


A Study of Unified Framework for Extremism Classification, Ideology Detection, Propaganda Analysis, and Flagged Data Detection Using Transformers

R S Lakshmi Balajia¹, C S Thiruvengkataswamy², Malathy Batumalay³, N. Duraimutharasan⁴,
Amar Dev Thirukulam Devadas⁵, Thaweesak Yingthawornsuk^{6,*}

¹Department of Advanced Computing Sciences, Academy of Maritime Education and Training University (AMET), Chennai, India

²Program Director, Govt of India, Tamil Nadu, India

³Faculty of Data Science and Information Technology, INTI International University, Nilai, Malaysia

⁴Professor, School of Computer Science and Applications, Reva University, Bangalore, India

⁵Faculty of Innovation Design Management, University of Europe for Applied Sciences, Potsdam, Germany

⁶Department of Media Technology, King Mongkut's University of Technology Thonburi, Bangkok, Thailand

(Received: November 28, 2024; Revised: January 5, 2025; Accepted: March 10, 2025; Available online: July 10, 2025)

Abstract

The rise of extremism and its rapid dissemination through propaganda channels have become pressing global challenges, threatening peace, security, and social cohesion. This study aligns with the United Nations Sustainable Development Goal 16 by proposing a unified framework leveraging advanced machine learning and large language models to combat extremism through extremism classification, ideology detection, propaganda analysis, and flagged word recognition. This framework introduces process innovation by integrating state-of-the-art transformer models such as BERT, RoBERTa, DistilBERT and XLNet to streamline the analysis process and overcome traditional limitations in extremism detection with exceptional performance: 90.00% accuracy for extremism classification, 98.82% accuracy for ideology detection, and 99.71% accuracy for flagged word recognition. While the proposed approach demonstrates high precision and recall, it faces challenges such as potential data bias, ethical concerns in dataset usage and the risk of false positives, which could lead to misclassification of benign content. The inclusion of multilingual capabilities broadens the applicability of the framework but variations in linguistic structures and cultural contexts introduce complexities in model generalization. Additionally, ethical considerations in handling extremist content, especially in social media data collection, necessitate stringent privacy safeguards to prevent unintended harm. By providing actionable insights, this research contributes to counter-extremism efforts in areas such as online content moderation, law enforcement and intelligence analysis, laying a foundation for future advancements in safeguarding global security which enhance the process innovation.

Keywords: Extremism Classification, Ideology Detection, Machine Learning, Propaganda Analysis, Flagged Word Recognition, Large Language Models, BERT, Counter Extremism, Natural Language Processing, Process Innovation

1. Introduction

In the modern era, the ease of disseminating extremist ideas and related propaganda poses a significant threat to social order and integration [1]. This phenomenon began to escalate in the early 2000s, coinciding with the rapid expansion of the internet and social media, which allowed harmful ideologies to reach broader and more vulnerable audiences [2]. These platforms enable individuals and groups to propagate radical messages efficiently, often targeting those susceptible to such influence [3]. The current digital landscape has amplified opportunities for the promotion of extreme beliefs, leading to societal divisions [4]. Extremism—marked by an unwillingness to compromise and the adoption of

*Corresponding author: Thaweesak Yingthawornsuk (thaweesak.yin@kmutt.ac.th)

 DOI: <https://doi.org/10.47738/jads.v6i3.702>

This is an open access article under the CC-BY license (<https://creativecommons.org/licenses/by/4.0/>).

© Authors retain all copyrights

radical viewpoints and tactics—has become a pressing global issue. It manifests in various forms, including religious violence, political assassinations, and antisocial ideologies that frequently result in aggression and instability [5].

Over the past two decades, the threat has intensified as extremist groups increasingly shift their activities online. They exploit digital tools to incite hatred, recruit followers, and spread their narratives, contributing to hate crimes, violent radicalization, and public unrest [6]. This undermines democratic values, social equity, and human rights. A particular challenge lies in distinguishing between legitimate political discourse and extremist rhetoric [7], [8], which complicates detection and response efforts. Misjudging such communications may result in either overreaction or inaction due to the complex and evolving nature of radical content [9].

Moreover, the dynamic nature of extremist ideologies, which constantly adapt to societal responses and countermeasures, complicates identification and intervention [10]. Current detection methods are insufficient in isolating and interpreting radicalized content, highlighting a need for more advanced technological solutions [11]. Modern Machine Learning (ML) models, particularly those based on transformer architectures such as BERT, RoBERTa, and XLNet, offer promising tools for this task. Unlike traditional models like Naïve Bayes or Support Vector Machines, which struggle with contextual understanding, transformer-based models process text bidirectionally and capture long-range dependencies, enabling superior performance in identifying implicit and coded extremist language [7].

BERT utilizes contextual embeddings to enhance extremism classification; RoBERTa refines this through improved training strategies, while XLNet employs permutation-based learning to generalize better across linguistic variations [7]. These models not only improve our grasp of radicalization processes but also support educators, policymakers, and community leaders in countering extremist narratives.

This study aims to explore a machine learning–driven framework to address these critical issues. Such an approach has the potential to enable proactive measures that prevent the emergence of extremist content, thereby fostering mutual respect, social harmony, and peaceful coexistence.

2. Literature Review

The rise of extremist ideologies and the threats they pose have prompted multidisciplinary investigations in fields such as psychology, sociology, and computer science [12]. Scholars have explored the factors driving radicalization, the characteristics of extremist content, and the effectiveness of various detection strategies. Early models of radicalization emphasized individual and collective grievances as primary motivators of extremism, drawing from qualitative assessments of ideological shifts and social pressures [13], [14].

A significant development in the field was the shift toward analyzing extremist activity in online environments. Empirical research using large-scale data analyses revealed how extremist groups exploit digital platforms, particularly social media, to spread their ideologies and recruit new followers [15]. This approach offered a more comprehensive understanding of behavioral patterns compared to earlier studies based solely on case studies or interviews.

The complexity of online extremism has led to the growing adoption of computational methods for detection and classification [16]. Traditional psychological and sociological models, while insightful, often lack scalability and fail to provide timely responses in dynamic digital spaces. Machine learning and natural language processing techniques address this limitation by leveraging large volumes of textual data to automate the identification of extremist content.

Transformer-based deep learning models such as BERT and RoBERTa have demonstrated particular effectiveness in this context. These models can capture contextual meaning and ideological nuances within online narratives, making them well-suited for real-world counter-extremism applications [17]. Studies have shown that accurate classification of extremist content requires diverse and well-constructed datasets to ensure robustness and generalizability [18].

The introduction of transformer models marked a turning point in natural language processing. These models significantly enhanced performance in tasks such as text classification, including the detection of radical ideologies, due to their ability to interpret contextual and semantic features [19]. Despite their effectiveness, concerns remain regarding

the ethical implications of using artificial intelligence in this domain. Issues such as algorithmic bias and lack of transparency have been raised, emphasizing the need for responsible AI development [20].

Recent efforts to improve classification systems have led to the creation of more balanced and multi-class datasets. One such dataset enabled classification of extremist content across ideologies and categories, including propaganda, radicalization, and recruitment, using models like BERT, RoBERTa, and DistilBERT. These systems achieved f1-scores as high as 0.72, demonstrating the feasibility of accurate multi-ideology classification [21]. However, the constantly evolving nature of extremist discourse still presents significant challenges.

This study builds on these findings by proposing a unified machine learning framework for detecting and classifying extremist content. The framework incorporates several transformer-based models—BERT, RoBERTa, DistilBERT, T5, and XLNet—and demonstrates high performance across multiple tasks: 90.00% accuracy for extremism classification, 98.82% for ideology detection, 99.71% for flagged word recognition, and 88.03% for propaganda analysis. Among these, RoBERTa achieved the highest classification accuracy at 99.91%.

The inclusion of multilingual capabilities further extends the applicability of this framework across different languages and cultural contexts. These results underscore the effectiveness of transformer-based models in supporting counter-extremism initiatives and contribute to more proactive and targeted efforts against radicalization [13].

3. Methodology

As illustrated in figure 1, This study employs a multi-faceted methodological approach to develop a unified framework for extremism classification, ideology detection, propaganda analysis, and flagged word recognition. The methodology is structured into several key components: data collection, pre-processing, feature extraction, model development, and evaluation.

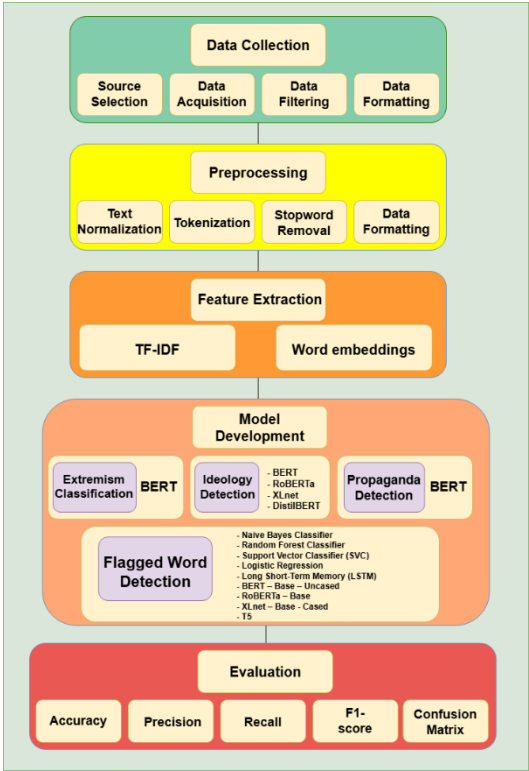


Figure 1. Methodological Approach for Developing a Unified Framework

3.1. Dataset Construction and Description

A comprehensive and ethically curated dataset was constructed to capture extremist content from a wide range of online sources. As illustrated in figure 2, data was collected from Twitter, Reddit, 4chan, religious texts, academic research

articles, white supremacist platforms, and news outlets. The dataset encompasses ideological, political, and religious narratives across both contemporary and historical contexts. All collection activities adhered to strict ethical protocols, ensuring compliance with data protection regulations such as the GDPR. Personally identifiable information (PII) was anonymized, and content involving vulnerable groups was excluded. Institutional ethical clearance was obtained to ensure the responsible handling of sensitive material.

