Exploring Multi-Level Threats in Telegram Data with AI-Human Annotation: A Preliminary Study

Kamalakkannan Ravi†, Adan Ernesto Vela†, Elizabeth Jenaway§, Steven Windisch§
†University of Central Florida, Orlando, USA
§Temple University, Philadelphia, Pennsylvania, USA
{kamalakkannan.ravi, adan.vela}@ucf.edu
{elizabeth.jenaway, steven.windisch}@temple.edu

2023 International Conference on Machine Learning and Applications (ICMLA) | 979-8-3503-4534-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICMLA58977.2023.00229

***Abstract*—This research addresses the crucial challenge of effectively measuring threats in social media comments targeting voting, public officials, and institutions in the United States. Our understanding of these online threats and their links to real-world risks is limited, making it difficult to assess their seriousness. To overcome these limitations, we propose a comprehensive threat level scale from 0 to 5 and collect a dataset of 1.3 million Telegram responses for developing and rigorously testing these threat levels. Additionally, we explore OpenAI-human annotation to efficiently label this vast dataset. Our innovative two-step transfer learning approach initially employs a pre-existing, pre-trained model for labeling, followed by expert validation. Next, we use the AI-annotated samples to develop independent models, and expert annotators verify their predictions. Notably, our findings demonstrate that the GPT-2 model, despite its fewer annotated training set, performs comparably to OpenAI's annotations, showcasing its potential for cost-effective threat detection with more annotated samples. With the long-term objective of establishing continuous threat-level monitoring, we identify the strengths and limitations of our current approach and propose a roadmap for enhancing threat detection.**

***Index Terms*—social networking (online), Telegram, learning (artificial intelligence), human-in-the-loop, radicalization behavioral indicators, text analysis, predictive models, natural language processing**

I. Introduction

The widespread use of social media and instant messaging has revolutionized communication, but it has also led to rising threats and hate speech. This poses challenges for public figures and organizations, making it crucial to monitor and moderate online discourse effectively. This requires proactive monitoring, strict moderation policies, and advanced content analysis to create a safer digital environment.

One platform that has gained attention in recent years is Telegram, known for its encryption and user anonymity features. While its encrypted communication creates challenges for content moderation [1, 2], it also offers a chance to study harmful content. Telegram stands out by allowing group users to have threaded discussions and comment on messages [3], like on Twitter or Reddit. Analyzing these comments on Telegram can help us understand harmful communication and how to keep users safe.

Previous research has categorized harmful communication into broad and narrow groups. The broader categories include hate speech, which has been studied [4–7]. Hate speech involves using hurtful words to target specific groups and show hostility and bias. It can be complicated, with varying levels of harm. Offensive language [8, 9], on the other hand, covers things like sexist or racist slurs, attacks on minorities, baseless criticism, and spreading harmful content. Lastly, cyberbullying [10, 11] is the most aggressive type, including threats of violence, hurtful messages, and personal insults.

In contrast, the narrow categories of hate speech delve into specific subtypes that target distinct groups or behaviors. These subtypes include hate speech directed towards refugees and Muslims [7, 12], racism [9], homophobia [12], sexism [9], toxicity [13], aggression [14], and harassment [10, 15]. These categories encompass various manifestations of harmful communication, but they often overlap and highlight the complexity of the issue.

Common approaches to categorizing hate speech rely on basic labels like positive, negative, or neutral [16–19], or they assess it based on violence or dehumanization [20]. However, these techniques may not fully grasp the extensive and often significant real-world consequences of harmful discourse against election and public officials. While these methods have their uses, they fail to establish a connection between online conversations and real-life risks in the United States. This underscores the necessity for a more comprehensive method to evaluate the harm conveyed in social media comments [**Research Gap 1**].

Moreover, the existing datasets for hate speech have mainly concentrated on various social media platforms such as websites [4, 6], Twitter [5, 9, 10, 21, 22], Facebook [23], Wikipedia [14, 24], newspapers [13, 25], Reddit [15, 25], Telegram [26], and YouTube [27]. However, these datasets often lack detailed categorization of harmful content.

Precisely, in the case of Telegram, a widely-used communication platform, there exists a notable dearth of comprehensive data. Even the most extensive dataset, like the Pushshift dataset [28], falls short as it excludes responses to messages shared on Telegram Channels. This missing piece in the dataset necessitates the development of a more encompassing dataset that accurately records the complete context and interplay of hate speech on Telegram [**Research Gap 2**].

