IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 July 2022, accepted 3 September 2022, date of publication 12 September 2022, date of current version 7 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3205744

RESEARCH ARTICLE

Multi-Ideology Multi-Class Extremism Classification Using Deep Learning Techniques

MAYUR GAIKWAD[1], SWATI AHIRRAO[1], KETAN KOTECHA[2], AND AJITH ABRAHAM[3,4], (Senior Member, IEEE)
[1]Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune 412115, India
[2]Symbiosis Centre for Applied Artificial Intelligence (SCAAI), Symbiosis International (Deemed University), Pune 412115, India
[3]Machine Intelligence Research Laboratories (MIR Labs), Auburn, WA 98071, USA
[4]Center for Artificial Intelligence, Innopolis University, 420500 Innopolis, Russia

Corresponding authors: Swati Ahirrao (sahirrao4@gmail.com) and Ketan Kotecha (head@scaai.siu.edu.in)

This work was supported by the Analytical Center for Government of the Russian Federation under Grant 70-2021-00143 dd. 01.11.2021 (IGK000000D730321P5Q0002).

ABSTRACT Social media is an integral part of today's social communication. Social media platforms have a global reach with immense popularity among the young generation. This reach and influencing power of social media has attracted extremist and terrorist organizations to social media platforms. Numerous terrorist organizations like ISIS, Taliban, Al-Qaeda, and Proud Boys and conspiracy theory groups like Alt-Right and QAnon spread their propaganda, radicalize and recruit youths via social media platforms. Thus, online extremism research is imperative to monitor extremists' influence and their spread of hate on social media. The existing research is limited to the specific ideology, which results in a bias towards a particular ideology. The classification of extremism is presented only in binary or tertiary classes with no further insights. This research work presents the development of a seed dataset and balanced multi-ideology extremism text dataset with multi-class labels. Recently natural language processing with deep learning has gained significant attention in extremism detection research. Thus this research focuses on collecting, cleaning and classifying the extremist tweets. This study presents a multi-class classification of the balanced multi-ideology dataset. This dataset is termed Merged Islamic State of Iraq and Syria (ISIS) /Jihadist-White Supremacist (MIWS). The MIWS dataset is evaluated using pre-trained Bidirectional Encoder Representation for Transformers (BERT) and variants like Robustly Optimized BERT Pretraining Approach (RoBERTa) and DistilBERT and achieves the highest f1-score of 0.72. RoBERTa and DistilBERT provide f1-score of 0.68 and 0.71, respectively. Thus, deep learning can be effectively used to identify extremism from multiple ideologies and segregate them into propaganda, radicalization and recruitment.

INDEX TERMS Extremism, hate, propaganda, radicalization, recruitment, deep learning.

I. INTRODUCTION

Social media has become an unstoppable force in recent years. Facebook, Twitter, and WhatsApp are the market leaders in social networks with a high userbase [1], [2]. Twitter has 206 million daily users, most in the 25-34 age group [2]. Facebook-owned services: Facebook, Messenger, WhatsApp, and Instagram have 2.6 billion daily users [1]. Thus, these social media platforms have extensive reach among various nationalities, races, and ages. This has given different anti-social elements to spread their propaganda, radicalize and recruit people. Terrorist organizations such as ISIS, Taliban, and Al–Qaeda employ social media to attract vulnerable youth [3], [4], [5]. Similarly, far-right organizations like Proud Boys spread their propaganda via social media [6] Christchurch Mosque Tragedy [7] further cemented

The associate editor coordinating the review of this manuscript and approving it for publication was Patrizia Grifoni.

the role of social media in extremism propagation where attackers live-streamed the attack. In recent, the capture of Kabul by the Taliban garnered much support on Twitter [8]. Various hashtags like #westandwithtaliban, #wesupporttaliban, and #talibanourguardians were trending on Twitter [8]. Online radicalization and hate speech influenced the Capitol Riots [9] and Poway Synagogue Shootings [10] In the recent time Buffalo Massacre 2022 [11], [12], the perpetrator believed in conspiracy theories like 'Great Replacement' spread on forums like 4chan [13] In addition, the Buffalo shooter was reported to be influenced by Christchurch and Oslo attackers [14].

The effect of online extremism can be felt in regions where extremist organizations don't even exist [15] Lone wolf attacks are the primary example of online radicalization [16] ISIS/Jihadist and White Supremacist ideologies are spread through social media platforms using the same strategies [17] Anti-government propaganda [18], hate and perpetuating violence against targeted people [19], community and nations are recurring themes in social media posts from ISIS/Jihadist and White Supremacist ideologies. Extremists also use anti-vaccine and COVID news to radicalize people [20] This proves extremists adapt their strategies to spread propaganda, radicalization, and recruit people based on current events. This makes online extremism research challenging. Online extremism research is the only way to analyze, predict, and restrict extremist ideologies on social media. Thus, it is necessary to understand propaganda, radicalization, and recruitment texts to counter the rise of extremism on social media platforms.

Propaganda is 'content, which can be biased and exploited for personal or political gains' [21] Misinformation is also a part of propaganda. Radicalization is a 'change in behaviour, attitude or perception towards person or community [22]'. Extremists radicalize people by misquoting religious texts, citing political uprisings, etc. Recruitment in the context of extremism is 'inciting people to join the terrorist cause or commit a violent attack' [23]. Organizations like ISIS, the Taliban, and Al-Qaeda recruit by glorifying the death of terrorists. At the same time, right-wing white supremacists use 'anti-government themes,' 'anti-Semitism,' and recently 'coronavirus themes' for recruitment [24].

As social media use increases exponentially, extremist organizations have spread their reach to every corner of the globe. Thus, it is imperative to develop automatic detection of propaganda, radicalization, and recruitment texts to stop the spread of online extremism.

The motivation behind this study is to propose an extremist identification system free from ideological bias. Thus any extremist content should be identified without any judgement on political and religious leanings. Another motivation is to devise a system which can monitor social media so violent events like US Capitol Riots, Christchurch Attack or France Teacher attack [25].

Machine Learning has been used for extremism detection and analysis since the 2010s. Multiple researchers efficiently used Machine Learning (ML) with different features for extremism detection on various social media and websites. This is discussed in Section 2. But these ML algorithms and features face critical problems like data limitations and context identification issues.

Researchers used Deep Learning techniques in every possible data science field. Deep learning techniques that use transformer modules can recognize context better than other algorithms. So, to better identify the context related to propaganda, radicalization, and recruitment, Bidirectional Encoder Representations for Transformers (BERT), RoBERTa, and DistilBERT are used in this study.

Following are the contributions of this study:

- Construction of balanced multi-ideologies, multi-class, extremism seed dataset gathered from journal papers, newspapers, and websites.
- Development of balanced multi-ideology extremism dataset, which consists of recent tweets.
- Development of a framework using natural language processing (NLP) techniques for online extremism dataset creation, labeling, and validation of online extremism tweets and classifying them into Propaganda, Radicalization, and Recruitment.
- Evaluation of balanced multi-ideology extremism dataset using pre-trained neural networks such as BERT and its variants such as RoBERTa and DistilBERT.
- Comparative analysis of BERT with variants of BERT such as RoBERTa and DistilBERT.

This study is divided into different sections. Section II discusses the literature analysed for this work. The methodology is described in Section III. Section IV, Empirical Analysis, provides experimentation details for this study. Section V deals with Results and Discussion about the experiments performed. Sections VI, VII and VIII discuss Limitations, Conclusion and Future Work, respectively.

II. LITERATURE REVIEW

Extremism on social media has been discussed in literature since the early 2000s [26] Studies from 2015 to 2021 are considered to keep the research relevant and up to date. Most studies discuss extremism for a single ideology [5], [27], [28], [29]. There are different extremist organizations with numerous ideologies. This makes the classification of extremism biased to a particular ideology. So, there is a need for analysis and classification research that incorporates multiple extremist ideologies.

There are only a handful of standard datasets. These datasets [30] and [31] deal with ISIS Tweets propagated during its heydays from 2014-2017. ISIS standard dataset contains nearly 15,000 extremist tweets. White Supremacist ideology standard datasets [28], [32] are from Stormfront and Gab websites, respectively. The problem with standard datasets is that they are outdated and unbalanced. Therefore, most extremist researchers prefer to gather their data. Most extremism research is carried out on data collected from Twitter [27], [33], [34] This is because of the popularity of Twitter among young people and its microblogging format. Few researchers have also explored Facebook [35], Stormfront [28], and Gab [32] for extremism detection. Twitter policies restrict researchers from publishing tweet data. In addition, extremist accounts may get suspended. Thus, the collection of extremism data is challenging.

Most recent studies use various machine learning (ML) algorithms [35], [36], [37], [38] and deep learning (DL) techniques [33], [39], [40], [41] to classify extremism texts. The studies use Support Vector Machine (SVM) [42], [29] Random Forest [42], [43], and XGboost [44], [45] for extremism text classification. DL techniques such as Long Short Term Memory (LSTM) [28], [33], LSTM + Convolutional Neural Network (CNN) [41], and Bidirectional Encoder Representations for Transformers (BERT) [33] are applied by recent studies for extremism classification.

Studies like [46] and [47] provide comparable results on SVM with a precision of 0.92 and an accuracy of 0.95, respectively. Random Forest in a few studies [43], [48] gives better results with an f1-score of 0.84 and 0.93, respectively. Deep learning techniques [28], [39] provide better results than machine learning algorithms. A study using BERT with Word2Vec [33] gives the f1-score of 0.79 and a precision of 0.80. Similarly, a study using LSTM [39] provides a precision of 0.8596, and LSTM + CNN [41] also offers an accuracy of 0.9266. Thus, it is observed that deep learning techniques are better than machine learning algorithms in performance without using explicit feature extraction methods.