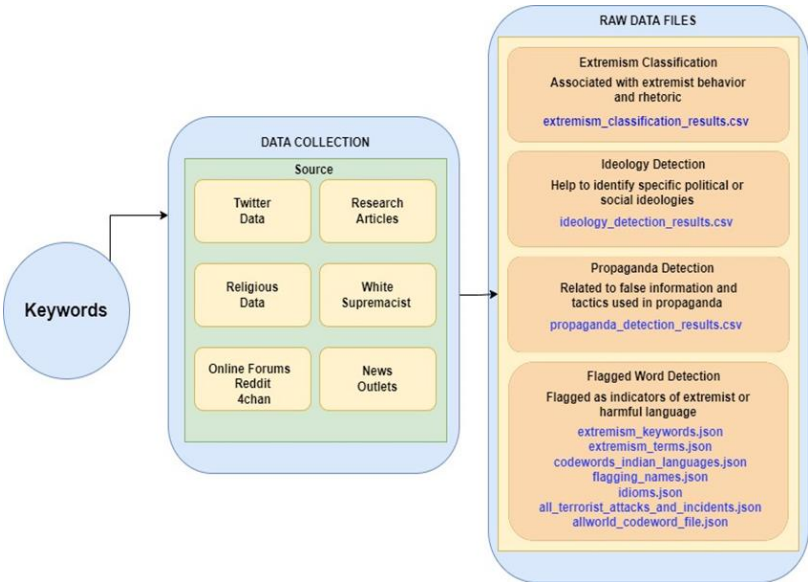


Figure 2. Methodology of Data Collection

Keyword-based filtering techniques were applied using terms such as “jihad,” “radicalization,” “recruitment,” and “propaganda,” augmented with language-specific keywords drawn from domain expertise. The dataset supports nine languages—Tamil, Hindi, Malayalam, Telugu, Urdu, Arabic, Thai, Malay, and English. To maintain contextual richness, texts were preserved in their original form or translated, when necessary, particularly for idiomatic and culturally embedded expressions. The use of a hybrid flagged-word detection approach—combining multilingual NLP tools with expert validation—further ensured the accuracy and cultural relevance of the identified extremist content. Priority was given to high-engagement posts, defined as those receiving over 50 retweets or 100 likes. Data collection spanned from 2018 to 2024 to reflect current extremist trends and global developments.

The dataset was organized into key themes such as propaganda, recruitment, and ideological rhetoric. Each tweet is documented with detailed attributes including Tweet ID, User ID, Username, Tweet Content, Timestamp, Language, Retweet Count, Like Count, Hashtags, Mentions, and Geo-location. Crucially, additional fields include Flagged Words, Ideology Label, Extremism Classification, and Propaganda Classification—enabling nuanced analysis for various machine learning tasks. This structured metadata is summarized in [table 1](#).

Table 1. Dataset Attributes

Attribute	Type	Description
Tweet ID	String	Unique identifier for each tweet
User ID	String	Unique identifier for the user
Username	String	Display name of the user
Tweet Content	Text	The actual text of the tweet
Timestamp	DateTime	Date and time when the tweet was posted
Language	String	Language of the tweet
Retweet Count	Integer	Number of times the tweet has been retweeted
Like Count	Integer	Number of likes received by the tweet
Hashtags	List of Strings	List of hashtags used in the tweet

Mentions	List of Strings	List of usernames mentioned in the tweet
Geo-location	String (or null)	Geographical location from which the tweet was posted, if available
Flagged Words	List of Strings	Words identified as extremist or flagged
Ideology Label	String (or null)	Detected ideology label associated with the tweet
Extremism Classification	Boolean	Whether the tweet is classified as extremist
Propaganda Classification	Boolean	Whether the tweet contains propaganda content

The dataset was further enriched by incorporating various supporting files. For instance, Idioms.json includes 2,154 idioms across 9 languages; Event.json contains over 500 records of national holidays and cultural events; and Flagging_names.json lists 497 names of terrorist leaders, operatives, and flagged cities. The Extremism_terms_codeword_indian_languages.json contributes 3,750 relevant terms in Indian languages, while All_terrorist_attacks_and_incidents.json documents 540 global terrorist events. The Mcodewords.json file includes 1,458 vocabulary items linked to terrorism and religious extremism. The Extremism_keywords.json provides 900 keywords related to radical discourse and methods. Core Twitter data includes 28,411 posts related to extremism, and an additional 21,347 entries comprise recruitment, propaganda, and radicalization content. The ISIS/Jihadist subset adds 2,900 records detailing propaganda and mobilization strategies. Collectively, this rich, multilingual, and thematically diverse dataset forms a solid foundation for detecting and analyzing extremist content using advanced machine learning techniques.

3.2. Dataset Preparation and Pre-processing

To support accurate and scalable extremism analysis, a comprehensive dataset was compiled from diverse sources, including Twitter, Reddit, 4chan, religious texts, news media, and extremist-affiliated platforms. The collected data reflects political, religious, and ideological content across both historical and contemporary contexts. To enrich the linguistic and thematic coverage, several structured datasets were incorporated, offering cultural, geographical, and temporal diversity. These include idioms, named entities, ideology-linked keywords, and historical terrorism events. The integration of such datasets supports multilingual analysis and provides contextual grounding for machine learning tasks. Table 2 show overview of dataset used in this research.

Table 2. Supplementary Dataset Overview

File Name	Description	Languages	Data Count
Event.json	National and religious festivals, keynote events	Not applicable	500
Idioms.json	Collection of idioms from various cultures	Multiple (9 languages)	2,154
Flagging_names.json	Names of terrorist leaders, operatives, and flagged cities	Not applicable	497
Extremism_terms_codeword_indian_languages.json	Terrorism-related terms in Indian languages	Multiple (9 languages)	3,750
All_terrorist_attacks_and_incidents.json	Historical data on terrorist incidents with names and descriptions	Not applicable	540
Mcodewords.json	General extremist vocabulary including insurgency and radicalism	Multiple (9 languages)	1,458
Extremism_Keywords.json	Propaganda terms, radicalization methods, and geographical keywords	Multiple (9 languages)	900
Twitter Dataset	Tweets related to terrorism and extremism	Not applicable	28,411
Recruitment, Radicalization, Propaganda Dataset	Content focused on radicalization and extremist propaganda activities	Not applicable	21,347
ISIS/Jihadist Dataset	Specific dataset on ISIS and jihadist propaganda and mobilization	Not applicable	2,900

Once compiled, the dataset underwent a comprehensive pre-processing phase to ensure cleanliness, linguistic consistency, and suitability for training advanced machine learning models. The complete preprocessing workflow is illustrated in [figure 3](#), which outlines each step of the pipeline from raw data intake to structured and normalized output. The process was designed to address challenges specific to extremist discourse, such as ideological ambiguity, multilingual noise, and cultural nuance.

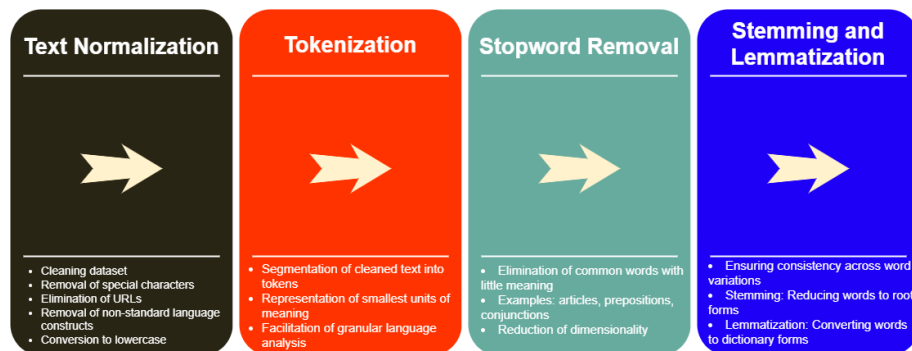


Figure 3. Preprocessing Steps for Text Data

Language-specific processing techniques were applied to account for structural and cultural variations. Tokenization was implemented using advanced tokenizers such as the BERT tokenizer, which is optimized for complex scripts including Arabic, Tamil, and Hindi. For languages with high idiomatic density, such as Tamil and Hindi, idiomatic expressions were translated to English using domain glossaries and pre-trained language models to preserve contextual integrity. Posts written in hybrid forms, such as Hinglish and Tanglish, were segmented into dominant language components and processed accordingly.