To address these research gaps, we also explore the possibility of an AI-human annotation system that facilitates effective labeling of threat levels [**Research Gap 3**]. In our research, we aim to bridge these research gaps by developing a compre-

1946-0759/23/$31.00 ©2023 IEEE
DOI 10.1109/ICMLA58977.2023.00229

hensive approach that goes beyond simplistic classifications of hate speech.

II. Problem Statement

Our overarching goal is to develop a multi-level threat scale capable of quantifying the threat levels directed toward voting and public officials in the United States. To achieve this, we must address three research gaps identified in our introduction.

Firstly, to address **Research Gap 1**, we propose the creation of a nuanced threat-level classification system. This system will move beyond simplistic categorizations and strive to encompass the nuanced and severe aspects of harmful communication.

In addressing **Research Gap 2**, our strategy involves the collection and analysis of responses to messages posted on public Telegram channels to construct a robust dataset that spans a wide spectrum of harmful content. This effort will significantly enhance our online threat scale.

Lastly, to tackle **Research Gap 3**, we introduce a human-centric, two-step transfer learning approach to explore the possibility of an AI-human annotation system for the Telegram corpus we gather. Initially, we employ an existing pre-trained model to label a subset of the corpus based on threat scales. Once these labels are validated by experts, this labeled subset is utilized in the subsequent phase to create independent models. The predictions of these models are then validated by expert annotators.

III. Dataset

TABLE I: Telegram data

Channel name	Messages	Replies
absoluteTruth1776	1,816	27,797
AlexJones	1,676	12,820
AlexjonesInfoWars	64	639
DonaldTrumpJr	3,344	103,089
FreedomFighters	2,460	13,233
InfoWars.com	4,083	20,074
PrayingMedic	9,683	65,133
RTM	5,593	320,005
ThePatriotVoice	20,114	181,152
TheTrumpRepublicans	6,891	164,302
TrumpSupportersChannel	1,531	151,911
WeTheMedia	8,263	294,954
WhiteLivesMatter	29	38

Motivation. Our motivation stems from events like Jan 6 [29] and Pizzagate [30], where online discussions led to real-world violence. We aim to create a dataset linking online discourse to real-world risks or violence. We initiated our effort by searching Telegram channels using keywords associated with these events such as *Jan 6, Proud, Patriot, Freedom, Donald Trump supporters, MAGA,* and *Conspiracies*. We observed a high rate of grievances and threatening language in these channels. We identified 13 Telegram channels listed in Table I and collected all messages and replies from these channels, starting from their inception dates up to April 8, 2023, using the chat export [31]. The resulting dataset comprised the extracted messages and their corresponding replies.

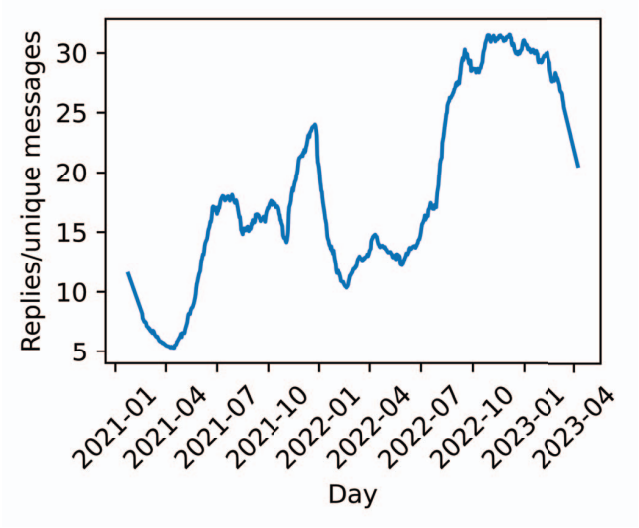

Fig. 1: Replies and messages ratio over time

TABLE II: Word and Sentence Count

	Word		Sentence	
	μ	σ	μ	σ
Messages	69.8	159.4	2.4	3.7
Replies	16.1	25.8	1.7	1.6

We ensured the data quality by cleaning it up, which included removing website links and empty messages and replies. After these steps, the dataset contained 65,547 messages and 1,355,147 replies, as shown in Table I. The collected data exhibits an uneven distribution of messages and replies from 2021 to 2023. Figure 1 visually displays the daily posting rates of messages and replies between January 24, 2021, and April 8, 2023. Although the dataset covers a consistent time span across all channels, the start dates of individual channels vary, leading to differences in posting rates. The increase in Telegram engagement following the election is also mentioned in [32], discussing the rise of far-right groups on the platform due to its isolated channel feature. However, our focus remains on public Telegram channels.

We want to highlight that while each channel has its own community, sometimes messages and replies get shared across channels and get different responses. This makes labeling during our research more complex. We also analyzed the number of words and sentences in the messages and replies which indicates a long-tailed skewed distribution, suggesting there are some messages with very high word or sentence counts along with many that have lower counts, as seen in Table II.