The classification of extremism texts is either in binary [29], [42], [43] or tertiary [35], [47], [49] classes. These classes are 'extremist'-'non-extremist' [42], 'hate'-'non-hate' [50], etc. Class labels like 'neutral' [5], [51] or 'irrelevant' [5], [39] are also used to determine other texts. Existing research on the binary classification of extremism detection doesn't provide insight into the extremist text. The valuable analysis like targets of extremists, radicalization methods, and recruitment tactics are lost with just binary or tertiary classification. So there is a need for multi-class classification, which detects extremist texts into Propaganda, Radicalization, and Recruitment [52].

Few studies use Cohen's Kappa (McHugh, 2012) and Fleiss's Kappa (Fleiss, 1971) for validation. These methods are used to validate the labels manually. Most works [39], [53] use two experts to verify that Cohen's Kappa ranges from 0.10 to 0.76. Fleiss's Kappa in [28] gives kappa around 0.4 to 0.6, with three experts for validation. The issues plaguing manual validation are expert bias, a small sample for validation, and low metrics for validation.

Recently more extremism detection studies have been published. This study [54] is focused on radicalization detection using sentiment analysis. The authors focus on ISIS/ Jihadist ideology. The authors use SenticNet, and AffectiveSpace as features while Logistic Regression and LibSVM as classifiers. This study also classifies text into binary classes: extremism or non-extremism and hate or non-hate.

Religious extremism in the Kazakh language is identified by [55] The extremism text is collected from Vkontakte [56] The authors identified extremism text from keywords like kafir, kill, etc. Thus, limiting the study to ISIS/Jihadist ideology. The annotation is performed by observing extremists' keywords in the collected corpus. This annotation is binary. This study used Logistic Regression, SVM, and Random Forest for extremism text classification.

Reference [57] another study used CNN and LSTM for extremism classification. The authors use the previously collected Vkontakte dataset for classification in this study. The authors got an AUC of 0.99 for CNN and LSTM extremism text classification.

Table 1 provides details of the literature studied for extremism detection. Table 1 compares studies based on the source from which data is collected (Data Source), algorithms or techniques used, classes or labels with size, and different performance metrics used. Thus, it can be observed from Table 1 that Twitter is the most studied platform for extremism research. The binary classification dominates the extremism detection research. The performance metrics primarily used are precision, f1-score, and ROC-AUC. Similarly, these conclusions can also be obtained from the systematic literature review [58], covering multiple online extremism detection studies and their current issues.

III. METHODOLOGY

This section presents the data collection, labeling, and validation process. Figure 1 shows the entire process flow for creating, labeling, and classifying the ideologically balanced MIWS dataset. The process can be divided into five steps: Collection of Data, Validation of Seed Data, Merging, and Classification. The following sections explain the steps mentioned above:

A. COLLECTION OF DATA

1) SEED DATA

Four hundred examples within the seed dataset are selected from varied sources like newspapers, journal articles, websites, blogs, and reports. There are two hundred posts of ISIS/Jihadists and two hundred for White Supremacists.

2) SEED DATA COLLECTION

Examples from research articles and websites that identify influential propagandists, violent radicals, and extremist recruiters are collected for the seed dataset.

a: SOURCES

The first ideology considers for seed data collection is ISIS/Jihadist. The reason for selecting ISIS/Jihadist ideology is their vast extent of propaganda, targeted recruitment, and multiple violent acts. White Supremacist ideology is also considered for seed data collection for similar reasons. The motive of the seed dataset is to identify text

TABLE 1. Literature review.

Study	Data Source	Technique with Algorithms	Classes	Metrics
[29]	Twitter	ML: Linear SVM, Logistic Regression	Positive (Radical): 619,861, Negative (Non-Radical): 1,566,570	f1-score: 0.94
[28]	StormFront	DL: Convolutional Neural Network (CNN), LSTM	Hate: 1,119, non-Hate: 8,537	Accuracy: 0.78
[61]	Twitter	Relative Entropy, Adoption Probability	Pro-ISIS: 602,511, non-ISIS: 1,368,827	ROC-AUC: 0.60
[62]	ISIS Kaggle, StormFront	DL: CNN	Terrorism related: 17,000, Hate and Non-hate [28].	f1-score: 0.93, AUC: 0.99
[33]	Twitter	DL: LSTM, BERT	White Supremacist: 2294, Non-White Supremacist: 2294	f1-score: 0.79
[42]	Vkontakte	ML: Support Vector Classifier (SVC), Random Forest (RF), Gradient Boost	Extremist:Not Available (NA),Non-Extremist: NA	f1-score: 0.86
[35]	Facebook	ML: K-Nearest Neighbour (KNN), SVC	Neutral:4315, Moderate: 5279, Low Extreme: 2991 and High Extreme: 6912	Accuracy: 0.82
[43]	Twitter	ML: SVM, RF	Extremist: 17,350, Neutral: 122,000	f1-score: 0.87
[47]	Twitter	ML: Logistic Regression, SVM	Terrorism Supporting: 13,369, Terrorism Non-supporting: 16,506, Random: 38,617	Accuracy: 0.95
[41]	Twitter	DL: CNN, LSTM, Gated Recurrent Network (GRU)	Extremist: 12,754, Non-Extremist: 8,432	Accuracy: 0.92
[5]	Twitter	ML: RF, SVM	Pro-Afghan: NA, Pro-Taliban: NA, Neutral: NA, Irrelevant: NA	Precision: 0.84
[40]	Twitter	ML + DL: Random Forest, SVM, Neural Network	Known Bad: 17,000, Random Good: 8,000, Non-radical: 122,000.	Accuracy: 1
[63]	Twitter	Bag of Words	RightWing: 15,911, Non-RightWing: 29,836.	f1-score:0.95
[34]	Twitter	ML: Single Layer Perceptron	RightWing: 50,000, Safe: 50,000	Precision: 0.84
[53]	Twitter	ML: Naïve Bayes	Extremist: 17,350, Neutral: 197,743	Precision: 0.90
[39]	Multiple	DL: LSTM	NA	Precision: 0.85
[55]	Vkontakte	ML: SVM, Random Forest	NA	AUC: 0.99
[64]	Reddit	Recurrence Quantification Analysis	NA	Low Frequency Users, Median: 25.23, Standard Deviation: 0.97

for propaganda, radicalization, and recruitment. Different sources like newspapers and journal articles and counter-extremism websites are selected. Snowballing technique is also used to gather text examples of propaganda, radicalization, and recruitment from verified sources.

b: JOURNAL PAPERS AND REPORTS

The seed data in journal papers or research articles explicitly is small in numbers. Thus multiple reports are also considered. Details about journal papers and reports are provided in this study [58] The search for journal papers was limited

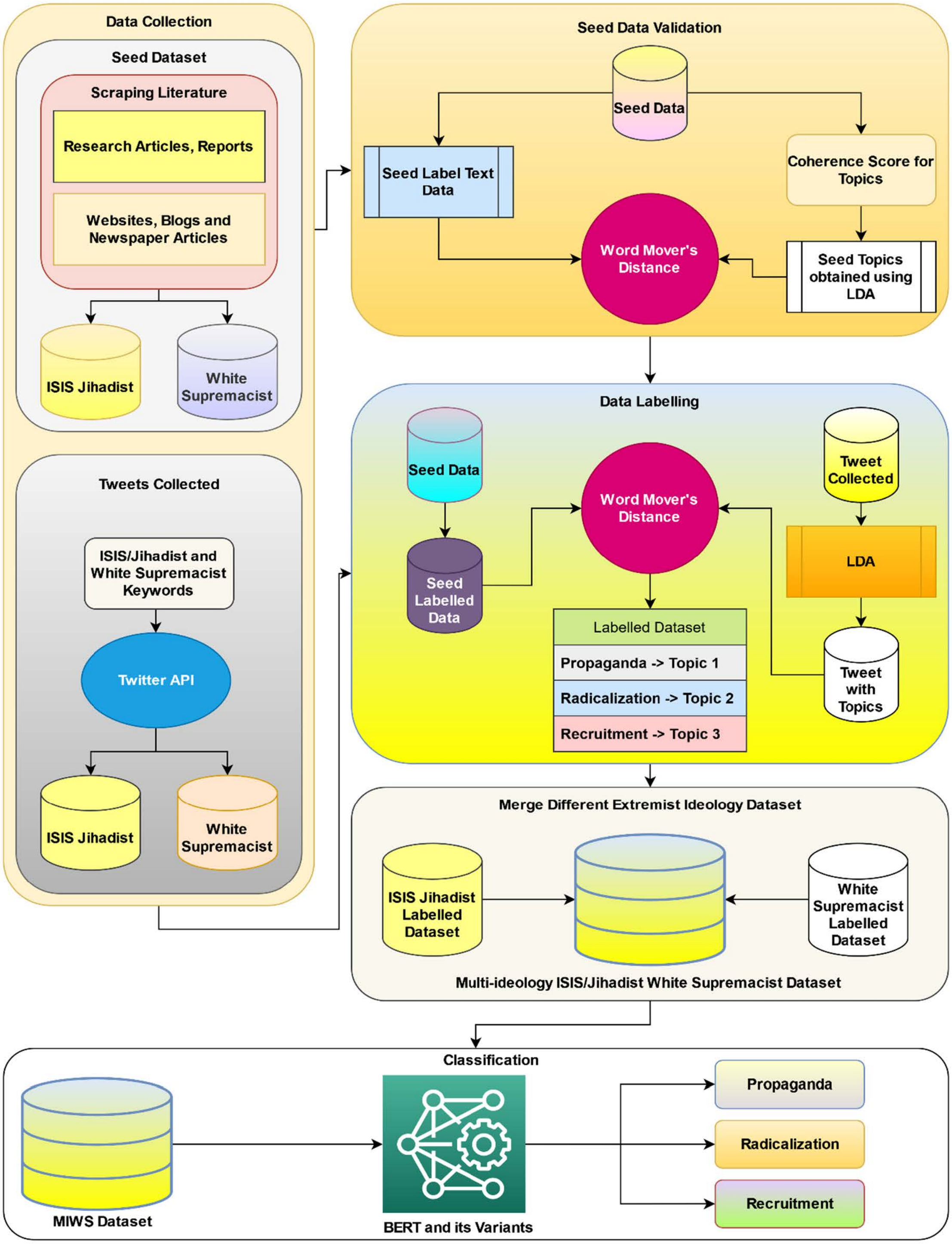

FIGURE 1. Process flow for collection, labelling, and classification of MIWS dataset.

from January 2015 to December 2021. Snowballing technique ensured relevant older studies get selected.

c: NEWSPAPERS, BLOGS, AND WEBSITES

Newspaper articles, blogs, and websites are collected using Snowballing technique. Most seed examples come from newspaper stories, blogs [59], or websites [60] These websites labeled social media users as propagandists or recruiters. Thus their posts are labeled as propaganda or recruitment based on the user.

