashboard between October 2018 and June 2020 that contain any words from our hate lexicon. The downloaded data contained 43,996 posts.

As a preprocessing step, we removed posts containing only URLs or blank shares of other posts. Next, we removed duplicate posts. This was a majority of the posts because CrowdTangle periodically provided updated interaction metrics for the same post leading to duplicates. We retained only the most recent versions of each post. We then randomized the posts in the dataset to remove any consecutive posts from the same account, group, or page. This process would help decrease anchoring bias in the annotations, which emerges when an annotator reads different posts from the same source and may become likely to label them similarly [89]. Finally, we removed posts with less than three syllables

[1] https://github.com/TID-Lab/myanmarhsc

using the Myanmar word segementer available as part of the Myanmar language tools package [2]. These posts will be too short and lack sufficient context for accurate labeling and feature engineering. After these preprocessing and filtering steps, 5,646 posts remained, which we proceeded to label.

3.1.5 Labelling Plan. Our recruited context experts understood Myanmar's socio-political terrain but were not necessarily hate speech experts. We were concerned that one remote training session might be insufficient to prepare them adequately, so we adopted a pairing strategy to boost annotator agreement. In this strategy, each annotator initially receives the same set of posts to label as one other annotator. Each annotator is asked to label the posts assigned to them independently. After labeling their initial batch of posts, they are then asked to schedule a call with their partner and the training facilitator to discuss their experience labeling the posts and address areas in which they disagreed. Each annotator was provided with a laptop and an internet connection to facilitate these conversations. There were eight annotators in our setup, resulting in four pairs. Each annotator received 100 posts per day over four days. After the fourth day, the remainder of the posts were divided equally amongst the annotators for individual labeling. Table 1 shows the number of posts labeled as Yes or No for each annotator, grouped with their paired annotator.

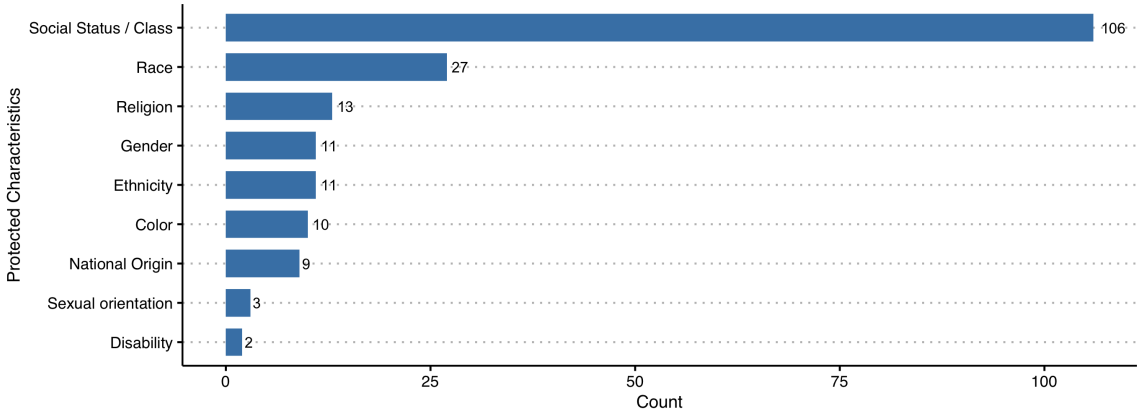

Fig. 1. Protected characteristics invoked in posts labelled as hate speech.

3.1.6 Labelling Result. After labeling, the final dataset contained 225 unique instances of posts labeled as hate speech. Figure 1 shows the distribution of protected characteristics identified by the annotators as invoked in the post. We find that this is consistent with our expectations given Myanmar's socio-political scene. To measure the reliability of annotations, researchers rely on a measure of inter-annotator agreement to quantify the level of overlap among annotators on the labels they have chosen for each sentence. However, these scores such as Fleiss' [35] and Cohen's [18] kappas have been known to be affected by annotator bias and class imbalance. There is also no widely accepted kappa level for determining sufficient reliability [77]. As a result, most hate speech research report very low agreement scores or do not report this value. For example, a batch of 100 posts given to one pair of annotators may include only one hate speech and 99 non-hate speech posts. If one annotator returns all 100 as not hate speech while the other accurately returns 99 not hate speech and one hate speech, their Cohen's kappa is 0.0 even though they agree 99% of the time. Though percent agreement has been critiqued on the basis that it does not account for random agreements or guesses

[2]https://github.com/MyanmarOnlineAdvertising/myanmar_language_tools

due to inadequate annotator training [63], we adopt it as our measure of annotator improvement for our use case. We limit annotator guessing by training annotators extensively and setting up iterative peer feedback sessions at the end of each cycle.