IV. Methodology

A. Defining Threat Levels.

Threat levels. Our primary objective is to establish a comprehensive system for capturing online threats that could pose real-world risks to voting and public officials in the United States. Each level is underpinned by distinct definitions rooted in three categories protected by the First Amendment [33, 34]: fighting words, incitement, and true threats. In our study, we posit that the mere presence of online violent threats constitutes a form of violence in itself.

Initially, we categorized harmful comments into two main groups, along with a category for comments with no threat (0). These groups were Judicial harm (4) and Non-judicial harm (5), with the possibility of implied threats (3). However, these categories might not cover comments that explicitly or implicitly endorse harm while targeting individuals or groups with criminal, illegal (1), or morally (2) reprehensible actions intended to defame them. This led us to introduce two more threat levels for a more precise evaluation of harmful comments.

To establish these threat levels, we conducted a comprehensive analysis of randomly selected replies from our Telegram corpus. We meticulously refined the definitions of each category until we were satisfied with their precision, utilizing a subset of approximately 163 replies for this purpose. This framework allows for a more nuanced and precise evaluation of the severity of online comments. The threat levels are as follows:

0. None of the above: No indication, suggestion, or desire of physical harm, imprisonment, or threat towards another a person, group, or organization.

1. Claims of illegal/criminal actions/behavior: Statements that allege or assert that a person, persons, group, or organization committed or is engaging in illegal or criminal activities without explicitly calling for arrest, prosecution, punishment, harm, or other action. This category excludes particularly heinous crimes like child abuse, molestation, sexual assault/rape, sex trafficking, torture, terrorism, murder, and mass murder/genocide.

2. Claims of immoral/evil behaviors: Statements that assert that a person, persons, group, or organization is immoral or evil, and statements that allege or assert that a person, group, or organization are committing acts or engaged in behavior that is considered particularly immoral or evil. Examples include grooming and pedophilia, in addition to particularly heinous crimes like child abuse, molestation, sexual assault/rape, sex trafficking, torture, terrorism, murder, and mass murder/genocide.

3. Implied support/ambiguous call/threats for action/harm: Statements that suggest or imply support for arrest, prosecution, punishment, or other harmful or threatening actions, but without explicit instructions or clear intent. Such implied or ambiguous statements are often in the form of rhetorical questions or make reference to violence or harmful action through double meanings. This category can include protest/political chants and battle cries (e.g., '1776!', 'Live free or die!').

4. Calls/threats for judicial action/harm: Statements that explicitly call for or threaten legal action against a person, group, or organization in the form of civil action, arrest, or criminal prosecution. May include calls for punishment within standard legal norms, including prison, while excluding excessive, or cruel and unusual punishments. This category includes statements in the form of protest/political chants (e.g. 'Lock her up').

5. Calls/threats for non-judicial or extra-judicial action/harm: Statements that explicitly call for or threaten non-judicial or extra-judicial action or harm against a person, group, or organization, such as physical violence or vigilante justice. This includes actions or punishments considered outside of standard legal norms for being excessive, disproportionate, or cruel and unusual. This category also includes battle cries (e.g. 'Time to start a civil war!') and protest/political chants (e.g. 'Hang Mike Pence!') that refer to non-judicial or extra-judicial action.

B. Evaluating GPT-3.5 as a potential annotator.

After establishing our threat level framework, we applied it to guide the annotation of a subset of 800 responses from our Telegram corpus using OpenAI's GPT-3.5 model. Table III provides examples of replies and their respective annotations. We utilized the OpenAI API (OA) to prompt the model to classify each text into one of the six classes outlined in section IV-A. The annotation cost, using OpenAI's GPT-3.5, is 7.27 USD for 11,326 samples. However, this cost increases to an estimated 8672 USD when annotating 1,355,147 samples. If we opt for the use of GPT-4 from OpenAI, the cost would more than double, as indicated by the OpenAI's pricing structure [35]. This process resulted in a collection of 180 samples (30 from each class), which were then submitted for expert validation.

Considering that our threat scale varies from 0 to 5, we needed a methodology to measure the concurrence between the two human annotators and the OpenAI model known as inter-rater reliability. However, given the characteristics of our threat scale, it was crucial to consider the relative distance between ratings. For instance, a difference in ratings of 3 and 5 is less significant than a difference between ratings of 3 and 0. Traditional weighting methods such as 'quadratic' or 'linear' were not suitable as they assume ordered disagreement, which is not the case in our scale.

