An Exploration of Online Toxic Content Against Refugees

Arjun Arunasalam[†*], Habiba Farrukh[‡*], Eliz Tekcan[†*], and Z. Berkay Celik[†]
[†] Purdue University, {aarunasa, etekcan, zcelik}@purdue.edu
[‡] University of California, Irvine, habibaf@uci.edu

***Abstract*—Refugees form a vulnerable population due to their forced displacement, facing many challenges in the process, such as language barriers and financial hardship. Recent world events such as the Ukrainian and Afgan refugee crises have centered this population in online discourse, especially in social media, e.g., TikTok and Twitter. Although discourse can be benign, hateful and malicious discourse also emerges. Thus, refugees often become targets of toxic content, where malicious attackers post online hate targeting this population. Such online toxicity can vary in nature; e.g., toxicity can differ in scale (individual vs. group), and intent (embarrassment vs. harm), and the varying types of toxicity targeting refugees largely remain unexplored. We seek to understand the types of toxic content targeting refugees in online spaces. To do so, we carefully curate seed queries to collect a corpus of ∼3 M Twitter posts targeting refugees. We semantically sample this corpus to produce an annotated dataset of $1{,}400$ posts against refugees from seven different languages. We additionally use a deductive approach to qualitatively analyze the motivating sentiments (reasons) behind toxic posts. We discover that trolling and hate speech are the predominant toxic content that targets refugees. Furthermore, we uncover four main motivating sentiments (e.g., perceived ungratefulness, perceived fear of safety). Our findings synthesize important lessons for moderating toxic content, especially for vulnerable communities.**

I. Introduction

The UNHCR defines a refugee as *"someone who is unable or unwilling to return to their country of origin owing to a well-founded fear of being persecuted for reasons of race, religion, nationality, membership of a particular social group, or political opinion."* [33] Due to the hardships this population faces, which range from facing language barriers to overcoming discrimination, refugees are a vulnerable population.

The refugee crisis, most recently in Afghanistan and Ukraine, has centered this vulnerable population in online discourse. Such discourse can vary in nature, from being benign and supportive of refugees to promoting hatred and toxicity towards the population. Toxic content refers to a broad scope of attacks involving media, perpetrated by an attacker that does not require advanced capabilities (e.g., privileged access) [27]. Not requiring advanced capabilities allows attackers to have a low barrier of entry, especially on social media platforms (e.g., Facebook, Twitter). This low barrier of entry and the centering of refugees in online discourse allow for the widespread proliferation of toxic content targeting this population.

Interestingly, toxic content targeting *any* person varies in characteristics. Attacks can differ in intended viewers, (e.g., to be seen by a target or wide audience), harms (e.g., damage reputation, coerce), and scale (e.g., individual vs. groups) [27]. Thus, it becomes imperative to understand what types of toxicity target the refugee population.

In this work, we explore the types of toxic content targeting refugees on online platforms. To do so, we first curate an annotated dataset of toxic posts against refugees in seven different languages, Arabic, English, German, Italian, Spanish, Turkish, and Urdu (selected languages are spoken in regions with high refugee populations). We use a lexicon of toxic keywords from the HateBase API [14], a collaborative library of hate words, augmented with toxic unigrams (words), bigrams, and hashtags targeting refugees to collect a dataset of ∼3 million posts from Twitter. We then leverage a semantic textual similarity technique (SBERT) [22] to sample semantically diverse and representative posts. To identify the types of toxic content targeted at refugees, five annotators manually label the sampled dataset via our annotation guide, designed based on a taxonomy of toxic content [27], (e.g., bullying, hate speech). As a result, we curate a labeled dataset of $1{,}400$ posts comprising toxic content against refugees from seven different languages. We further analyze this dataset with qualitative analysis to uncover motivating sentiments behind toxic content - specific reasons attackers target refugees.

We discover that trolling and hate speech are, on average, the predominant toxic content targeting the refugee population on Twitter. We also uncover four main motivating sentiments behind these toxic posts – attackers' reasons for perpetrating attacks. Attackers either have a (1) perceived fear of safety, (2) are worried about the invasion of cultural/religious values, (3) have concerns about the economic implications, and (4) perceive refugees to be ungrateful.

Our exploratory analysis synthesizes important takeaways and provides direction for future work. Our study shows that toxic content against refugees is prevalent. We also argue that the moderation of toxic content should be adjusted depending on the toxic content type. We propose four extensions to our current study in our future work. First, we hope to expand our dataset and leverage it to build automated classifiers to conduct large-scale analysis. Second, with an expanded dataset, our objective is to develop a better understanding of toxicity against refugees through inferential statistical tests. Third, we hope to conduct

[*]Authors Arunasalam, Farrukh, and Tekcan have made equal contributions.