We incorporated this peer learning model as part of the annotator training process to facilitate knowledge sharing where some annotators have some historical or current affairs context of some posts and build annotator's confidence from learning how much they are in sync with a peer. Figure 2 shows how percentage agreement among pairs of annotators increased after the first peer feedback session. Annotators were especially encouraged to discuss areas of disagreement. To avoid a situation where a pair is consistently wrong, we asked each pair to meet remotely with the training facilitator to discuss what they learned from that batch and confirm their results. By the fourth iteration, annotator agreement had increased from an average of 68.75% to 93%.

Table 1. Number of posts labelled as Yes or No by each annotator.

Step	Pair 1				Pair 2				Pair 3				Pair 4			
	A1		A2		A3		A4		A5		A6		A7		A8	
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No
1	60	40	76	24	59	41	13	87	10	90	17	83	17	83	34	66
2	19	81	4	96	3	97	2	98	5	95	2	98	0	100	13	87
3	2	98	4	96	1	99	1	99	4	96	1	99	5	95	1	99
4	12	88	4	96	4	96	6	96	9	91	6	94	2	98	0	100

3.1.7 Text classification. We designed a four-stage data preprocessing pipeline: first, we converted the Zawgyi character to Unicode. We used an open-source Myanmar language tools library [49] for this step. Next, we removed posts predominantly in languages that are not Burmese. We removed emojis. We retained instances of code-switching between English and Burmese. In addition, since the Burmese language does not consistently use white spaces to mark word boundaries, we used the Myanmar language tools [49] library to perform word-level segmentation on the text splitting each post into tokens. Finally, we used an openly available Burmese stop words list [62] to remove stop words from each remaining post in our dataset.

3.1.8 Classification Models. We employed classical machine learning classification algorithms (Support Vector Machines [20], Balanced Random Forests [17], and FastText [11]) to conform to similar constraints in contexts where models such as pretrained Transformers are unavailable. We ran several feature combinations of n-grams, term-frequency weighting, and hyper-parameter searches to select the best combination. We show precision, recall, and F-1 scores for the models in Table 2.From Table 2, we observe that the FastText model (with oversampling) performed best across all three metrics.

3.2 Context Experts as Model Validators

Beyond the cross-validation techniques discussed in the previous section, we sought to validate our model qualitatively to understand which cases led to model error and why. There are limited tools for performing these kinds of model validation [40]. For our analysis, we collected live data from Facebook different from those in our training and tests set and passed these data for inference on our best-performing model. We provided a sample of the data and asked our context experts to label them. We then show the model decisions to the context experts and discuss areas where they disagreed with the model.