Thus, we formulated a customized weight matrix aligned with our threat scale. Given the complexities of our threat scale and data, we employed a custom weight matrix to compute Cohen's Kappa, developed with expert input [36] based on the context of the threat levels. This matrix, detailed in Table IV, highlights the intricacies and difficulties involved in annotating replies, as demonstrated in Table III. It brings to the forefront the distinctive labeling approaches and the inherent subjectivity embedded in the process.

TABLE III: Sample replies with annotation. O - OpenAI, E - Expert, F - fastText, L - LSTM, G - GPT-2 and S - SVM.

Telegram reply	O	E	F	L	G	S
Don't see the term "censorship" anywhere in lib Bari Weiss' commentary. It's because other libs need her antiseptic description to accept what has been done. It's like describing a murder scene as "someone who did something bad to someone else". I suppose that it has some value to know the code words that Twitter insiders used but really, "Visibility filtering"? Really? A rose by any other name, illegal censorship.	1	1	1	1	1	1
After all this fuck has said and done to the American people, we the American people should have went to DC and thrown every traitor to the American constitution in the fucking ocean! Dragged them all to the east coast and thrown them in! This so called president has violated every right we have and still we sit back and do "NOTHING"!!!! This piece of dog shit has called domestic terrorists, has threatened us on live tv, and still we sit and do nothing! Him and his son highly invested in China, Ukraine, and who knows what else, and still he or his son not held accountable? Really? Grow some balls america! Stand the hell up! Protect our Constitution! It was written for us, not them! They work for us, not the other way around!	5	5	4	3	5	3
No way that judge would have ruled in her favor. he is corrupt as well and if she would have won her case most in power in Arizona would be tried for treason. Either that or Hobbs threatened the judge and his family with death .. time for us to take up our arms and start doing something to save our country … it is almost to the point of no return now. Soon they will forcefully take our guns by our woke military and the mercenaries that have crossed our border. The military is not the answer especially now that they have crossed over to extreme wokeness … there is no plan… we r the plan… and their plan is to kill us and our families so they have the planet to themselves. Wake up people before the bus shows up at your front door to take u off to the camps and then it is too late	5	3	4	0	2	5

TABLE IV: Custom weight matrix

Threat Labels	0	1	2	3	4	5
0	0	0.2	0.2	0.3	0.5	0.5
1	0.2	0	0.1	0.2	0.4	0.4
2	0.2	0.1	0	0.2	0.4	0.4
3	0.3	0.3	0.2	0	0.2	0.1
4	0.5	0.4	0.4	0.2	0	0.3
5	0.5	0.4	0.4	0.2	0.3	0

Cohen's Kappa inter-rater agreement score is detailed in Table VII. Our annotators, graduate students with expertise in violent extremism studies, demonstrated an average agreement of 0.44 (moderate agreement) with the OpenAI model, and an inter-rater concurrence of 0.49 (moderate agreement). The complexities and challenges inherent in annotating replies, as evidenced by the confusion matrix between our annotators and OpenAI shown in Tables V and VI.

TABLE V: Confusion matrix: Expert 1 and OpenAI

		OpenAI					
		0	**1**	**2**	**3**	**4**	**5**
Expert 1	**0**	61	4	21	10	2	3
	1	7	4	1	3	0	0
	2	0	0	7	1	0	0
	3	3	0	2	10	4	3
	4	3	0	1	7	4	0
	5	2	0	1	4	2	10

TABLE VI: Confusion matrix: Expert 1 and Expert 2

		Expert 2					
		0	**1**	**2**	**3**	**4**	**5**
Expert 1	**0**	47	21	12	10	8	3
	1	0	6	1	2	6	0
	2	0	0	6	1	0	1
	3	3	0	1	8	4	6
	4	0	0	0	1	14	0
	5	0	2	0	2	3	12

Based on Landis and Koch's [37] scale, a kappa score of 0.44 falls within the 'moderate' agreement range. This denotes a moderate level of agreement on threat perception between human annotators and the OpenAI model, paving the way for the latter's use in generating labeled seed data for our model training.

C. Utilizing OpenAI for high-volume annotation and training our model.

After establishing the annotation process with OpenAI, we labeled 11,136 samples, dividing them into a 10,000-sample training set and a 1,136-sample test set. All the experiments were performed on a workstation with 128 GB RAM and 48 cores. We selected fastText, ULMFit, SVM, and GPT-2 models due to their fast training times and cost-effectiveness.

