IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 4 November 2022, accepted 25 November 2022, date of publication 9 December 2022, date of current version 27 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3227962

RESEARCH ARTICLE

Detecting Extremism on Twitter During U.S. Capitol Riot Using Deep Learning Techniques

ARUNDARASI RAJENDRAN[1], VATTIKUTI SREE SAHITHI[1], CHHAVI GUPTA[1], MADHURI YADAV[1], SWATI AHIRRAO[1], KETAN KOTECHA[2], MAYUR GAIKWAD[1], (Member, IEEE), AJITH ABRAHAM[3], (Senior Member, IEEE), NADA AHMED[4], AND SARAH M. ALHAMMAD[4]

[1]Symbiosis Institute of Technology, Symbiosis International University (Deemed University), Lavale, Pune, Maharashtra 412115, India
[2]Symbiosis Centre for Applied Artificial Intelligence (SCAAI), Symbiosis International (Deemed University), Pune, Maharashtra 412115, India
[3]Machine Intelligence Research Laboratories, Auburn, WA 98071, USA
[4]Department of Computer Sciences, College of Computer Science and Information, Princess Nourah Bint Abdulrahman University, Riyadh 11564, Saudi Arabia

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

ABSTRACT In the 21st century, social media platforms have become famous for communicating ideas, opinions, and emotions. These platforms are influential in reaching out to youth, recruiting, and spreading propaganda. Extremist groups are now active users of social media platforms; therefore, it is necessary to monitor their activities. Therefore, there is an urgent need to detect extremism on social media platforms. Existing research on extremism lacks a dedicated extremism dataset and provides minimal insights into extremism texts. This study introduces the development of an extremism dataset containing tweets collected from Twitter and classifying extremism texts as propaganda, recruitment, radicalization, and non-extremism. The proposed extremism dataset is evaluated using different Artificial Intelligence approaches such as Bi-LSTM, BERT, RoBERTa, and DistilBERT. Among the four models, RoBERTa proved to be the most suitable for detecting extremism on social media, with an accuracy of 95%.

INDEX TERMS Deep learning, extremism detection, machine learning, sentiment analysis, social media analysis, twitter, U.S. capitol riot.

I. INTRODUCTION

Social media platforms are used to make friends online, chat or share informal information. Currently, it is used as an application in business marketing, discussion on the latest events, debates on political events, news, entertainment, and so on. Today there are over 3.78 billion users of social media worldwide [1] and 330 million monthly active users on Twitter [2], with an average of 500 million tweets posted every day [3]. A study by Maryville University suggests that almost 72% of adults in the United States are users of social media platforms [4], which will increase in the coming years. Therefore, social networking services such as Facebook, YouTube, Instagram, and Twitter are used by the public to influence election results and mislead the public through false information. False information through social sites sometimes causes uproar among the masses giving rise to protests and riots. Riots result from dissatisfaction with a particular event, including violence and vandalism. Many political or social movements are intended to gain either political or social advantage, yet they do not aim to disturb the nation's harmony. Usually, extremists target the emotional and moral sentiments of the masses to disrupt the peace within the communities, and one of their ways of doing that is through social media. Therefore, social media texts can be examined to identify the perpetrators influencing and engaging people to support the riot. The agenda of such riots are planned through social media platforms by creating groups to fight for a common cause or sometimes through the spread of misinformation. Individuals can spread extremism by discussing their views and opinions on the outcome of an event or through conspiratorial theories posted by subversive

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

groups. The language of such individuals and groups can be devoted to a single person, misleading the public and being supportive of one side. Hence, extremism detection on social media is vital for the detection of extremist users as well as for preventing extremist content from being posted.

The latest event certifying the negative impact of social media platforms was the 2020 U.S. presidential election which gave rise to the infamous Capitol Riot. Around 300k (i.e., 3 million) tweets associated with the presidential elections in the United States in 2020 contained deceptive content [3]. With the impending outcome of the presidential election in 2020, users discussed the 2020 U.S. election, Donald Trump, voter fraud, and election fraud across social media platforms such as Parler [5], Facebook, Instagram, and Telegram [6]. The paper [7] claimed that on 6th January 2020, the pro-Trump crowd stormed the U.S. Capitol. The attack on the Capitol was planned using social media platforms where 800 people stormed the Capitol [8] with the intent of causing harm and disharmony among the citizens. Such posts on social media enraged other users to cause violence; therefore, thorough attention and identification of such content are possible using deep learning techniques.

Some papers have explored social media texts about the U.S. Capitol riot, especially from Parler and Twitter. The studies like [9] and [10] have discussed the U.S. Capitol riot and its theoretical aspects, focusing on Twitter posts and the users' reactions. However, there is a lack of research on practical insights using state-of-the-art techniques to help improve and control extremist activities through online platforms for political riots.

Researchers have explored the political causes of the U.S. Capitol riot attack; however, few studies experimented to find extremism in tweets related to this riot. Besides, the publicly available Twitter datasets on the U.S. Capitol riot and U.S. presidential elections in 2020 are also limited. The existing research talks about detecting online extremism and extremism text sentiment analysis. Furthermore, most extremism research focuses on detecting radical ISIS accounts, sentiment analysis of an event, or the prediction of the possibility of the occurrence of a protest, utilizing machine learning and deep learning approaches. However, advanced models such as BERT and RoBERTa are efficient for text categorization, but they have not experimented with much. Besides, the identification of multiclass extremism has not been a popular choice among many authors. Thus, social media text analysis for identification of the type of extremism was absent in existing research. The datasets required for the experiment did not have the necessary features to train and test the models. Besides, there was a lack of post-classification and dataset training for performing model testing. The use of traditional machine learning models is prevalent in many extremism studies.

