# ALERT: Active Learning and Explainable AI for Robust Threat Detection in Telegram

KAMALAKKANNAN RAVI and JIANN-SHIUN YUAN, University of Central Florida, USA

The increasing regulatory scrutiny of social media, particularly regarding extremist content and misinformation, underscores the need for advanced threat detection systems. This paper presents ALERT (Active Learning and Explainable AI for Robust Threat Detection in Telegram), a novel framework that enhances threat classification by introducing refined categories and creating tailored datasets[1]. ALERT processes 2,301,110 replies from 17 Telegram channels, focusing on extreme content, with a dataset that predominantly reflects far-right discourse, consistent with activity trends on the platform. By leveraging an iterative active learning approach, it reduces labeling efforts by 86.5%, yielding a labeled dataset of 15,076 replies. ALERT's RoBERTa+ model, pre-trained on domain-specific data, achieved over 90% in precision, recall, accuracy, and F1-score, demonstrating strong generalization for threat detection. The integrated explainable AI (XAI) modules highlight key text features driving model predictions, ensuring transparency while maintaining performance. ALERT offers significant improvements in classification precision and user confidence, providing a critical tool[2] for addressing digital threats while navigating regulatory and privacy challenges.

CCS Concepts: • **Computing methodologies** → **Neural networks**; *Natural language processing*; *Knowledge representation and reasoning*; • **Information systems** → *Sentiment analysis*; • **Social and professional topics** → *Privacy policies*; • **Applied computing** → *Sociology*; • **Security and privacy** → Social aspects of security and privacy; • **Human-centered computing** → *Empirical studies in collaborative and social computing*.

Additional Key Words and Phrases: active learning, cyberbullying, explainable AI, extremism, Learning (artificial intelligence), natural language processing, online radicalization, political violence, social media, telegram, user-generated content

## 1 INTRODUCTION

The relationship between governments and major tech platforms like Telegram, TikTok, and X (formerly Twitter) has intensified, driven by mounting regulatory scrutiny [46]. High-profile events, such as the arrest of Telegram's founder Pavel Durov and ongoing legal battles in Brazil and the U.S., underscore the growing concern over platforms that enable the spread of harmful content [3, 46]. These platforms play an increasingly significant role in shaping real-world issues like misinformation and harmful trends, which has prompted a heightened demand for greater accountability [53]. Yet, achieving this accountability must be carefully managed to avoid infringing on free speech and digital privacy rights, making it essential to find a balanced regulatory approach [12, 15].

This challenge is particularly pronounced on platforms like Telegram, where extremist content and behaviors are pervasive, creating a pressing need for effective threat detection [40, 50]. The rise of political polarization through social media further exacerbates the problem, reinforcing the importance of advanced systems capable

[1] Dataset repository: https://data.mendeley.com/datasets/tm9s68vgxd/1

[2] GitHub repository: https://github.com/kamalravi/TelegramThreatLevel

Website: https://kamalravi.com.
Authors' Contact Information: Kamalakkannan Ravi, kamalakkannan.ravi@ucf.edu; Jiann-Shiun Yuan, jiann-shiun.yuan@ucf.edu, University of Central Florida, Orlando, Florida, USA.

ACM 2576-5337/2025/8-ART
https://doi.org/10.1145/3758967

of nuanced threat analysis [30]. Despite ongoing efforts to develop refined threat categorization [38, 39], there remains a critical need for more specific and focused categorization methods to accurately detect and prevent violence.

A key obstacle in this effort is the scarcity of comprehensive datasets designed specifically for threat detection on emerging platforms like Telegram. This issue parallels challenges in fields such as medical imaging and computer vision, where the lack of well-labeled datasets hinders model performance [2, 6, 17, 21, 27, 36]. While platforms like Twitter and Reddit have benefited from the availability of diverse datasets [37, 43], Telegram's data remains limited in both coverage and specificity, impeding the development of effective threat detection systems.

Another significant challenge lies in the high costs associated with data labeling. Accurately identifying threats, especially those that involve subtle or complex language, requires specialized expertise, making manual annotation both time-consuming and expensive [19, 40]. The large volume of daily messages on platforms like Telegram only adds to the complexity of this task. At the same time, there is a growing need for threat detection systems that incorporate explainable AI (XAI) to improve transparency and build user trust [18]. Current models often lack the refinement required to distinguish different types of threats, such as judicial versus non-judicial threats, which limits their effectiveness [16, 40].

This paper addresses these interconnected challenges by proposing a series of advancements in threat categorization, the creation of tailored datasets, the development of cost-effective labeling strategies, and the integration of XAI to enhance the transparency and precision of threat detection systems.

## 2 RELATED WORKS

Research in social media threat detection spans several key areas: threat categorization systems, tailored datasets, cost-effective labeling strategies, and explainable AI (XAI). These elements are critical to addressing extremism, misinformation, and harmful content across platforms like Twitter, Reddit, and Telegram.

### 2.1 Threat Categorization Systems

Threat categorization has evolved from basic keyword-based methods to more complex machine learning models. Early efforts by Davidson et al. [16] and Zampieri et al. [55] categorized posts into hate speech, offensive language, and neutral content. Melton et al. [33] expanded this with the Offensive Language Identification Dataset (OLID), distinguishing hate speech from offensive and neutral content. Waseem et al. [52] focused on labeling Twitter posts as Racist, Sexist, or Harmless, enhancing the granularity of harmful language classification.

More recent work by Oussalah et al. [34] developed models to score radicalization, and Behzadan et al. [10] focused on identifying cyber threat indicators in tweets. Vaidya et al. [48] classified toxic versus non-toxic comments, while Mahata et al. [32] separated targeted from untargeted offenses. Ashraf et al. [5] applied binary classification to threats aimed at individuals or groups. Ravi et al. [40] extended these approaches, introducing a threat level scale for quantifying risks to public figures and institutions, though many models still struggle with distinguishing fine threat categories, leading to high false-positive rates **(Research Gap 1)**.

### 2.2 Dataset Availability and Telegram

Well-labeled datasets are essential for training effective models. Twitter, YouTube, and Reddit datasets are well-established, with Waseem's dataset [52] providing 15,216 labeled tweets, Davidson et al.'s [16] HON dataset contributing around 25,000 tweets, and OLID [55] offering approximately 13,000 tweets labeled for hate speech detection. Specific threat-related datasets, such as Behzadan et al.'s [10] 21,000 cyber-related tweets, and Mahata et al.'s [32] dataset on targeted versus untargeted offenses, add to this pool. On YouTube, Hammer et al. [20] compiled a dataset of 28,643 sentences and 9,845 comments, including 1,387 violent threats, while the Jigsaw dataset [48] contains over 1.8 million comments.

However, Telegram remains under-researched [41]. Datasets like Pushshift [9] do not capture its unique discussion threads, which are key to understanding extremist content. This gap is notable, given Telegram's role in promoting violence, as observed by U.S. Deputy Attorney General Lisa Monaco [47]. While fields like computer vision and medical diagnostics benefit from large, annotated datasets [23, 45], comparable initiatives for Telegram are needed **(Research Gap 2)**.

### 2.3 Cost-Effective Data Labeling

Manual labeling, while accurate, is resource-intensive. For example, Waseem et al. [52] created a dataset of 15,216 labeled Twitter posts using a mix of amateur and expert annotators. In contrast, Davidson et al. [16] labeled 25,000 tweets through Amazon Mechanical Turk, using multiple reviewers per tweet to ensure consistency. However, manual annotation becomes less scalable as data volumes grow on platforms like Twitter, Reddit, and Telegram. Platforms like Amazon Mechanical Turk and Figure Eight have been commonly used [32, 55] but remain time-consuming and costly. The Gab dataset [33] and Ashraf et al.'s [5] work on binary threat classification highlight the challenge of labeling even small datasets with expert review.

Recent studies have explored semi-supervised and active learning techniques to mitigate costs. These methods focus on uncertain instances and integrate small labeled datasets with larger unlabeled datasets for improved performance [19]. Despite these advancements, scalable approaches to reduce labeling efforts for nuanced platforms like Telegram remain underdeveloped **(Research Gap 3)**.