Training and Hyperparameters. As anticipated, the fastText model exhibited exceptional speed, capitalizing on word embeddings and n-gram features. We adopted the optimal parameters as suggested by Joulin et al. [38], which encompassed: learning rate (lr) = 0.1, epochs = 1000, wordNgrams

TABLE VII: Cohen's Kappa on 180 Telegram replies

Expert 1 & OpenAI	Expert 2 & OpenAI	Expert 1 & 2
0.45	0.43	0.49

= 2, bucket = 200000, dimensions (dim) = 50, and loss = 'hs'. The model's training took a remarkable 18.41 seconds, plus an additional testing time of 1.32 seconds.

For the ULMFiT model, we implemented the advised parameters [39] and began the process with the pre-trained AWD-LSTM model. This approach involved using an LSTM encoder to process the input text, thereby capturing the contextual information and semantic interrelations between words. The training requires 108 minutes for fine-tuning and training and 14.7 seconds for testing.

We configured the SVM model utilizing the recommended parameters [25] and applied 1-gram and 2-gram features with TF-IDF weighted embeddings. The optimal parameters were: C = 10, Gamma = 0.1, kernel = rbf, lower case = False, and class weight = 'balanced'. The training took 83.80 seconds, while the testing phase required 5.64 seconds.

Lastly, for the GPT-2 Transformer model, we adhered to the guidelines put forth by Wolf et al. [40]. The model utilizes a transformer encoder, an effective neural network architecture for processing sequential data such as text. The model's batch size and gradient accumulation were set to 1, and it underwent 100 epochs of training using the pre-trained 'gpt2' model. The GPT-2 model's training, took only 15.25 minutes per epoch for the entire training process, while the testing phase took a brief 22.45 seconds, illustrating the speed and efficacy of our chosen models.

V. Results and Discussion

A. Multi-Level Threat Scale

The establishment of threat levels in our research represents a significant step toward capturing the prevalence of harmful comments with the potential to harm voting processes and public officials in the United States. These threat levels are rooted in distinct definitions based on categories protected by the First Amendment, recognizing that online threats often lead to tangible effects, including psychological distress and potential offline violence. Our framework encompasses a range of threat categories, allowing for nuanced evaluations of the severity of online comments.

B. A large scale data

In our pursuit of constructing a comprehensive dataset of 1.3 million replies, we identified and collected messages and replies from 13 public Telegram channels, as outlined in Table I. This extensive dataset, encompassing a substantial volume of messages and replies, serves as a valuable resource for our research.

TABLE VIII: Trained model performance on 1,136 Telegram replies

Model	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)
ULMFit	58.2	54.1	58.2	55.7
fastText	60.8	60.0	60.8	60.2
SVM	65.5	63.6	65.5	64.3
GPT-2	67.0	66.6	67.0	**66.2**

C. Examining Inter-rater Agreement Between OpenAI and Experts

In the pursuit of mitigating challenges and costs associated with human annotators, we sought to measure the agreement between expert human annotations and those generated by OpenAI on a sample set. We interpreted our kappa values using the scale proposed by Landis and Koch [37]. With a kappa value of 0.44, falling within the 'moderate' agreement category, this classification aligns with Viera and Garrett's [41] work, which also categorizes a kappa value between 0.41 and 0.60 as indicative of 'moderate' agreement. In this context, a kappa value of 1 means perfect agreement, while 0 suggests any agreement is likely due to chance. The assessment highlights the complexities and subjectivity in the annotation process. Achieving perfect alignment between human annotators and automated systems is challenging due to subjective aspects, as discussed by Viera and Garrett. While our kappa scores suggest moderate agreement, they remind us of the ongoing challenges in reaching complete harmony between human and machine annotations.

D. Standalone model and test results

While OpenAI's model proves advantageous for short-term annotation tasks, its cost-effectiveness diminishes due to the financial implications tied to the number of API requests per minute, the number of tokens per minute, and the escalating price per token. The costs are amplified considering the need to feed the model with both threat label prompts and input text to classify, as detailed in methodology section IV-B. Moreover, the use of a third-party pay-per-use API raises concerns regarding privacy, security, and data portability when handling sensitive data. Consequently, creating a standalone model, informed by seed labels obtained through OpenAI, becomes essential.

As a result, we developed models using the seed data annotated by OpenAI, presenting the corresponding test results (on 1,136 samples) in Table VIII. The GPT-2 model exhibited robust performance with a weighted F1-score of 66.2%. Although this score may not seem impressively high, it's crucial to acknowledge that the training data originates from another machine-annotated set, carrying its inherent misclassification errors.