Text was standardized by removing URLs, emojis, HTML tags, and converting all content to lowercase. A refined stopwords list was used to eliminate generic non-informative words while retaining domain-relevant terms with ideological significance, such as "freedom" and "resistance". WordPiece tokenization was employed to manage rare or compound terms, followed by lemmatization to reduce inflected words to their root forms for greater analytical coherence.

To address data imbalance in categories like recruitment and propaganda, synthetic oversampling techniques such as SMOTE were used. This was further reinforced with manual augmentation involving the insertion of synonymous keywords and equivalent phrases. Synonym replacement techniques and back-translation were also applied to improve generalizability and robustness, especially in low-resource languages like Malayalam and Thai. As demonstrated in [figure 3](#), the pre-processing workflow ensured that the dataset is linguistically diverse, ethically sound, and technically robust. This allows for high-quality training of machine learning models to perform complex tasks such as extremism classification, ideology detection, and propaganda recognition with accuracy and cultural sensitivity.

3.3. Feature Extraction

After completing the pre-processing phase, the next crucial step was feature extraction, which involved transforming the cleaned textual data into structured numerical representations suitable for machine learning and deep learning models. Feature extraction plays a vital role in bridging the gap between unstructured text and computational models by encoding semantic, lexical, and contextual information that can be learned during training. This study adopted a hybrid approach that integrates both traditional natural language processing techniques and modern deep learning-based representations to effectively capture the multi-dimensional characteristics of extremist content.

Traditional methods were first applied to provide a simple yet interpretable baseline for textual feature extraction. Two widely used approaches—Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF)—were utilized to represent word-level characteristics. The Bag-of-Words model treated each text as an unordered collection of words, focusing on term frequency without accounting for word order or semantic dependencies. Despite its simplicity,

BoW enabled comparative analysis based on raw term occurrence, which was helpful in identifying dominant thematic patterns.

To address BoW's limitations in recognizing term significance, the TF-IDF method was employed. TF-IDF calculates the importance of a word in a given document relative to its frequency across the entire corpus. This technique prioritized domain-specific vocabulary such as "jihad" or "radicalization," which appear less frequently but carry substantial ideological weight. Words with high document frequency, such as "the" or "and," were naturally de-emphasized. Together, BoW and TF-IDF provided a strong and interpretable foundation for understanding term-level salience, especially in the context of detecting rhetoric common to extremist narratives.

While traditional models were useful for capturing lexical patterns, they lacked the ability to encode semantic meaning and contextual nuance. To address these limitations, the study adopted transformer-based word embeddings, specifically using Bidirectional Encoder Representations from Transformers (BERT). Unlike earlier models that read text sequentially, BERT processes entire sentences bidirectionally, which enables it to interpret a word in relation to both its preceding and succeeding terms. This characteristic is especially important in detecting subtle or ambiguous expressions in extremist discourse.

BERT embeddings allowed the model to differentiate the meaning of terms like "radical" depending on context, such as distinguishing between radical thought in academic discussion versus radical ideologies in extremist content. This contextualization significantly enhanced the model's ability to identify hate speech and propaganda that may otherwise go undetected by classical approaches. By examining the broader linguistic environment, BERT improved the accuracy of classifying content as ideological or propagandistic, regardless of whether the language was explicit or implicit.

The combination of traditional NLP techniques with deep contextual embeddings resulted in a rich and balanced feature space. This integration enabled the models to learn both term-specific patterns and higher-level semantic relationships, enhancing the classification of extremist and radical content across multilingual and culturally diverse contexts. Ultimately, this hybrid feature extraction approach contributed to more precise and context-aware detection of harmful narratives in the dataset.

3.4. Model Development

The classification framework for extremist content detection was developed using a combination of traditional machine learning and advanced deep learning models. This hybrid approach aimed to balance interpretability, efficiency, and contextual understanding across a diverse and multilingual dataset. The objective was to build a robust, scalable, and ethically applicable system for identifying radical, ideological, and propagandistic narratives.

Several machine learning algorithms were explored for their classification performance on high-dimensional text data. Support Vector Machines (SVM) were chosen for their effectiveness in managing complex, non-linear decision boundaries, especially in high-dimensional feature spaces like those generated through TF-IDF vectors. Logistic Regression was included for its speed and reliability in binary classification and its ease of interpretability through probability estimates. Naïve Bayes, a probabilistic model, proved effective for handling large sparse text datasets, thanks to its reliance on word frequency distributions. Despite its assumption of feature independence, it performed well in preliminary tests. Random Forest was introduced to enhance accuracy and robustness by combining the outputs of multiple decision trees. This ensemble technique helped reduce overfitting and allowed the model to capture more nuanced patterns within noisy and variable social media text.

To capture contextual and semantic nuances in extremist discourse, transformer-based models were employed. BERT served as the foundation due to its strong performance in text classification tasks and its bidirectional attention mechanism, which enables it to interpret words within broader sentence contexts. BERT significantly improved the identification of subtle language shifts and implicit ideological content. RoBERTa, an optimized variant of BERT, was also used for its superior pre-training methodology that omits the Next Sentence Prediction objective and relies on larger batch sizes and corpora. This model was especially useful in handling deeply contextualized narratives.

DistilBERT, a lightweight version of BERT, was included to support tasks requiring faster inference, such as real-time flagged word identification. Although it is computationally more efficient, it retains approximately 97% of BERT’s accuracy. XLNet was selected for its autoregressive permutation-based architecture, enabling it to learn bidirectional dependencies while preserving sequence order. This made it particularly suitable for longer and complex sequences of extremist text. T5, a text-to-text transfer model, was employed for its flexibility in converting any NLP task into a unified format, enabling integration across classification, translation, summarization, and generation tasks.

All models were trained using a stratified sample of the dataset, preserving the balance among extremist, non-extremist, and propaganda classes. The dataset was split into 70% for training, 15% for validation, and 15% for testing. A 5-fold cross-validation strategy was used to evaluate generalizability and reduce overfitting. Performance metrics including precision, recall, and F1-score were computed per fold and averaged to assess classification effectiveness. Hyperparameter optimization was conducted using grid search and randomized search techniques. [Table 3](#) summarizes the final hyperparameter configurations for each model.

Table 3. Hyperparameters for Machine Learning and Deep Learning Models

Model	Key Hyperparameters
Naive Bayes	Smoothing Parameter (alpha): 1.0
Random Forest	Number of Trees: 200, Max Depth: 20, Min Samples Split: 2, Min Samples Leaf: 1, Bootstrap: True, Class Weight: 'balanced'
SVC	Kernel: RBF, Regularization Parameter (C): 1.0, Gamma: 'scale', Max Iterations: 1000, Class Weight: 'balanced'
Logistic Regression	Solver: 'lbfgs', Regularization Penalty: L2, Regularization Parameter (C): 1.0, Max Iterations: 500
LSTM	Embedding Dimension: 300, Hidden Units: 128, Dropout Rate: 0.3, Recurrent Dropout: 0.2, Batch Size: 32, Learning Rate: 1e-3, Optimizer: Adam, Epochs: 20
BERT	Pretrained Model: bert-base-multilingual-cased, Learning Rate: 3e-5, Batch Size: 16, Max Sequence Length: 128, Dropout Rate: 0.1, Optimizer: AdamW, Weight Decay: 0.01, Epochs: 10
RoBERTa	Pretrained Model: roberta-base, Learning Rate: 3e-5, Batch Size: 16, Max Sequence Length: 128, Dropout Rate: 0.1, Optimizer: AdamW, Epochs: 10
DistilBERT	Pretrained Model: distilbert-base-uncased, Learning Rate: 5e-5, Batch Size: 32, Max Sequence Length: 128, Epochs: 8
XLNet	Pretrained Model: xlnet-base-cased, Learning Rate: 2e-5, Batch Size: 16, Max Sequence Length: 128, Epochs: 8
T5	Pretrained Model: t5-small, Learning Rate: 3e-5, Batch Size: 16, Max Sequence Length: 128, Epochs: 10

While transformer-based models like BERT, RoBERTa, and XLNet offered high contextual sensitivity and interpretability, GPT-based models were deliberately excluded from this study. This decision was based on several critical limitations of GPT, including uncontrolled output generation, lack of transparency in decision-making, and the risk of producing unverified content. In national security and defense contexts—where reliability, control, and ethical accountability are paramount—these limitations present unacceptable risks. In contrast, the selected transformer models are better suited for high-stakes applications due to their structured architecture, controllable outputs, and alignment with mission-critical requirements.

3.5. Evaluation Metrics

The performance of the developed models was rigorously evaluated using a range of standard classification metrics to ensure a comprehensive assessment of their effectiveness. The following metrics were employed. Accuracy was calculated to determine the overall proportion of correctly classified instances out of the total number of instances in the

dataset. While useful, accuracy alone can be misleading in imbalanced datasets, so it was supplemented with other metrics.

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \tag{1}$$

TP = True Positives (correctly classified positive instances), TN = True Negatives (correctly classified negative instances), FP = False Positives (incorrectly classified positive instances), FN = False Negatives (incorrectly classified negative instances).