### 2.4 Explainable AI for Threat Detection

Explainable AI (XAI) aims to improve the transparency of machine learning models. For example, Bunde [13] showed that adding model explanations to ULMFiT for hate speech detection boosted user confidence. Similarly, Babaeianjelodar et al. [7] applied Shapley Additive Explanations (SHAP) to XGBoost models, enhancing interpretability compared to black-box models like LSTM.

Post-hoc XAI techniques, such as SHAP and Local Interpretable Model-Agnostic Explanations (LIME), help explain model predictions. Kibriya et al. [25] applied these methods to clarify how features influence outcomes in a Conv-BiRNN-BiLSTM model. Ravi and Vela [38] demonstrated the potential of XAI for threat detection on social media by analyzing attention weights in their model, showing how individual words shape ideological orientation.

Despite progress, most threat detection systems remain opaque, limiting their usefulness in applications where trust and interpretability are vital. As demand grows for models that balance accuracy with explainability, particularly for platforms like Telegram, further advancements in XAI frameworks are needed **(Research Gap 4)**.

## 3 PROBLEM STATEMENT

This research aims to develop a comprehensive, scalable, and explainable threat detection framework tailored for social media platforms, particularly in underexplored areas like Telegram within the United States. For effective threat detection, we identify and tackle several critical research gaps:

**RG 1 - Enhanced Threat Categorization Systems.** Our proposal emphasizes explicit calls for harm, providing greater clarity and precision in categorizing threats. This approach, elaborated in sub-section 5.1, ensures that subtle distinctions are recognized, reducing ambiguity in threat classification.

**RG 2 - Comprehensive Telegram Dataset.** We curate a large and diverse dataset that accurately captures the nuances of threaded discussions and extremist rhetoric. As discussed in Section 4, this dataset will serve as a foundation for developing precise threat categorization and models, addressing the need for high-quality data in threat detection.

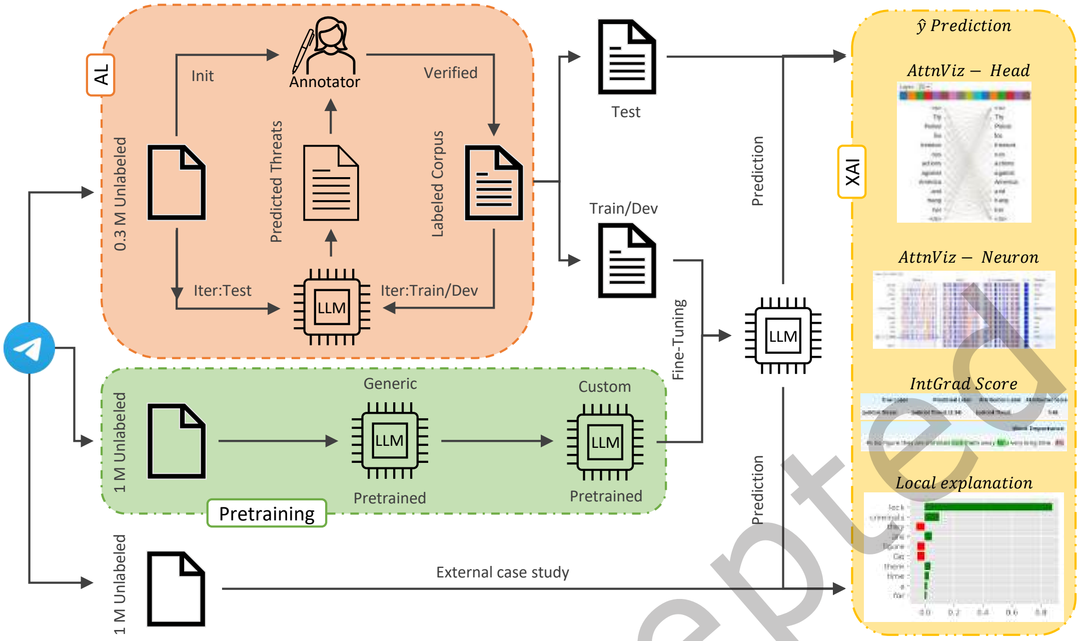


Fig. 1. ALERT flowchart illustrating the progression of our research. Key stages include: Active Learning (AL block), pretraining on custom data (Pretraining block), fine-tuning, testing, and external case studies. The Explainable AI (XAI block) represents Prediction, Attention Visualization, Integrated Gradients, and local explanations. Post-hoc evaluations of reliability, safety, and fairness are reflected through the external case study and XAI analysis components.

**RG 3 - Cost-Effective Labeling Strategy.** Our proposed cost-effective strategy, grounded in active learning, focuses on identifying and labeling the most informative instances. As detailed in Section 5, this approach reduces manual effort and enhances the scalability of our ALERT framework.

**RG 4 - Scalable Explainable AI (XAI) Framework.** We present a scalable XAI framework that balances interpretability and accuracy, as detailed in Section 6. By integrating both intrinsic techniques like Attention Visualization and post-hoc methods such as Integrated Gradients and LIME, our model offers interpretable insights into its decision-making process, addressing the need for clarity in automated threat classification.

Upon addressing these research gaps, we evaluate the model's reliability, safety, and fairness through post-hoc analysis and external case studies, as reflected in the corresponding blocks in Figure 1. Additionally, we will outline the model's limitations and future directions in Section 7, followed by a comprehensive summary in Section 8.

## 4 DATA COLLECTION

We identified 25 Telegram channels representing both far-right and far-left ideologies using targeted keywords like "jan 6," "anarchy," "proud," "riot," "patriot," "freedom," "crime," "maga," "late-stage capitalism," and "conspiracies" [50]. While our initial sampling aimed to include a balance of ideological viewpoints, the final dataset is predominantly composed of far-right discourse, reflecting the actual activity distribution observed on Telegram. The data collection process involved creating a Telegram account using a Google Voice virtual number. By joining

Table 1. Chosen Telegram channels' user, message, and reply count

| Telegram Channel Name | Users | Messages | Replies |
|---|---|---|---|
| AlexJones | 33,496 | 1,636 | 12,675 |
| Analyzing America | 59,482 | 2,658 | 345,595 |
| Anticapitalist Surrealism | 6,124 | 45,768 | 97,211 |
| Black Crimes Matter | 3,075 | 212 | 1,405 |
| COVID VACCINE VICTIMS | 82,999 | 10,941 | 56,751 |
| DonaldTrumpJr | 459,768 | 7,059 | 168,212 |
| FreedomFighters | 4,290 | 6,961 | 17,288 |
| InfoWars.com | 28,376 | 4,098 | 20,190 |
| Lacan's Wh*re House | 8,455 | 50,279 | 95,317 |
| Patriot Streetfighter | 98,163 | 20,598 | 133,508 |
| PrayingMedic | 135,834 | 12,738 | 84,624 |
| Resist the Mainstream | 180,147 | 11,659 | 555,424 |
| Riot Dogs | 1,974 | 62,913 | 95,763 |
| ThePatriotVoice | 47,578 | 23,123 | 192,407 |
| TheTrumpRepublicans | 41,136 | 9,768 | 184,570 |
| TrumpSupportersChannel | 185,684 | 1,943 | 171,712 |
| United Anarchists | 3,945 | 45,519 | 68,458 |
| **Total** | 1,380,526 | 317,873 | 2,301,110 |

public channels and exporting chat histories in .json format through the *view discussions* feature. No interaction with users or collection of personally identifiable information occurred, adhering to ethical standards as outlined in Section 8.

Unlike most prior Telegram datasets, including Pushshift Telegram [9], the U.S. Election 2024 dataset [11], and TGDataset [28], our collection preserves the threaded structure of public discussions. We leveraged Telegram's View Discussions feature to extract linked replies associated with each channel post, enabling us to capture multi-turn interactions and conversational context. This structure is particularly valuable for threat detection, as it reflects how users respond to and build on each other's messages, which is not typically available in flat Telegram message dumps.

To focus on meaningful activity, we prioritized channels exhibiting frequent grievances and threatening language. Channels with fewer than 1,000 replies were excluded due to low activity [37], such as "late-stage capitalism," "Lin Wood," "Donald J. Trump," "White Lives Matter Official," "absoluteTruth1776," "WeTheMedia," "AlexJones InfoWars," and "Patriot Front." Ultimately, 17 channels were selected, as detailed in Table 1, which outlines user, message, and reply counts. The data was collected from the inception of each channel until February 6, 2024. After filtering out non-text content such as URLs and empty responses, the final dataset consisted of 2,301,110 replies.