Nevertheless, deep learning models are also implemented to compare and discuss their performance. Therefore, this research aims to overcome all these limitations by gathering data with the required attributes. This curated dataset can help observe the trends prevalent on Twitter during the U.S. Capitol riot, classify tweets into propaganda, recruitment, radicalization and non-extremism, and compare performance of trained deep learning models on the collected dataset. The significant research gaps found in the literature survey are as listed below.

- Lack of publicly available dataset on U.S. Capitol riot.
- Limited discussion on techniques helpful in combating online extremism.
- Limited research on identifying extreme political discourse-related social media posts.

A. CONTRIBUTIONS

This research has contributed significant work keeping the research gaps and critical objectives in mind. The significant contributions of this work are shown below.

- Development of seed dataset and online extremism dataset consists of Tweets collected from Twitter.
- Evaluation of online extremism dataset using LSTM, BERT and its variants are known as Roberta and DistilBERT.
- Evaluation of extremism dataset and classification of tweets into multiple labels such as propaganda, recruitment, radicalism, and non-extremism.
- Analysis of trending hashtags of before and after the U.S. Capitol riot incident.

This research paper is organized into six sections. The second section presents the literature review of the previous work performed, followed by the third section, which discusses the proposed architecture of this work. Then sections four, five, and six mention the data collection, preprocessing, and visualization. The seventh section covers the technical aspect of the work, explaining the experimental setup, and the observed results during the experiment. Limitations in this research are mentioned in section eight. The ninth section discusses the future scope and conclusion of this study.

II. RELATED WORK

This section presents an overview of relevant research papers addressing the classification and detection of social media-based extremist associations and predicting public demonstrations.

Today many researchers investigate the automatic detection of extremists on social media texts using various methods [11], [12]. Not only that, research has also been done on public protests and elections for sentiment analysis [13], [14], [15]. Another study [16] extracts the sentiment or emotion behind social media texts in specific contexts. Machine learning techniques have been utilized extensively in extremism research since 2013 [17]. The prominent machine learning algorithms used for extremism detection research are Logistic Regression, Naive Bayes, Decision Tree Classifier, K-Neighbors Classifier, Random Forest Classifier, and Support Vector Machine [18]. On the other hand, the implemented deep learning techniques for extremism detection research are Convolutional Neural Networks (CNN) and

Recurrent Neural Networks (RNN) [18]. Automated detection of extremism is more focused on areas related to social movements, presidential elections, political issues, and terrorism by researchers.

A. MACHINE LEARNING CLASSIFIERS

Specific research papers have tested different machine learning algorithms to detect extremism content related to terrorism. Hamidreza [11] designed an automatic detection scheme to detect extremists based on three features of a social media user. These features are the textual content of the user, profile information and usernames. The effectiveness of the automatic detection scheme is demonstrated by testing the trained model on a realistic ISIS dataset collected from Twitter. The dataset contains messages by ISIS terrorist groups for recruitment and propaganda [19]. The authors used a set of 3000 Twitter handles in the experiment, divided into 150 suspended ISIS-related Twitter accounts and 150 regular Twitter user accounts [11]. Various semi-supervised and supervised machine learning algorithms were implemented to predict extremist users. These algorithms are SVM, Char-LSTM, LabelSpreading (RBF), Laplacian SVM, LabelSpreading (KNN), Co-Training (SVM), KNN, Gaussian NB, L.R., AdaBoost, and Random forest. The semi-supervised, LabelSpreading and Char-LSTM achieved the highest F1 score. Compared to others in the positive class, Char-LSTM has a high precision of 77% and a high recall of 76%.

Abd-Elaal et al. [20] article presented an intelligent system for detecting ISIS online communities on the social media platform Twitter. The dataset for this study was obtained by looking at extremist accounts on Twitter that used the most common hashtags in ISIS propaganda. Around 21,000 tweets were collected for each Pro-ISIS, Anti-ISIS, and non-ISIS user. The dataset underwent various transformations, dividing them into text features vectoring, text feature analytics, and behavioural features organization. The suggested system examines linguistic and behavioural characteristics such as hashtags, mentions, and followers. This system features a crawling subsystem that establishes an ISIS account detector using previously identified ISIS-related accounts. Using the crawling subsystem, anyone can invade ISIS's online Twitter community. It also features an inquiry subsystem for detecting Pro-ISIS accounts. The user can use inquiring subsystems to look up a specific Twitter account using the Twitter ID as input. The studies utilized six distinct machine learning algorithms: Bernoulli Naive Bayes, Decision Tree Classifier, K Neighbors Classifier, Linear Support Vector Classifier, Logistic Regression, and Random Forest Classifier.