E. Model fidelity

Following our methodology, we curated 30 samples from each OpenAI prediction from the test set of 1,136 samples, culminating in 180 samples for expert annotation. This step

TABLE IX: Model fidelity measured on 180 test samples

Model	Cohen's Kappa	Strength of Agreement
OpenAI	0.51	**Moderate**
ULMFit	0.26	Fair
fastText	0.36	Fair
SVM	0.33	Fair
GPT-2	0.43	**Moderate**

ensured the accuracy and reliability of our model's performance. To gauge the extent of agreement, we computed Cohen's Kappa score, comparing the results generated by our standalone models with those produced by our expert annotator. The detailed results for various models can be found in Table IX.

Our standalone GPT-2 model showcased its capacity to faithfully reproduce prediction results, achieving a Kappa score of 0.43, a result that closely parallels the agreement recorded by OpenAI (0.51). This alignment underscores that our GPT-2 model, despite being trained on a more modest set of 10,000 OpenAI annotated samples using a transfer learning approach, can produce predictions on par with OpenAI's performance. Consequently, these findings point towards the potential benefits of enhancing OpenAI predictions with improved instructions.

VI. Limitations and Future Work

In this project, we acknowledge certain limitations in our approach and outline areas for future improvement.

Annotation Approach. Our method, combining OpenAI models with human expert verification, showed moderate agreement, primarily for non-threatening comments. To enhance our ability to identify nuanced and potentially harmful comments, we plan to refine both our model and instructions via lexical analysis [42, 43]. This refinement aims to improve annotation accuracy, potentially reducing annotation costs and enhancing threat detection.

Comparison with Gold Standard. To establish a robust benchmark, we intend to annotate a substantial subset of Telegram reply samples. This step will allow us to assess the performance of our improved AI-generated annotations against ground truth annotations. This process will address the annotation limitation and align our model with prior binary hate speech classification works.

Data Collection. Our dataset exhibits temporal imbalances due to variations in channel start dates, which may affect the representativeness of our findings. Cross-posted messages across channels introduce complexities and potential duplicates, impacting the annotation process. Additionally, the skewed distribution of word and sentence counts highlights data heterogeneity. Future research could explore analyzing messages and comparing human and AI annotations, providing a broader perspective.

VII. Conclusion

In this paper, we tackle the pivotal challenge of identifying and quantifying the risk level of threats posed against public figures and institutions in the United States, discerned from social media discourse. The inherent complexity of understanding threats necessitates a more refined mechanism to assess their gravity and interconnectedness.

To bridge this gap in knowledge, we introduce an innovative threat-level grading system, extending from 0 to 5. This nuanced system, built from an analysis of selected replies from our Telegram dataset, provides an intricate look into threats, going beyond conventional classifications. It thereby contributes to a deeper understanding of the complexities and intensity of harmful communications.

In addition, our approach involves data collection from 13 public Telegram channels that are likely to contain harmful commentary, resulting in a robust dataset of over 1.3 million responses. While the dataset exhibits inherent challenges, such as distribution imbalances, potential overlaps, and skewed word and sentence counts, it remains an invaluable resource for exploring the intricate nature of online threats.

We further employ a two-step transfer learning strategy to efficiently annotate this extensive dataset. Initially, we harness the power of OpenAI's pre-trained GPT-3.5 model to label a subset of the Telegram corpus according to the threat-level scale. Once expert-validated, this subset serves to train multiple models, including fastText, ULMFiT, SVM, and GPT-2. Notably, our standalone GPT-2 model exhibits promising capability in replicating predictions, achieving a Kappa score closely mirroring OpenAI's. This suggests the potential to enhance our model's performance and mitigate annotation and development costs by refining OpenAI's predictions.

Our standalone model is specifically designed to analyze and categorize harmful language in social media, providing proactive threat detection. It overcomes privacy and security issues tied to third-party APIs, using OpenAI-sourced seed labels to handle sensitive data securely. Beyond that, it offers potential aid in content moderation by identifying harmful phenomena such as violent extremism [44], anti-minority sentiment [45], and other risk indicators [46].

Furthermore, with the long-term objective of establishing continuous threat-level monitoring, we acknowledge limitations and outline future directions. These include improving annotation accuracy by incorporating lexical analysis [42, 43], benchmarking our AI-generated annotations against gold standards, addressing data collection imbalances and complexities, and exploring message-level analysis for comprehensive threat detection enhancement.

In a final step to encourage further studies, reproducibility and democratize research, we will make our code, models, and data available to the public, upholding a transparent approach to community research and language modeling.

Ethical Considerations

We recognize that quantifying the threat level of public Telegram replies using expert annotation is not representative

of the broader Telegram communities within the United States, especially since the popularity of Telegram in the United States is relatively lower compared to many other countries [47]. In 2023, Telegram recorded approximately 20 million downloads and has around 10 million monthly users in the United States. Although these figures may be significant for other applications, they represent less than 2% of Telegram's global user base. Therefore, when considering Telegram as a platform, it is essential to acknowledge its relatively smaller user presence in the United States compared to other regions.