B. BALANCED MIWS DATASET

This section informs about the collection, annotation, and validation of ideologically balanced MIWS.

TABLE 2. Ideologies and keywords.

ISIS/Jihadist Keywords	White Supremacist Keywords
Munafiq	Anti-white
Kuffar	Whitepower
Kufr	Whitegenocide
Murtad	ZOG
Talibanourguardians	goy
Wesupporttaliban	Globalists

1) TWEET COLLECTION

A total of 60,000 tweets are included in the MIWS dataset, of which 30,000 belong to ISIS/Jihadist, and 30,000 belong to White Supremacist ideology.

Different journals, newspapers, blogs, and reports collected extremist tweets based on manual identification or detection of extremists [26], [52], [59] In this research, extremist tweets are compiled based on selected keywords that extremists use. Some popular extremist keywords used are ''munafiq,'' ''kuffar'', ''anti-white,'' ''ZOG,'' ''goy,'' goyim,'' kufr.'' These keywords are obtained from [52], [65], [66] studies. Some recent keywords like ''talibanourguardians'', ''wesupporttaliban'', and ''globalists'' were identified. Table 2 provides popular extremist keywords within literature and on social media that denote extremism.

2) DATA LABELLING

NLP techniques like topic modelling and word distances are used to label the data. Latent Dirichlet Allocation (LDA) [67] is used for topic identification. Similarly, Word Mover's Distance (WMD) [68] is employed to determine the similarity between topics and classes.

3) LDA ON COLLECTED TWEETS

LDA provides topic-wise distribution of tweets. LDA determines these topics based on influencing words within the documents. Thus important words get highlighted within a specific topic. ISIS/Jihadist and White Supremacist tweets were divided into three LDA topics. Hyperparameter tuning was performed to produce the best model.

4) COMPARISON OF LABELED SEED DATASET AND LDA TOPICS OF COLLECTED TWEETS USING WORD MOVER'S DISTANCE

The LDA topics from collected tweets are compared with seed labels using Word Mover's Distance. This process is performed individually on each ideology. This is because words in ISIS /Jihadist and White Supremacist extremists significantly differ in the seed dataset. Word Mover's Distance (WMD) is applied to validate and label the topics from seed labels. WMD provides purposeful differentiation between terms from their co-occurrences within the corpus. So similarity increases when the distance between words decreases. WMD was applied using word2vec, which was pretrained on Google News Vector. Algorithm 1 describes the process in detail. Here seed dataset (Sd) contains three labels Propaganda (P), Radicalization (Rd), and Recruitment (Rc). Ct denotes collected tweets. Ct is passed to the LDA algorithm with the number of classes (n). LDA segregates tweets into T1, T2 and T3 topics. Each label text from Sd is compared with each tweet with the topic in Ct. This comparison is performed using WMD, which provides distances between texts. The computational cost of the algorithm depends upon the number of labels in Sd and the number of topics in Ct. Therefore, the algorithm has O(N^2) time complexity.

The average distance between label and topic was recorded, and the corresponding label whose average distance is least was assigned to that topic. Table 3 and Table 4 provide WMD distance between topics and seed labels; only the least average distances are considered as words should be near each other. Thus the lowest distance points to the label assigned to the topic. The radicalization label was given to Topic 0, which has the lowest WMD distance of 0.8575; Propaganda was assigned to Topic 1 as the lowest distance is 0.8455, and Recruitment was assigned to Topic 2 as the lowest distance of 0.8464 for ISIS/Jihadist collected tweets as seen in Table 3. Similarly, the Propaganda label was given to Topic 2 with the lowest distance of 0.7924, Radicalization to Topic 1 with the lowest distance is 0.8021, and Recruitment for Topic 0 with the lowest distance is 0.8032 for White Supremacist collected tweets, as seen in Table 4. This whole process is explained in this study [69].

The correctness labels were manually verified by taking random samples and checking the words belonging to Propaganda, Radicalization and Recruitment, as mentioned in [58].

5) MIWS DATASET

Topic 0 from Table 2 and Topic 1 from Table 4 are labeled as Radicalization. Radicalization has 18,120 tweets which are usually politically aligned. Topic 1 from Table 2 and Topic 2 from Table 3 are assigned Propaganda labels. Propaganda with 24,987 tweets is the largest of all three classes and is oriented towards religion, achievements, and glorification of ideology. Topic 2 from Table 3 and Topic 0 from Table 4 are labeled Recruitment. Recruitment has 16,893 tweets tilting towards hate, degrading conservative ways, and instigating violence against a group or individual.

To analyse the difference between created classes, we use paired t-test on individual classes. Term Frequency Inverse Document Frequency (TFIDF) score for each word in

ALGORITHM 1: Distance Between Seed Dataset Labels and Collected Tweets' Topics

```
Sd ← {P, Rd, Rc}
Ct ← {Tweets}
LDA (Ct, n=3):
{T1, T2, T3 } ← Ct
for all labels in Sd do
        for all topics in Ct do
                distance ← WMD (labels, topics)
        end for
end for
```

TABLE 3. Comparing ISIS/Jihadist tweet LDA topics and ISIS/Jihadist seed labels with word mover's distance.

Seed Labels ISIS/Jihadist ISIS/Jihadist Topics	Propaganda	Radicalization	Recruitment
Topic 0	0.8598	0.8575	0.8591
Topic 1	0.8455	0.8588	0.8494
Topic 2	0.8490	0.8584	0.8464

TABLE 4. Comparing white supremacist tweet LDA topics and white supremacist seed labels with word mover's distance.

Seed Labels WS WS Topics	Propaganda	Radicalization	Recruitment
Topic 0	0.8038	0.8039	0.8032
Topic 1	0.8028	0.8021	0.8041
Topic 2	0.7924	0.8029	0.8035

Propaganda, Radicalization and Recruitment was calculated. These TFIDF scores were used as paired input to t-test algorithm. Two hypotheses were generated:

H0 – There no significant difference between classes

H1 – There is significant difference between classes

Table 5 provides details about classes and their p-value. As per hypotheses, if $p < 0.05$, H0 is reject, hence there is significant difference between classes.

TABLE 5. p-value obtained from t-test on classes.

Class 1	Class 2	p-value
Propaganda	Radicalization	6.14e-5
Propaganda	Recruitment	8.63e-5
Radicalization	Recruitment	1.04e-4

The p-value after t-test between Propaganda, Radicalization and Recruitment are less than 0.05. So **H0** is rejected. Thus the difference between classes is statistically significant.

After this annotation, both ideologies, ISIS/Jihadist and White Supremacist are merged. This new dataset is called Merged ISIS/Jihadist-White Supremacist (MIWS).

6) SEED AND MIWS DATA DESCRIPTION

Table 5, provides descriptions of Seed and MIWS datasets. Table 5 shows the timespan of data collection, number of classes, features, word count, and size of classes.

7) BALANCING DATASETS

Propaganda tweets are more than Radicalization and Recruitment tweets. This affects the overall accuracy of the model. So, to balance the data, Randomized Under Sampling was used [70] Randomized Under Sampling ensures similar examples for every class by randomly removing instances from oversampled classes. So, this under-sampled data was divided into the train, validation, and test sets. Different splitting ratios like 60:20:20, 70:15:15, 80:10:10, and 90:05:05 were used. The best results were obtained by 90 % training, 05 % validation, and 05 % testing examples. In addition to splitting, 5-fold cross-validation was used to ensure the model is efficiently trained on the entire dataset.

IV. EMPIRICAL ANALYSIS

This section provides information about deep learning classifiers, hyperparameters, and results for balanced MIWS extremism classification into propaganda, radicalization, and recruitment.

TABLE 6. Dataset information.

Details	Seed Dataset	MIWS
Timespan	2015 - 2021	June 2021 – Oct. 2021
Classes	3	3
Count of Features	7	6
Word Count / Row (Min, Max)	Min : 1, Max : 291	Min :1, Max : 32
Examples	400	60K
Size of Classes	Propaganda - 225 Recruitment - 100 Radicalization - 71	Propaganda - 24,987 Recruitment - 16,893 Radicalization - 18,120
Ideologies	2	2

TABLE 7. Hyperparameters used for BERT.