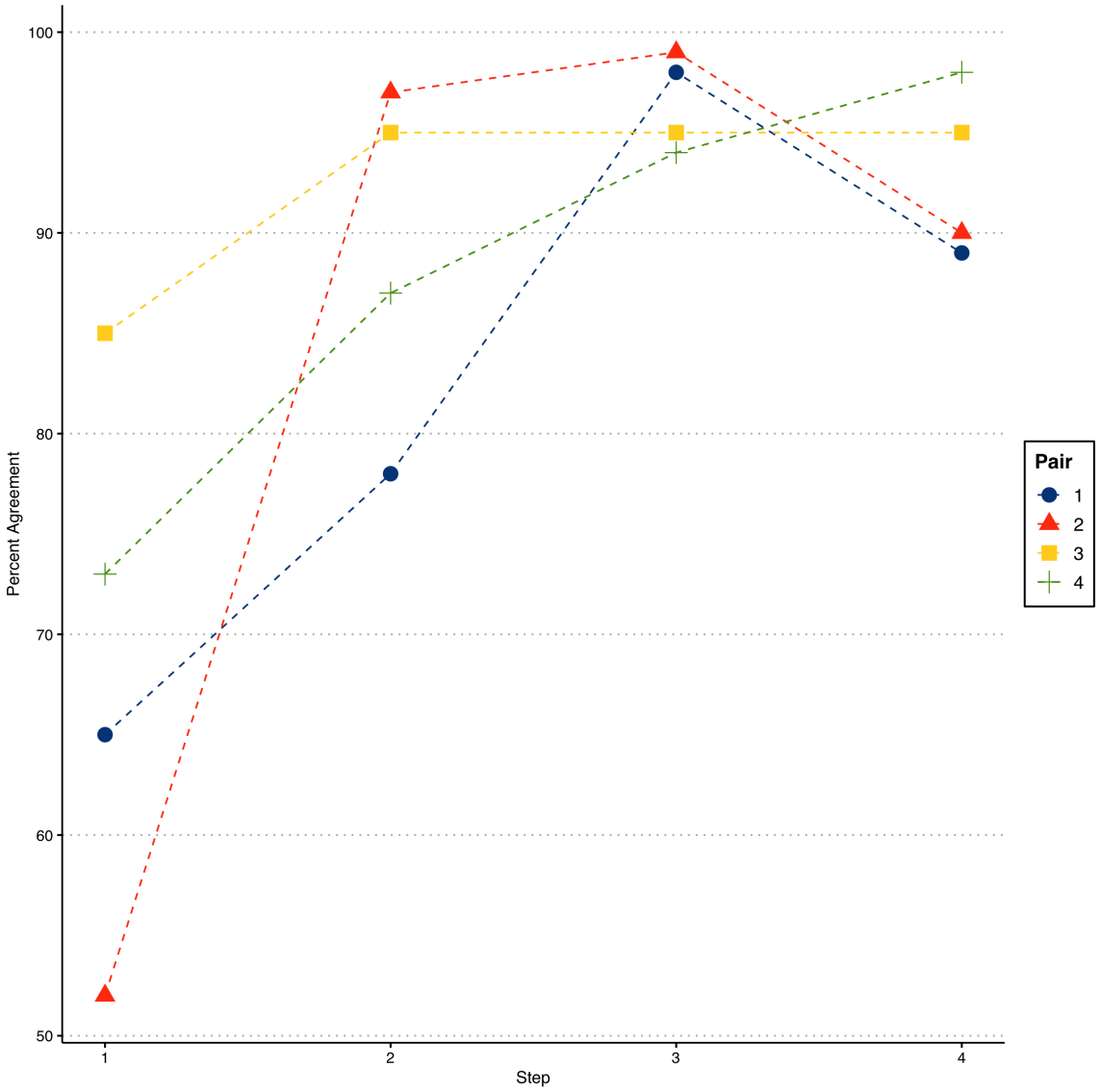

Fig. 2. Percentage agreement within pairs of annotators for each training stage.

We identified two major error types from the model from this step. First, false-positive cases where benign posts containing terms in the hate lexicons or words typically used in hate speech posts and highly represented in the training data were wrongly flagged as hate speech by the model. Second, false-negative cases were hateful posts that did not contain the archetypal hate terms were not classified as hate speech by the model.

Table 2. Model performance for hate speech classification

Model	Precision	Recall	F1-score
SVM	0.48	0.50	0.49
SVM (with oversampling)	0.88	0.87	0.87
BRF	0.53	0.70	0.45
BRF (with oversampling)	0.88	0.87	0.87
FastText	0.84	0.63	0.69
FastText (with oversampling)	0.93	0.92	0.92

4 STUDY FINDINGS & DISCUSSION

In this section, we discuss our findings from the collaborative model with context experts, which we have described in the previous section. While we drew most of these observations from our process in Myanmar, they demonstrate similar issues in other low-resource language settings where machine learning may be applied. An appreciation of these findings is crucial for designing efficient machine learning systems. We have identified three central issues: (i) exploring avenues to boost ground truth data, (ii) incentivizing data work for machine learning, and (iii) engendering open data sharing practices.

4.1 Exploring avenues to boost ground truth data

A major challenge with hate speech detection tasks in low-resource language settings is limited training data in target languages. This limitation is often due to the rarity of hate speech in proportion to the amount of available social media data and the infeasibility of labeling the entire dataset on a given platform [59]. Much of the progress in the field of machine learning has been driven by the availability of benchmark datasets such as ImageNet [26] for computer vision, and GLUE [93] for natural language processing. While such datasets have sprung up for hate speech detection in the English language [61, 97], researchers often filter out data not in English, leaving more work to do for very low-resource languages. This lack of training data can hamper machine learning research within these contexts. The ML4D community can adopt the process described in this work to bootstrap the development of hate speech training datasets in diverse contexts.