## 5 ACTIVE LEARNING FOR DATA ANNOTATION

Manual annotation for large datasets requires considerable effort from domain experts, limiting scalability. To mitigate this, we implemented an active learning framework that prioritizes labeling the most informative samples [42]. Specifically, we selected samples with high model confidence, particularly those predicted as threats, and paired them with an equivalent number of non-threat messages to maintain a relative class balance. These samples were then manually reviewed and labeled, allowing us to correct any model misclassifications before retraining. This approach significantly reduces labeling burden while preserving human oversight and improving data quality with each iteration.

## 5.1 Definition of Threat Labels

Previous classification methods, such as Davidson et al.'s broad categorization of harmful content into normal speech, hate speech, and offensive speech [16], offer limited granularity for detecting violent extremism [8]. Ravi et al. advanced this by proposing a six-level threat classification system, but it encountered issues with category overlap and subtle differences [40]. To address these limitations, we introduce a refined system that simplifies threat classification into three main categories:

*5.1.1* ***No Threat.*** This category encompasses statements that do not suggest harm or violence toward individuals, groups, or organizations. Examples include: "Hillary was just seething. That look on her face was priceless," "You'll regret this,", and "Live free or die." These statements, while potentially charged or ambiguous, lack explicit harmful language.

*5.1.2* ***Judicial Threat.*** The "Judicial Threat" category includes statements advocating for legal action, such as arrest or prosecution. Examples include: "Lock her up" and "Put that POS in jail‼" Although these statements operate within legal frameworks, they convey hostility and imply severe consequences.

*5.1.3* ***Non-Judicial Threat.*** This category encompasses the most severe threats, advocating for unlawful or violent actions. Statements such as "Time to start a civil war!" and "Hang Mike Pence!" exemplify explicit calls for violence, raising significant concerns in the monitoring of violent extremism. Any statement exhibiting elements of both judicial and non-judicial threats is classified as non-judicial due to its heightened severity.

## 5.2 Iterative Active Learning Process

To manage a large dataset with minimal manual annotation, we developed an iterative active learning process (Algorithm 1).

To assess annotation quality, we conducted an inter-rater agreement analysis on an initial set of 10,000 samples labeled by two trained annotators; one graduate and one undergraduate student. Using three threat classes, we computed Cohen's Kappa score of 0.89, indicating high agreement. Label distributions were: 191 (No Threat), 108 (Judicial Threat), and 153 (Non-Judicial Threat). Based on this strong reliability, the graduate annotator proceeded to label the remaining data using the iterative active learning framework. The total annotation process spanned approximately 6 weeks.

Starting with a small, balanced subset of 300 samples, manually derived from 452 labeled samples in the initial 10,000 (drawn from 301,110 unlabeled samples), we fine-tuned a RoBERTa model. The model then predicted labels for a new batch of 10,000 samples, prioritizing "Judicial Threat" and "Non-Judicial Threat" categories, and added 300 new labeled samples to the dataset.

This process was repeated iteratively, progressively increasing the size of the labeled dataset. By the second iteration, 900 samples were labeled. The dataset grew to 1,789 samples by the third iteration, and by the fourth iteration, it reached 2,262 labeled samples. A significant jump occurred in the fifth iteration, with 7,605 samples labeled after testing a larger batch of 100,000 samples. The final iteration added 7,472 new labeled samples after testing a larger batch of 151,110 samples, resulting in a total of 15,076 labeled data points.

This iterative approach substantially reduced the manual labeling effort by 86.50%, with only 40,629 samples manually labeled out of the 301,110 originally selected for the process. Model performance, evaluated through F1 scores, showed steady improvement throughout the iterations and reached saturation by the sixth iteration. Full details of each iteration, including the number of labeled samples and F1 scores, are presented in Table 2.

**Algorithm 1** Active Learning for Data Annotation

```
**Input:** Unlabeled dataset $U$, initial labeled dataset $L = \emptyset$
**Output:** Labeled dataset $L$
**Step 1: Initial Labeling**
Randomly select a subset of samples from $U$ for manual annotation
Manually label the selected samples into predefined classes
Add labeled samples to $L$, update $U = U \setminus L$
**Step 2: Iterative Active Learning**
**while** stopping criterion not met **do**
    **Model Training**
    Split $L$ into training and validation sets
    Fine-tune a model (e.g., RoBERTa) on the training set
    Evaluate the model on the validation set
    **Model Testing and Sample Selection**
    Select a new subset of samples from $U$
    Test the fine-tuned model on the new subset of samples
    Identify samples that are predicted with the highest uncertainty or belong to the most informative classes
    **Manual Annotation**
    Manually label the selected samples
    Add newly labeled samples to $L$, update $U = U \setminus L$
**end while**
**End of Algorithm**
```

Table 2. Performance of Iterative Active Learning

| Iter. | Total Train/Val | Val F1 Score (%) | Test F1 Score (%) | Labeled Samples Accumulated (Test) |
|---|---|---|---|---|
| 1 | 300 | 95.00 | 74.07 | 600 |
| 2 | 600 | 88.16 | 69.59 | 900 |
| 3 | 900 | 88.88 | 73.34 | 1,789 |
| 4 | 1,789 | 88.79 | 81.50 | 2,262 |
| 5 | 2,262 | 86.10 | 82.05 | 7,605 |
| 6 | 7,605 | 87.66 | 80.64 | **15,076** |

The RoBERTa model was trained using established parameters [54], with a batch size of 2 and training typically requiring 100 epochs, although performance plateaued by 10 epochs. All experiments were conducted on a system with 12 CPUs, 32 GB GPU, and 64 GB RAM. By selectively focusing expert annotation efforts on the most informative samples, we significantly reduced labeling costs and created a robust labeled dataset of 15,076 samples, while leaving over 2 million samples unlabeled. Table 3 further details the word and sentence counts for each class and the unlabeled data, illustrating the scope of our labeling process.

## 6 XAI FOR THREAT DETECTION

This section presents a model for detecting threats in user-generated content on Telegram, accompanied by explanations for its classifications. Using 15,076 labeled replies and 2 million unlabeled replies, we selected

Table 3. Word and sentence count

| | Word | | Sentence | |
|---|---|---|---|---|
| | $\mu$ | $\sigma$ | $\mu$ | $\sigma$ |
| **No Threat** | 21.75 | 24.63 | 1.95 | 1.54 |
| **Judicial Threat** | 18.59 | 17.07 | 1.90 | 1.32 |
| **Non-Judicial Threat** | 16.93 | 19.33 | 1.73 | 1.35 |
| **Unlabeled** | 14.47 | 23.95 | 1.55 | 1.49 |

Table 4. Threat Detection Performance Metrics on 2,261 test Telegram replies

| Model | Acc. (%) | Prec. (%) | Recall (%) | F1 (%) |
|---|---|---|---|---|
| LightGBM [24] | 73.91 | 75.57 | 73.91 | 74.13 |
| fastText [22] | 75.86 | 76.44 | 75.90 | 76.04 |
| Pretrained GPT-2 [35] | 81.07 | 81.07 | 81.07 | 81.05 |
| Pretrained RoBERTa [31] | 90 | 89.98 | 90 | 89.97 |
| Pretrained RoBERTa with Additional Pretraining **(RoBERTa+)** | **90.28** | **90.27** | **90.27** | **90.27** |

and trained models suited for large-scale and imbalanced datasets. The selection includes both state-of-the-art transformers and traditional machine learning models, chosen for their performance and scalability.

## 6.1 Model Selection and Training

The labeled dataset was split into 10,554 training samples, and 2,261 each for validation and testing. Transformer models such as RoBERTa and GPT-2 were explored. RoBERTa, an optimized version of BERT, excels in understanding context and generalizing across NLP tasks [31]. GPT-2, designed for text generation, provides robust contextual analysis despite being originally built for shorter texts [35].