BERT Hyperparameters	Values
Learning Rate	2e-6
Learning Rate decay	2e-4
Optimizer	AdamW
Dropout	0.3
Batch Size	8
Max sentence length	60
Training Epochs	20
Warmup steps	100

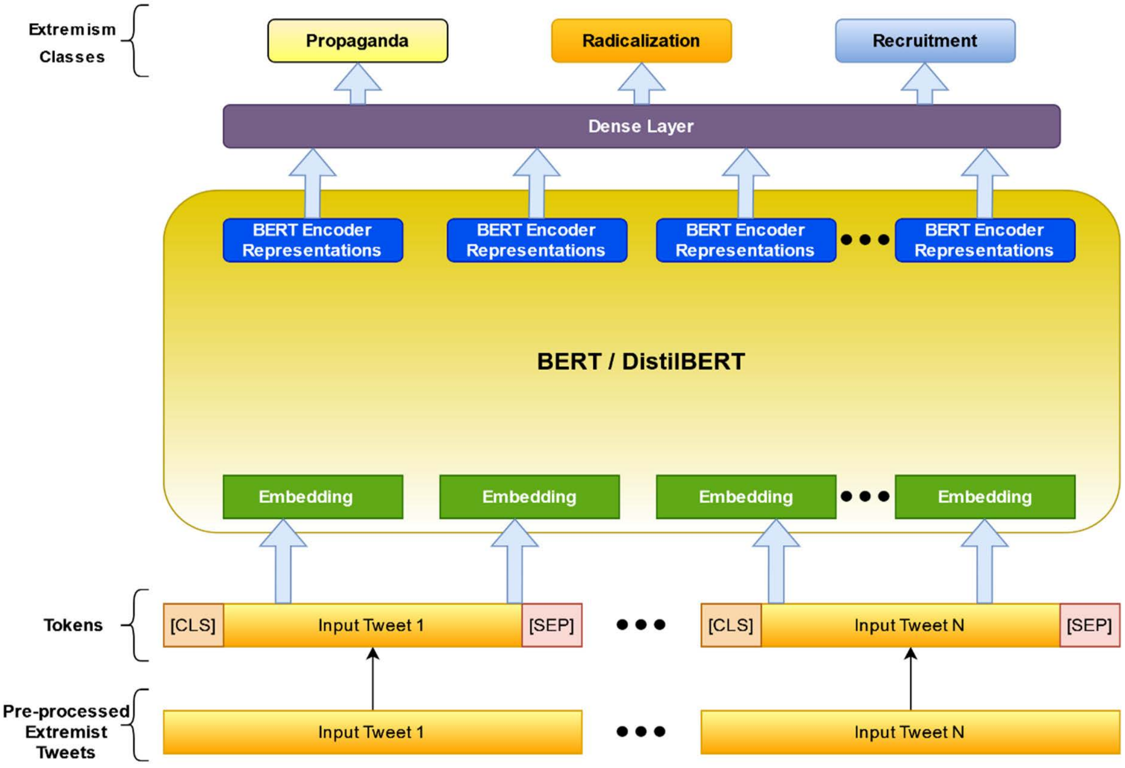

FIGURE 2. BERT / DistilBERT architecture.

A. EXPERIMENTAL SETUP

The experiments were conducted on two different systems. The initial experiments were conducted on HP Workstation Z8 G4 machine with Xeon 3GHz processor, 128 GB RAM, and Nvidia Quadro P400 2GB GPU. Hyperparameter tuning experiments were performed on Google Colab Pro.

B. DEEP LEARNING CLASSIFIER

1) BERT

Bidirectional Encoder Representations for Transformers (BERT) use the transformer attention model with the help of encoders [71] BERT processes text sequences in both left-right and right-left ways. This makes BERT efficient and accurate. BERT's stellar performance in NLP problems was the reason to choose BERT as a classifier for online extremism detection research. Figure 2 describes the general architecture of BERT used in this study. Table 6 shows hyperparameters used to obtain results using BERT. Different pre-trained networks were utilized along with BERT, such as Bert-base-uncased, Bert-large-uncased, Bert-base-cased, Bert-base-twitter, and Bert-large-twitter. The 'base' pre-trained networks usually consist of 12 layers and 768 hidden units, while 'large' pre-trained networks have 24 layers and 1024 hidden units.

2) DISTILBERT

DistilBERT is another variant of BERT [72] The DistilBERT model was created using the knowledge distillation process. This makes DistilBERT smaller and faster than BERT. For this study, DistilBERT is used in combination with pre-trained Bert-base-uncased. As seen in Figure 2, the architecture of DistilBERT is similar to BERT only difference between them is that DistilBERT performs knowledge distillation while pre-training. As seen in Table 7, fewer hyperparameters were used during DistilBERT training due to early convergence to an acceptable result.

TABLE 8. Hyperparameters used for DistilBERT.

DistilBERT Hyperparameters	Values
Learning Rate	2e-5
Optimizer	AdamW
Dropout	0.3
Batch Size	16

3) ROBERTA

RoBERTa [73] enhances BERT's masking strategy and removes the next sentence prediction embedded in BERT. This can be observed in Figure 3. RoBERTa was modeled by training BERT on large mini-batches and different learning rates. RoBERTa was also trained on larger datasets for a longer time, making RoBERTa more generalized for downstream tasks. As seen in Table 8, RoBERTa was finetuned using the same hyperparameters as BERT. The pre-trained network used is Roberta-base-uncased, which has 12 layers and 768 hidden units.

V. RESULTS AND DISCUSSIONS

Earlier literature has used ROC-AUC, Precision, Recall, f1-score, and Accuracy as the performance metric. We have selected Precision, Recall, and f1-score as performance metrics. The reason behind choosing these performance metrics is to achieve higher performance using class balance and obtain the correct representation of misclassified elements. Seven combinations of Deep Learning Classifier and pre-trained network were performed to achieve significant results in Table 9. BERT and variants give acceptable results for MIWS classification. The precision for BERT and variants lies between 0.69 to 0.72, recall lies between 0.68 to 0.72, and the f1-score ranges from 0.68 to 0.72. The most important results can be seen when BERT with bert-base-twitter and BERT with bert-large-twitter are used. BERT with bert-base-twitter gives precision, recall, f1-score and accuracy of about 0.71, while BERT with bert-large-twitter gives precision, recall, f1-score and accuracy of about 0.72 as seen in Figure 4, Figure 5, and Figure 6. DistilBERT with bert-base-uncased offers similar results to BERT with bert-base-Twitter but doesn't hold up in label-wise comparison. BERT results are better than other combinations because pre-trained networks are trained on Twitter datasets.

TABLE 9. Hyperparameters used for RoBERTa.

RoBERTa Hyperparameters	Values
Learning Rate	2e-5
Learning Rate decay	2e-4
Optimizer	AdamW
Dropout	0.3
Batch Size	16
Max sentence length	60
Training Epochs	20
Warmup steps	100

In Table 10, the rank significance of the results is provided. The ranks were calculated in descending order; thus, the lower the rank higher the significance. The Friedman Rank Test shows a p-value = 0.0006, which is less than 0.05. Therefore, the results obtained are significant for the collected dataset. The BERT model with Twitter pretrained has a lower rank compared to other models. Thus Table 11 and Table 12 show label-wise evaluation for BERT. It can be seen in both cases; performance is similar for propaganda and radicalization. But performance suffers for recruitment class for both experiments. This can be because recruitment may have more similar words to other classes.

VI. LIMITATIONS

Few limitations can be addressed in the future.

- The ideologically balanced MIWS dataset is labeled from the seed dataset with only 400 examples of ISIS and White supremacist ideology. In the future, the size of the seed dataset can be increased.
- Seed dataset contain event specific texts such as ISIS's threats to America after bombing etc. Thus only limited extremist keyword were present in seed dataset.
- The custom data collected was based on keywords. These keywords are prominently used by extremists but could have been used by critics, journalists, or news

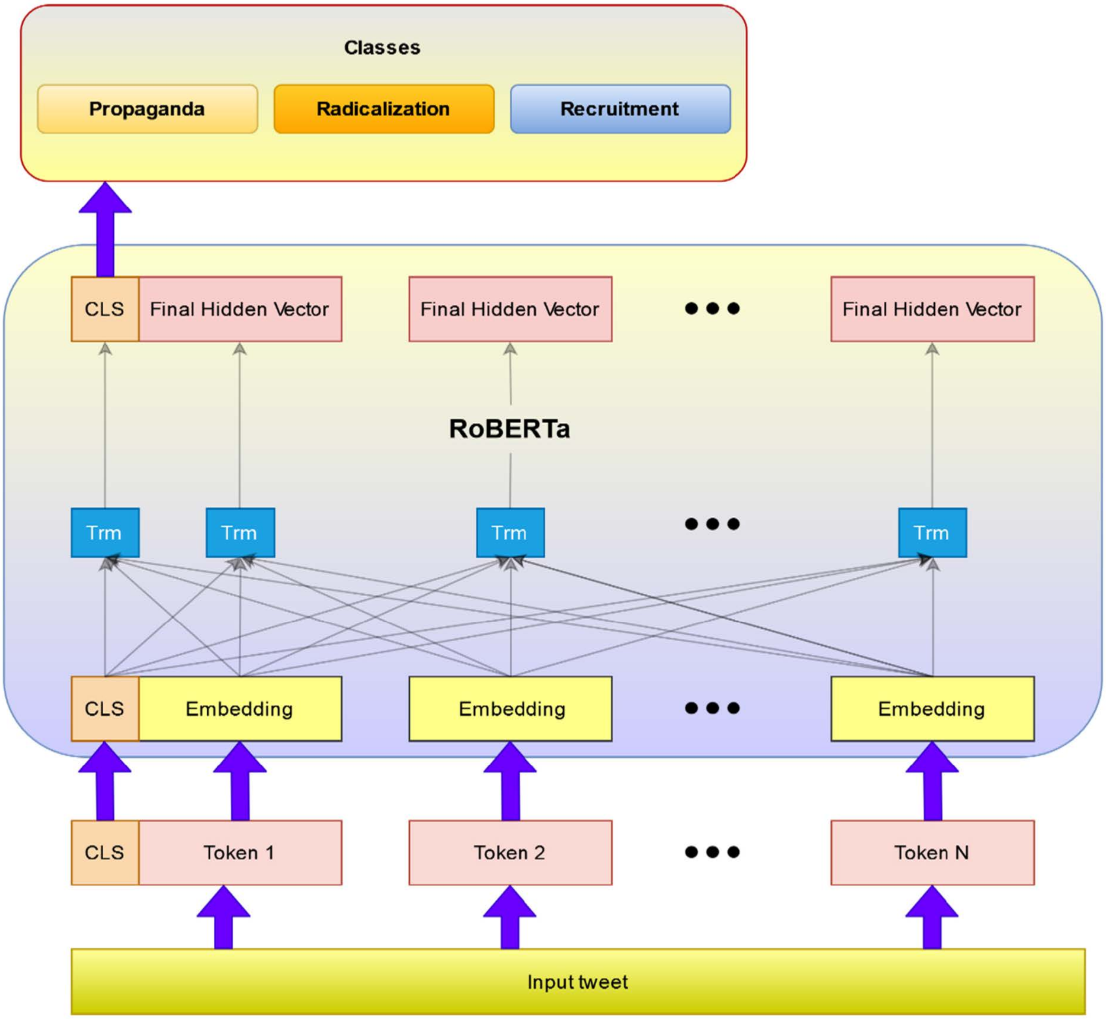

FIGURE 3. RoBERTa architecture.