A caveat to note is that engaging in elaborate data collection and labeling efforts do not always guarantee significant outcomes. We found that only less than 4% of the entire Facebook posts in our dataset were labeled by annotators as hate speech. To boost training data, practitioners have relied on curated hate speech lexicons in the target languages. One example is the PeaceTech Lab Lexicons, a series of hate speech terms explaining inflammatory social media keywords and offering counter-speech suggestions to combat the spread of hate speech[**?**]. The PeaceTech Lab has curated hate lexicons for languages in conflict-affected countries such as the Democratic Republic of the Congo, Sudan, and Lebanon. Though keywords approaches are ineffective when used alone, they can help researchers select a subset of data to work with. Organizing civil society workshops or other avenues for crowdsourcing hate terms can be beneficial for supporting hate speech detection work.

Our findings reveal how quickly new hate terms emerge on social media. This dynamism is primarily due to changes in political and social concerns. For instance, online actors may derive new hate terms as social media discourse changes from talking about a local election to focusing on a pandemic [100]. Context experts can help identify when social media topics drift or when users find creative ways to guise hate speech.

4.2 Incentivizing data work for machine learning

Data work is "any human activity related to creating, collecting, managing, curating, analyzing, interpreting, and communicating data" [12]. This fundamental component of machine learning is often undervalued in research and practice [92] and can lead to negative outcomes because data is critical for effective machine learning systems [74]. Sambasivan et al. [79] discuss the challenges with data cascades in high-stakes AI, which they define as compounding events over time that result from the undervaluing of data quality by researchers and practitioners. These data cascades often have significant effects in low-resource language settings due to the lack of existing ground data and supporting infrastructure for data collection and processing. We identified some of the cascade factors identified by Sambasivan et al. [79] in our work, namely, incentives in AI and data education. First, we observed that while stakeholders and partners might identify with and value the role of AI in addressing the problem of low-resource hate speech, they often did not place similar consideration on the invisible and challenging data work. Second, context experts and partners lack experience in creating the kinds of complex and quality datasets that hate speech research requires. Stakeholders ought to see the value in data work to appreciate the time spent on training context experts.

At the onset of communications with sponsors and partners, it is crucial to emphasize the importance of data work for machine learning and the vital role that context experts will play. This step may imply a more costly engagement with context experts, both in time and financial compensation. We learned that a comprehensive data training plan is helpful at the start of the project to address the data education challenge.

4.3 Supporting open data sharing practices

Social media researchers in low-resource contexts do not often have the freedom to choose what platforms to work on. This decision is mainly driven by which platform is widely used and most likely to have relevant data within a given context. With over 21 million active users in 2019 [48], our work in Myanmar primarily focused on Facebook. Like any other platform, Facebook offers unique affordances and constraints for data access and sharing. For example, the company grants privileged data access to researchers via its CrowdTangle API service. Yet, its terms of use do not allow researchers to share any data with people outside CrowdTangle. Such practices make it challenging to freely collaborate with other teams who may also have access to data but are bound by the platform's terms.

These limitations posed by platforms motivate community-based data sharing workflows led by context experts and built on trust, ethics, and shared values. Data sharing has been discussed as an effective means for scientific progress, especially within developing countries [16, 81]. However, there are lots of structural, organizational, cultural, and ethical complexities that undermine data sharing [5]. Exploring data sharing practices that actively address power imbalances, understands potential benefits and risks, and comply with local norms and cultures [3] will help produce a critical mass of relevant data that can be useful for tackling low-resource hate speech.

Furthermore, significant aspects of a country's history are often undocumented digitally, only existing as paper documents within locked-out cabinets or as informal knowledge passed on from one generation to another. In some cases, only a fraction of the context experts know the several historical and cultural issues that underlie hateful speech posted online. Our annotation model addresses this knowledge imbalance with the peer feedback model discussed as part of our process.

5 CONCLUSION

There is an urgent need to identify ways to tackle hate speech on social media in low-resource language settings. We have argued that addressing this problem will require community-based approaches that combine a deep understanding of social and political contexts with automated tools to process the vast amount of content produced online. In this paper, we presented findings from our remote study on the automatic detection of hate speech on Facebook in Myanmar. We have developed a systematic process for collaborating with context experts covering critical stages of data collection, annotation, and model validation strategies. Our work offers insights for researchers and practitioners in machine learning for development (ML4D) and highlights challenges stemming from small and imbalanced datasets, the need to balance non-glamorous data work and stakeholder priorities, and closed data sharing practices. These findings motivate further research exploring strategies for data augmentation, non-text-based detection models, multimodal hate speech detection, and best practices for working with non-machine learning experts on machine learning-powered projects.