Mussiraliyeva et al. [21] the study discusses the identification of religious extremism on social networks using Machine learning models on the dataset curated from the V.K. social network in the Kazakh language. The authors tested the dataset with six machine learning models: Support Vector Machine, Naive Bayes, Logistic Regression, Random Forest, Decision Trees, and K Nearest Neighbors. The author has used multiple feature techniques such as Statistical Features, LIWC, POS, and TF-IDF to improve the model's accuracy. Also, they applied oversampling and under-sampling methods to compare the respective performance of the models. The best result was achieved by Naive Bayes, having 94% accuracy, which showed its efficiency in detecting extremist content on the web. Likewise, Aldera et al. [22] paper focused on classifying extremist posts in the Arabic language using 89816 tweets annotated manually. The models used to test the dataset included Support Vector Machine, Logistic Regression, Multinomial Naive Bayes, Random Forest, and BERT with TF-IDF as one of the feature extraction methods. Among the Machine learning models, SVM achieved the highest accuracy of 97.29%, and BERT outperformed it by 0.20% accuracy, proving its efficiency in text classification over machine learning models.

Meanwhile, some research papers investigated the probability of a public protest happening. Bahrami et al. [23] aimed to predict protests through machine learning algorithms. It first searches Twitter's Trending Topics for hashtags that call for protests and downloads the associated tweets. Four machine learning algorithms are used to predict the tweets. Their findings show that Twitter can effectively forecast future protests, with an average prediction accuracy of over 75%. Different classifiers are examined in this study, including C4.5, Naive Bayes, Logistic Regression, and SVM, with Logistic Regression yielding the best overall results. This research focuses on predicting violent public protests. Few papers have experimented with predicting and analyzing mild forms of political protest, for example, as seen in [24]. In addition, some studies explore violent protests; for instance, another study [25] forecasts when a protest in China will take place by identifying protest-related articles and negative propaganda. There are some limitations to using machine learning algorithms in this research. Machine learning algorithms cannot take the overall dependencies associated with a sentence; thus, machine learning models do not efficiently categorize the text as extremist or non-extremist. Another limitation is that machine learning algorithms require feature extraction to achieve a better performance score [17]. Also, it cannot consider extensive data because of predefined features, and context analysis is challenging using machine learning [17].

B. DEEP LEARNING CLASSIFIERS

Natural Language Processing is being used to perform analysis on extracted text data. Recently, pre-trained deep learning models that produce word embeddings were used to analyze text data. Pre-trained models such as BERT and LSTM are becoming increasingly popular as a result of their excellent accuracy when compared to other standard machine learning models. Sentiment detection was also done with the help of machine learning and deep learning models. Previous work has shown using LSTM, GRU and BERT models for sentiment analysis. Most research papers have shown how they have used machine learning models for sentiment detection, such as SVM, Naive Bayes, and Logistic Regression.

TABLE 1. Machine learning classifiers.

Source	ML classifier(s)	Techniques used	Hyperparameters	Evaluation metrics
[11]	SVM, Char-LSTM, Random forest, Adaboost, Logistic Regression, Gaussian Naive Bayes, KNN, and Laplacian SVM	Feature engineering	SVM: C = 1, kernel = linear, tolerance = 0.001. Char-LSTM: maximum, username length = 10, dimension = 16, single layer units = 30. RF: Estimators = 200. Adaboost: estimators = 200, learning rate = 0.01. LR: penalty = l2, C = 1, tolerance = 0.01. GNB: no parameter to tune KNN: neighbors= 5. L-SVM: kernel = linear, Cl = 0.6 and Cs = 0.6.	**SVM**: P-96%, R-50%, F1-65% **Char-LSTM**: P-77%, R-76%, F1-76% **RF**: P-79%, R-71%, F1-74% **Adaboost**: P - 88%, R - 58%, F1-69% **LR**: P-76%, R-61%, F1-67% **GNB**: P-89%, R-56%, F1-69% **KNN**: P-81%, R-70%, F1-74% **L-SVM**: P-89%, R-60%, F1-70%
[20]	Decision tree classifier, k-neighbors classifier, logistic regression, random forest classifier, linear SVC, bernoulli N.B.	Word embedding	N.A.	**Bernoulli NB**: A-86% **DT**: A-79% **KNN**: A-77% **Linear SVC**: A-84% **LR**: A-84% **RF**: A-80%
[21]	SVM,Decision Tree,Random Forest,KNN,Naive Bayes,Logistic Regression	Statistical Features,POS,TF-IDF,LIWC	NA	**SVM (Statistical Features+TF-IDF+POS)**: A-84.12%, P-5.12%, R-66.25%, F1-36.43%, AUC-82.63% **DT (Statistical Features+TF-IDF)**: A-94.44%, P-95.29%, R-20.1%, F1-33.2%, AUC-64.72% **RF(Statistical Features+TF-IDF+POS)**: A-93.69%, P-100%, R-8.19%, F1-15.14%, AUC-91.51% **KNN (Statistical Features+TF-IDF+POS)**: A-93.54%, P-81.58%, R-7.69%, F1-14.06%, AUC-61.05% **NB (Statistical Features+TF-IDF)**: A-96.81%, P-89.42%, R-