Given our dataset's characteristics, including imbalanced class (with class counts of 5025 No Threat, 4658 Judicial Threat, and 5393 Non-Judicial Threat) and long-tail word distribution (Table 3), we fine-tuned two RoBERTa models: one fine-tuned directly on labeled data and another (RoBERTa+) pretrained on 1 million unlabeled Telegram replies and then fine-tuned. We also employed traditional models such as LightGBM, which efficiently handles large datasets using gradient-based optimization [24], and fastText, known for its scalability via hierarchical softmax and n-gram embeddings [22].

**Training Details.** RoBERTa was trained for 100 epochs (168 hours), and GPT-2 for 62 hours. Pretraining RoBERTa+ required 12,260 steps (137 hours). LightGBM was fine-tuned using grid-search cross-validation, optimizing the learning rate (0.1) and maximum depth (7). fastText training took 8 seconds. Performance metrics are reported in Table 4.

**Training Details.** Models were trained on a workstation with 12 CPUs, a 32 GB GPU, and 64 GB RAM. RoBERTa was trained for 100 epochs ( 168 hours) and GPT-2 for 62 hours, following established guidelines [54]. RoBERTa+ pretraining involved 12,260 steps (137 hours) with a batch size of 256. LightGBM used 5-fold cross-validation, optimizing parameters like a 0.1 learning rate and depth of 7. FastText training took 8 seconds with recommended settings. Performance metrics (accuracy, precision, recall, F1-score) are in Table 4.

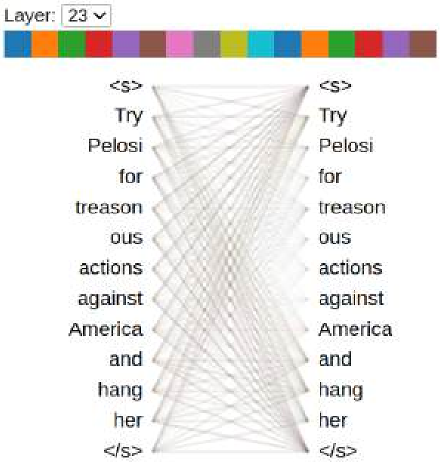


Fig. 2. Attention-head view of RoBERTa+ for layer 24, displaying the input text *"Try Pelosi for treasonous actions against America and hang her."* The figure represents attention patterns across different attention heads within the layer.

## 6.2 Explainable AI

Given the stakes of automated threat detection, ensuring transparency and interpretability through XAI is essential. As models grow in complexity, interpretability becomes crucial, especially in content with social and cultural nuances. Our XAI framework combines intrinsic and post-hoc techniques for model transparency while preserving accuracy [1, 4].

*6.2.1* ***Intrinsic Explainability.*** Intrinsic techniques are built into the model architecture to provide insights during the decision-making process. **Attention Visualization.** In transformer-based models like RoBERTa+, attention visualization aids in interpreting which parts of the input text influence the model's classification decisions. This is particularly useful for understanding how the model prioritizes specific words or phrases within complex, context-rich content, revealing the rationale behind its predictions such as threats.

To facilitate this, BertViz provides an interactive tool for visualizing attention weights across different layers and heads of Transformer models [49]. The attention-head view (Figure 2) illustrates the attention patterns learned by RoBERTa+, showing how tokens such as *"hang"* in the final layer receive increased attention. In this view, lines connecting tokens represent self-attention, where color indicates the attention head and thickness reflects the attention score. This observation is consistent with prior findings that deeper Transformer layers tend to focus on semantically meaningful or task-relevant tokens, while earlier layers emphasize structural markers like <s> and </s> for positional anchoring and classification context [14]. The neuron view (Figure 3) offers a more granular perspective by visualizing individual neurons in the query and key vectors, showcasing their interactions. While the head view captures broader attention patterns, such as model biases, the neuron view links specific neurons to decision-making processes, providing deeper insights into how particular predictions are formed.

*6.2.2* ***Post-hoc Explainability.*** Post-hoc techniques explain model decisions without modifying the architecture. **Integrated Gradients.** This technique traces each input feature's contribution to the output, clarifying how each word affects classification. We used Captum's TokenReferenceBase and LayerIntegratedGradients to compute word-level attribution scores in RoBERTa+ [26]. As illustrated in Figure 5, the True Label indicates human-annotated classifications, while the Predicted Label reflects the model's output with associated confidence scores

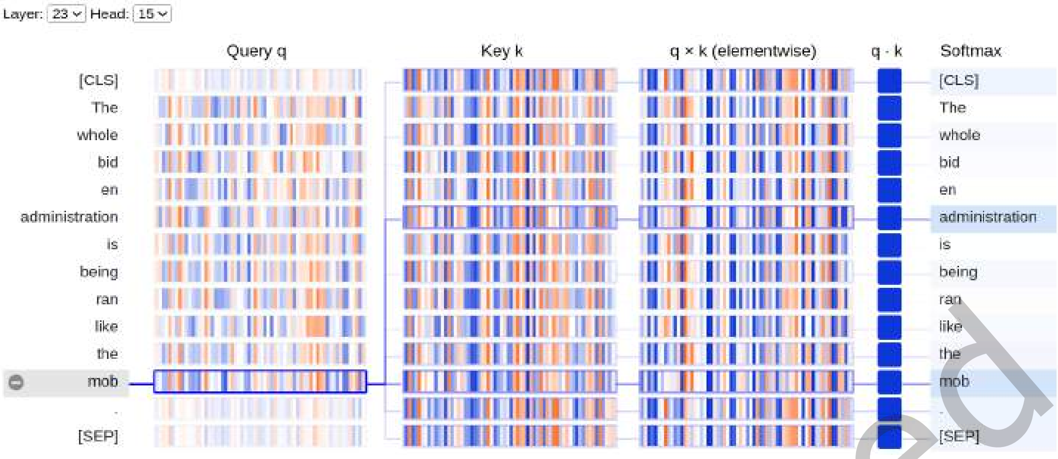


Fig. 3. Neuron view of RoBERTa+ for layer 24, head 16, showing the input text *"The whole Biden administration is being ran like the mob."* Positive values are shown in blue and negative values in orange, with color intensity reflecting the magnitude. Similar to the attention-head view, line thickness represents the strength of attention between words.

(e.g., Judicial Threat with a confidence score of 3.54). The Attribution Label and Attribution Score quantify the contribution of key terms such as *"criminals"* and *"lock them away"* (with a score of 1.46). This process elucidates how individual words drive the model's decision, enhancing the transparency and interpretability of automated threat detection.

**Local Explanations.** Local Interpretable Model-Agnostic Explanations (LIME) provide instance-specific explanations by perturbing input data and evaluating the impact on predictions, particularly valuable for understanding edge cases and complex predictions. For example, given the text *"Go figure they are criminals; lock them away for a very long time,"* LIME highlights critical words contributing to the model's classification (Figure 4). The bar chart presents word contribution values on the X-axis, with positive values indicating words that increase the likelihood of the predicted class, and negative values showing words that decrease it. Words like *"lock"* and *"criminals"* are identified as the most influential features, positively contributing to the classification as a threat. LIME's visual outputs facilitate quick identification of key features, enhancing transparency in model reasoning and improving the interpretability of automated threat detection systems.

## 7 RESULTS AND DISCUSSION

After applying Active Learning (Section 5) and training classifiers with explainability methods (Section 6), we evaluated the reliability, safety, and fairness of the threat detection models.

### 7.1 Reliability Evaluation

*7.1.1* ***Performance Metrics.*** We assessed the performance of the data labeling process and threat detection models. Our Active Learning approach minimized manual annotation while maintaining high threat detection accuracy. By manually labeling only the most informative samples, those identified by the model as predicted threats along with balanced non-threat samples, we enhanced efficiency and reduced costs. By iteration five, we achieved a test F1 score of 82.05% with only 40,629 samples manually labeled out of 301,110, reflecting an 86.5% reduction in annotation requirements, as shown in Table 2.