Ethics statement: We hope that this work can support efforts to protect the integrity of social media platforms and defend democracies against the threats of hate speech, disinformation, and misinformation. We understand that automated tools such as what we propose in this work could be misused to target and oppress dissenting voices, especially within authoritarian regimes. We unequivocally state that such use will be contrary to our core goal of offering ideas for combating harmful content online.

6 FUTURE WORK

We now outline some promising directions for future research work in this area addressing issues of limited ground truth data, non-text-based techniques, multimodality, and non-expert collaboration:

6.1 Data augmentation for low-resource hate speech detection

Data augmentation involves strategies for increasing training examples for machine learning tasks without explicitly collecting new data. Data augmentation has received recent attention in natural language processing research due to increased work in new domains with limited data and the need for substantial amounts of training data for large neural networks [33]. As our findings have shown, running entire data labeling schemes may not necessarily lead to more data, especially for tasks where true labels are scarce. Low-resource hate speech detection work can benefit from data augmentation strategies that generate new data to augment the sparsity in new contexts. This technique can take the form of generating entirely new corpora from validated ground truths in high resource settings or creating new variations of existing small data in the target language.

6.2 Exploring network-based modeling approaches

Current hate speech detection tools in low-resource language settings mostly rely on text-based user-generated data and less on other contextual information such as linked news sources, parallel comments from local news websites, platform metadata, and other social network data. Incorporating these additional contexts into existing models could improve single-instance hate speech detection and coordinated hate attacks that can be difficult to detect in real-time. One idea is to focus on understanding problematic actors within these contexts and identifying relationships that increase the likelihood of spreading hateful content. Some recent works have explored graph-based models for this task [7, 24, 66] but little is known about how these methods translate to low-resource contexts. A practical solution

could help address the challenges posed by limited training data and other vulnerabilities of text-based methods such as topic drift and adversarial attacks.

6.3 Multimodal hate speech detection

A vast amount of hate speech posts in low-resource language settings are multimodal, often containing a combination of images, audio content, videos, live streams, external links, etc. Models trained with data on a single modality might be insufficient to address the problem effectively. Presently, content moderators have to watch hours of videos in their native language to identify hateful rhetoric embedded in the videos before recommending take-downs. This process is challenging to scale, takes lots of moderation hours, and videos tend to spread faster than platform action. Image-text datasets such as MMHS150K [41] and the Hateful Memes [50] have been released in English for this task. Further work is needed for hate speech detection for multimodal content for low-resource contexts.

6.4 Active learning and uncertainty-aware predictions

Given established concerns over the scarcity of true labels and the prevalence of noisy data in the domain of hate speech detection, an active learning strategy might help boost model performance in practice. The idea is that a machine learning model can perform better with fewer labelled training instances if it can choose a reasonable sample to learn from [82]. This is especially useful for systems that are deployed in the wild where the machine learning model can provide strategic queries to human annotators and retrain on a valuable set of new training data. An uncertainty-aware sampling strategy can help ensure that the uncertainly resulting from the data imbalance is well-calibrated [57].

6.5 Working with non-experts on machine learning deployment projects

Machine learning work with non-experts requires a degree of care to avoid the trap of participation-washing where their involvement is either outrightly performative or only aimed at extracting free labor and consultation [83]. We are only beginning to scratch the surface of what forms effective AI collaboration with context experts may take. In this work, we have shown how we defined and scoped the problem with context experts, collected and labeled the data, and validated the model. Nonetheless, the technical gaps in the context experts' knowledge of machine learning implied that some parts of the modeling process were inadvertently black-boxed from them. This gap may present challenges for machine learning project sustainability, and further work is needed to understand ways to identify and mitigate the critical (people, process, and cultural [52]) failure factors that might hinder the long term success of AI deployments in low-resource contexts.

ACKNOWLEDGMENTS

Special thanks to our partners at The Carter Center for collaborating with us in Myanmar and establishing our partnership with the New Myanmar Foundation (NMF); to the NMF collaborating with us and serving as context experts despite challenges with the COVID-19 pandemic; and to Koe Koe Tech for their review and feedback on an initial version of the hate speech annotation guideline.

REFERENCES

[1] 2021. Confidence to Deploy AI with World-Class Training Data. https://appen.com/

[2] 2021. Myanmar Innovation Lab. https://phandeeyar.org/

[3] Rediet Abebe, Kehinde Aruleba, Abeba Birhane, Sara Kingsley, George Obaido, Sekou L Remy, and Swathi Sadagopan. 2021. Narratives and Counternarratives on Data Sharing in Africa. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 329–341.