TABLE 1. ***(Continued.)*** **Machine learning classifiers.**

				60.79%, F1-72.38%, AUC-97.39% **LR (Statistical Features+TF-IDF)**:A-96.01%, P-95.68%, R-43.92%, F1-60.2%, AUC-97.59%
[22]	Logistic Regression, SVM, Random Forest,BERT,Multinomial Naive Bayes	Feature Extraction N-grams,TF-IDF,Word2Vec	BERT: hidden layers = 24, batch size = 16,3 , learning rate = 2e-5, optimiser = AdamBERT.	**LR (TF-IDF)**: A-97.23%, F1-97.24%, AUC-99.19% **MNB (TF-IDF + Bigrams)**: A-90.52%, F1-90.37%, AUC-98.90% **SVM (TF-IDF)**: A-97.29%, F1-97.30%, AUC-99.09% **RF (TF-IDF)**: A-96.71%, F1-96.53%, AUC-99.02% **BERT**: A-97.49%, F1-97.49%, AUC-99.48%
[23]	C4.5, Naïve Bayes, Logistic Regression, and SVM	Feature selection, Text analysis techniques	N.A.	**LR (Twitter features and event specific features)**: A-90% AUC-84.42%

To understand the public's reaction, some researchers look into the sentiment behind the social movement, political, and terrorism-related social media texts.

Deep learning techniques are employed for classification and prediction in extremism research but are mainly used for sentiment analysis. The deep learning techniques for extremism detection are mostly LSTM, GRU, Random Embedding, FastText, and CNN. Some research papers have used LSTM as a deep learning technique for tweet sentiment analysis. Ahmad et al. [12] proposed a terrorism-related content analysis methodology that categorizes tweets into extremist and non-extremist classes. The study uses Twitter posts to create a tweet classification system that uses a deep learning-based sentiment analysis technique called LSTM + CNN to classify tweets as extremist or non-extremist classes. The data was gathered utilizing a Twitter streaming API and other Dark Web forums such as Al-Firdaws, Montada, alokab, and Islamic Network. There are 12,754 tweets labelled "extremist" and 8,432 tweets marked "non-extremist" in the training dataset. It compares word embedding learned with CNN, LSTM, FastText, and GRU to conventional feature sets like n-grams, bag-of-words, TF-IDF, and bag-of-words (BoW) for extremism classification. After experimenting with various parameter values for eight LSTM + CNN models, it was discovered that the performance of the LSTM + CNN models was superior to the other models. It had a 92.66% accuracy rate. The precision score in LSTM + CNN is 90%, the recall score is 88%, and the F1 score is 88%.

Even the BERT deep learning algorithm is used in many extremism research for sentiment analysis. Chiorrini et al. [26] investigate using BERT models for sentiment analysis and recognizing emotions in tweets. They have evaluated the performance of these models on real-world tweet data. This model is created by fine-tuning the BERT model on specific tweet datasets. The sentiment analysis was done by training the model with 1,600,000 tweets, and testing with 430 manually annotated tweets as positive, negative, or neutral. For the emotion analysis, they considered the tweet emotion intensity dataset consisting of 6755 tweets labelled as anger, fear, happiness, or sadness. The models had an accuracy of 92% on sentiment analysis and 90% on emotion recognition, according to the findings of the studies.

Meanwhile, Alatawi et al. [27] experimented with detecting white supremacists using the BERT model. This paper identifies white supremacists by hate speech on Twitter, as it is imperative to interpret the spread of such hateful content to prevent it [28]. They used both Bi-LSTM and BERT models, where the BERT model showed the highest F1 score. They paired a Twitter dataset with a Stormfront dataset compiled from a white nationalist site.

A few recently published studies on extremism research utilizing deep learning models are detecting Islamic radicalism in Arabic tweets using NLP and detecting extreme sentiments on social networks with BERT. Mursi et al. [29] presented research on detecting Islamic extremism in Arabic tweets using machine learning algorithms and conducted sentiment analysis on the dataset. For this experiment, they curated their dataset, which was manually labelled as extremism and non-extremism by cybersecurity specialists. Two machine learning algorithms, Super Vector Machine and Multi-Layer Perceptron, are trained using the curated dataset that was converted into a matrix of token count through CountVectorizer and TFIDF. Both models have achieved high accuracy; however, the super vector machine achieved a greater accuracy of 92% than the multi-layer perceptron.

Jamil et al. [30] proposed the detection of extreme sentiments on social media posts with the help of a semi-supervised algorithm known as BERT to reevaluate the accuracy of their prior research. The extreme sentiment is a kind of sentiment analysis, which identifies any negative or positive opinion, evaluation or judgment relevant to a particular thing or person. The former research used an unsupervised approach known as ExtremeSentiLex that automatically detects extreme sentiments on social media posts. Based on their previous work, this research is extended by taking the classified social media posts and validating it using the BERT model. In their experiment, they implemented this methodology on five sets of the dataset; one of them is TurntoIslam dataset relevant to extremism. In the TurntoIslam dataset, the texts were classified as positive extreme, negative extreme, positive non-extreme, negative non-extreme and inconclusive, with negative extreme and inconclusive labels having the highest precision, recall, and F1-score being above 85%.

In deep learning, the advantage is that it is not required to perform feature extraction because most deep learning libraries have an in-built embedding layer that performs feature extraction. Therefore, there is always some advantage with deep learning models when working with large datasets.