Furthermore, we wish to underscore that our data collection method specifically targets public Telegram groups or channels that are accessible without the need for logins or invitations. Private groups, which are restricted and require an invitation link or owner approval for access, are beyond the scope of our research. This intentional exclusion is a preventive measure designed to mitigate the potential misuse of our data collection method in private digital spaces by individuals with malicious intent.

ACKNOWLEDGEMENTS

This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 20STTPC00001-01-02. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security. We are grateful to Jacqueline Petasne and Adrienne Brookstein for their contributions to the annotation process.

REFERENCES

[1] J. Mayer, "Content moderation for end-to-end encrypted messaging," *Princeton University*, 2019.

[2] S. Kamara, M. Knodel, E. Llansó, G. Nojeim, L. Qin, D. Thakur, and C. Vogus, "Outside looking in: Approaches to content moderation in end-to-end encrypted systems," Center for Democracy and Technology, Aug 2021.

[3] (2018, January) Group chats on telegram 2018. Telegram. Telegram groups are a powerful tool for building communities and each can have up to 200,000 members. [Online]. Available: https://telegram.org/tour/groups

[4] W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web," in *Proceedings of the second workshop on language in social media*, 2012, pp. 19–26.

[5] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the international AAAI conference on web and social media*, vol. 11, no. 1, 2017, pp. 512–515.

[6] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, "Hate speech detection with comment embeddings," in *Proceedings of the 24th international conference on world wide web*, 2015, pp. 29–30.

[7] Z. Zhang, D. Robinson, and J. Tepper, "Detecting hate speech on twitter using a convolution-gru based deep neural network," in *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*. Springer, 2018, pp. 745–760.

[8] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Predicting the type and target of offensive posts in social media." Association for Computational Linguistics, Jun. 2019.

[9] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on twitter," in *Proceedings of the NAACL student research workshop*, 2016, pp. 88–93.

[10] J. Golbeck, Z. Ashktorab, R. O. Banjo, A. Berlinger, S. Bhagwan, C. Buntain, P. Cheakalos, A. A. Geller, Q. Gergory, R. K. Gnanasekaran *et al.*, "A large labeled corpus for online harassment research," in *Proceedings of the 2017 ACM on web science conference*, 2017, pp. 229–233.

[11] M. M. Islam, M. A. Uddin, L. Islam, A. Akter, S. Sharmin, and U. K. Acharjee, "Cyberbullying detection on social networks using machine learning approaches," in *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*. IEEE, 2020, pp. 1–6.

[12] S. Assimakopoulos, R. Vella Muskat, L. van der Plas, and A. Gatt, "Annotating for hate speech: The MaNeCo corpus and some input from critical discourse analysis." European Language Resources Association, May 2020.

[13] V. Kolhatkar, H. Wu, L. Cavasso, E. Francis, K. Shukla, and M. Taboada, "The sfu opinion and comments corpus: A corpus for the analysis of online news comments," *Corpus Pragmatics*, vol. 4, pp. 155–190, 2020.

[14] E. Wulczyn, N. Thain, and L. Dixon, "Ex machina: Personal attacks seen at scale," in *Proceedings of the 26th international conference on world wide web*, 2017, pp. 1391–1399.

[15] H. Almerekhi, H. Kwak, B. J. Jansen, and J. Salminen, "Detecting toxicity triggers in online discussions," in *Proceedings of the 30th ACM conference on hypertext and social media*, 2019, pp. 291–292.

[16] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," in *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, 2002, pp. 79–86.

[17] F. N. Ribeiro, M. Araújo, P. Gonçalves, M. André Gonçalves, and F. Benevenuto, "Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods," *EPJ Data Science*, vol. 5, pp. 1–29, 2016.

[18] F. H. Khan, U. Qamar, and S. Bashir, "esap: A decision support framework for enhanced sentiment analysis and polarity classification," *Information Sciences*, vol. 367, pp. 862–873, 2016.