TABLE 10. Aggregated results.

DL Technique	Pretrained Network	Precision	Recall	f1-score	Accuracy
BERT	Bert-based-uncased	0.69	0.69	0.69	0.69
BERT	Bert-based-cased	0.71	0.69	0.70	0.70
BERT	Bert-large-uncased	0.71	0.69	0.71	0.71
RoBERTa	roberta-base	0.69	0.68	0.68	0.68
BERT	Bert-base-twitter	0.71	0.71	0.71	0.71
DistilBERT	Bert-based-uncased	0.71	0.71	0.71	0.71
BERT	Bert-large-twitter	0.72	0.72	0.72	0.72

TABLE 11. Rank significance of results.

DL Technique	Pretrained Network	Ranks
BERT	Bert-based-uncased	5.8
BERT	Bert-based-cased	4.5
BERT	Bert-large-uncased	3.8
RoBERTa	roberta-base	6.8
BERT	Bert-base-twitter	3.0
DistilBERT	Bert-based-uncased	3.0
BERT	Bert-large-twitter	1.0

agencies to report the news. Thus, some false positives are there in the MIWS dataset.

- The tweet data collected is also event specific. Most jihadist tweets were about the Taliban takeover of

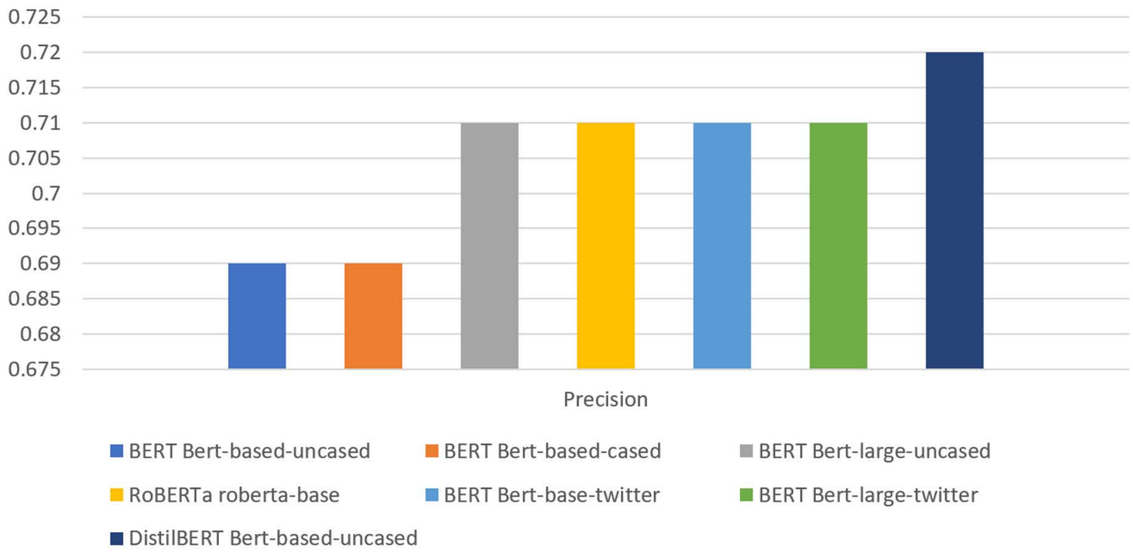

FIGURE 4. Precision of all deep learning techniques used.

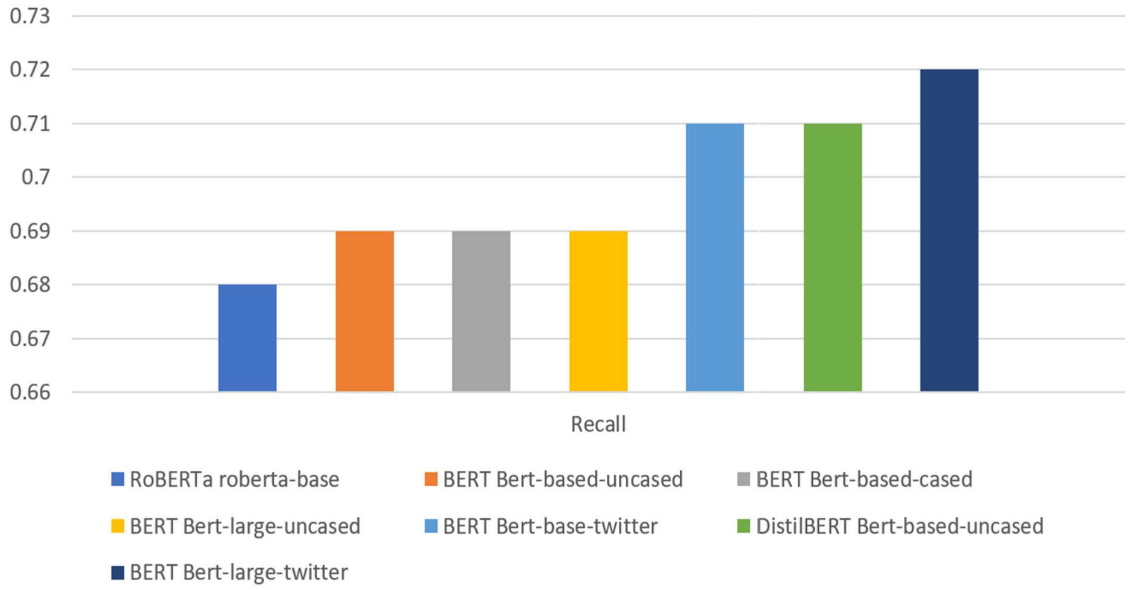

FIGURE 5. Recall of all deep learning techniques used.

TABLE 12. BERT with bert-base-twitter results.

BERT, bert-base-twitter	Labels	Precision	Recall	f1-score	Support
	Propaganda	0.71	0.73	0.72	506
	Radicalization	0.75	0.68	0.71	506
	Recruitment	0.68	0.72	0.70	506

Afghanistan, while white supremacists are antisemitic and anti-vaccine. This may cause event-related classification issues in the trained model.

- Comparison using WMD gives close distances between classes; this can be due to similar keywords across categories. So, the misclassification of some custom-collected tweets may have happened.
- Some tweets collected were unavailable on Twitter due to their extremist nature. Thus tweet recollection or user account monitoring was hampered.

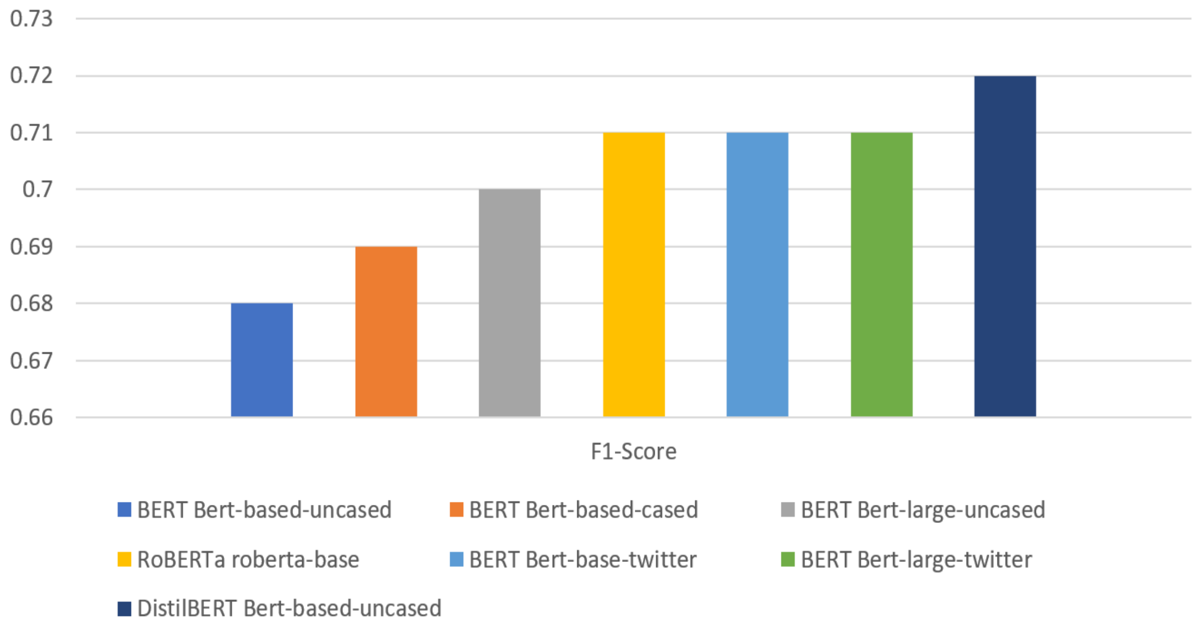

FIGURE 6. f1-score of all deep learning techniques used.