[4] Ika Alfina, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. 2017. Hate speech detection in the Indonesian language: A dataset and preliminary study. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. IEEE, 233–238.

[5] Evelyn Anane-Sarpong, Tenzin Wangmo, Claire Leonie Ward, Osman Sankoh, Marcel Tanner, and Bernice Simone Elger. 2018. "You cannot collect data using your own resources and put It on open access": Perspectives from Africa about public health data-sharing. *Developing world bioethics* 18, 4 (2018), 394–405.

[6] Lisa Attygalle. 2017. The context experts. *Tamarack Institute* (2017).

[7] Matthew Beatty. 2020. Graph-Based Methods to Detect Hate Speech Diffusion on Twitter. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 502–506.

[8] Darina Benikova, Michael Wojatzki, and Torsten Zesch. 2017. What does this imply? Examining the Impact of Implicitness on the Perception of Hate Speech. In *International Conference of the German Society for Computational Linguistics and Language Technology*. Springer, 171–179.

[9] Michael L Best. 2014. The internet that Facebook built. *Commun. ACM* 57, 12 (2014), 21–23.

[10] Michael L Best. 2016. Mobile computing and political transformation. *Commun. ACM* 59, 10 (2016), 21–23.

[11] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.

[12] Claus Bossen, Kathleen H Pine, Federico Cabitza, Gunnar Ellingsen, and Enrico Maria Piras. 2019. Data work in healthcare: An Introduction. , 465–474 pages.

[13] Pete Burnap and Matthew L Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & internet* 7, 2 (2015), 223–242.

[14] Pete Burnap and Matthew L Williams. 2016. Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data science* 5 (2016), 1–15.

[15] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on web science conference*. 13–22.

[16] Winner Dominic Chawinga and Sandy Zinn. 2019. Global perspectives of research data sharing: A systematic literature review. *Library & Information Science Research* 41, 2 (2019), 109–122.

[17] Chao Chen, Andy Liaw, Leo Breiman, et al. 2004. Using random forest to learn imbalanced data. *University of California, Berkeley* 110, 1-12 (2004), 24.

[18] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[19] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116* (2019).

[20] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

[21] UN Human Rights Council. 2018. Report of the Independent International Fact-Finding Mission on Myanmar, A/HRC/39/64. (2018).

[22] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 11.

[23] Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444* (2018).

[24] Marco Del Tredici, Diego Marcheggiani, Sabine Schulte im Walde, and Raquel Fernández. 2019. You shall know a user by the company it keeps: Dynamic representations for social media users in nlp. *arXiv preprint arXiv:1909.00412* (2019).

[25] Fabio Del Vigna12, Andrea Cimino23, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*. 86–95.

[26] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[27] Chenchen Ding, Hnin Thu Zar Aye, Win Pa Pa, Khin Thandar Nwet, Khin Mar Soe, Masao Utiyama, and Eiichiro Sumita. 2019. Towards Burmese (Myanmar) morphological analysis: Syllable-based tokenization and part-of-speech tagging. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 19, 1 (2019), 1–34.

[28] Chenchen Ding, Ye Kyaw Thu, Masao Utiyama, and Eiichiro Sumita. 2016. Word Segmentation for Burmese (Myanmar). *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 15, 4 (2016), 1–10.

[29] Chenchen Ding, Masao Utiyama, and Eiichiro Sumita. 2018. NOVA: A feasible and flexible annotation system for joint tokenization and part-of-speech tagging. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 18, 2 (2018), 1–18.

[30] Chenchen Ding, Sann Su Su Yee, Win Pa Pa, Khin Mar Soe, Masao Utiyama, and Eiichiro Sumita. 2020. A Burmese (Myanmar) treebank: Guideline and analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 19, 3 (2020), 1–13.

[31] Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.

[32] Facebook. 2021. *Objectionable Content | Community Standards*. Retrieved November 30, 2021 from https://www.facebook.com/communitystandards/objectionable_content

[33] Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075* (2021).

[34] Christina Fink. 2018. Dangerous speech, anti-Muslim violence, and Facebook in Myanmar. *Journal of International Affairs* 71, 1.5 (2018), 43–52.

[35] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[36] Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)* 51, 4 (2018), 1–30.

[37] Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

[38] Iginio Gagliardone, Matti Pohjonen, Zenebe Beyene, Abdissa Zerai, Gerawork Aynekulu, Mesfin Bekalu, Jonathan Bright, Mulatu Moges, Michael Seifu, Nicole Stremlau, et al. 2016. Mechachal: Online debates and elections in Ethiopia-from hate speech to engagement in social media. *Available at SSRN 2831369* (2016).