Symposium on Usable Security and Privacy (USEC) 2024
26 February 2024, San Diego, CA, USA
ISBN 979-8-9894372-5-2
https://dx.doi.org/10.14722/usec.2024.23014
www.ndss-symposium.org, https://www.usablesecurity.net/USEC/

additional qualitative analysis employing methods such as triangulation to dive deeper into motivating sentiments. Finally, we hope to conduct semi-structured interviews with content moderation practitioners for better transparency on the status quo of tackling online hate against vulnerable populations.

Our research contributions are as follows:

- Curate the first multilingual dataset of toxic content targeting refugees, which is labeled with respect to toxic content types.
- Expose four sentiments that motivate attackers to proliferate toxic content.
- Synthesize lessons and future work to improve the moderation of toxic content for the refugee population.

II. Background and Related Work

Toxic Content. The security community has long studied toxic content as an S&P issue, perceiving it as a form of online hate and harassment where an attacker can target individuals or groups of people [27]. Prior work in this domain is diverse, ranging from building automated classification models [8], [5], [10] to understanding sentiments against specific communities (e.g., women [6] and immigrants [20]). Recent work explores what motivates attackers to post toxic content against refugees on online platforms [35], [17], [28], [1]. Yet, the findings of this previous research are highly specific to a region of focus.

In contrast to prior works, our study focuses on analyzing toxic content across seven different languages, representing regions with high populations of refugees. Additionally, the motivating sentiments we discover (the reasons attackers post toxic content) are consistent across all languages.

Vulnerable Populations and Refugees. The security community has also investigated the S&P needs of vulnerable and at-risk communities [36], [3], [25], [13], [19]. In the context of refugees, one work has investigated refugees' S&P in regard to the digital technology they use [24], while another work has studied the implications of toxic content on refugees [2].

To our knowledge, our exploratory results are the first to curate a dataset of toxic content targeting refugees, labeled based on the type of toxic content. Furthermore, our proposed work, which involves expanding our dataset, extending additional quantitative and qualitative analysis, and a proposed study to understand content moderator practices for vulnerable communities, remains largely unexplored. Both our current results and future work aim to extend to the growing body of S&P research in online hate and vulnerable populations.

III. Research Question and Methodology

Recent events, especially the Ukrainian and Afghanistan refugee crises, have propelled discourse on refugees on various online mediums, such as news and social media. Given that toxic content can vary in nature (e.g., target, purpose), we aim to unpack and understand online toxic content targeting refugees, seeking to answer the following research question:

> **What types of toxic content are targeted at refugees on social media?**

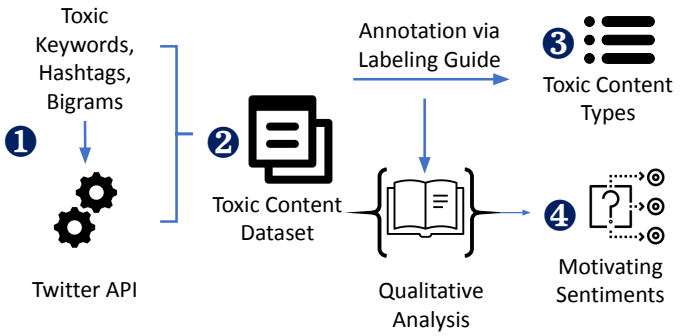

Fig. 1: Overview of identifying toxic content against refugees.

Figure 1 shows an overview of our approach to answering our research question. We first curate an annotated dataset of toxic online posts in seven languages (❶-❷). We then perform data annotation using a meticulously developed guide to identify the types of toxic content targeted at refugees (❸). We also perform qualitative analysis to study the motivating sentiment for posting toxic content against refugees (❹).

A. Discovering Toxic Content in Online Spaces

We investigated the existence of toxic content against refugees on social media by analyzing popular social media platforms. In our initial analysis, we found toxic content targeting refugees on Twitter, Facebook, Instagram, and TikTok, as well as postings with comparable toxic hashtags on all of the aforementioned platforms. To showcase the types of toxic content and the motivating sentiments for such content, we curated a dataset of social media posts, including toxic content targeted at refugees. For our dataset collection, we chose the *Twitter* platform as it provides a platform for open dialog and represents the broad public [26][1].