Precision was used to measure the proportion of true positive predictions relative to all positive predictions. This metric is particularly important in the context of extremist content detection, as it helps assess the model's ability to minimize false positives, ensuring that non-extremist content is not misclassified as extremist.

$$Precision = \frac{TP}{(TP+FP)} \tag{2}$$

Recall was employed to measure the proportion of actual positive cases that were correctly identified by the model. High recall is essential in identifying all potential extremist content, even at the risk of a higher false-positive rate, ensuring that harmful content is not overlooked.

$$Recall \text{ (Sensitivity)} = \frac{TP}{(TP+FN)} \tag{3}$$

The F1-score, the harmonic mean of precision and recall, was used as a balanced metric to evaluate both the precision and recall of the models. This metric provides a more nuanced assessment, especially when dealing with imbalanced data where one class might be underrepresented.

$$F1 \text{ Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

As per [table 4](#), a confusion matrix was generated to visualize the performance of each model across different categories of extremism. This matrix provided deeper insights into the model's classification behavior, revealing patterns in misclassification and helping to diagnose potential weaknesses in the model's predictions. Additionally, the models were tested against a hold-out test set, composed of data not used during training or cross-validation. This allowed for an assessment of the models' generalizability and their ability to perform accurately on unseen data. By employing these evaluation metrics, the study ensured a robust and reliable analysis of the models' performance in detecting extremist content.

Table 4. Confusion Matrix for Model Evaluation

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

4. Results and Discussion

The results presented in this section highlight the performance of the developed models for detecting and classifying extremism, ideologies, propaganda, and flagged words associated with extremist content [\[42\]](#). A detailed analysis of each model’s accuracy and effectiveness provides insights into their capabilities and areas for improvement. [Table 5](#) summarizes the accuracy achieved by each model.

4.1. Extremism Classification Model

The Extremism Classification Model uses BERT which further enhances the traditional word embedding’s through addressing the context and semantic aspects of the text. The construction of this model permits the training in both directions. This is very useful in that it helps in understanding the intragroup bonding of these data structures rather than an individual unit as observed in old models. This deep learning paradigm is specifically beneficial to applications

encompassing natural language dimension including extremism detection since context in these cases is fundamental for proper discrimination. As per [table 5](#), The classification report firstly illustrates the class-wise performance: The "Not Extremism" class has the highest precision of 0.88, a recall of 0.92 and an F1 score of 0.90 based on 1090 N instances. On the other hand, the "Extremism" class achieves precision and recall of 0.90 and 0.84 while the F1 score of the same class based on 1116 samples was 0.90 (see [figure 4](#) for confusion matrix). We see a combined model having a global mean accuracy of 0.90, signifying its efficiency in distinguishing the extremist and non-extremist content.

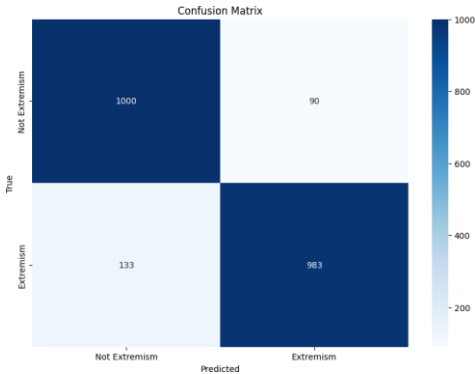


Figure 4. Confusion Matrix of Extremism Classification Model

The model exhibits exceptional performance metrics, highlighted by an Area under the Receiver Operating Characteristic Curve (ROC AUC) of 0.95 as illustrated in [figure 5](#), indicating a remarkable ability to distinguish between extremist and non-extremist content effectively.

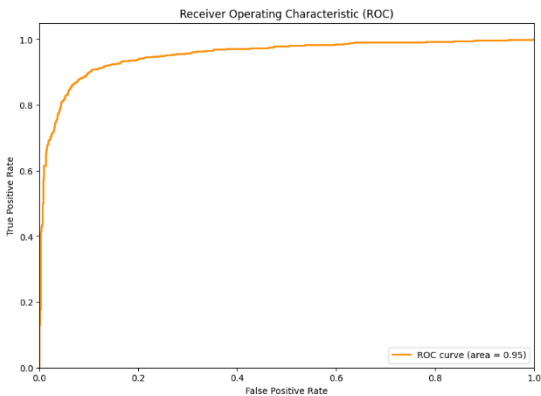


Figure 5. Receiver Operating Characteristic (AUC) of Extremism Classification Model

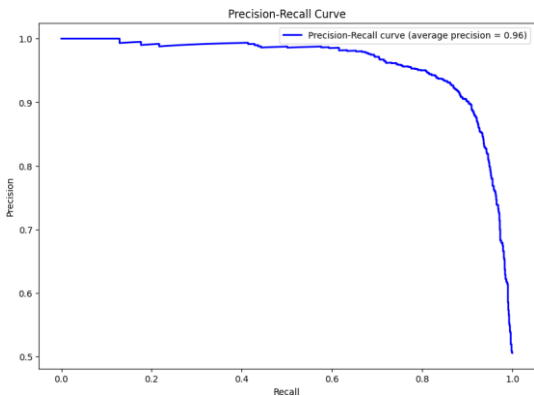


Figure 6. Precision Recall Curve of Extremism Classification Model

Furthermore, the Precision-Recall Curve reveals a striking Average Precision of 0.96 as illustrated in [figure 6](#), emphasizing the model's reliability in identifying true positives while minimizing false alarms.

4.2. Ideology Detection Model

The Ideology Detection Model was developed to classify online content into Radical and Non-Radical categories using transformer-based architectures including BERT, RoBERTa, DistilBERT, and XLNet. Each model was evaluated based on accuracy, precision, recall, F1-score, and AUC to assess its reliability in distinguishing ideological content. The performance metrics for each model are detailed in [table 5](#), while the confusion matrices and ROC curves are illustrated in [figure 7](#) and [figure 8](#), respectively.

Table 5. Ideology Detection Metrics

Model	Accuracy	Precision (Non-Radical)	Recall (Non-Radical)	Precision (Radical)	Recall (Radical)	F1-Score (Non-Radical)	F1-Score (Radical)	AUC
BERT	98.82%	0.98	1.00	1.00	0.98	0.99	0.99	1.00
RoBERTa	99.91%	1.00	1.00	1.00	1.00	1.00	1.00	1.00
DistilBERT	99.54%	0.99	1.00	1.00	0.99	1.00	1.00	1.00
XLNet	99.36%	0.99	1.00	1.00	0.99	0.99	0.99	1.00

Among all models, RoBERTa achieved the best results, with 99.91% accuracy and perfect scores (1.00) across all evaluation metrics. This indicates flawless classification performance with zero misclassifications. BERT-base-uncased also performed strongly, achieving 98.82% accuracy, with F1-scores of 0.99 for both classes and a perfect AUC of 1.00. Its uncased format improved generalizability across languages and writing styles. DistilBERT, a lighter version of BERT, reached 99.54% accuracy, offering a balance between speed and precision. Its F1-scores were also perfect, making it ideal for real-time or resource-limited applications. XLNet followed closely with 99.36% accuracy, demonstrating strong contextual performance, particularly in longer sequences of ideological text.

As shown in [figure 7](#), all models maintained very low false positive and false negative rates. For example, BERT produced 539 true negatives, 2 false positives, 11 false negatives, and 552 true positives, reinforcing its reliability. The ROC curves in [figure 8](#) confirm this further, with all models achieving an AUC of 1.00, indicating perfect class separability between radical and non-radical ideologies.

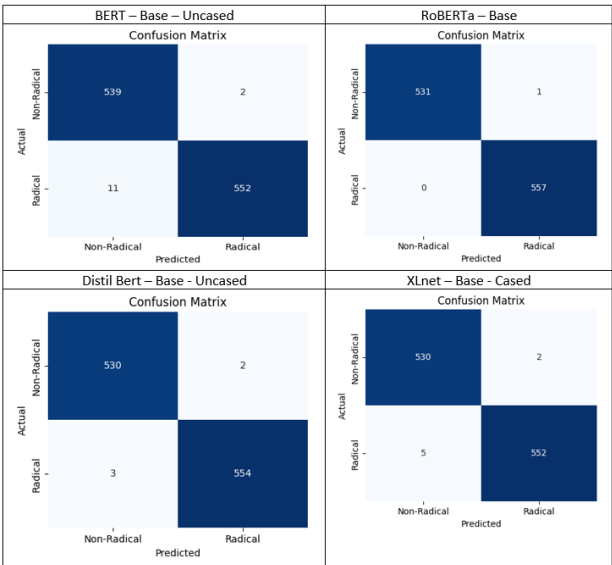


Figure 7. Confusion Matrix of Ideology Detection Model

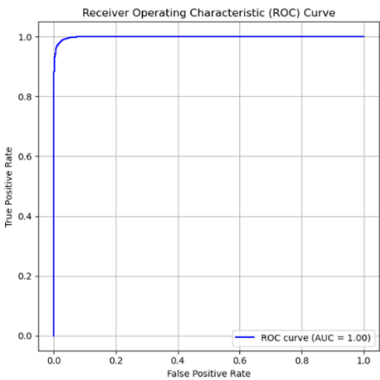


Figure 8. Receiver Operating Characteristic (AUC) of Ideology Detection Model

These results confirm the high effectiveness of the models, particularly RoBERTa, in accurately identifying ideological content with minimal error—critical for use in applications such as online content moderation, intelligence analysis, and narrative monitoring.