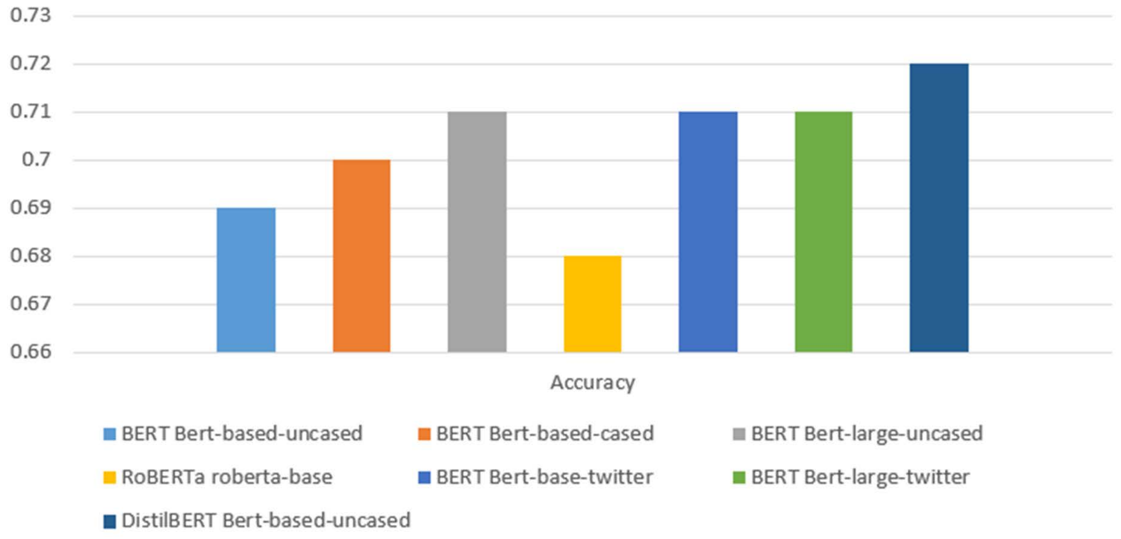

FIGURE 7. Accuracy of all deep learning techniques used.

TABLE 13. BERT with bert-large-twitter results.

BERT, bert-large-twitter	Labels	Precision	Recall	f1-score	Support
	Propaganda	0.72	0.73	0.72	506
	Radicalization	0.75	0.69	0.72	506
	Recruitment	0.68	0.72	0.70	506

VII. CONCLUSION

This research presents the development of a seed dataset and balanced multi-ideologies, multi-class, extremism tweet dataset and its evaluation using BERT and its variants. This work is the first to classify extremist tweets into propaganda, radicalization, and recruitment using deep learning to the best of our knowledge.

Examples of propaganda, radicalization and recruitment texts are collected from research articles, blogs, and websites. Thus, varied samples of seeds are collected to reduce researcher bias. Custom tweets are gathered using important extremism-related keywords from research articles, blogs, and websites. A total of 60,000 tweets from ISIS/Jihadist and White Supremacist ideologies were collected to create a balanced multi-ideology online extremism dataset. NLP techniques such as topic modelling, i.e., LDA and word distances, i.e., WMD are used for labelling collected tweets. These methods ensure tweets are segregated into propaganda, radicalization and recruitment.

This research work evaluates the multi-ideology, multi-class extremism MIWS dataset using BERT and its variants such as RoBERTa and DistilBERT. RoBERTa offers a precision of 0.69, recall of 0.68, and f1-score of 0.69. DistilBERT gives a precision of 0.71, recall of 0.71, and f1-score of 0.71.

In this research work, two different BERT pre-trained networks are employed. The first pre-trained network is trained on Wikipedia data, while the second is trained on Twitter data. Pre-trained network BERT trained on Twitter data provides the best results showing precision of 0.72, recall of 0.72, and an f1-score of 0.72.

This study sets objectives to detect extremism with insights. The automatic detection of extremism with insights can help to analyze extremist threats rapidly. This study can be used by researchers, social media networks, social media analysts, and government agencies to detect extremist propaganda and recruitment. Thus, social media networks or government agencies can take preventive measures to curb the spread of online extremism.

VIII. FUTURE WORK

This research has opened multiple avenues to classify better and detect online extremism. Following are some areas that can be enhanced in the future:

- Increasing the size of the Seed Dataset: More seed examples identified as propaganda, radicalization, and recruitment can be included.
- Geographical Location as Feature: SpaCy can extract location from tweets. These geographical locations can be used as a feature to classify extremist tweets.
- Real-Time Classification: Tweets can be fetched in real-time and classified into propaganda, radicalization, and recruitment.

ACKNOWLEDGMENT

The authors would like to thank Symbiosis International (Deemed University) for permitting us to conduct our research and use resources to accomplish the objectives. They would also like to thank the anonymous reviewers who helped them to bring out the best version of this article.

REFERENCES

[1] B. Dean. (2021). *Facebook Demographic Statistics: How Many People Use Facebook in 2021?* Backlinko. Accessed: Oct. 20, 2021. [Online]. Available: https://backlinko.com/facebook-users

[2] B. Dean. (2021). *How Many People Use Twitter in 2021? [New Twitter Stats]* Backlinko. Accessed: Oct. 20, 2021. [Online]. Available: https://backlinko.com/twitter-users

[3] A. Bermingham, M. Conway, L. McInerney, N. O'Hare, and A. F. Smeaton, "Combining social network analysis and sentiment analysis to explore the potential for online radicalisation," in *Proc. Int. Conf. Adv. Social Netw. Anal. Mining*, Jul. 2009, pp. 231–236, doi: 10.1109/ASONAM.2009.31.

[4] J. P. Farwell, "The media strategy of ISIS," *Survival*, vol. 56, no. 6, pp. 49–55, Nov. 2014, doi: 10.1080/00396338.2014.985436.

[5] W. Sharif, S. Mumtaz, Z. Shafiq, O. Riaz, T. Ali, M. Husnain, and G. S. Choi, "An empirical approach for extreme behavior identification through tweets using machine learning," *Appl. Sci.*, vol. 9, no. 18, p. 3723, Sep. 2019, doi: 10.3390/app9183723.

[6] N. MacFarquhar, A. Feuer, M. Baker, and S. Frenkel. (2020). *The Proud Boys, Who Trade in Political Violence, Get a Boost From Trump—The New York Times*. The New York Times. Accessed: Mar. 15, 2021. [Online]. Available: https://www.nytimes.com/2020/09/30/us/proud-boys-trump.html

[7] J. Coaston. (2019). *The New Zealand Shooter's Manifesto Shows How White Nationalist Rhetoric Spreads*. Vox. Accessed: Oct. 10, 2020. [Online]. Available: https://www.vox.com/identities/2019/3/15/18267163/new-zealand-shooting-christchurch-white-nationalism-racism-language

[8] R. R. Habib. (2021). *Taliban's Takeover of Afghanistan Should Not be Celebrated*. The Express Tribune. Accessed: Aug. 25, 2021. [Online]. Available: https://tribune.com.pk/article/97460/talibans-takeover-of-afghanistan-should-not-be-celebrated

[9] K. Duffy. (2021). *Facebook Failed to Prevent Far-Right Groups From Planning the US Capitol Siege, According to an Internal Report*. Business Insider. Accessed: Oct. 5, 2021. [Online]. Available: https://www.businessinsider.in/tech/news/facebook-failed-to-prevent-far-right-groups-from-planning-the-us-capitol-siege-according-to-an-internal-report/articleshow/82230545.cms

[10] J. Gage. (2019). *California Police Investigate Hate-Filled 8chan Manifesto That Could Link Synagogue Shooting to Mosque Attack*. Washington Examiner. Accessed: Oct. 10, 2020. [Online]. Available: https://www.washingtonexaminer.com/news/california-police-investigate-hate-filled-8chan-manifesto-that-could-link-synagogue-shooting-to-mosque-attack

[11] S. Prokupecz, C. Maxouris, D. Andone, S. Beech, and A. Vera. (Jun. 1, 2022). *What we Know About Buffalo Supermarket Shooting Suspect Payton Gendron*. CNN. Accessed: Jun. 1, 2022. [Online]. Available: https://edition.cnn.com/2022/05/15/us/payton-gendron-buffalo-shooting-suspect-what-we-know/index.html

[12] S. Levin. (2022). *How the Buffalo Massacre is Part of US Tradition: 'We'll Continue to See Killings*. The Guardian. Accessed: Jun. 1, 2022. [Online]. Available: https://www.theguardian.com/us-news/2022/may/18/buffalo-shooting-us-tradition-history-white-supremacist-violence

[13] C. Duffy and S. O'Brien. (2022). *Following Buffalo Shooting, 4chan Shows How Some Platforms are Accountable Only to Themselves*. CNN. Accessed: Jun. 1, 2022. [Online]. Available: https://edition.cnn.com/2022/05/18/tech/4chan-buffalo-shooting-accountability/index.html

[14] B. Collins. (2022). *The Buffalo Supermarket Shooting Suspect Allegedly Posted an Apparent Manifesto Repeatedly Citing 'Great Replacement' Theory*. NBC News. Accessed: Jun. 1, 2022. [Online]. Available: https://www.nbcnews.com/news/us-news/buffalo-supermarket-shooting-suspect-posted-apparent-manifesto-repeate-rcna28889

[15] L. Buckingham and N. Alali, "Extreme parallels: A corpus-driven analysis of ISIS and far-right discourse," *Kotuitui: New Zealand J. Social Sci.*, vol. 15, no. 2, pp. 310–331, Jul. 2020, doi: 10.1080/1177083X.2019.1698623.