To comprehensively represent refugees from different regions of the globe, we curated a multilingual dataset. We selected the languages used by top refugee-hosting countries based on the United Nations High Commissioner for Refugees (UNHCR) statistics [21]. According to UNHCR data, the top refugee-hosting nations' official languages include Arabic, English, German, Italian, Spanish, Turkish, and Urdu.

Generating Seed Queries. To collect toxic content against refugees in each of the seven languages, we used the Twitter API [29]. This API provides metadata of posts (or tweets) on Twitter and can be queried using a set of keywords. We leverage the HateBase API [14] to generate a lexicon of hate speech in these languages and identify 1,919 keywords. To enrich this set with additional refugee-related words, we collected ∼45K tweets over six months (June 1st 2021 - December 1st 2021) by searching for the keywords "refugee(s)" and "asylum seekers" in the seven languages (using the Twitter API).

Two of the authors manually analyzed these tweets to identify the most frequent hashtags, words, and bigrams and selected ones with toxic or hateful meanings based on two criteria: (*a*) *Is this hashtag/word/bigram targeting refugees?* and (*b*) *Is it intended to express any negative sentiment (e.g., hate*

[1]Since data collection, the Twitter platform has rebranded to *X* [31]

TABLE I: Toxic keywords (hashtags, words, bigrams).

Language	Toxic Hashtags	Toxic Words	Toxic Bigram
English	18	12	2
Turkish	28	16	6
Italian	1	18	11
Urdu	3	11	11
Spanish	0	16	11
German	3	41	0
Arabic	25	70	10

and threats) about or against refugees? We verified the meaning and interpretation of identified toxic keywords with native speakers of each language. Through this, we created a final lexicon of 2,232 unigram/bigrams/hashtags pertaining to toxic content against refugees. Table I shows the summary statistics of keywords discovered with this process. The top unigrams, bigrams, and hashtags can be found in the Appendix Table IV.

Collecting and Sampling Toxic Content. We use our lexicon of toxic keywords to query the Twitter API. We leverage the API's feature to limit the search results to be within the period July 1st 2021 to December 31st 2021. In total, we collect $\sim$3 million tweets in seven languages that include toxic content targeted at refugees. We then leverage Sentence-BERT (SBERT) [22], a large-scale semantic similarity comparison network, to subsample semantically diverse and representative instances of toxic content against refugees. Through this process, we gather 1,500 tweets for each of the seven languages, resulting in a dataset of 10,500 tweets in total.

B. Dataset Annotation

We use our dataset of $10,500$ tweets to identify the various types of toxic content targeted at refugees. Initially, we explored using NLP techniques to automatically label tweets with the appropriate type of toxic content. However, automated NLP models typically require large amounts of labeled training data, and currently, to the best of our knowledge, there is no labeled dataset for classifying toxic content types. Therefore, we curated a labeled dataset of tweets containing toxic content against refugees using manual annotation.

Taxonomy Selection. For our manual labeling process, we constructed an initial taxonomy of toxic content categories using the types of toxic content outlined in prior work [27]. This taxonomy includes 10 types of toxic content: trolling, hate speech, profane or offensive, threats of violence, purposeful embarrassment, incitement, sexual harassment, unwanted explicit content, bullying, and hard to classify. We include an additional category (out of context) to represent tweets that do not contain toxic content against refugees (e.g., collected tweets may report hateful hashtags or attacks, not perpetrating a hateful attack). Based on this taxonomy, we created an annotation guide that includes the definition of the 10 toxic attack types (and the out-of-context category) that serve as our labels, examples of annotated tweets, and a series of questions to determine the labels for a single tweet. Our complete labeling guide is presented in Appendix B.

Translation, Annotator Training, and Agreement. Before annotation, we translated all non-English tweets into English using the Google translation API [7]. Two authors examined translations and observed that many translated tweets in Urdu, Arabic, and Turkish were difficult to interpret due to lack of coherence. To address this, we recruited three volunteers who were fluent in each of these languages to annotate the respective languages. We provided our annotation guide to volunteer annotators and conducted a two-hour training session with them to explain the types of toxic content and their attributes.

Each of the five annotators (two authors, three volunteers) then started to label the 1,500 tweets for each language. Annotators assigned one or more labels to each tweet. Given that a proportion of the tweets did not include any toxic content (out of context), we instructed the annotators to stop labeling when they assigned 200 tweets to any of the 10 toxic content categories (a total of 1,400 labeled tweets).