[39] Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online*. 85–90.

[40] Jerry Gao, Chuanqi Tao, Dou Jie, and Shengqiang Lu. 2019. What is AI software testing? and why. In *2019 IEEE International Conference on Service-Oriented System Engineering (SOSE)*. IEEE, 27–2709.

[41] Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1470–1478.

[42] António Guterres et al. 2019. United Nations Strategy and Plan of Action on Hate Speech. *no. May* (2019), 1–5.

[43] Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The unreasonable effectiveness of data. *IEEE Intelligent Systems* 24, 2 (2009), 8–12.

[44] HateBase20. 2020. . Retrieved June, 2020 from www.hatebase.org

[45] Alvi Md Ishmam and Sadia Sharmin. 2019. Hateful speech detection in public facebook pages for the bengali language. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. IEEE, 555–560.

[46] Mathias Jenny et al. 2017. *Burmese: A comprehensive grammar*. Routledge.

[47] Shengyi Jiang, Xiuwen Huang, Xiaonan Cai, and Nankai Lin. 2021. Pre-trained Models and Evaluation Data for the Myanmar Language. In *International Conference on Neural Information Processing*. Springer, 449–458.

[48] Simon Kemp. 2019. *We Are Social*. Retrieved October 19, 2021 from https://datareportal.com/reports/digital-2019-myanmar

[49] Aye Hnin Khine, Lynn Nyan Htut, Mon Ye Kyaw, Htet Hein Aung, and Moe Seth. 2020. Myanmar Language Tools. https://github.com/MyanmarOnlineAdvertising/myanmar_language_tools.

[50] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *arXiv preprint arXiv:2005.04790* (2020).

[51] April Kontostathis, Kelly Reynolds, Andy Garron, and Lynne Edwards. 2013. Detecting cyberbullying: query terms and techniques. In *Proceedings of the 5th annual acm web science conference*. 195–204.

[52] Rajendra Kumar and Michael L Best. 2006. Impact and sustainability of e-government services in developing countries: Lessons learned from Tamil Nadu, India. *The Information Society* 22, 1 (2006), 1–12.

[53] Nyi Nyi Kyaw. 2019. Facebooking in Myanmar: From hate speech to fake news to partisan political communication. (2019).

[54]]peacetechlab PeaceTech Lab. [n. d.]. *Lexicons*. https://www.peacetechlab.org/toolbox-lexicons

[55] Ronan Lee. 2016. The Dark Side of Liberalization: How Myanmar's Political and Media Freedoms Are Being Used to Limit Muslim Rights. *Islam and Christian–Muslim Relations* 27, 2 (2016), 195–211.

[56] Ronan Lee. 2019. Extreme speech| extreme speech in Myanmar: The role of state media in the Rohingya forced migration crisis. *International Journal of Communication* 13 (2019), 22.

[57] David D Lewis and William A Gale. 1994. A sequential algorithm for training text classifiers. In *SIGIR'94*. Springer, 3–12.

[58] Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one* 14, 8 (2019), e0221152.

[59] Kosisochukwu Madukwe, Xiaoying Gao, and Bing Xue. 2020. In data we trust: A critical analysis of hate speech detection datasets. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*. 150–161.

[60] Shervin Malmasi and Marcos Zampieri. 2018. Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence* 30, 2 (2018), 187–202.

[61] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. *arXiv preprint arXiv:2012.10289* (2020).

[62] Zin Maung Maung and Yoshiki Mikami. 2008. A rule-based syllable segmentation of Myanmar text. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.

[63] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica* 22, 3 (2012), 276–282.

[64] Timothy McLaughlin. 2018. How Facebook's rise fueled chaos and confusion in Myanmar. *Retrieved* 2, 21 (2018), 2020.

[65] Yashar Mehdad and Joel Tetreault. 2016. Do characters abuse more than words?. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 299–303.

[66] Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. 2018. Author profiling for abuse detection. In *Proceedings of the 27th international conference on computational linguistics*. 1088–1098.

[67] Aye Myat Mon, Chenchen Ding, Hour Kaing, Khin Mar Soe, Masao Utiyama, and Eiichiro Sumita. 2020. A Myanmar (Burmese)-English Named Entity Transliteration Dictionary. In *Proceedings of the 12th Language Resources and Evaluation Conference*. 2980–2983.