C. RELATED DATASETS

The research on detecting and classifying extremist affiliations on social media was conducted using a custom dataset compiled from Twitter and Dark Web forums such as Al-Firdaws, Montada, alokab, and Islamic Network [12]. One of the limitations mentioned in the research paper is that the dataset lacked visual and social context features. In addition, the dataset was imbalanced, as the number of extremist labels was higher than non-extremist labels. Five distinct datasets were used in another research paper about detecting radical content on Twitter [31]. A combination of standard and custom datasets was used to make them. One of the limitations of these different datasets was that they were collected in different periods. Also, the number of tweets collected for each type of dataset varies from each other. This research paper mentioned a limitation that they should take more extensive samples of data for better prediction and also extend the collection of data in other languages, especially the Pashto language. The research on disruptive event detection collected their data from Twitter and gnip using hashtags such as #Ramadi, #Aleppo, #Cairo, and #Dubai for one of the datasets, and the England riot dataset was purchased online [32]. The limitation of this dataset was that the data was imbalanced, as there were more event tweets than non-event tweets. The dataset was only in English.

Meanwhile, in the research on predicting public protests, they collected tweets about the protests against the Trump presidency after the announcement of the presidential election results in November 2016 [23]. Some of the limitations seen in the dataset are that more event-specific features could have been used to bring better performance scores. They could have collected the dataset before the presidential election was almost over to check the likelihood of public protest.

The research paper on violent extremist detection in social media used a custom dataset collected from Twitter [11]. The limitation of their dataset is that they only managed it in a particular year, i.e. 2019, so they did not use a large sample of data. Also, the dataset only focuses on tweets in English. The second limitation in their dataset is that they have not used more user-specific features. If we see the research paper on radicalization detection based on emotion signals and semantic similarity, they have only collected their entire dataset from magazines. Their dataset lacks radical accounts as they are banned. Also, the dataset does not consist of other languages, so that is another limitation. The research paper on detecting violent radical accounts on Twitter could not find an Arabic ISIS-related dataset because of the lack of proper Arabic resources [20]. Therefore, they had to use two different datasets. The first dataset is a collected dataset that was extracted from Pro-ISIS, Anti-ISIS, and non-ISIS Arabic-speaking Twitter accounts. The first dataset's limitations are that the labels Pro-ISIS, Anti-ISIS, and non-ISIS were not balanced. The second dataset is a translated dataset collected from published non-Arabic ISIS-related datasets found in online data science communities. The limitations in the second dataset were that the sample size was not equivalent and the time in which these datasets were extracted is not uniform. Lastly, the paper that researched linguistic cues for analyzing social movements collected their data in two places: one using hashtags like #blacklivesmatter from June 2014 to June 2015 on Twitter and the other dataset from news articles using the same hashtag [33]. The dataset could have been made using even more different hashtags, which is one of the limitations in the dataset. Plus, the dataset could have used more text-specific or news-related features.

Researchers were able to collect data from various sources, create their datasets, use standard datasets found online, or use both types, as seen in Table 3 under dataset type. One of the prominent sources where researchers could gather large amounts of data is social media platforms such as Twitter and Facebook. Also, through available online repositories, for example, Kaggle and UCI. However, there are limitations in these datasets used in their research. Some research uses different sets of datasets, while others only use a particular data collection. These different sets of datasets were combined from standard or custom datasets, which is better for predicting extremism. But the studies in which only a standard dataset was used did not give more accurate results.

TABLE 2. Deep learning classifiers.

Source	DL classifier	Techniques used	Hyperparameters	Evaluation metrics
[12]	LSTM with CNN	Word embedding Feature extraction Dropout layer applied to avoid overfitting Bilingual sentiment lexicon	LSTM with CNN: Dropout layer rate = 0.5, Pooling size = 2 x 2, Units = 100, Padding = same, Kernel size = 2 x 2, Number of filters are 2,4,6,8,10,12,14,16, Vocabulary size = 2000, Size of input vector = 50, Embedding dimension = 128, Batch size = 8, Number of epochs= 7.	**LSTM WITH CNN**: A-92.66%, P-88.32%, R-89.47%, F1-90.71%
[26]	BERT	Word embedding Sentiment analysis Emotion recognition	BERT for emotion recognition: learning rate = 2e-5, train batch size = 8, eval batch size = 8, max seq. length = 95, adam epsilon = 1e-8. BERT for sentiment analysis: learning rate = 1e-5, train batch size = 8, eval batch size = 8, max seq length = 82, adam epsilon = 1e-7.	**BERT (sentiment analysis)**: A-92% **BERT (emotion recognition)**: A-90% **UNCASED BERT (emotion recognition)**: P-93%, R-96% **CASED BERT (emotion recognition)**: P-91%, R-96% **UNCASED BERT (sentiment analysis)**: P-96%, R-91% **CASED BERT (sentiment analysis)**: P-96%, R-91%
[27]	Bi-LSTM and BERT	Word embedding Local Interpretable Model-agnostic Explanations	BERT: learning rate = 2e-5, num train epochs = 3.0, size of batch = 16,8 for training and testing.	**Bi-LSTM**: A-80.25%, F1-79.25% **BERT large**: A-81.573%, P-77.73%, F1-79.605%
[29]	Super Vector Machine and Multi-Layer Perceptron	CountVectorizer, and Term Frequency Inverse Document Frequency	SVM and MLP: Training size = 3000, Folds = 10.	**SVM**: A-92%, P-89%, R-95%, F1-92% **MLP**: A-91%, P-90%, R-90%, F1-91%
[30]	pretrained BERT-base	ExtremeSentiLex SentiWordNet 3.0 SenticNet50	BERT base: layers = 12, hidden units = 768, self-attention heads = 12, input of sequence tokens = 512.	**BERT**: A-82%, P-89%, R-98%, F1-88%

Some studies have shown that they have collected data from an extensive range of periods, which in some cases is reasonable. For this research, it was difficult to retrieve a dataset during the U.S. presidential election in 2020 and the U.S. Capitol riot in 2020. The available dataset had too many or fewer attributes necessary to implement the

TABLE 3. Datasets of relevant extremism research papers.