To verify that annotators would produce reliable results, we measured inter-coder agreement using Krippendorff's alpha metric. This inter-coder agreement metric accounts for an arbitrary number of annotators labeling any number of instances [16]. We evaluated this metric over 100 English tweets that had been labeled individually by each annotator. These tweets were independent of our dataset. The level of agreement among five annotators assessing toxic content types ranged from 0.7 to 0.9. These values imply substantial to almost perfect agreement [16]. The second round of labeling was performed to reconcile disagreements before the annotators coded each of their datasets independently.

We produced the first labeling guide and labeled dataset for toxic content against refugees, annotated based on toxic content type. Our dataset consists of 200 tweets per language for a total of 1,400 tweets. Our multilingual dataset (which comprises seven languages) highlights the different toxic content attacks refugees face[2]. After dataset curation, we held debriefing sessions to get feedback from annotators. All annotators acknowledged the unpleasantness of reading toxic content but felt positive overall due to the research contribution the annotation would produce.

Qualitative Analysis of Motivating Sentiment. To better understand the type of toxic content posted against refugees, we subsequently conducted a qualitative analysis of the annotated 1,400 tweets using deductive coding. We grounded our coding on motivating sentiments – specific justifications for why attackers perpetrate toxic content on social media. All annotators were involved in the process, generating codes in joint sessions to resolve conflicts together.

IV. Evaluation Results

A. Toxic Content Types Targeting Refugees

Table II presents a detailed breakdown of the toxic content types for each of the seven languages. Rows 1-7 are proportions of 200 Tweets in each language; the final row is a proportion of 1,400 Tweets. More than 50% of the annotated tweets have two labels. This prevalence suggests that it is common for toxic content posters to perpetrate different types of toxicity in a single tweet. Overall, the most prominent toxic content type is hate speech, which comprises 31.52% of the entire

[2]Our labeled dataset is available to researchers upon request.

TABLE II: Percentage (%) of toxic content type by language.

Language	Trolling	Hate speech	Profane or offensive	Threats of violence	Purposeful embarrassment	Incitement	Sexual harassment	Unwanted explicit content	Hard to classify	Bullying
English	35.42	27.68	14.02	1.48	6.27	6.27	0.37	0	5.17	3.32
Spanish	22.82	14.29	19.25	1.39	17.46	2.78	3.17	3.37	11.9	3.57
Italian	50.38	21.21	4.17	0.38	2.27	14.39	0.38	0	6.44	0.38
German	26.13	26.27	20.8	0.4	14.8	2.27	0.13	0.27	8.67	0.27
Arabic	1.68	64.29	12.61	0.42	1.68	0.84	0	0	13.45	5.04
Turkish	31.39	35.9	17.48	0	5.64	2.63	0.75	0	4.89	1.32
Urdu	31.94	48.89	5.28	0.28	0.28	7.78	0	0	5.56	0
Total	28.3	31.52	15.21	0.58	8.8	4.45	0.79	0.65	8.02	1.68

Percentages in cells are with respect to total tweets for the language. Sum across language (a single row) may exceed 100% as a tweet can be multi-labeled.

TABLE III: Motivating sentiment behind toxic content attacks and example hashtags.

Reason	Example Hashtags†
Invasion of religious/cultural values	#return_refugees_home
Benefits and economic implications	#Refugees_Take_Rights_of_Egyptians
Refugees being ungrateful	#boycott_refugee_products
Fear of safety	#NoRapeFugees, #Refugees_Ticking_bombs

† Non-English hashtags translated to English

dataset, with trolling (28.3%) coming a close second. Threats of violence, sexual harassment, and unwanted explicit content are much less prevalent (0.58%, 0.79%, 0.65%, respectively).

We note that the bullying label denotes attacks that specifically target individuals. Attacks denoted by all other labels can target individuals or groups. A toxic content attack can be multi-labeled (e.g., an attack of purposeful embarrassment directed at an individual refugee is also labeled as bullying). We refer to attacks targeting individuals as *directed* attacks and attacks targeted at the refugee community as *general* attacks. We observe a low percentage of bullying (*directed* attacks) - 1.68%. Noticeably, these trends exist across all languages in our dataset (e.g., trolling and hate speech represent the highest proportion in each language). Our analysis of toxic content on a social media platform exposes how *general* attacks are more common than *directed* attacks targeting individual refugees.