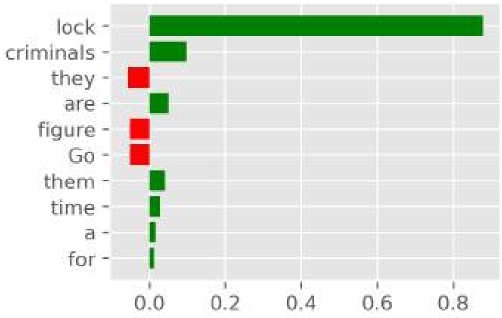
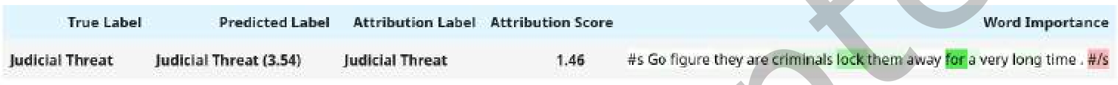


Fig. 4. Bar chart illustrating feature contributions for the RoBERTa+ model with the input text *"Go figure they are criminals; lock them away for a very long time."* The X-axis shows feature contribution values, indicating each word's influence on the model's prediction.



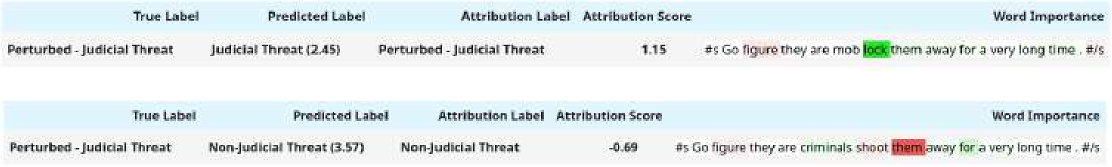
| True Label | Predicted Label | Attribution Label | Attribution Score | Word Importance |
|---|---|---|---|---|
| Judicial Threat | Judicial Threat (3.54) | Judicial Threat | 1.46 | #s Go figure they are criminals lock them away for a very long time . #/s |

Fig. 5. Visualizing token importance attributions using Integrated Gradients for the RoBERTa+ model with the input text *"Go figure they are criminals lock them away for a very long time."*

On threat detection, model performance varied from 73.91% to 90.28% accuracy, with the RoBERTa+ model achieving the highest metrics, exceeding 90% in precision, recall, and F1-score. Pre-trained on domain-specific Telegram data, RoBERTa+ demonstrated superior generalization in threat detection, supporting previous findings on pretraining benefits for linguistic generalization [51]. While models like LightGBM and fastText offered faster computation with lower F1 scores, Transformer models, including RoBERTa+ and GPT-2, provided higher F1 scores at the expense of longer computation times, highlighting the trade-offs between computational efficiency (sub-section 6.1) and accuracy (Table 4). In addition, our active learning framework complements LLMs by providing a scalable way to build domain-robust training data tailored for high-risk, nuanced settings like Telegram.

*7.1.2* ***Robustness.*** We evaluated robustness through adversarial inputs, external case study, and temporal analysis. **Adversarial Inputs.** We tested the model's robustness by altering key tokens in a small set of manually selected threat-related text samples to illustrate model vulnerabilities. For instance, in the sentence "Go figure they are criminals; lock them away for a very long time," the tokens "lock" and "criminals" were identified as the most significant (Figures 5 and 4), with "lock" having the greatest influence. Minor adjustments, such as replacing "criminals" with "mob," caused only a small drop in confidence, while changing "lock" to "shoot" led to a major shift in the model's prediction (Figure 6). This demonstrates that while RoBERTa+ is stable against minor alterations, it adapts to more significant changes, aligning with the transparency discussed in sub-section 6.2.2.

**External Case Study.** We tested the RoBERTa+ model on 1 million out-of-distribution (OOD) Telegram replies, categorizing them as "No Threat," "Judicial Threat," or "Non-Judicial Threat." As shown in Table 5, channels like TrumpSupportersChannel, ResisttheMainstream, InfoWars, and TheTrumpRepublicans exhibited significant threat-related content. For instance, TrumpSupportersChannel had 2.91% judicial and 3.30% non-judicial threats, often tied to legal matters and ideological engagement. ResisttheMainstream showed similar patterns, with 2.46% judicial and 3.78% non-judicial threats. InfoWars had 2.01% judicial and 4.38% non-judicial threats, reflecting



| True Label | Predicted Label | Attribution Label | Attribution Score | Word Importance |
|---|---|---|---|---|
| Perturbed - Judicial Threat | Judicial Threat (2.45) | Perturbed - Judicial Threat | 1.15 | #s Go figure they are mob lock them away for a very long time . #/s |

| True Label | Predicted Label | Attribution Label | Attribution Score | Word Importance |
|---|---|---|---|---|
| Perturbed - Judicial Threat | Non-Judicial Threat (3.57) | Non-Judicial Threat | -0.69 | #s Go figure they are criminals shoot them away for a very long time . #/s |

Fig. 6. Adversarial perturbations for RoBERTa+: (top) The word *"criminals"* is perturbed to *"mob"* with no change in prediction. (bottom) The word *"lock"* is perturbed to *"shoot"*, causing the prediction shift to "Non-Judicial Threat."

Table 5. Classification of Threat Types by Telegram Channel

| Channel | Replies | No Threat (%) | Judicial Threat (%) | Non-Judicial Threat (%) |
|---|---|---|---|---|
| LacansWhreHouse | 41399 | 99.28 | 0.10 | 0.62 |
| RiotDogs | 41479 | 98.98 | 0.12 | 0.90 |
| PrayingMedic | 36625 | 98.92 | 0.50 | 0.58 |
| AnticapitalistSurrealism | 42042 | 98.89 | 0.14 | 0.97 |
| UnitedAnarchists | 29993 | 98.87 | 0.16 | 0.97 |
| COVIDVACCINEVICTIMS | 24709 | 98.26 | 0.92 | 0.81 |
| PatriotStreetfighter | 58240 | 97.90 | 0.74 | 1.36 |
| BlackCrimesMatter | 600 | 97.83 | 0.33 | 1.83 |
| DonaldTrumpJr | 73258 | 97.54 | 1.27 | 1.19 |
| ThePatriotVoice | 83482 | 96.43 | 1.48 | 2.09 |
| FreedomFighters | 7607 | 96.08 | 0.78 | 3.14 |
| AlexJones | 5509 | 95.63 | 1.76 | 2.61 |
| TrumpSupportersChannel | 74522 | 93.78 | 2.91 | 3.30 |
| ResisttheMainstream | 241033 | 93.76 | 2.46 | 3.78 |
| InfoWars | 8774 | 93.62 | 2.01 | 4.38 |
| TheTrumpRepublicans | 80297 | 92.71 | 3.58 | 3.70 |
| AnalyzingAmerica | 150431 | 90.07 | 4.87 | 5.06 |

resistance narratives. The highest levels were found on AnalyzingAmerica, with 4.87% judicial and 5.06% non-judicial threats, highlighting the model's effectiveness in high-stakes, ideologically charged discussions.

**Temporal Analysis:** Figure 7 illustrates the model's adaptation to evolving language patterns and ideologies, maintaining robust classification accuracy amid key events like the January 6 insurrection and U.S. elections. For this analysis, the model was trained on the full labeled dataset across all time periods. Only the test set varied by month, allowing us to assess how well the model generalized to messages from different points in time.