Source	Dataset Type	Source	Ideology	Keywords and Hashtags used	Language	Data collection period	Labels and Size/Percentage in Dataset
[11]	Custom dataset ISIS-related Twitter handles	Twitter	ISIS	#AbuBakralBaghdadi, #ISIL, #ISIS, #Daesh, and #IslamicState	English	2019	suspended Twitter handles (positive extremism label): 150 normal user handles (negative extremism label): 150
[12]	Custom dataset ISIS-related tweets scraped	Twitter and Dark Web Forums , which are Al-Firdaws, Montada, alokab, and Islamic Network	ISIS	ISIS, bomb, suicide	English and Arabic	2016-2017, 2019	Extremist: 12,754 tweets Non-extremist: 8,432 tweets
[20]	Custom dataset Arabic twitter accounts that are Pro-ISIS, Anti-ISIS and non-ISIS plus, non-Arabic ISIS datasets	Twitter and Kaggle	ISIS	Pro-ISIS: وأعدوا, الدولة_الإسلامية, باقية, تتمدد Anti-ISIS: جرائم داعش, داعش تجار الدم, مسلمون ضد داعش, فضائح داعش, بلغ,عش ع Non-ISIS: News, sports, religion, art	English and Arabic	2015, 2020	Pro-ISIS : 21,000 tweets Anti-ISIS: 21,000 tweets Non-ISIS: 21,000 tweets
[23]	Custom dataset Tweets related to U.S. presidential election in 2016	Twitter	Presidential elections, political riots	#NotMyPresident a, #muslimban and #travelban	English	9th November to 15, 2016 and 27th January to 31, 2017	1 - protest in state 0 - no protest in state
[31]	Standard and Custom dataset Radical corpus, Neutral corpus,	Twitter , Kaggle , ISIS English Magazines	ISIS, Jihadist, Islamist	Pro-ISIS: Amaq, Dawla, and Wilyat Anti-ISIS: isil, isis, Islamicstate, Mosul, raqqa, islamic state. Religious texts: selecting based on type	English and Arabic	Pro-ISIS: period of 3 months Anti-ISIS: collected on 7-11-2016 and 7-4-2016	Pro-ISIS users: 17,350 tweets collected from 112 pro-ISIS accounts Anti-ISIS users: 122,000 tweets from 95,725 anti-

TABLE 3. ***(Continued.)*** **Datasets of relevant extremism research papers.**

	Religious corpus, newly collected dataset, and random dataset			and sources. The texts are filtered based on Jihadist and Islamist type New dataset: Neutral users:- #Al-baghdadi, #baghdadikilled #abubakaralbaghdadi #ISIS Random dataset: hashtags related to current events, sports, and more topics		Religious texts: 15 Dabiq issues (2014-2016) and 9 Rumiyah issues (2016-2017) New dataset: Pro-ISIS users collected in 2019 from September to November Neutral users collected in 2019 from 28th October to 30th October Random dataset: Collected in 2019 from 15th October to 20th October	ISIS accounts New Pro-ISIS users: 9,000 tweets from 13 suspended accounts Random users: 7,000 random tweets
[32]	Standard and Custom dataset Middle East Disruptive Events dataset and England riots 2011 dataset	Twitter and Gnip	Political riots and protests	Keywords such as Iraq, Syria, Egypt and hashtags used are #Ramadi, #Aleppo, #Cairo, #Dubai	English	Middle East dataset collected in 2015 from 1st October to 30th November England riot dataset purchased from 6th August, 2011 to 12th August, 2011	3,100 event tweets 1,900 non-event tweets

TABLE 3. ***(Continued.)*** **Datasets of relevant extremism research papers.**

[33]	Standard dataset Black Lives matter tweet dataset and news articles	Twitter, Center for Media and Social Impact, Lexis-Nexis Academic	Social movements	#blacklivesmatter, #ferguson, #blackman, #gunman, #shooting, #blackpeople, and #blackperson	English	1st June, 2014, to 1st June, 2015.	Positive, neutral and negative