B. Motivating Sentiment and Justification

Table III represents four main reasons/narrative justification behind toxic content perpetrated by attackers and sample hashtags associated with these reasons. Broadly, toxic content posters are motivated by perceptions of refugees themselves and the perceived effects of their presence in the hosting country. For instance, toxic content posters often exhibit hate speech and trolling abuse when criticizing refugees for their appearance and differing cultural/religious values, implying that refugees intend to invade a nation through the spread of culture or religious influence. One attacker expressed disdain, suggesting

> *"[politician] makes u.s more refugee country packed with more Muslims... [Muslim politician] can become your president."*

Online toxic content is also driven by perceived economic effects, such as the notion that refugees receive benefits that are inaccessible to non-refugee citizens. For example, one abuser stated

> *"The USA must [IMMEDIATELY] cut off all welfare ... education benefits to all ... refugees"*,

while another mentioned that

> *"people going to get it in their tiny thick [expletive] heads that [refugees] are BENEFIT seekers."*

Claims that such benefits are not available to locals also fuel toxicity, e.g., one user exhibiting hate speech justifies their abuse by explaining

> *"the money, housing, education; the benefits are needed for our vets &; our own people."*

Similarly, refugees' perceived lack of gratitude is another motivating sentiment for posting toxic content. One abuser declared that

> *"the bad thing about some refugees who arrive in the country, they are ungrateful to the country that gave them asylum ... [and] dedicate themselves to looting the nation."*

Another used an expletive to express a similar sentiment, stating

> *"Go back to your [expletive] Somalia you [expletive] ungrateful refugee."*

Finally, the perceived fear of safety represents one aspect of narrative justification. We observe that themes of perceived fear are often accompanied by stories of refugees enacting harm, which are likely to spread due to fake news [23]. Despite the lack of veracity in these stories, toxic content posters use them to make claims, such as

> *"refugee influx in the east.. [will lead to] rape all over the country."*

It is integral to note that the motivating sentiments uncovered in our analysis may interconnect with other themes. For instance, toxic content attackers often conflate refugees' religion and culture when, in fact, these factors are separate. We find that this theme is especially raised in the context of Islam and refugees from Muslim-majority countries. Similarly, the sentiment of lack of gratitude often results from a perceived fear of safety. For instance, toxic content attackers cite fake news of a crime perpetrated by refugees and subsequently call them ungrateful. More work is required to better understand these motivating sentiments and how they are interconnected.

V. Discussion, Limitations and Future Work

We now synthesize the key takeaways of our study, its limitations and also future work that we plan to conduct.

A. Key Takeaways

Prevalence of Toxic Content. Our dataset curation highlights how toxic content is still prevalent on the Twitter platform. Toxic content also proliferates in various languages (not limited to the seven we select for labeling). In addition, we also discover hateful "hashtags" accompanying hateful posts. These findings suggest the need for improved content moderation on Twitter (and, by extension, other social media platforms).

Prior work has shown that social media may benefit from the proliferation of toxic content due to user engagement, which may lead to more revenue [11]. However, we have recently seen that the proliferation of toxic content has also negatively impacted platform engagement and revenue, encouraging social media companies to actively combat such content. First, we increasingly see companies are unwilling to advertise products or services on toxic platforms [30], [9], [4]. These companies do not want to be seen as endorsing toxic platforms and consequently terminating existing ad relationships. This can negatively impact social media companies, which generate a large portion of revenue through advertisements. Second, toxic content has negatively impacted user experience - with users shown to like social media less due to toxic content [9] and also perceive companies less favorably if they advertise on toxic platforms [15]. This erosion of trust from users and advertisers can arguably have a more long-term negative impact on revenue.

We also note that our data was collected prior to Twitter's change in ownership [31], highlighting how previous methods did not effectively moderate toxicity against refugees. Since the ownership change, Twitter has pivoted towards community-enforced notes. Here, groups of users are able to provide a supplemental note on a post [32], in the hopes that a community-based approach can mitigate problematic content (e.g., fake new, toxic content). Despite such shifts, hate speech continues to proliferate on Twitter. We propose that stronger transparency into how social media moderates toxic content can help improve the status quo of content moderation, especially when it comes to vulnerable populations, such as refugees.

Moderation for Different Toxic Content Types. Interestingly, prior work has exposed how bullying is the most common toxic content that surveyed/interviewed refugees are exposed to [2]. This contrasts our findings, where, on average, only 1.68% of our data was flagged for bullying. This finding highlights how the prevalence of one toxic content type on a social media platform is not necessarily a strong indicator of what the target interacts with most. Consequently, it is important to investigate whether different toxic content types against refugees warrant varying levels of content moderation. To illustrate, given that bullying is the most common type of toxic content with which refugees interact, it is reasonable to assume that there would be strong support within the population to enforce stricter content moderation (compared to other types).