*7.1.3* ***Compare Against Human Interpretations.*** We assessed the alignment between the model's explainability methods and human reasoning by comparing its outputs to assessments from a domain expert graduate student. Attention Visualization (sub-section 6.2.1) highlighted terms like "hang" in threat predictions (Figure 2), reflecting human focus on harmful language. Similarly, Neuron Visualization (Figure 3) indicated that specific neurons responded to words like "mob," reinforcing the model's credibility. In Post-hoc Explainability (sub-section 6.2.2), Integrated Gradients (Figure 5) identified terms such as "criminals" as critical for classifying judicial threats, which aligned with human perceptions. This alignment was further supported by LIME (Figure 4), demonstrating that these terms consistently influenced predictions, even after perturbation. Adversarial Robustness (sub-section 7.1.2)

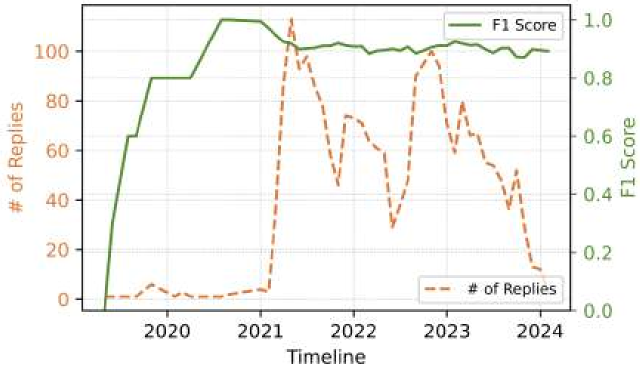


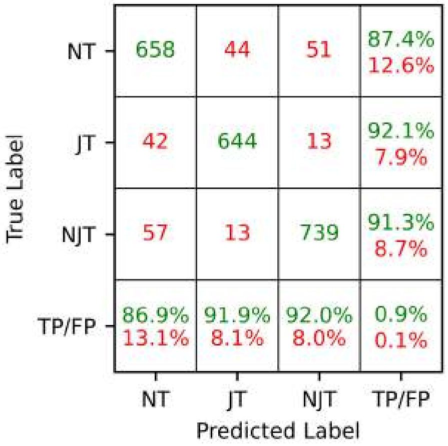
Fig. 7. Test results by month for RoBERTa+

showed that minor perturbations (Figure 6) maintained threat classifications, while more substantial changes resulted in shifts, mirroring human interpretation. Overall, the explainability methods aligned with human evaluations and enhanced trust in the model's threat detection capabilities.

## 7.2 Safety Evaluation

*7.2.1* ***Error Analysis in Critical Scenarios.*** We conducted an error analysis of 2,261 test samples to evaluate the RoBERTa+ model's safety, focusing on false positives (FP) and false negatives (FN). The confusion matrix (Figure 8) revealed that in the "No Threat" category, 658 instances were correctly classified, while 44 were misclassified as judicial threats and 51 as non-judicial threats. For "Judicial Threats," 644 were accurately identified, but 42 genuine threats were missed (FN), and 13 were incorrectly labeled as non-judicial threats. In the "Non-Judicial Threats" category, 739 samples were correct, with 57 misclassified as "No Threat" and 13 as judicial threats. False positives often stemmed from misinterpreting inflammatory language, while false negatives posed risks by failing to detect real threats. Manual review indicated that subtle or domain-specific language contributed to these errors, highlighting the need for improved contextual understanding and more diverse training data. Precision-Recall Curves (Figure 9) demonstrated high precision scores—0.915 for "No Threat," 0.940 for "Judicial Threat," and 0.953 for "Non-Judicial Threat." While the model achieved 90.28% accuracy, the presence of false negatives emphasizes the need for ongoing refinement to enhance safety and reliability.

*7.2.2* ***Impact Assessment.*** Evaluating the consequences of misclassifications is crucial in high-stakes applications like threat detection. Of the 2,261 predictions, 220 were misclassified, highlighting the need for safeguards. This section examines three examples of misclassifications, as shown in Figure 10. The statement *"Freeze and seize all assets and return to American taxpayers!"* was incorrectly labeled as "No Threat," neglecting its legal implications. In another case, *"Hope that teen spends the rest of their miserable existence in prison. POS,"* was misclassified as "No Threat" by an expert, while the model correctly identified it as a "Judicial Threat." Lastly, the statement *"Arrest him, put a mask on him in solitary, and inject him with five doses,"* was classified as a "Judicial Threat" by the expert despite its violent undertones. These examples illustrate how ambiguous language can lead to misclassifications, emphasizing the need for additional training data to enhance threat differentiation. While some human labeling errors occurred, as discussed in sub-section 5.1, the RoBERTa+ model generally captures nuances and delivers accurate predictions. Continuous refinement with more data will further reduce errors and improve safety and reliability in real-world threat detection scenarios.



|  | NT | JT | NJT | TP/FP |
|---|---|---|---|---|
| NT | 658 | 44 | 51 | 87.4%<br>12.6% |
| JT | 42 | 644 | 13 | 92.1%<br>7.9% |
| NJT | 57 | 13 | 739 | 91.3%<br>8.7% |
| TP/FP | 86.9%<br>13.1% | 91.9%<br>8.1% | 92.0%<br>8.0% | 0.9%<br>0.1% |

True Label
Predicted Label

Fig. 8. Confusion matrix for RoBERTa+

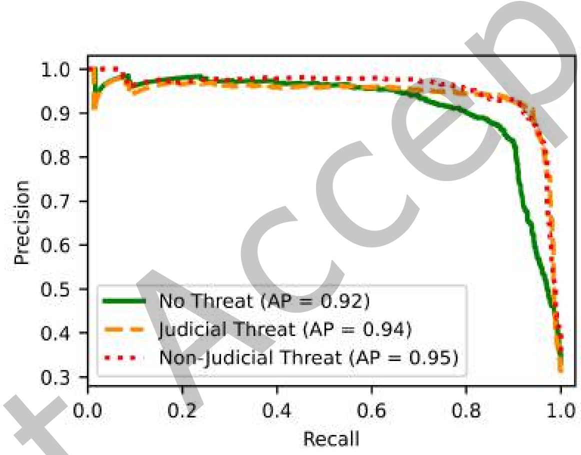
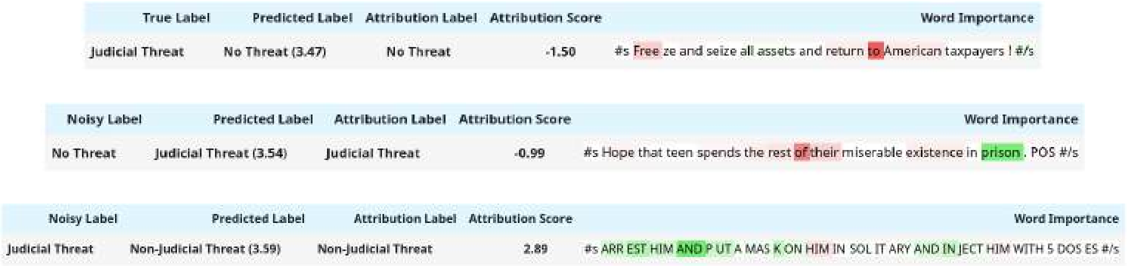


Fig. 9. Precision-Recall curve for RoBERTa+

## 7.3 Fairness Evaluation

**Data Collection, Labeling, and Prediction.** To ensure fair representation, we collected data from various far-right and far-left Telegram channels using diverse keywords to capture a range of political discourse. However, potential bias remains, as low-activity channels were excluded and the dataset is dominated by far-right viewpoints. This imbalance may skew the model's predictions, leading to higher threat detection rates for certain ideological groups. For instance, judicial threat predictions for far-right channels like DonaldTrumpJr (28.57% predicted vs. 28.57% actual) align closely with actual labels, while far-left channels like Riot Dogs show an underprediction of non-judicial threats (23.33% predicted vs. 36.67% actual). The model overpredicts non-judicial threats for far-right channels like Freedom Fighters (52.63% predicted vs. 42.11% actual) and underpredicts for far-left channels like Anticapitalist Surrealism (29.63% predicted vs. 33.33% actual). Despite these differences, the model focuses on message content, not political affiliations, which are not included in the feature set. As discussed in the Adversarial



| True Label | Predicted Label | Attribution Label | Attribution Score | Word Importance |
|---|---|---|---|---|
| Judicial Threat | No Threat (3.47) | No Threat | -1.50 | #s Free ze and seize all assets and return to American taxpayers ! #/s |

| Noisy Label | Predicted Label | Attribution Label | Attribution Score | Word Importance |
|---|---|---|---|---|
| No Threat | Judicial Threat (3.54) | Judicial Threat | -0.99 | #s Hope that teen spends the rest of their miserable existence in prison . POS #/s |

| Noisy Label | Predicted Label | Attribution Label | Attribution Score | Word Importance |
|---|---|---|---|---|
| Judicial Threat | Non-Judicial Threat (3.59) | Non-Judicial Threat | 2.89 | #s ARR EST HIM AND P UT A MAS K ON HIM IN SOL IT ARY AND IN JECT HIM WITH 5 DOS ES #/s |

Fig. 10. Impact of incorrect classifications by RoBERTa+: (top) Misclassified as No Threat despite terms indicating a Judicial Threat (*"Freeze"* and *"Seize"*). (middle) Noisy label of No Threat; model correctly identified it as a Judicial Threat due to imprisonment reference. (bottom) Noisy label of Judicial Threat; model correctly classified it as a Non-Judicial Threat due to the statement's violent nature.