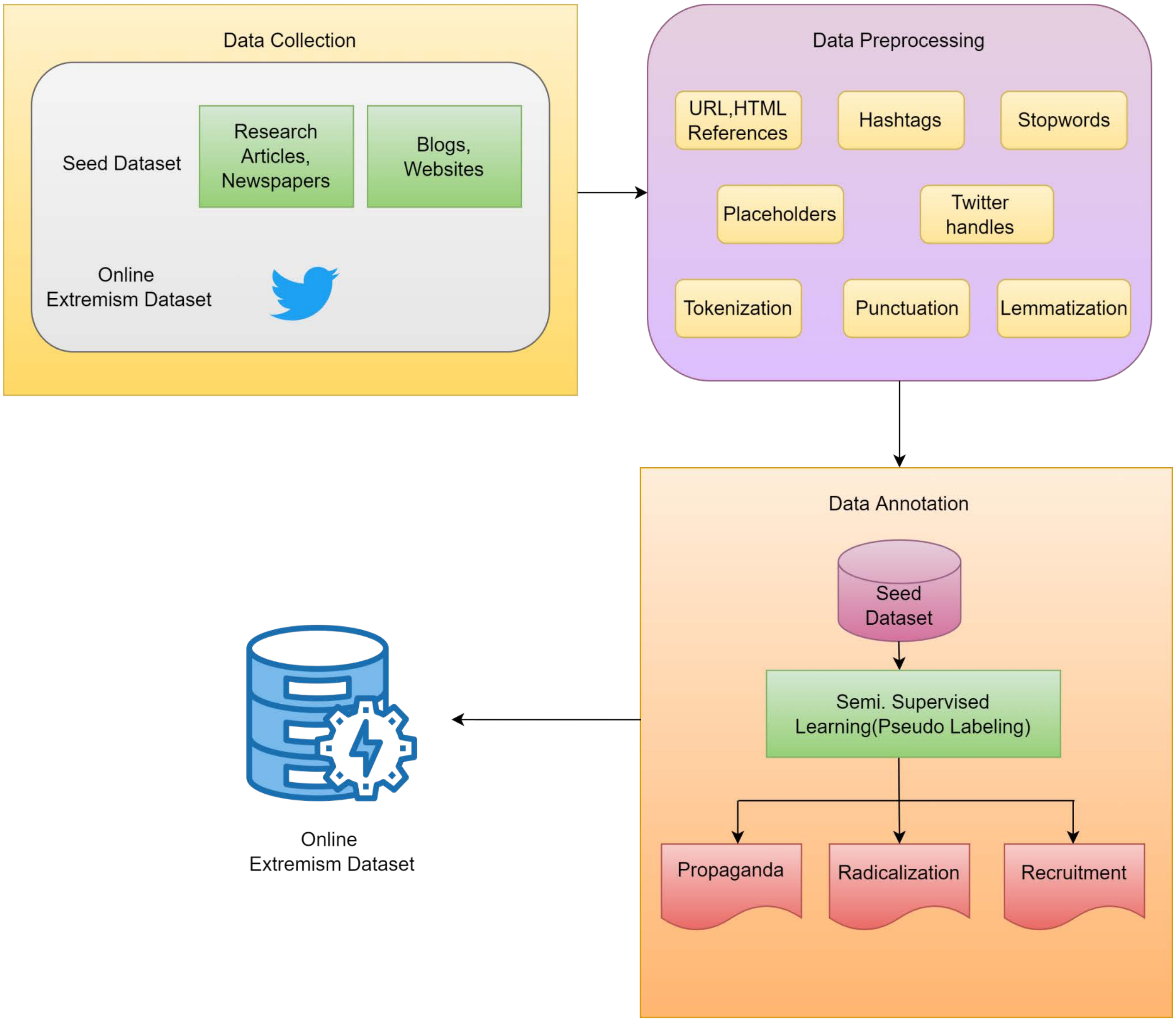

FIGURE 1. **Creation of Extremism Dataset.**

models. Furthermore, the classification of the dataset into categories such as propaganda, recruitment, and radicalization was required, which was not present in the available datasets. The existing datasets found online were tweets based on the U.S. presidential elections, and there were not many datasets containing tweets about the U.S. Capitol riot. This research