4.3. Propaganda Detection Model

The Propaganda Detection Model was designed to classify textual content into three categories: Propaganda, Recruitment, and Radicalization. While the model achieved an overall accuracy of 88.03%, its performance varied across classes, reflecting the inherent complexity of propaganda detection. As shown in [table 7](#), the model performed exceptionally well in identifying propaganda and radicalization content, but showed weaker results in detecting recruitment-related material. Propaganda is often conveyed through subtle persuasive techniques, emotional framing, and indirect language, making it more challenging to detect than explicit extremist statements. Despite these challenges, the model demonstrated high precision (0.993) and recall (0.995) for the Propaganda class, resulting in an excellent F1-score of 0.994. This indicates strong reliability in detecting propaganda narratives, likely due to the presence of more consistent linguistic patterns in this class.

The Radicalization class also showed strong performance, with precision, recall, and F1-score all above 0.96, highlighting the model’s capacity to detect ideologically charged content. However, for the Recruitment class, the model struggled, achieving a lower recall of 0.618 and an F1-score of 0.689, which suggests it frequently misclassified recruitment content as either propaganda or radicalization. These results suggest a need for improved data representation and contextual modeling, particularly for recruitment language, which may be less explicitly structured.

Table 7. Summary of model predictions for Propaganda Detection Model

Class	Precision	Recall	F1-Score
Propaganda	0.993	0.995	0.994
Recruitment	0.778	0.618	0.689
Radicalization	0.969	0.966	0.967

The confusion matrix in [figure 9](#) provides further insight into the model’s predictions across the three categories. In the Propaganda class, the model identified 3347 true positives, with only 10 false positives and 13 false negatives, indicating precise detection. In contrast, the Recruitment class showed a weaker signal, with only 21 true positives, 3 false positives, and 15 false negatives, suggesting the model often confused recruitment with other categories. The Radicalization class performed better, with 504 true positives and just 3 misclassifications in both false positive and false negative directions.

The ROC AUC scores, visualized in [figure 10](#), further highlight the model’s class-level performance. The AUC for Propaganda and Radicalization was high at 0.98, confirming the model’s excellent discriminative power in these categories. However, the AUC for Recruitment was only 0.81, indicating that this class is not as clearly separable in the feature space and may require additional data or contextual refinement.

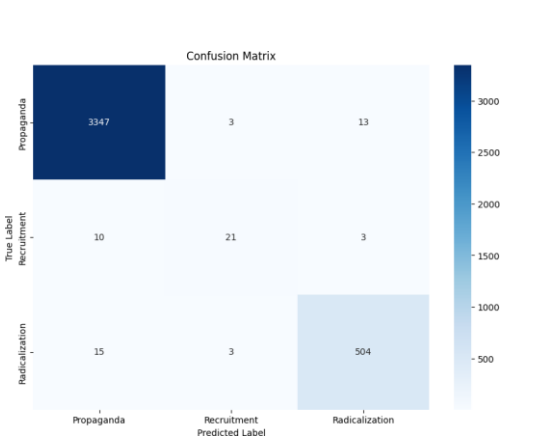


Figure 9. Confusion Matrix of Propaganda Detection Model

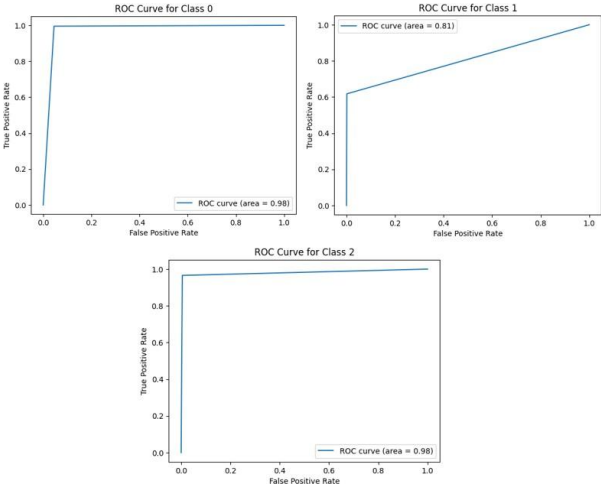


Figure 10. Receiver Operating Characteristic (AUC) of Ideology Detection Model

While the Propaganda Detection Model performs well overall—particularly in identifying propaganda and radicalization—it exhibits clear limitations in handling recruitment content. Improving recruitment detection will be critical for enhancing the model’s operational effectiveness in real-world applications. Future improvements may include the addition of annotated recruitment-specific data, multimodal learning (e.g., text + imagery), and the incorporation of attention-based mechanisms to better capture subtle contextual signals in recruitment narratives.

4.4. Flagged Word Detection Models

Detecting flagged words is a critical task in countering extremism and propaganda. This section evaluates the performance of multiple machine learning and transformer-based models in classifying flagged and non-flagged content, using metrics such as precision, recall, F1-score, and accuracy. The results are summarized in [table 8](#), while model-specific confusion matrices and ROC curves are visualized in [figure 11](#) to [figure 17](#).

Table 8. Summary of model predictions for detecting flagged words.

Models	Class	Precision	Recall	F1-Score
Naive Bayes Classifier	Non- Flagged	1.00	0.07	0.14
	Flagged	0.95	1.00	0.98
Random Forest Classifier	Non- Flagged	1.00	0.55	0.71
	Flagged	0.98	1.00	0.99
Support Vector Classifier (SVC)	Non- Flagged	1.00	0.71	0.83

Logistic Regression	Flagged	0.98	1.00	0.99
	Non- Flagged	1.00	0.13	0.23
	Flagged	0.95	1.00	0.98
Long Short-Term Memory (LSTM)	Non- Flagged	0.93	0.75	0.83
	Flagged	0.99	1.00	0.99
Transformers (BERT)	Non- Flagged	0.98	0.97	0.97
	Flagged	0.97	0.98	0.97
RoBERTa	Non- Flagged	0.98	0.99	0.98
	Flagged	0.99	0.98	0.98
XLNet	Non- Flagged	0.97	0.99	0.98
	Flagged	0.99	0.97	0.98
T5	Non- Flagged	0.98	0.99	0.99
	Flagged	0.99	0.98	0.99

Traditional machine learning models show varied performance. Naive Bayes achieved perfect precision (1.00) for non-flagged words, but had extremely low recall (0.07), leading to a weak F1-score (0.14) and a high number of false negatives (see [figure 11](#)). While it performs strongly on flagged content (F1-score: 0.98), this imbalance renders it unsuitable for real-world applications where missing unflagged instances is critical. Random Forest improved recall to 0.55 for non-flagged terms, raising its F1-score to 0.71, but still misclassified a significant number of non-flagged entries ([figure 12](#)). It achieved high performance on flagged terms with an F1-score of 0.99. Similarly, Support Vector Classifier (SVC) showed stronger recall (0.71) and F1-score (0.83) for non-flagged instances and consistent flagged detection ([figure 13](#)), making it more balanced among classical methods.

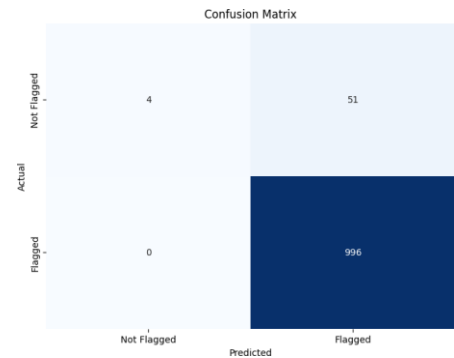


Figure 11. Confusion Matrix of Naïve Bayes (Flagged Word Detection)

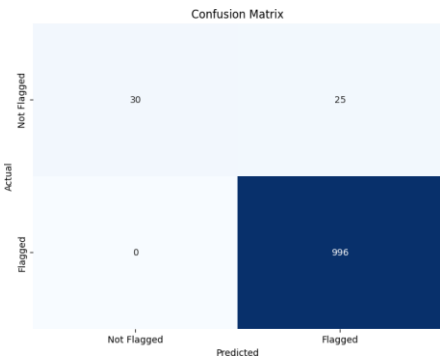


Figure 12. Confusion Matrix of Random Forest Classifier (Flagged Word Detection)

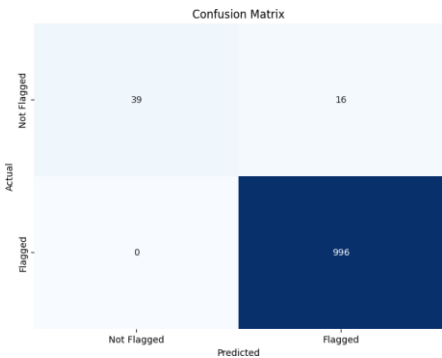


Figure 13. Confusion Matrix of Support Vector Classifier (Flagged Word Detection)

Logistic Regression, despite its perfect precision, had very low recall (0.13) for non-flagged words, as shown in [figure 14](#), with only 7 true positives and 48 false negatives—highlighting a severe limitation in detecting unflagged content. LSTM models offered a more balanced approach, with an F1-score of 0.83 for non-flagged and 0.99 for flagged content (see [figure 15](#)). Its ability to capture temporal dependencies makes it superior to simpler classifiers.