[16] A. Loewenstein. (2019). *White Supremacy in Australia Set the Stage for the Christchurch Massacre*. The Nation. Accessed: Aug. 20, 2020. [Online]. Available: https://www.thenation.com/article/archive/christchurch-massacre-australia-racism-white-nationalism-media/

[17] J. M. Berger, "Nazis vs. ISIS on Twitter: A comparative study of white nationalist and ISIS online social media networks," George Washington Univ., Washington, DC, USA, Tech. Rep., 2016. [Online]. Available: https://extremism.gwu.edu/sites/g/files/zaxdzs2191/f/downloads/JMB%20Diminishing%20Returns.pdf

[18] K. Ong, "Ideological convergence in the extreme right," *Counter Terrorist Trends Analyses*, vol. 12, no. 5, pp. 1–7, Nov. 2020.

[19] V. A. Auger, "Right-wing terror," *Perspect. Terrorism*, vol. 14, no. 3, pp. 87–97, Nov. 2020.

[20] G. Davies, E. Wu, and R. Frank, "A Witch's brew of grievances: The potential effects of COVID-19 on radicalization to violent extremism," *Stud. Conflict Terrorism*, vol. 2021, pp. 1–24, May 2021, doi: 10.1080/1057610X.2021.1923188.

[21] B. L. Smith. (1999). *Propoganda Encyclopedia*. Britannica. [Online]. Available: https://www.britannica.com/topic/propaganda

[22] C. McCauley and S. Moskalenko, "Mechanisms of political radicalization: Pathways toward terrorism," *Terrorism Political Violence*, vol. 20, no. 3, pp. 415–433, Jul. 2008, doi: 10.1080/09546550802073367.

[23] M. S. Kimmel, "Globalization and its Mal(e)contents," *Int. Sociol.*, vol. 18, no. 3, pp. 603–620, Sep. 2003, doi: 10.1177/02685809030183008.

[24] A. Kingdon. (2020). *The Gift of the Gab: The Utilisation of COVID-19 for Neo-Nazi Recruitment*. Global Network on Extremism and Technology. Accessed: Oct. 10, 2020. [Online]. Available: https://gnet-research.org/2020/05/07/the-gift-of-the-gab-the-utilisation-of-covid-19-for-neo-nazi-recruitment/

[25] BBC. (2020). *France Teacher Attack: Four Pupils Held Over Beheading*. BBC. Accessed: Oct. 10, 2021. [Online]. Available: https://www.bbc.com/news/world-europe-54598546

[26] B. Ray and G. E. Marsh, "Recruitment by extremist groups on the internet," *1st Monday*, vol. 6, no. 2, Feb. 2001, doi: 10.5210/fm.v6i2.834.

[27] L. Nizzoli, M. Avvenuti, S. Cresci, and M. Tesconi, "Extremist propaganda tweet classification with deep learning in realistic scenarios," in *Proc. 10th ACM Conf. Web Sci.*, Jun. 2019, pp. 203–204, doi: 10.1145/3292522.3326050.

[28] O. de Gibert, N. Perez, A. García-Pablos, and M. Cuadros, "Hate speech dataset from a white supremacy forum," Sep. 2018, *arXiv:1809.04444*.

[29] O. Araque and C. A. Iglesias, "An approach for radicalization detection based on emotion signals and semantic similarity," *IEEE Access*, vol. 8, pp. 17877–17891, 2020, doi: 10.1109/ACCESS.2020.2967219.

[30] Fifth Tribe. (2015). *How ISIS Uses Twitter*. Kaggle. [Online]. Available: https://www.kaggle.com/fifthtribe/how-isis-uses-twitter

[31] FifthTribe. (2017). *ISIS Religious Text*. Kaggle. Accessed: Oct. 10, 2020. [Online]. Available: https://www.kaggle.com/fifthtribe/isis-religious-texts

[32] K. Brendan. (2020). *The Gab Hate Corpus: A Collection of 27k Posts Annotated for Hate Speech*. Psyarxiv. Accessed: Oct. 10, 2020. [Online]. Available: https://psyarxiv.com/hqjxn/

[33] H. S. Alatawi, A. M. Alhothali, and K. M. Moria, "Detecting white supremacist hate speech using domain specific word embedding with deep learning and BERT," *IEEE Access*, vol. 9, pp. 106363–106374, 2021, doi: 10.1109/ACCESS.2021.3100435.

[34] S. Jaki and T. De Smedt, "Right-wing German hate speech on Twitter: Analysis and automatic detection," Oct. 2019, *arXiv:1910.07518*, doi: 10.48550/arxiv.1910.07518.

[35] M. Asif, A. Ishtiaq, H. Ahmad, H. Aljuaid, and J. Shah, "Sentiment analysis of extremism in social media from textual information," *Telematics Informat.*, vol. 48, May 2020, Art. no. 101345, doi: 10.1016/j.tele.2020.101345.

[36] S. Agarwal and A. Sureka, "A focused crawler for mining hate and extremism promoting videos on YouTube.," in *Proc. 25th ACM Conf. Hypertext Social Media*, Sep. 2014, pp. 294–296, doi: 10.1145/2631775.2631776.

[37] S. Agarwal and A. Sureka, "Investigating the potential of aggregated tweets as surrogate data for forecasting civil protests," in *Proc. 3rd IKDD Conf. Data Sci. (CODS)*, Pune, India. New York, NY, USA: Association for Computing Machinery, 2016, pp. 1–6, Art. no. 8, doi: 10.1145/2888451.2888466.

[38] M. Benigni, *Detection and Analysis of Online Extremist Communities*. Pittsburgh, PA, USA: Carnegie Mellon Univ., 2017.

[39] A. Kaur, J. K. Saini, and D. Bansal, "Detecting radical text over online media using deep learning," Jul. 2019, *arXiv:1907.12368*, doi: 10.48550/arXiv.1907.12368.

[40] M. Nouh, J. R. C. Nurse, and M. Goldsmith, "Understanding the radical mind: Identifying signals to detect extremist content on Twitter," in *Proc. IEEE Int. Conf. Intell. Secur. Informat. (ISI)*, Jul. 2019, pp. 98–103, doi: 10.1109/ISI.2019.8823548.

[41] S. Ahmad, M. Z. Asghar, F. M. Alotaibi, and I. Awan, "Detection and classification of social media-based extremist affiliations using sentiment analysis techniques," *Hum.-Centric Comput. Inf. Sci.*, vol. 9, no. 1, Dec. 2019, doi: 10.1186/s13673-019-0185-6.

[42] S. Mussiraliyeva, M. Bolatbek, B. Omarov, and K. Bagitova, "Detection of extremist ideation on social media using machine learning techniques," in *Proc. 12th Int. Conf. Comput. Collective Intell. (ICCCI)*, Da Nang, Vietnam. Berlin, Germany: Springer-Verlag, Nov./Dec. 2020, pp. 743–752, doi: 10.1007/978-3-030-63007-2_58.

[43] Z. U. Rehman, S. Abbas, M. A. Khan, G. Mustafa, H. Fayyaz, M. Hanif, and M. A. Saeed, "Understanding the language of ISIS: An empirical approach to detect radical content on Twitter using machine learning," *Comput., Mater. Continua*, vol. 66, no. 2, pp. 1075–1090, 2021, doi: 10.32604/cmc.2020.012770.

[44] M. Petrovskiy and M. Chikunov, "Online extremism discovering through social network structure analysis," in *Proc. IEEE 2nd Int. Conf. Inf. Comput. Technol. (ICICT)*, Mar. 2019, pp. 243–249, doi: 10.1109/INFOCT.2019.8711254.

[45] I. V. Mashechkin, M. I. Petrovskiy, D. V. Tsarev, and M. N. Chikunov, "Machine learning methods for detecting and monitoring extremist information on the internet," *Program. Comput. Softw.*, vol. 45, no. 3, pp. 99–115, May 2019, doi: 10.1134/S0361768819030058.

[46] H. Saif, T. Dickinson, L. Kastler, M. Fernandez, and H. Alani, "A semantic graph-based approach for radicalisation detection on social media," in *Proc. Eur. Semantic Web Conf.*, 2017, pp. 571–587, doi: 10.1007/978-3-319-58068-5_35.

[47] M. F. Abrar, M. S. Arefin, and M. S. Hossain, "A framework for analyzing real-time tweets to detect terrorist activities," in *Proc. Int. Conf. Electr., Comput. Commun. Eng. (ECCE)*, Feb. 2019, pp. 1–6, doi: 10.1109/ECACE.2019.8679430.

[48] U. Kursuncu, M. Gaur, C. Castillo, A. Alambo, K. Thirunarayan, V. Shalin, D. Achilov, I. B. Arpinar, and A. Sheth, "Modeling Islamist extremist communications on social media using contextual dimensions," in *Proc. ACM Hum.-Comput. Interact.*, vol. 3, Nov. 2019, pp. 1–22, doi: 10.1145/3359253.

[49] M. Moussaoui, M. Zaghdoud, and J. Akaichi, "A possibilistic framework for the detection of terrorism-related Twitter communities in social media," *Concurrency Comput., Pract. Exper.*, vol. 31, no. 13, p. e5077, Jul. 2019, doi: 10.1002/cpe.5077.

[50] T. De Smedt, G. De Pauw, and P. Van Ostaeyen, "Automatic detection of online Jihadist hate speech," Feb. 2018, *arXiv:1803.04596*, doi: 10.48550/arXiv.1803.04596.

[51] B. S. Iskandar, "Terrorism detection based on sentiment analysis using machine learning," *J. Eng. Appl. Sci.*, vol. 12, no. 3, pp. 691–698, 2017, doi: 10.36478/jeasci.2017.691.698.