Inputs sub-section (7.1.2), the model's threat detection is based on the substance of the text, ensuring predictions reflect message content rather than group identity. In addition, while we present the outputs of Integrated Gradients, Attention Visualization, and LIME individually, we acknowledge the value in analyzing agreement and disagreement among these methods. Understanding where interpretability techniques converge or diverge, particularly in samples with high uncertainty, could further inform annotators and model evaluators. We propose exploring this as a direction for future work.

### 7.4 Limitations

The custom pretraining of the RoBERTa+ model on 1 million unlabeled Telegram replies resulted in only marginal performance improvements, suggesting that additional computational resources and extended training could yield better results. While RoBERTa was not used to assign ground truth labels, its role in sample selection may introduce subtle bias in downstream performance evaluation, which we mitigate through comparisons with alternative models. Additionally, future comparisons with SOTA LLMs (e.g., GPT-4, LLaMA) could benchmark ALERT, though such models may need prompt tuning for Telegram's nuanced language. Further, fairness evaluations based on demographic parity and equalized odds across political groups revealed disparities in classifying ideologically charged legal discussions and radical narratives. While qualitative XAI methods provided valuable insights, their subjectivity underscores the need for more objective interpretability measures. We also acknowledge the need for a more comprehensive evaluation of adversarial robustness, beyond the small set of examples tested in this study, and highlight this as a direction for future work. Therefore, we recommend exploring bias-correction methods, utilizing larger labeled datasets, and developing interfaces to track metrics such as time, accuracy, and satisfaction. Improving these aspects could enhance the transparency and consistency of model explanations, ultimately fostering interpretability and trust [18, 29].

## 8 CONCLUSION

In this paper, we introduced ALERT, a novel framework integrating active learning and XAI to enhance the detection of threats within Telegram's underexplored ecosystem. By addressing four key research gaps: improved threat classification, a comprehensive Telegram data set of 2.3 million responses, cost-effective labeling strategies, and a scalable XAI framework, we provided a robust solution to identify and explain potential risks on the platform.

Our experiments demonstrated the effectiveness of ALERT, which significantly reduced labeling efforts by 86.5% while maintaining high classification performance, achieving over 90% in precision, recall, accuracy, and

F1-score. The custom RoBERTa+ model, pre-trained on domain-specific data, showed strong generalization capabilities, particularly in capturing extremist rhetoric common within far-right Telegram channels, which were dominant in the dataset. Moreover, the integration of explainability techniques, such as Attention Visualization, Integrated Gradients, and LIME, allowed us to provide interpretable insights, ensuring both model transparency and user trust.

Despite these advancements, our findings indicate areas for improvement, including the marginal gains observed from custom pertaining to the RoBERTa+ model and the need to address fairness disparities in classifying ideologically charged content. We also recognize the limitations of qualitative XAI methods, suggesting the exploration of bias-correction techniques and larger labeled datasets to enhance the reliability of explanations.

Future work will focus on further refining these models, enhancing interpretability through more objective measures, comparisons with emerging SOTA LLMs using prompting-based approaches, and developing user interfaces to better assess the impact of explanations on accuracy, satisfaction, and trust, including comparing interpretability outputs across XAI methods to assess agreement on token-level contributions and developing an integrated interface to visualize these outputs for human annotators and auditors. To support reproducibility and future research, the dataset and source code developed for this work will be released upon publication. Dataset: https://data.mendeley.com/datasets/tm9s68vgxd/1, Code: https://github.com/kamalravi/TelegramThreatLevel. Ultimately, ALERT offers a critical step forward in creating scalable, transparent, and reliable AI-driven threat detection systems for social media platforms, particularly those like Telegram, which remain at the forefront of regulatory scrutiny for extremist content.

## ETHICAL CONSIDERATIONS

Our research analyzes public Telegram channels, excluding private groups to avoid misuse. While this focuses on accessible conversations, it does not fully represent U.S. Telegram users, given the platform's smaller user base in the U.S. [44]. We implemented a standardized annotation framework [40], developed with social science and criminology experts, to ensure consistent threat labeling. The "Judicial Threats" classification identifies content that may incite legal actions, without implying punitive judgments. Although we strove to reduce bias, the reliance on expert annotations presents limitations, particularly in socially and politically sensitive contexts. Ethical clearance was granted by the University (IRB ID: STUDY00006200), exempting the study under secondary research and ensuring adherence to ethical standards.

## REFERENCES

[1] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion* 99 (2023), 101805.

[2] Mohammad Alijanpour and Abolghasem Raie. 2021. Video Event Recognition using Two-Stream Convolutional Neural Networks. In *5th International Conference on Pattern Recognition and Image Analysis*. IEEE, 1–5.

[3] Bobby Allyn. 2024. *Telegram CEO Pavel Durov Arrested in France, Reports Say*. https://www.npr.org/2024/08/25/nx-s1-5088676/telegram-ceo-pavel-durov-arrested-france

[4] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[5] Noman Ashraf, Rabia Mustafa, Grigori Sidorov, and Alexander Gelbukh. 2020. Individual vs. group violent threats classification in online discussions. In *Companion Proceedings of the Web Conference 2020*. 629–633.

[6] SeyedArmin Azizi, Reza Soleimani, Mohsen Ahmadi, Ali Malekan, Laith Abualigah, and Fatemeh Dashtiahangar. 2022. Performance enhancement of an uncertain nonlinear medical robot with optimal nonlinear robust controller. *Computers in Biology and Medicine* 146 (2022), 105567.

[7] Marzieh Babaeianjelodar, Gurram Poorna Prudhvi, Stephen Lorenz, Keyu Chen, Sumona Mondal, Soumyabrata Dey, and Navin Kumar. 2022. Interpretable and high-performance hate and offensive speech detection. In *International Conference on Human-Computer*

*Interaction*. Springer.

[8] Babak Bahador. 2023. Monitoring hate speech and the limits of current definition. In *Challenges and perspectives of hate speech research*. Digital Communication Research, Vol. 12. Digital Communication Research, Berlin, 291–298. Primary publication; peer reviewed.

[9] Jason Baumgartner, Savvas Zannettou, Megan Squire, and Jeremy Blackburn. 2020. The pushshift telegram dataset. In *Proceedings of the international AAAI conference on web and social media*, Vol. 14. 840–847.

[10] Vahid Behzadan, Carlos Aguirre, Avishek Bose, and William Hsu. 2018. Corpus and deep learning classifier for collection of cyber threat indicators in twitter stream. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 5002–5007.

[11] Leonardo Blas, Luca Luceri, and Emilio Ferrara. 2024. Unearthing a Billion Telegram Posts about the 2024 US Presidential Election: Development of a Public Dataset. *University of Southern California, HUMANS Lab – Working Paper No. 2024.5* (October 2024). https://ssrn.com/abstract=5018893

[12] Antony J. Blinken. 2024. United States International Cyberspace & Digital Policy Strategy: Towards an Innovative, Secure, and Rights-Respecting Digital Future. https://www.state.gov/united-states-international-cyberspace-and-digital-policy-strategy/

[13] Enrico Bunde. 2021. AI-Assisted and Explainable Hate Speech Detection for Social Media Moderators - A Design Science Approach. In *54th Hawaii International Conference on System Sciences, HICSS 2021, Kauai, Hawaii, USA, January 5, 2021*. ScholarSpace, 1–10. https://hdl.handle.net/10125/70766

[14] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. [n. d.]. What Does BERT Look at? An Analysis of BERT‘s Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. https://doi.org/10.18653/v1/W19-4828

[15] Federal Communications Commission. 2024. Safeguarding and Securing the Open Internet; Restoring Internet Freedom. https://www.federalregister.gov/documents/2024/05/22/2024-10674/safeguarding-and-securing-the-open-internet-restoring-internet-freedom

[16] 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. 512–515.

[17] 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.

[18] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[19] Ismail Elezi, Zhiding Yu, Anima Anandkumar, Laura Leal-Taixe, and Jose M Alvarez. 2022. Not all labels are equal: Rationalizing the labeling costs for training object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14492–14501.