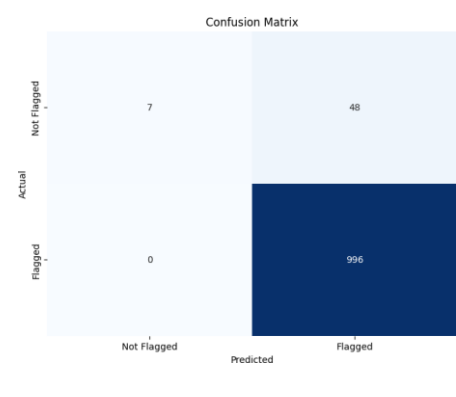


Figure 14. Confusion Matrix of Logistic Regression (Flagged Word Detection)

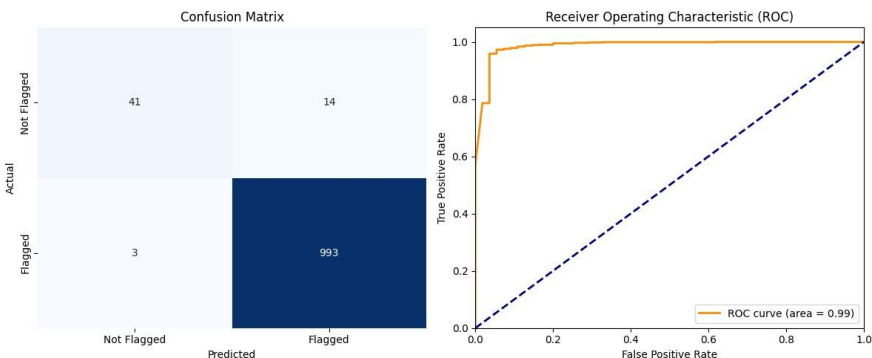


Figure 15. Confusion Matrix and ROC of LSTM (Flagged Word Detection)

Transformer-based models consistently outperformed traditional methods. T5 achieved the highest accuracy (99%) and balanced performance, with both flagged and non-flagged F1-scores at 0.99. As shown in [figure 16](#), its confusion matrix reported only 9 false positives and 23 false negatives, making it the most robust among all models tested. RoBERTa and XLNet also performed well, with accuracies of 98% each, though they recorded slightly more misclassifications ([figure 16](#)). BERT, with 97% accuracy, remained consistent but showed room for improvement due to 37 false positives and 26 false negatives. The ROC curves in [figure 17](#) confirm these findings, with all transformer models achieving AUC scores of 0.99, demonstrating excellent class separation. This validates their high reliability for both detection and classification tasks in real-world environments.

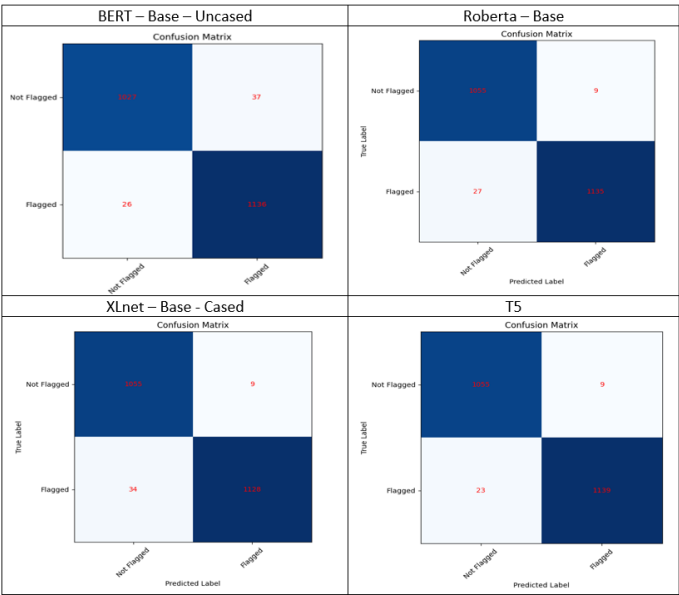


Figure 16. Confusion Matrix of Transformers (Flagged Word Detection)

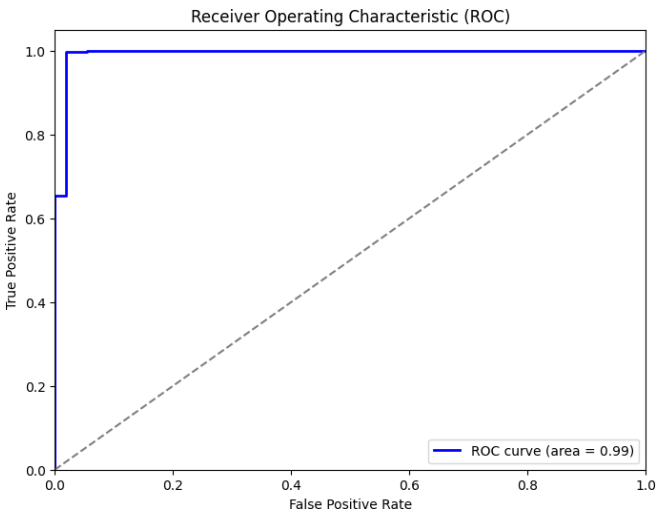


Figure 17. ROC of Transformers (Flagged Word Detection)

While traditional models offer interpretable and computationally light solutions, they suffer from lower recall, especially in non-flagged content detection. In contrast, transformer-based models—particularly T5, RoBERTa, and XLNet—consistently deliver high accuracy and balanced metrics, making them more suitable for high-risk applications such as extremism monitoring and content moderation. Future improvements may include domain-specific fine-tuning and hybrid model integration to combine interpretability with state-of-the-art performance.

4.5. Model Generalization

To evaluate the models' robustness, they were tested on a holdout dataset that had not been used during training or validation. The models, particularly those based on BERT, demonstrated strong generalization capabilities, maintaining high accuracy when applied to novel, unseen text data. This highlights the effectiveness of the models in real-world applications where they must contend with diverse and previously encountered content.

4.6. Discussion

The results of this study reinforce the effectiveness of both machine learning and deep learning models in analyzing and classifying extremist content across tasks such as extremism classification, ideology detection, propaganda recognition, and flagged word identification. Among these, the Ideology Detection Model performed best, achieving an accuracy of 98.82%, followed by the Extremism Classification Model at 90.00%, and the Propaganda Detection Model at 88.03%. These figures confirm that transformer-based models like BERT, RoBERTa, and XLNet are particularly well-suited to handling complex, context-dependent language commonly found in extremist content.

Despite strong results overall, propaganda detection remains the most challenging task due to its use of implicit language, emotional framing, and subtle persuasion techniques. While models like BERT showed solid contextual comprehension, improvements are still needed—especially in classifying recruitment-related content, where performance lagged significantly. Enhancing model generalization in this area will require more diverse, annotated, and multimodal datasets that capture the nuanced nature of propaganda.

The application of transformer models across tasks enabled improved detection in multilingual and culturally diverse content. For example, languages like Arabic and Tamil posed structural challenges due to their morphological richness, whereas Thai's agglutinative nature demanded specialized tokenization. Subword models like BERT handled these complexities effectively, but additional domain-specific fine-tuning remains necessary to adapt to linguistic variability and the evolution of extremist narratives.

The study demonstrated practical value for various stakeholders. Law enforcement agencies can use these models for intelligence gathering and threat assessment, while social media platforms may deploy them for content moderation. Policy makers and researchers can analyze ideological trends over time, shaping counter-extremism initiatives. However, real-world implementation also requires ethical safeguards to minimize false positives, protect civil liberties, and build public trust. Classification thresholds must be carefully calibrated, with transparency mechanisms and community engagement integrated into deployment protocols.

The use of classical models (Naive Bayes, Random Forest, Logistic Regression, SVC) alongside deep learning methods (LSTM, BERT, RoBERTa, XLNet, T5) highlighted the relative strengths and limitations of each. Classical models offered interpretability but lower recall for nuanced content, while transformers achieved higher accuracy and F1-scores, particularly for detecting radical ideologies and flagged language. For example, T5 and RoBERTa were particularly effective in classifying complex extremist expressions with minimal misclassification.

Future research should pursue ensemble and hybrid models to combine the strengths of multiple classifiers. For instance, integrating BERT with SVM or Random Forest could balance interpretability and contextual depth. Hybrid approaches, combining deep learning with rule-based methods, may enhance performance, reduce false positives, and improve interpretability in high-stakes applications.

Moreover, a dynamic data framework is critical. Since extremist rhetoric evolves over time and adapts to political, social, and technological contexts, models must support continuous learning and draw from real-time data sources—such as social media and underground forums. Emerging forms of extremism, such as eco-terrorism, further illustrate the need for updated datasets and fine-tuned models capable of distinguishing between genuine activism and radicalized narratives. RoBERTa and XLNet, with their contextual capacity, are well-positioned for such future adaptations.

Finally, ethical considerations must remain central. The risk of misclassifying non-extremist content as extremist carries serious implications. A human-in-the-loop framework, transparency in algorithmic decisions, and collaboration with

sociologists, ethicists, and affected communities are necessary for ensuring that these technologies serve the public good without compromising individual rights.

4.7. Models Examples and Predictions

The integrated system developed in this study demonstrates strong performance across all major tasks: extremism classification, ideology detection, propaganda detection, and flagged word identification. Each model contributes to a cohesive framework designed to detect radicalized, violent, and harmful narratives while preserving non-extremist content that promotes peace and tolerance. The extremism classification model effectively identifies violent or radical expressions by detecting subtle linguistic cues and flagged terminology, even when embedded in culturally or emotionally charged contexts. This capability is essential in recognizing ideologies aimed at influencing or mobilizing vulnerable groups. The model has shown consistent accuracy in isolating texts that promote violence and intolerance.