B. Limitations of Data

It is important to note that toxic content against refugees on Twitter may not be generalized to all social media platforms and languages not considered in our analysis. Similarly, toxic content posted in public settings, instead of private settings, such as through direct message features, is also likely to differ. In addition, accounts used to perpetrate toxic content are not guaranteed to be from genuine users. Prior work has shown that it is common for bots to be used to leverage such attacks [34]. Despite these limitations, our findings suggest the prevalence of toxic content against refugees on social media.

C. Future Work

Dataset Expansion. We intend to expand our dataset by increasing the number of toxic posts labeled and diversifying the language of posts. Increasing the dataset's diversity may expose intricacies specific to a language/region and allow us to train a model for automated classification. An automated method to annotate collected posts would make large-scale measurement studies feasible.

Inferential Statistics on Data. Our dataset curation is intended to be exploratory. However, once we expand our corpus to a larger sample size, we aim to conduct statistical tests to test/validate the hypothesis surrounding our data and develop a better understanding of toxic content targeting refugees. For example, we can test the association (e.g., using a Chi-square test) between the types of toxic content and the motivating sentiment or language. Similarly, we can employ *ANOVA* tests (and also post hoc tests) to test for significant differences between the types and languages of toxic content.

Expanded Qualitative Analysis. We also aspire to expand our qualitative analysis to refine our understanding of motivating sentiments. For example, methods such as data triangulation [12] would help to better understand motivating sentiments and map sentiments that are related to each other (e.g., how perceived fear of safety relates to perceived lack of gratitude).

Understanding Content Moderation within Industry. As previously outlined, we advocate for better transparency surrounding content moderation for toxic content on social media. One way to achieve this goal is through research. Here, we propose to conduct semi-structured interviews with industry practitioners involved with content moderation for toxic content. Through this study, we hope to unpack the status quo of content moderation of toxic content as it pertains to vulnerable populations. For instance, we aim to uncover whether practitioners apply a different set of considerations when handling toxic content against vulnerable populations such as refugees. Similarly, prior work has shown that allowing users to fine-tune automated toxic content classification based on user preferences can improve user experiences [18]. Whether or not any platform implements such approaches or if there are challenges in enacting such a change remains unexplored.

VI. Conclusions

The refugee population has become the target of online discourse, with toxic content against refugees proliferating in many online spaces such as Facebook and Twitter. In our exploratory work, we curate the first dataset of toxic content targeting refugees labeled with respect to toxic content type. We also qualitatively analyze posts for the reasons behind the attackers' posts. Our results and study synthesize important lessons and future directions to better moderate toxic content and online experiences for the refugee population.

REFERENCES

[1] C. Arcila-Calderón, D. Blanco-Herrero, M. Frías-Vázquez, and F. Seoane-Pérez, "Refugees welcome? online hate speech and sentiments in twitter in spain during the reception of the boat aquarius," *Sustainability*, 2021.

[2] A. Arunasalam, H. Farrukh, E. Tekcan, and Z. B. Celik, "Understanding the security and privacy implications of online toxic content on refugees," in *USENIX Security Symposium*, 2024.

[3] R. Bellini, E. Tseng, N. Warford, A. Daffalla, T. Matthews, S. Consolvo, J. P. Woelfer, P. G. Kelley, M. L. Mazurek, D. Cuomo *et al.*, "Sok: Safer digital-safety research involving at-risk users," *arXiv preprint arXiv:2309.00735*, 2023.

[4] "List of companies boycotting facebook," https://wraltechwire.com/2020/06/29/list-of-companies-boycotting-facebook-social-media-over-hate-speech-swells/, 2020, [Online; accessed 15-December-2023].

[5] H. Chen, S. McKeever, and S. J. Delany, "The use of deep learning distributed representations in the identification of abusive text," in *International AAAI Conference on Web and Social Media*, 2019.

[6] S. Chess and A. Shaw, "A conspiracy of fishes, or, how we learned to stop worrying about# gamergate and embrace hegemonic masculinity," *Journal of Broadcasting & Electronic Media*, 2015.

[7] "Cloud translation api," https://cloud.google.com/translate, 2021, [Online; accessed 14-December-2023].

[8] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying," in *International AAAI Conference on Web and Social Media*, 2011.

[9] "Hate speech hurts social media sites, brands, and the digital economy," https://www.project-disco.org/competition/new-research-hate-speech-hurts-social-media-sites-brands-and-the-digital-economy/, 2023, [Online; accessed 15-December-2023].