[68] Zewdie Mossie and Jenq-Haur Wang. 2018. Social network hate speech detection for Amharic language. *Computer Science & Information Technology* (2018), 41–55.

[69] Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2019. A BERT-based transfer learning approach for hate speech detection in online social media. In *International Conference on Complex Networks and Their Applications*. Springer, 928–940.

[70] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on BERT model. *PloS one* 15, 8 (2020), e0237861.

[71] Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive language detection on Arabic social media. In *Proceedings of the first workshop on abusive language online*. 52–56.

[72] Karsten Müller and Carlo Schwarz. 2021. Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association* 19, 4 (2021), 2131–2167.

[73] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*. 145–153.

[74] Neoklis Polyzotis, Sudip Roy, Steven Euijong Whang, and Martin Zinkevich. 2018. Data lifecycle challenges in production machine learning: a survey. *ACM SIGMOD Record* 47, 2 (2018), 17–28.

[75] Srimal Isaac Ranasinghe. 2018. *The Engaged Community: Trust-Building within Public Engagement toward Community Development*. Master's thesis. Environmental Design.

[76] Hammam Riza, Michael Purwoadi, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thai, Vichet Chea, Sethserey Sam, et al. 2016. Introduction of the asian language treebank. In *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*. IEEE, 1–6.

[77] Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118* (2017).

[78] Haji Mohammad Saleem, Kelly P Dillon, Susan Benesch, and Derek Ruths. 2017. A web of hate: Tackling hateful speech in online social spaces. *arXiv preprint arXiv:1709.10159* (2017).

[79] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[80] Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*. 1–10.

[81] Malete Daniel Sebake et al. 2012. *Assessing the motivators and barriers of interorganizational GIS data sharing for address data in South Africa*. Ph. D. Dissertation. University of Pretoria.

[82] Burr Settles. 2009. *Active Learning Literature Survey*. Computer Sciences Technical Report 1648. University of Wisconsin–Madison.

[83] Mona Sloan, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2020. Participation is not a design fix for machine learning (pp. 1–7). In *Proceedings of the International Conference on Machine Learning, Vienna, Austria*.

[84] Martin Smith. 1991. *Burma: Insurgency and the politics of ethnicity*. Zed Books.

[85] K Sreelakshmi, B Premjith, and KP Soman. 2020. Detection of hate speech text in Hindi-English code-mixed data. *Procedia Computer Science* 171 (2020), 737–744.

[86] CrowdTangle Team. 2020. CrowdTangle. Facebook, Menlo Park, California, United States.

[87] Ardeth Maung Thawnghmung. 2011. *Beyond armed resistance: ethnonational politics in Burma (Myanmar)*. Honolulu: East-West Center.

[88] Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. 2016. A dictionary-based approach to racism detection in dutch social media. *arXiv preprint arXiv:1608.08738* (2016).

[89] Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science* 185, 4157 (1974), 1124–1131.

[90] Twitter. 2021. *Hateful conduct policy | Help Center*. Retrieved November 30, 2021 from https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

[91] Tin Van Huynh, Vu Duc Nguyen, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen, and Anh Gia-Tuan Nguyen. 2019. Hate speech detection on vietnamese social media text using the bi-gru-lstm-cnn model. *arXiv preprint arXiv:1911.03644* (2019).

[92] Kiri Wagstaff. 2012. Machine learning that matters. *arXiv preprint arXiv:1206.4656* (2012).

[93] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461* (2018).

[94] William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*. 19–26.

[95] Alex Warofka. 2018. An independent assessment of the human rights impact of Facebook in Myanmar. *Facebook Newsroom, November* 5 (2018).

[96] Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*. 138–142.

[97] Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*. 88–93.
[98] Yudhanjaya Wijeratne. 2018. The control of hate speech on social media: Lessons from sri lanka. *CPR South* (2018).
[99] Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European semantic web conference*. Springer, 745–760.
[100] Caleb Ziems, Bing He, Sandeep Soni, and Srijan Kumar. 2020. Racism is a virus: Anti-asian hate and counterhate in social media during the covid-19 crisis. *arXiv preprint arXiv:2005.12423* (2020).
[101] Min Zin. 2015. Anti-Muslim Violence in Burma: Why Now? *social research* 82, 2 (2015), 375–397.

A APPENDIX

This Github link contains data on the Burmese hate lexicons (Hatebase.org and Phandeeyar), target protected characteristics, annotation guideline and plan: https://github.com/TID-Lab/myanmarhsc.