[19] L. Yue, W. Chen, X. Li, W. Zuo, and M. Yin, "A survey of sentiment analysis in social media," *Knowledge and Information Systems*, vol. 60, pp. 617–663, 2019.
[20] B. Bahador, "Monitoring hate speech and the limits of current definition," 2023.
[21] T. Wijesiriwardene, H. Inan, U. Kursuncu, M. Gaur, V. L. Shalin, K. Thirunarayan, A. Sheth, and I. B. Arpinar, "Alone: A dataset for toxic behavior among adolescents on twitter," in *Social Informatics: 12th International Conference, SocInfo 2020, Pisa, Italy, October 6–9, 2020, Proceedings 12*. Springer, 2020, pp. 427–439.
[22] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. R. Pardo, P. Rosso, and M. Sanguinetti, "Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter," in *Proceedings of the 13th international workshop on semantic evaluation*, 2019, pp. 54–63.
[23] R. Kumar, A. K. Ojha, S. Malmasi, and M. Zampieri, "Benchmarking aggression identification in social media," in *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*, 2018, pp. 1–11.
[24] G. Jigsaw, "Toxic comment classification challenge: identify and classify toxic online comments," 2018.
[25] K. Ravi, A. E. Vela, and R. Ewetz, "Classifying the ideological orientation of user-submitted texts in social media," in *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2022, pp. 413–418.
[26] V. Solopova, T. Scheffler, and M. Popa-Wyatt, "A telegram corpus for hate speech, offensive language, and online harm," *Journal of Open Humanities Data*, vol. 7, 2021.
[27] R. Ottoni, E. Cunha, G. Magno, P. Bernardina, W. Meira Jr, and V. Almeida, "Analyzing right-wing youtube channels: Hate, violence and discrimination," in *Proceedings of the 10th ACM conference on web science*, 2018, pp. 323–332.
[28] J. Baumgartner, S. Zannettou, M. Squire, and J. Blackburn, "The pushshift telegram dataset," in *Proceedings of the international AAAI conference on web and social media*, vol. 14, 2020, pp. 840–847.
[29] J. W. Jackson and V. B. Hinsz, "Group dynamics and the us capitol insurrection: An introduction to the special issue." *Group Dynamics: Theory, Research, and Practice*, vol. 26, no. 3, p. 169, 2022.
[30] M. Fisher, J. W. Cox, and P. Hermann, "Pizzagate: From rumor, to hashtag, to gunfire in dc," *Washington Post*, vol. 6, pp. 8410–8415, 2016.
[31] (2018, August) Chat export tool, better notifications and more. [Online]. Available: https://telegram.org/blog/export-and-more
[32] S. Zihiri, G. Lima, J. Han, M. Cha, and W. Lee, "Qanon shifts into the mainstream, remains a far-right ally," *Heliyon*, vol. 8, no. 2, 2022.
[33] C. E. Ring, "Hate speech in social media: An exploration of the problem and its proposed solutions," Ph.D. dissertation, University of Colorado at Boulder, 2013.
[34] L. E. Beausoleil, "Free, hateful, and posted: rethinking first amendment protection of hate speech in a social media world," *BCL Rev.*, vol. 60, p. 2101, 2019.
[35] Pricing openai.com. [Accessed 20-Jul-2023]. [Online]. Available: https://openai.com/pricing
[36] J. Cohen, "Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit." *Psychological bulletin*, vol. 70, no. 4, p. 213, 1968.
[37] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *biometrics*, pp. 159–174, 1977.
[38] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, "Fasttext.zip: Compressing text classification models," *arXiv preprint arXiv:1612.03651*, 2016.
[39] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," *arXiv preprint arXiv:1801.06146*, 2018.
[40] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.
[41] A. J. Viera, J. M. Garrett *et al.*, "Understanding interobserver agreement: the kappa statistic," *Fam med*, vol. 37, no. 5, pp. 360–363, 2005.
[42] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.
[43] N. D. Gitari, Z. Zuping, H. Damien, and J. Long, "A lexicon-based approach for hate speech detection," *International Journal of Multimedia and Ubiquitous Engineering*, vol. 10, no. 4, pp. 215–230, 2015.
[44] H. Alvari, S. Sarkar, and P. Shakarian, "Detection of violent extremists in social media," in *2019 2nd international conference on data intelligence and security (ICDIS)*. IEEE, 2019, pp. 43–47.
[45] T. J. Holt, J. D. Freilich, and S. M. Chermak, "Examining the online expression of ideology among far-right extremist forum users," *Terrorism and Political Violence*, vol. 34, no. 2, pp. 364–384, 2022.
[46] B. W. Hung, S. R. Muramudalige, A. P. Jayasumana, J. Klausen, R. Libretti, E. Moloney, and P. Renugopalakrishnan, "Recognizing radicalization indicators in text documents using human-in-the-loop information extraction and nlp techniques," in *2019 ieee international symposium on technologies for homeland security (hst)*. IEEE, 2019, pp. 1–7.
[47] Telegram Users by Country 2023 — worldpopulationreview.com. [Accessed 09-Jun-2023]. [Online]. Available: https://worldpopulationreview.com/country-rankings/telegram-users-by-country