[10] M. ElSherief, V. Kulkarni, D. Nguyen, W. Y. Wang, and E. Belding, "Hate lingo: A target-based linguistic analysis of hate speech in social media," in *International AAAI Conference on Web and Social Media*, 2018.

[11] "The facebook whistleblower says its algorithms are dangerous," https://www.technologyreview.com/2021/10/05/1036519/facebook-whistleblower-frances-haugen-algorithms/, 2021, [Online; accessed 15-December-2023].

[12] U. Flick, "Triangulation in data collection," *The SAGE handbook of qualitative data collection*, 2018.

[13] C. Geeng, M. Harris, E. Redmiles, and F. Roesner, "Like lesbians walking the perimeter: Experiences of U.S. LGBTQ+ folks with online security, safety, and privacy advice," in *USENIX Security Symposium*, 2022.

[14] "Hatebase," https://hatebase.org/about, 2021, [Online; accessed 13-August-2023].

[15] "Hate speech on social media can significantly damage brands," https://martech.org/hate-speech-on-social-media-can-significantly-damage-brands-study/, 2023, [Online; accessed 15-December-2023].

[16] A. F. Hayes and K. Krippendorff, "Answering the call for a standard reliability measure for coding data," *Communication Methods and Measures*, 2007.

[17] R. Kreis, "# refugeesnotwelcome: Anti-refugee discourse on twitter," *Discourse & Communication*, 2017.

[18] D. Kumar, P. G. Kelley, S. Consolvo, J. Mason, E. Bursztein, Z. Durumeric, K. Thomas, and M. Bailey, "Designing toxic content classification for a diversity of perspectives," in *Symposium on Usable Privacy and Security (SOUPS)*, 2021.

[19] I. Munyaka, E. Hargittai, and E. Redmiles, "The misinformation paradox: Older adults are cynical about news media, but engage with it anyway," *Journal of Online Trust and Safety*, 2022.

[20] N. Pitropakis, K. Kokot, D. Gkatzia, R. Ludwiniak, A. Mylonas, and M. Kandias, "Monitoring users' behavior: Anti-immigration speech detection on twitter," *Machine Learning and Knowledge Extraction*, 2020.

[21] "Refugee data finder," https://www.unhcr.org/refugee-statistics/download/?url=3HMho5, 2021, [Online; accessed 15-December-2023].

[22] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Conference on Empirical Methods in Natural Language Processing*, 2019.

[23] J. Roozenbeek and S. Van Der Linden, "The fake news game: actively inoculating against the risk of misinformation," *Journal of Risk Research*, 2019.

[24] L. Simko, A. Lerner, S. Ibtasam, F. Roesner, and T. Kohno, "Computer security and privacy for refugees in the united states," in *IEEE Symposium on Security and Privacy (SP)*, 2018.

[25] J. Slupska, S. Cho, M. Begonia, R. Abu-Salma, N. Prakash, and M. Balakrishnan, "They look at vulnerability and use that to abuse you: Participatory threat modelling with migrant domestic workers," in *USENIX Security Symposium*, 2022.

[26] "Social media stats worldwide," https://gs.statcounter.com/social-media-stats#monthly-202106-202112, 2021, [Online; accessed 10-August-2023].

[27] K. Thomas, D. Akhawe, M. Bailey, D. Boneh, E. Bursztein, S. Consolvo, N. Dell, Z. Durumeric, P. G. Kelley, D. Kumar *et al.*, "Sok: Hate, harassment, and the changing landscape of online abuse," in *IEEE Symposium on Security and Privacy (SP)*, 2021.

[28] S. Tuncer *et al.*, "Online hate on youtube: Anti-immigrant rhetoric against syrian refugees in canada and turkey," *Humanities Commons*, 2020.

[29] "Twitter api," https://developer.twitter.com/en/docs/twitter-api, 2022, [Online; accessed 10-December-2023].

[30] "Harmful content has surged on twitter, keeping advertisers away," https://time.com/6295711/twitters-hate-content-advertisers, 2023, [Online; accessed 13-August-2023].

[31] "Twitter is now x. here's what that means." https://www.cbsnews.com/news/twitter-rebrand-x-name-change-elon-musk-what-it-means/, [Online; accessed 15-December-2023].

[32] "Twitter's lack of moderation leaves room for harmful communities," https://newuniversity.org/2023/05/05/twitters-lack-of-moderation-leaves-room-for-harmful-communities/, [Online; accessed 15-December-2023].