[52] A. T. Chatfield, C. G. Reddick, and U. Brajawidagda, "Tweeting propaganda, radicalization and recruitment: Islamic state supporters multi-sided Twitter networks," in *Proc. 16th Annu. Int. Conf. Digit. Government Res.*, May 2015, pp. 239–249, doi: 10.1145/2757401.2757408.

[53] M. Fernandez, M. Asif, and H. Alani, "Understanding the roots of radicalisation on Twitter," in *Proc. 10th ACM Conf. Web Sci.*, May 2018, pp. 1–10, doi: 10.1145/3201064.3201082.

[54] O. Araque and C. A. Iglesias, "An ensemble method for radicalization and hate speech detection online empowered by sentic computing," *Cognit. Comput.*, vol. 14, no. 1, pp. 48–61, Jan. 2022, doi: 10.1007/s12559-021-09845-6.

[55] S. Mussiraliyeva, B. Omarov, P. Yoo, and M. Bolatbek, "Applying machine learning techniques for religious extremism detection on online user contents," *Comput., Mater. Continua*, vol. 70, no. 1, pp. 915–934, 2022, doi: 10.32604/cmc.2022.019189.

[56] Vkontakte. (2014). *Vkontakte Social Network*. Accessed: Oct. 10, 2020. [Online]. Available: https://vk.com/topic-78863260_30603285

[57] S. Mussiraliyeva, "Applying deep learning for extremism detection," in *Proc. Int. Conf. Adv. Informat. Comput. Res. (ICAICR)*, 2021, pp. 597–605, doi: 10.1007/978-981-16-3660-8_56.

[58] 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," *IEEE Access*, vol. 9, pp. 48364–48404, 2021, doi: 10.1109/ACCESS.2021.3068313.

[59] B. Johnson. (2020). *Shared Themes, Tactics in White Supremacist and Islamist Extremist Propaganda—Homeland Security Today*. Homeland Security Today. Accessed: Mar. 10, 2021. [Online]. Available: https://www.hstoday.us/subject-matter-areas/counterterrorism/shared-themes-recruitment-tactics-in-white-supremacist-and-islamist-extremist-propaganda/

[60] (Jan. 1, 2021). *ISIS Recruiters, Propagandists, and Inciters to Violence Operating on Twitter | Counter Extremism Project*. Counter Extremism. Accessed: Mar. 22, 2021. [Online]. Available: https://www.counterextremism.com/content/isis-recruiters-propagandists-and-inciters-violence-operating-twitter

[61] M. Rowe and H. Saif, "Mining pro-ISIS radicalisation signals from social media users," in *Proc. ICWSM*, Cologne, Germany, 2016.

[62] O. Theodosiadou, K. Pantelidou, N. Bastas, D. Chatzakou, T. Tsikrika, S. Vrochidis, and I. Kompatsiaris, ''Change point detection in terrorism-related online content using deep learning derived indicators,'' *Information*, vol. 12, no. 7, p. 274, Jul. 2021, doi: 10.3390/info12070274.

[63] M. Hartung, R. Klinger, F. Schmidtke, and L. Vogel, ''Identifying right-wing extremism in German Twitter profiles: A classification approach,'' in *Proc. Int. Conf. Appl. Natural Lang. Inf. Syst.*, 2017, pp. 320–325, doi: 10.1007/978-3-319-59569-6_40.

[64] A. Necaise, A. Williams, H. Vrzakova, and M. J. Amon, ''Regularity versus novelty of users' multimodal comment patterns and dynamics as markers of social media radicalization,'' in *Proc. 32st ACM Conf. Hypertext Social Media*, Aug. 2021, pp. 237–243, doi: 10.1145/3465336.3475095.

[65] L. Kaati, E. Omer, N. Prucha, and A. Shrestha, ''Detecting multipliers of Jihadism on Twitter,'' in *Proc. IEEE Int. Conf. Data Mining Workshop (ICDMW)*, Nov. 2015, pp. 954–960, doi: 10.1109/ICDMW.2015.9.

[66] *Time to Stop Jew Hatred?: Submission to the Queensland Parliamentary Inquiry Into Serious Vilification and Hate Crime*, QJBD, Brisbane, QLD, USA, 2021.

[67] D. Kochedykov, M. Apishev, L. Golitsyn, and K. Vorontsov, ''Fast and modular regularized topic modelling,'' in *Proc. 21st Conf. Open Innov. Assoc. (FRUCT)*, Nov. 2017, pp. 182–193, doi: 10.23919/FRUCT.2017.8250181.

[68] L. Wu, I. E.-H. Yen, K. Xu, F. Xu, A. Balakrishnan, P.-Y. Chen, P. Ravikumar, and M. J. Witbrock, ''Word mover's embedding: From Word2Vec to document embedding,'' in *Proc. Conf. Empirical Methods Natural Lang. Process.* Brussels, Belgium: Association for Computational Linguistics, Oct./Nov. 2018, pp. 4524–4534. [Online]. Available: https://aclanthology.org/D18-1482, doi: 10.18653/v1/D18-1482.

[69] M. Gaikwad, S. Ahirrao, S. Phansalkar, and K. Kotecha, ''Multi-ideology ISIS/Jihadist white supremacist (MIWS) dataset for multi-class extremism text classification,'' *Data*, vol. 6, no. 11, p. 117, Nov. 2021, doi: 10.3390/data6110117.

[70] M. Buda, A. Maki, and M. A. Mazurowski, ''A systematic study of the class imbalance problem in convolutional neural networks,'' *Neural Netw.*, vol. 106, pp. 249–259, Oct. 2018, doi: 10.1016/j.neunet.2018.07.011.

[71] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ''BERT: Pre-training of deep bidirectional transformers for language understanding,'' in *Proc. Conf. North*, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[72] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, ''DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter,'' *CoRR*, vol. abs/1910.01108, 2019. [Online]. Available: http://arxiv.org/abs/1910.01108

[73] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, ''RoBERTa: A robustly optimized BERT pretraining approach,'' Tech. Rep., Jul. 2019.

MAYUR GAIKWAD received the master's degree in computer science and engineering from the Symbiosis Institute of Technology, Pune, where he is currently pursuing the Ph.D. degree with Symbiosis International (Deemed University). His research interests include machine learning, deep learning, and natural language processing.

SWATI AHIRRAO received the Ph.D. degree from Symbiosis International (Deemed University), Lavale, Pune, Maharashtra, India. She is currently working as an Associate Professor with SIT. She has published over 31 research papers in international journals and conferences. According to Google Scholar, her articles have 71 citations, with an H-index of three and an i10-index of two. Her research interests include big data analytics, machine learning, deep learning, natural language processing, and reinforcement learning.

KETAN KOTECHA received the Ph.D. and M.Tech. degrees from IIT Bombay. He is currently holding the positions as a head of the Symbiosis Centre for Applied AI (SCAAI), the director of the Symbiosis Institute of Technology, and the dean of the Faculty of Engineering, Symbiosis International (Deemed University). He has expertise and experience in cutting-edge research and AI and deep learning projects for the last more than 25 years. He has published more than 100 widely in many excellent peer-reviewed journals on various topics ranging from cutting-edge AI, education policies, teaching-learning practices, and AI for all. He has also published three patents and delivered keynote speeches at various national and international forums, including at Machine Intelligence Laboratory, USA; IIT Bombay under the World Bank Project; the International Indian Science Festival organized by the Department of Science Technology, Government of India; and many more. He was a recipient of the two SPARC projects worth INR 166 lacs from MHRD, Government of India in AI, in collaboration with Arizona State University, USA; and the University of Queensland, Australia. He was also a recipient of numerous prestigious awards, such as Erasmus+ Faculty Mobility Grant to Poland, DUO-India Professors Fellowship for research in responsible AI in collaboration with Brunel University, U.K., LEAP Grant at Cambridge University, U.K., UKIERI Grant with Aston University, U.K., and the Grant from Royal Academy of Engineering, U.K., under Newton Bhabha Fund. He is also an Academic Editor of the *PeerJ Computer Science* journal and an Associate Editor of IEEE ACCESS journal.

AJITH ABRAHAM (Senior Member, IEEE) received the M.S. degree from Nanyang Technological University, Singapore, in 1998, and the Ph.D. degree in computer science from Monash University, Melbourne, Australia, in 2001.

He is currently the Director of the Machine Intelligence Research Laboratories (MIR Laboratories), a Not-for-Profit Scientific Network for Innovation and Research Excellence Connecting Industry and Academia. The Network with HQ in Seattle, USA, is currently has over 1,500 scientific members from over 105 countries. As an Investigator/a Co-Investigator, he has won research grants worth over 100 Million USA Dollar. He currently holds two university professorial appointments. He works as a Professor in artificial intelligence with Innopolis University, Russia, and the Yayasan Tun Ismail Mohamed Ali Professorial Chair in artificial intelligence with UCSI, Malaysia. He works in a multi-disciplinary environment. He has authored/coauthored more than 1,400 research publications out of which there are more than 100 books covering various aspects of computer science. One of his books was translated into Japanese and a few other articles were translated into Russian and Chinese. He has more than 46,000 academic citations (H-index of more than 102 as Per Google Scholar). He has given over 150 plenary lectures and conference tutorials (in more than 20 countries). He was the Chair of IEEE Systems Man and Cybernetics Society Technical Committee on Soft Computing (which has over 200 members), from 2008 to 2021, and served as a Distinguished Lecturer for IEEE Computer Society representing Europe, from 2011 to 2013. He was the Editor-in-Chief of *Engineering Applications of Artificial Intelligence* (EAAI), from 2016 to 2021, and serves/served on the editorial board for over 15 international journals indexed by Thomson IS.