[20] Hugo L Hammer, Michael A Riegler, Lilja Øvrelid, and Erik Velldal. 2019. Threat: A large annotated corpus for detection of violent threats. In *IEEE Intl. Conference on Content-Based Multimedia Indexing*. 1–5.

[21] Sanne A Hoogenboom, Kamalakkannan Ravi, Megan M Engels, Ismail Irmakci, Elif Keles, Candice W Bolan, Michael B Wallace, and Ulas Bagci. 2021. Missed diagnosis of pancreatic ductal adenocarcinoma detection using deep convolutional neural network. *Gastroenterology* 160, 6 (2021), 00794–0. https://doi.org/10.1016/S0016-5085(21)00794-0

[22] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. FastText.zip: Compressing text classification models. *arXiv preprint* (2016). https://arxiv.org/abs/1612.03651

[23] R Kamalakkannan, R Rajkumar, M Madan Raj, and S Shenbaga Devi. 2014. Imagined speech classification using EEG. *Advances in biomedical science and engineering* 1, 2 (2014), 20–32. https://www.researchgate.net/publication/309967859_Imagined_Speech_Classification_using_EEG

[24] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* 30 (2017).

[25] Hareem Kibriya, Ayesha Siddiqa, Wazir Zada Khan, and Muhammad Khurram Khan. 2024. Towards safer online communities: Deep learning and explainable AI for hate speech detection and classification. *Computers and Electrical Engineering* 116 (2024), 109153.

[26] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. 2020. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896* (2020).

[27] Santhosh Kumar, Kamalakkannan Ravi, Supriti Mulay, Keerthi Ram, and Mohanasankar Sivaprakasam. 2018. Deep Residual Network based Automatic Image Grading for Diabetic Macular Edema. In *Research Poster Papers of the 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. https://doi.org/10.13140/RG.2.2.24611.02082/1

[28] Massimo La Morgia, Alessandro Mei, and Alberto Maria Mongardini. 2025. TGDataset: Collecting and Exploring the Largest Telegram Channels Dataset. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*. Association for Computing Machinery, New York, NY, USA, 2325–2334. https://doi.org/10.1145/3690624.3709397

[29] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2019. An Evaluation of the Human-Interpretability of Explanation. *ArXiv* abs/1902.00006 (2019). https://api.semanticscholar.org/CorpusID:263787636

[30] Simon A. Levin, Helen V. Milner, and Charles Perrings. 2021. The dynamics of political polarization. *Proceedings of the National Academy of Sciences* 118, 50 (2021), e2116950118. https://doi.org/10.1073/pnas.2116950118

[31] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019). arXiv:1907.11692 http://arxiv.org/abs/1907.11692

[32] Debanjan Mahata, Haimin Zhang, Karan Uppal, Yaman Kumar, Rajiv Shah, Simra Shahid, Laiba Mehnaz, and Sarthak Anand. 2019. MIDAS at SemEval-2019 task 6: Identifying offensive posts and targeted offense from twitter. In *13th International Workshop on Semantic Evaluation*. 683–690.

[33] Joshua Melton, Arunkumar Bagavathi, and Siddharth Krishnan. 2020. DeL-haTE: a deep learning tunable ensemble for hate speech detection. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 1015–1022.

[34] Mourad Oussalah, F Faroughian, and Panos Kostakos. 2018. On detecting online radicalization using natural language processing. In *Intelligent Data Engineering and Automated Learning–IDEAL 2018: 19th International Conference, Madrid, Spain, Proceedings, Part II 19*. Springer, 21–27.

[35] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[36] Kamalakkannan Ravi, Sakthivel Selvaraj, Supriti Mulay, Keerthi Ram, and Mohanasankar Sivaprakasam. 2018. Breast cancer histology classification using Deep Residual Networks. In *Research Poster Papers of the 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. https://doi.org/10.13140/RG.2.2.22094.43840

[37] Kamalakkannan Ravi and Adan Ernesto Vela. 2024. Comprehensive dataset of user-submitted articles with ideological and extreme bias from Reddit. *Data in Brief* 56 (2024), 110849. https://doi.org/10.1016/j.dib.2024.110849

[38] Kamalakkannan Ravi and Adan Ernesto Vela. 2024. RICo: Reddit ideological communities. *Online Social Networks and Media* 42 (2024), 100279. https://doi.org/10.1016/j.osnem.2024.100279

[39] Kamalakkannan Ravi, Adan Ernesto Vela, and Rickard Ewetz. 2022. Classifying the Ideological Orientation of User-Submitted Texts in Social Media. In *Proceedings of the 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 413–418. https://doi.org/10.1109/ICMLA55696.2022.00066

[40] Kamalakkannan Ravi, Adan Ernesto Vela, Elizabeth Jenaway, and Steven Windisch. 2023. Exploring Multi-Level Threats in Telegram Data with AI-Human Annotation: A Preliminary Study. In *2023 International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 1520–1527. https://doi.org/10.1109/ICMLA58977.2023.00229

[41] Kamalakkannan Ravi and Jiann-Shiun Yuan. 2024. Ideological orientation and extremism detection in online social networking sites: A systematic review. *Intelligent Systems with Applications* 24 (2024), 200456. https://doi.org/10.1016/j.iswa.2024.200456

[42] Burr Settles. 2009. Active Learning Literature Survey. University of Wisconsin-Madison Department of Computer Sciences. https://api.semanticscholar.org/CorpusID:324600

[43] Shuhao Shi, Kai Qiao, Jian Chen, Shuai Yang, Jie Yang, Baojie Song, Linyuan Wang, and Binghai Yan. 2023. MGTAB: A Multi-Relational Graph-Based Twitter Account Detection Benchmark. *ArXiv* abs/2301.01123 (2023). https://api.semanticscholar.org/CorpusID:255393896

[44] Galen Stocking, Amy Mitchell, Katerina Eva Matsa, Regina Widjaya, Mark Jurkowitz, Shreenita Ghosh, Aaron Smith, Sarah Naseer, and Christopher St Aubin. 2022. The role of alternative social media in the news and information environment. *Pew Research Center* (2022).

[45] Azwad Tamir, Milad Salem, and Jiann-Shiun Yuan. 2023. ProtEC: A Transformer Based Deep Learning System for Accurate Annotation of Enzyme Commission Numbers. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2023).

[46] The New York Times. 2024. *Has the Tide Turned for TikTok, Telegram and X?* https://www.nytimes.com/2024/09/10/opinion/telegram-tiktok-x-social-media.html

[47] Department of Justice United States. 2024. Leaders of Transnational Terrorist Group Charged with Soliciting Hate Crimes, Soliciting the Murder of Federal Officials, and Conspiring to Provide Material Support to Terrorists. https://www.justice.gov/archives/opa/pr/leaders-transnational-terrorist-group-charged-soliciting-hate-crimes-soliciting-murder

[48] Ameya Vaidya, Feng Mai, and Yue Ning. 2020. Empirical analysis of multi-task learning for reducing identity bias in toxic comment detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 683–693.

[49] Jesse Vig. 2019. A Multiscale Visualization of Attention in the Transformer Model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Florence, Italy, 37–42.

[50] Samantha Walther and Andrew Mccoy. 2021. US Extremism on Telegram: Fueling Disinformation, Conspiracy Theories, and Accelerationism. *Perspectives on Terrorism* 15, 2 (2021), 100–124. https://api.semanticscholar.org/CorpusID:234492283

[51] Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R. Bowman. 2020. Learning Which Features Matter: RoBERTa Acquires a Preference for Linguistic Generalizations (Eventually). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 217–235. https://doi.org/10.18653/v1/2020.emnlp-main.16

[52] 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. https://doi.org/10.18653/v1/W16-5618

[53] Kim L Withers, James L Parrish, Steven Terrell, and Timothy J Ellis. 2017. The relationship between the "dark triad" personality traits and deviant behavior on social networking sites. In *AMCIS 2017 Proceedings*. https://aisel.aisnet.org/amcis2017/SocialComputing/Presentations/14

[54] Thomas Wolf and Others. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online, 38–45.

[55] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 1415–1420. https://doi.org/10.18653/v1/N19-1144