[33] "What is a refugee," https://www.unhcr.org/what-is-a-refugee, 2023, [Online; accessed 13-December-2023].

[34] J. Uyheng, D. Bellutta, and K. M. Carley, "Bots amplify and redirect hate speech in online discourse about racism during the covid-19 pandemic," *Social Media + Society*, 2022.

[35] M. F. Vázquez and F. S. Pérez, "Hate speech in spain against aquarius refugees 2018 in twitter," in *International Conference on Technological Ecosystems for Enhancing Multiculturality*, 2019.

[36] N. Warford, T. Matthews, K. Yang, O. Akgul, S. Consolvo, P. G. Kelley, N. Malkin, M. L. Mazurek, M. Sleeper, and K. Thomas, "Sok: A framework for unifying at-risk user research," in *IEEE Symposium on Security and Privacy (SP)*, 2022.

APPENDIX A
TOXIC CONTENT LEXICONS

In Table IV, we provide the top unigrams, bigrams, and hashtags used to search for toxic content against refugees. In total, we created a final lexicon of 2,232 unigrams/bigrams/hashtags.

APPENDIX B
LABELING GUIDE

We detail the labeling guide annotators use to label toxic content posts (as introduced in Section III-B). Volunteer annotators are given this labeling guide during a 2 hour online training session and trained on how to use it.

1. Is the language of the tweet Turkish, German, Arabic, Urdu, Italian, English or Spanish?
Yes: Continue to next question
No: Label as out of context

2. Is the content of the tweet semantically meaningful?

TABLE IV: Top unigrams, bigrams, and hashtags used to search for toxic content.

Unigrams	Bigrams	Hashtags
terrorists	sac terrorists	#norapefugees
taliban	military terrorist leave	#knife_attack
illegal	we don't want refugees	#I_don't_want_a_refuge_in_my_country
deported	back to Kabul	#border tight
integration	invented criminals	#refugees_will_be_deported
violence	fake refugees	#blamegameonpakistan
we don't want *	persecuted southerners	#fencingforpeacefulpakistan
Ignorant	refugee deportation	#We reject_settlement_and naturalization_of_refugees
Fugitive	without refugees	#Enough_refugees_ in_ Egypt_ our_ disgrace
destroyed	#redistribution unsafe	#Refugees_Take_Rights_of_Egyptians

*Translated Non-English unigram/bigrams may contain more than one/two words respectively.

Yes: Continue to next question
No: Move to the next tweet

3. Does this tweet explicitly target refugee(s)?
Yes: Continue to next question
No: Label as out of context, move to next tweet

4. Does this tweet contain toxic content directed towards refugee(s)? (A "refugee" is a person who is unable or unwilling to return to his or her home country because of a "well-founded fear of persecution" due to race, membership in a particular social group, political opinion, religion, or national origin.)
Yes: Move on
No: Label as out of context, move to the next tweet

5. Does this tweet seek to harm or intimidate or coerce an individual perceived as vulnerable?
Yes: Label as bullying, continue to next question
No: Continue to next question

6. Does this tweet intentionally provoke someone/group of people with inflammatory remarks? (Inflammatory remarks are rousing or likely to rouse anger, violence, rioting)
Yes: Label as trolling, continue to next question
No: Continue to next question

7. Does this tweet contain abusive or threatening content that expresses prejudice targeting a group of people based on their race, gender, political/ideological affiliation, religion or a similar property?
Yes: Label as hate speech, continue to next question
No: Continue to next question

8. Does this tweet use profane or offensive language? (Eg. showing lack of respect to someone's religious beliefs, cursing, swearing, expletives, culturally offensive content)
Yes: Label as profane or offensive, continue to next question
No: Continue to next question

9. Does this tweet physically threaten someone?
Yes: Label as threats of violence, continue to next question
No: Continue to next question

10. Does this tweet try to purposely embarrass someone?
Yes: Label as purposeful embarrassment, continue to next question
No: Continue to next question

11. Does this tweet provoke unlawful behavior or urge someone to behave unlawfully?
Yes: Label as incitement continue to next question
No: Continue to next question

12. Does this tweet sexually harass someone? (Eg. Unwelcome sexual advances, requests for sexual favors)
Yes: Label as sexual harassment continue to next question
No: Continue to next question

13. Does this tweet contain unwanted explicit content? (Eg. sexting, violent and adult content)
Yes: Label as unwanted explicit content continue to next question
No: Continue to next question

14. Was it difficult to answer any of the questions regarding this tweet?
Yes: Label as hard to classify, move to the next tweet
No: Move to the next tweet