The ideology detection model excels at differentiating between radical and non-radical ideologies. It preserves non-violent discourse while filtering harmful ideological content with a high degree of accuracy. By comparing the thematic and semantic layers of texts, the model supports broader analysis of ideological polarization in digital environments. The propaganda detection model, leveraging models such as bert, categorizes content into propaganda, recruitment, and radicalization. It is especially useful in identifying persuasive or manipulative language that supports extremist agendas, including texts that encourage mobilization or glorify martyrdom. This contributes directly to early warning systems and preventive interventions.

The flagged word detection model is highly effective in identifying harmful and coded language. It supports multilingual detection, recognizing extremist expressions embedded in various languages, idioms, and indirect speech. Its ability to classify such expressions, as detailed in [table 9](#), enhances both the precision and adaptability of the system. Together, these models function as a comprehensive tool for moderating online content, informing policy, and guiding interventions.

Table 9. Results and predictions for detecting flagged words.

Text	Prediction	Classifier	Accuracy
"The cat is out of the bag; ISIS has set its sights on the crowded markets during the festival."	Flagged	BERT	99.71%
"They're using the festival as a cover for their recruitment efforts; it's a wolf in sheep's clothing."	Flagged	Random Forest	98.48%
"Makkal irangum manithargalum idhayathil kalangam seyyum enbathu marakkamudiyuma?" (Will people rise up in their hearts?)	Flagged	T5	95.15%
"Thamizhagaikku inriya naattukaar idhayam vilanguvathu yaar?" (Who will light the hearts of Tamils?)	Non-Flagged	Naive Bayes	98.48%
"The blast left an indelible mark on the community."	Flagged	SVC	98.48%
"Not begging King Salman or Erdogan for it."	Non-Flagged	Logistic Regression	95.43%
"Two back to back shots of tequila at work, and I'm drunkity drunk."	Non-Flagged	LSTM	98.38%
Bombs are the only solution to rid the world of infidels.	Flagged	RoBERTa	99.61%
The only way to defeat the enemy is through bloodshed and destruction.	Flagged	XLNet	99.70%
Bomb hi ekmatra solution hai jo duniya ko kafiron se mukt kar sakta hai. (Bombs are the only solution to rid the world of infidels.)	Flagged	T5	99.82%

The examples in [table 10](#) illustrate predictions from several tasks, highlighting how each model processes real-world input and assigns appropriate classifications with high accuracy. These examples confirm that the models not only classify overtly violent or ideological content but are also capable of interpreting nuanced messages that reflect extremist

discourse. The system’s performance across diverse inputs and languages demonstrates its potential for scalable deployment in counter-extremism efforts, automated content moderation, and digital policy enforcement.

Table 10. Results and predictions for three models

Text	Prediction	Accuracy
Eyewitnesses report hundreds of huge outraged crowds gathering from all over Baghdad and heading towards the Green Zone now...	Extremism	85%
I really hate how people diss my bands! Trace is clearly NOT ugly!	Not Extremism	8%
Radical Texts		
The Messenger (sallallahu ‘alayhi wa sallam) said, “Actions are judged according to their intentions, and every man shall have what he intended.”	Radical	92%
Seeking knowledge and understanding is a fundamental duty for all believers; education empowers us to make informed decisions and promotes harmony.	Non-Radical	8%
Our brothers and sisters have sacrificed their lives for this cause; their blood will not go in vain. Let their martyrdom inspire us to continue the struggle.	Propaganda	93%
Young men and women, the time has come to rise and defend your faith and your people. The enemy seeks to destroy us, but we will not stand down.	Recruitment	96%

5. Conclusion

This study presents a comprehensive framework for the detection and classification of extremist content using a combination of traditional machine learning and state-of-the-art deep learning models. The results demonstrate the high effectiveness of the proposed system, particularly the Ideology Detection Model, which achieved an accuracy of 98.82%, and the Extremism Classification Model, with a performance of 90.00%. The Flagged Word Detection Model, powered by T5, achieved a notable 99.71% accuracy, highlighting the potential of transformer-based architectures in identifying harmful language across linguistic variations.

Although the Propaganda Detection Model recorded a relatively lower accuracy of 88.03%, this is largely attributable to the limited size and variability of the available dataset. Nonetheless, the successful integration of multilingual support in the flagged word classification task demonstrates the system’s adaptability across diverse cultural and linguistic contexts. Overall, the framework lays a strong foundation for deploying AI-driven tools in real-world scenarios to identify and mitigate extremist narratives, contributing meaningfully to public safety and digital governance.

While the models developed in this study offer strong performance, several avenues remain for further exploration. Data augmentation is a critical need—especially in domains like propaganda detection, where existing datasets are limited. Expanding the training data with more diverse and nuanced examples will significantly improve classification robustness. Additionally, hybrid modeling approaches that combine classical classifiers with deep learning and transformer-based models could improve both accuracy and interpretability.

Another promising direction involves analyzing the interplay between user behavior, emotional tone, and extremist messaging, which may provide deeper insights into the underlying motivations and psychological triggers of radical content creation. Enhancing model capabilities to monitor real-time data streams from social platforms and encrypted channels would enable more proactive detection and intervention.

Interdisciplinary collaboration with psychologists, sociologists, and policy experts will be essential in refining model objectives, especially for ethically sensitive applications. Translating technical advancements into deployable solutions for law enforcement, content moderation teams, and regulatory bodies remains a high priority. As extremist discourse continues to evolve, ensuring that detection systems adapt dynamically—through continuous learning and real-time

updates—will be crucial for long-term impact. Ultimately, advancing this research will play a vital role in countering radicalization and promoting a safer digital environment.

6. Declarations

6.1. Author Contributions

Conceptualization: R.S.L.B., C.S.T., M.B., N.D., A.D.T.D., and T.Y.; Methodology: T.Y.; Software: R.S.L.B.; Validation: R.S.L.B., T.Y., and A.D.T.D.; Formal Analysis: R.S.L.B., T.Y., and A.D.T.D.; Investigation: R.S.L.B.; Resources: T.Y.; Data Curation: T.Y.; Writing Original Draft Preparation: R.S.L.B., T.Y., and A.D.T.D.; Writing Review and Editing: T.Y., R.S.L.B., and A.D.T.D.; Visualization: R.S.L.B. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- [1] C. McCauley and S. Moskalenko, "Understanding political radicalization: the two-pyramids model.", *American Psychologist*, vol. 72, no. 3, pp. 205-216, 2017.
- [2] J. M. Berger and J. Morgan, "*The ISIS Twitter census: Defining and describing the population of ISIS supporters on Twitter*," The Brookings Institution, Washington, DC, 2015.
- [3] H. Al-Garawi, A. Abed, and A. Dhiab, "A machine learning approach to detect extremist texts," *Journal of Cybersecurity*, vol. 4, no. 1, pp. 1-12, 2018.
- [4] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint*, vol. 2018, no. 1, pp. 1-12, 2018. arXiv:1810.04805.
- [5] M. Whittaker, "AI and the ethics of decision making: A study on bias in machine learning," *The AI Ethics Journal*, vol. 1, no. 1, pp. 23-38, 2019.
- [6] M. Gaikwad, S. Ahirrao, S. Phansalkar and K. Kotecha, "Online Extremism Detection: A Systematic Literature Review With Emphasis on Datasets, Classification Techniques, Validation Methods, and Tools," in *IEEE Access*, vol. 9, no. 1, pp. 48364-48404, 2021, doi: 10.1109/ACCESS.2021.3068313.
- [7] B. M. G, "Transformer-based Models for Language Identification: A Comparative Study," *2023 International Conference on System, Computation, Automation and Networking (ICSCAN)*, PUDUCHERRY, India, vol. 2023, no. 1, pp. 1-6, 2013, doi: 10.1109/ICSCAN58655.2023.10394757.
- [8] D. Subekti and D. Mutiarin, "Political Polarization in Social Media: A Meta-Analysis ", *TU Review*, vol. 26, no. 2, pp. 1–23, Dec. 2023.
- [9] H. Alghamdi and A. Selamat, "Techniques to detect terrorists/extremists on the dark web: a review," *Data Technologies and Applications*, vol. 56, no. 4, pp. 461–482, 2022. doi: 10.1108/DTA-07-2021-0177.

-
- [10] L. Gao and R. Huang, "Detecting online hate speech using context aware models," in *Proc. Int. Conf. Recent Advances in Natural Language Processing (RANLP)*, Varna, Bulgaria, vol. 2017, no. Sep., pp. 260–266, 2017. INCOMA Ltd. doi: 10.26615/978-954-452-049-6_036.
- [11] S. Jahnke, K. Abad Borger, and A. Beelmann, "Predictors of political violence outcomes among young people: A systematic review and meta-analysis," *Political Psychology*, vol. 43, no. 1, pp. 111–129, 2022. doi: 10.1111/pops.12743.
- [12] M. Mostafa, A. S. Almogren, M. Al-Qurishi, and M. Alrubaian, "Modality deep-learning frameworks for fake news detection on social networks: A systematic literature review," *ACM Comput. Surv.*, vol. 57, no. 3, Art. no. 77, pp. 1–50, Nov. 2024. doi: 10.1145/3700748.
- [13] D. Koehler, *Understanding Deradicalization: Methods, Tools and Programs for Countering Violent Extremism*, 1st ed. Routledge, 2016. doi: 10.4324/9781315649566.
- [14] Y. Liu, M. Ott, N. Goyal, J. Du, D. Joshi, C. Chen, D. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint*, vol. 2019, no. 1, pp. 1–12, arXiv:1907.11692.
- [15] H. Mohaouchane, A. Mourhir and N. S. Nikolov, "Detecting Offensive Language on Arabic Social Media Using Deep Learning," *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, Granada, Spain, vol. 2019, no. 1, pp. 466–471, 2019, doi: 10.1109/SNAMS.2019.8931839.
- [16] J. Sweeney, "Extremist groups and online radicalization: Strategies for countering recruitment through social media," *International Journal of Cybersecurity Intelligence and Cybercrime*, vol. 2, no. 1, pp. 32–50, 2019.
- [17] A. S. Parihar, S. Thapa and S. Mishra, "Hate Speech Detection Using Natural Language Processing: Applications and Challenges," *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, India, vol. 2021, no. 1, pp. 1302–1308, 2021, doi: 10.1109/ICOEI51242.2021.9452882.
- [18] R. Thompson, "Radicalization and the use of social media," *J. Strateg. Secur.*, vol. 4, no. 4, pp. 167–190, 2011.
- [19] A. B. Altinel, G. K. Baydogmus, S. Sahin and M. Z. Gurbuz, "So-haTRed: A Novel Hybrid System for Turkish Hate Speech Detection in Social Media With Ensemble Deep Learning Improved by BERT and Clustered-Graph Networks," in *IEEE Access*, vol. 12, no. 1, pp. 86252–86270, 2024, doi: 10.1109/ACCESS.2024.3415350.
- [20] S. Zimmerman, U. Kruschwitz, and C. Fox, "Improving hate speech detection with deep learning ensembles," in *Proc. 11th Int. Conf. Lang. Resour. Eval. (LREC)*, Miyazaki, Japan, vol. 2018, no. May, pp. 1–7, May 2018. European Language Resources Association (ELRA).
- [21] F. T. Boishakhi, P. C. Shill and M. G. R. Alam, "Multi-modal Hate Speech Detection using Machine Learning," *2021 IEEE International Conference on Big Data (Big Data)*, Orlando, FL, USA, vol. 2021, no. 1, pp. 4496–4499, doi: 10.1109/BigData52589.2021.9671955.
- [22] S. Malmasi and M. Zampieri, "Detecting hate speech in social media," *arXiv preprint*, vol. 2017, no. 1, pp. 1–12, 2017, arXiv:1712.06427.
- [23] A. Kumar, V. Tyagi and S. Das, "Deep Learning for Hate Speech Detection in social media," *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)*, Kuala Lumpur, Malaysia, vol. 2021, no. 1, pp. 1–4, 2021, doi: 10.1109/GUCON50781.2021.9573687.
- [24] A. Nandi, K. Sarkar, A. Mallick, et al., "A survey of hate speech detection in Indian languages," *Soc. Netw. Anal. Min.*, vol. 14, no. 70, 2024. doi: 10.1007/s13278-024-01223-y.
- [25] R. Alshalan and H. Al-Khalifa, "A deep learning approach for automatic hate speech detection in the Saudi Twittersphere," *Appl. Sci.*, vol. 10, no. 23, Art. no. 8614, pp. 1–12, 2020. doi: 10.3390/app10238614.
- [26] A. Al-Hassan and H. Al-Dossari, "Detection of hate speech in Arabic tweets using deep learning," *Multimedia Syst.*, vol. 28, no. 1, pp. 1963–1974, 2022. doi: 10.1007/s00530-020-00742-w.
- [27] M. Subramanian, V. E. Sathiskumar, G. Deepalakshmi, J. Cho, and G. Manikandan, "A survey on hate speech detection and sentiment analysis using machine learning and deep learning models," *Alexandria Eng. J.*, vol. 80, no. 1, pp. 110–121, 2023. doi: 10.1016/j.aej.2023.08.038.
- [28] M. S. S. Shah, A. M. Abuaieta, and S. S. Almazrouei, "Safeguarding online communications using DistilRoBERTa for detection of terrorism and offensive chats," *J. Inf. Secur. Cybercrimes Res.*, vol. 7, no. 1, pp. 93–107, Jun. 2024. doi: 10.26735/VNVR2791.

-
- [29] M. Gaikwad, S. Ahirrao, S. Phansalkar, K. Kotecha, and S. Rani, "Multi-ideology, multiclass online extremism dataset, and its evaluation using machine learning," *Comput. Intell. Neurosci.*, vol. 2023, no. 1, pp. 45-63, 2023. doi: 10.1155/2023/4563145.
- [30] Shynar Mussiraliyeva, Kalamkas Bagitova and Daniyar Sultan, "Social Media Mining to Detect Online Violent Extremism using Machine Learning Techniques" *International Journal of Advanced Computer Science and Applications(IJACSA)*, vol. 14, no. 6, pp. 1-12, 2023.
- [31] S. Aldera, A. Emam, M. Al-Qurishi, M. Alrubaian and A. Alothaim, "Online Extremism Detection in Textual Content: A Systematic Literature Review," in *IEEE Access*, vol. 9, no. 1, pp. 42384-42396, 2021, doi: 10.1109/ACCESS.2021.3064178.
- [32] U. Mittal, "Detecting Hate Speech Utilizing Deep Convolutional Network and Transformer Models," *2023 International Conference on Electrical, Electronics, Communication and Computers (ELEXCOM)*, Roorkee, India, vol. 2023, no. 1, pp. 1-4, 2023, doi: 10.1109/ELEXCOM58812.2023.10370502.
- [33] M. Irfan, Z. A. Almeshaland M. Anwar, "Unleashing transformative potential of artificial intelligence (AI) in countering terrorism online radicalisation extremism and possible recruitment". *University of Limerick*, vol. 2024, no. Mar., pp. 1-12, 22-Mar-2024, doi: 10.34961/researchrepository-ul.25451590.v1.
- [34] M. Hussain, "The role of artificial intelligence in countering violent extremism," *Global Security Studies*, vol. 13, no. 3, pp. 43-61, 2022.
- [35] S. Chalke and D. Mishra, "Studying the impact of social media algorithms on the spread of misinformation and its effects on society," *Int. J. Adv. Res. Sci., Commun. Technol.*, vol. 2023, no. 1, pp. 352-356, Sep. 2023. doi: 10.48175/IJAR SCT-13054.
- [36] R. A. Khan and M. Saeed, "Framework for detecting online radicalization through text mining and machine learning," *Computers and Security*, vol. 121, no. 1, p. 10-27, 2022.
- [37] A. Shaw, "Social media, extremism, and radicalization," *Science advances*, vol. 9, no. 35), pp. 20-031, 2023.
- [38] W. Ahmed, N. Semary, K. Amin, and M. A. Hammad, "Sentiment analysis on Twitter using machine learning techniques and TF-IDF feature extraction: A comparative study," *Int. J. Comput. Inf.*, vol. 10, no. 3, pp. 52-57, 2023. doi: 10.21608/ijci.2023.236052.1128.
- [39] J. Pérez-Rodríguez and J. Villalobos, "Machine learning models for detecting terrorism-related content on social media," *Applied Sciences*, vol. 12, no. 16, pp. 80-92, 2022.
- [40] S. Mussiraliyeva, M. Bolatbek, B. Omarov, Z. Medetbek, G. Baispay and R. Ospanov, "On Detecting Online Radicalization and Extremism Using Natural Language Processing," *2020 21st International Arab Conference on Information Technology (ACIT)*, Giza, Egypt, vol. 2020, no. 1, pp. 1-5, 2020, doi: 10.1109/ACIT50332.2020.9300086.
- [41] O. Berjawi, G. Fenza and V. Loia, "A Comprehensive Survey of Detection and Prevention Approaches for Online Radicalization: Identifying Gaps and Future Directions," in *IEEE Access*, vol. 11, no. 1, pp. 120463-120491, 2023, doi: 10.1109/ACCESS.2023.3326995.
- [42] M. Fernandez and H. Alani, "Artificial intelligence and online extremism: Challenges and opportunities," in *Predictive Policing and Artificial Intelligence*, J. McDaniel and K. Pease, Eds. Abingdon: Routledge, vol. 2021, no. 1, pp. 132-162, 2021. doi: 10.4324/9780429265365-7.
- [43] I. Ajala, M. Scharrow, M. Steimle, and M. E. Risse, "Combining artificial intelligence and expert content analysis to explore radical views on Twitter: Case study on far-right discourse," *J. Clean. Prod.*, vol. 362, no. 1, pp. 13-22, 2022.
- [44] R. Montasari, "Machine learning and deep learning techniques in countering cyberterrorism," in *Cyberspace, Cyberterrorism and the International Security in the Fourth Industrial Revolution: Threats, Assessment and Responses*, Cham: Springer International Publishing, 2024, pp. 135-158. doi: 10.1007/978-3-031-50454-9_8.
- [45] T. T. Tin, K. J. Xin, A. Aitizaz, L. K. Tiung, T. C. Keat, and H. Sarwar, "Machine learning based predictive modelling of cybersecurity threats utilizing behavioural data," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 9, pp. 1-12, 2023.