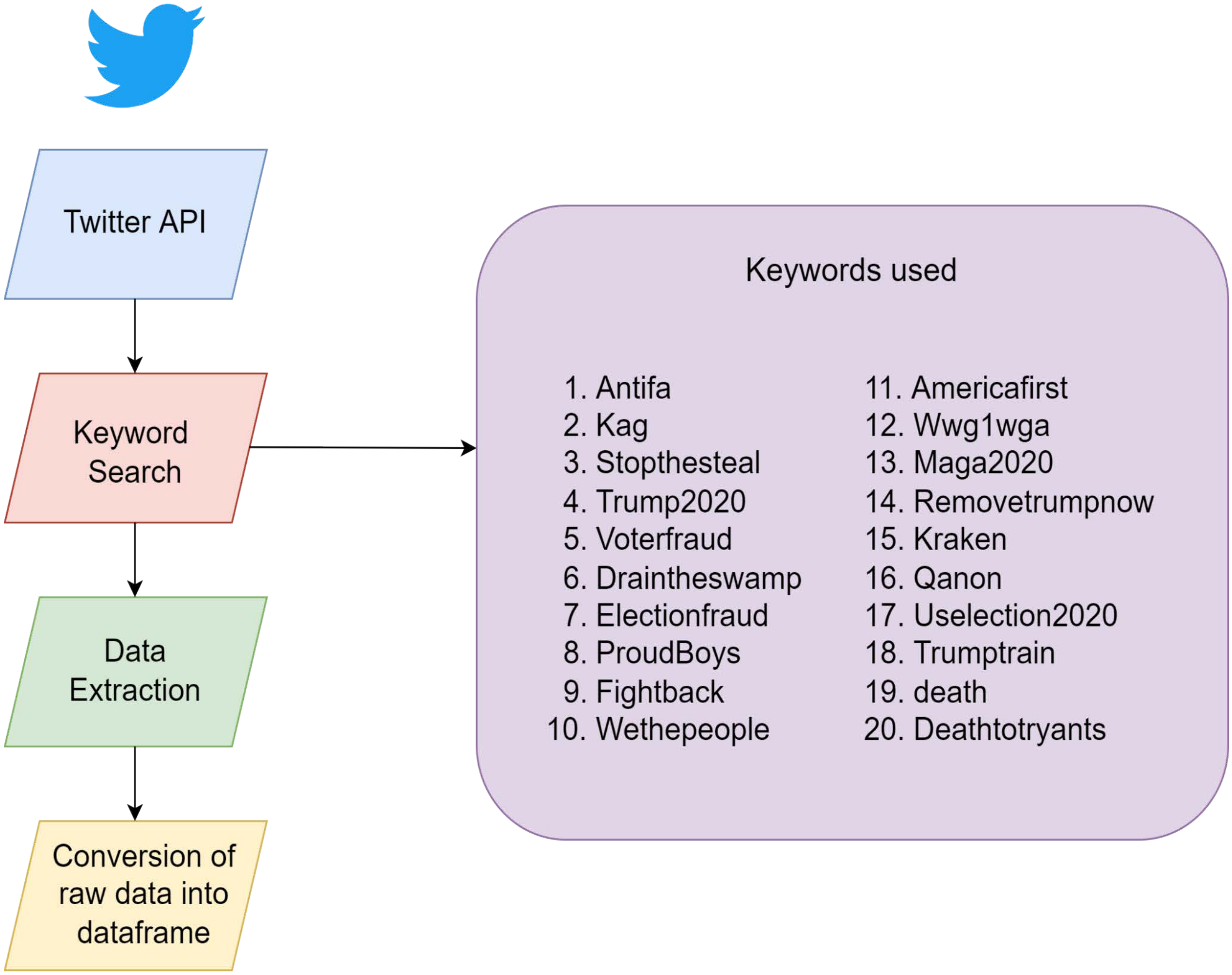

FIGURE 2. Data extraction steps.

required a dataset containing tweets about the U.S. Capitol riot. The customized dataset contains tweets from Twitter after a survey of popular keywords used during the U.S. Capitol riot. The dataset was manually labelled to classify different types of extremism in the text. Each label of the tweets is divided into 28.6%, 32%, 37.2%, and 2.3% recruitment, radicalism, propaganda, and non-extremism tweets. The customized dataset is larger in sample size than existing datasets, and the tweets are collected between the U.S. presidential election in 2020 and the U.S. Capitol riot in 2021. Existing datasets could not provide labels of extremism. Hence, a customized dataset is curated for this research work.

III. PROPOSED ARCHITECTURE

This section discusses the architecture followed during the study, which involves data collection, labelling, and implementation of deep learning models such as Bi-LSTM, BERT, RoBERTa, and Distill-BERT.

A. DATA COLLECTION

The collected data is from the Twitter platform, where many trending hashtags were related to the U.S. Capitol riot. The data collection is performed using the Twitterscraper library. The dataset for this study collected tweets from 25th December 2020 to 10th January 2021, including only the tweets posted in English. The dataframe prepared was used for further cleaning and preprocessing. Figure 2 shows the steps followed during the extraction of tweets.

B. U.S. CAPITOL RIOT KEYWORDS AND COMBINATIONS FOR TWEET COLLECTION

The hashtags are collected from various online sources (news articles, research papers). The listed keywords were the most used on social platforms and were part of many discussions significant to the U.S. Capitol riot. Hence, we have gathered these keywords to extract tweets that can recognize the extremism in the posts shared on Twitter. The dataset contains posts including these keywords. These keywords will help us

TABLE 4. Number of tweets per keyword.

Keyword	Count of tweets	Keyword	Count of tweets
#Antifa	8,442	#Americafirst	5,096
#Kag	8,326	#Wwg1wga	4,912
#Stopthesteal	7,723	#Maga2020	4,724
#Trump2020	7,683	#Removetrumpnow	4,712
#Voterfraud	7,127	#Kraken	2,937
#Draintheswamp	5,867	#Qanon	2,908
#Electionfraud	5,518	#Uselection2020	883
#ProudBoys	5,509	#Trumptrain	500
#Fightback	5,366	#Death	5
#Wethepeople	5,261	#Deathtotyrants	2

determine the sentiment in the tweets. Table 4 contains the list of all collected keywords used for extracting tweets to create the dataset and the count of tweets collected for each keyword used during data collection. Thus a total of 93,501 tweets were collected from 25th December 2020 to 10th January 2021.

The metadata obtained from Tweets using the Twitter API is listed below.

1. **Datetime:** Time and date at which tweet was created or posted.
2. **Tweet Id:**It is the unique identifier of a tweet.
3. **Text:**The tweet posted by the user (UTF-8 format).
4. **Username:** It contains the name of the user who posted the tweet.

C. SAMPLE TWEETS

Table 5 shows the final dataset prepared for training after the tweet collection. The data contains four columns, which include the date of the tweet posted, the unique id of the tweet posted, the tweet posted by the user, and the username. The text column is of utmost importance for this study as the extremism analysis, labelling, modelling, and testing are performed only on the text data. The text data is further cleaned and preprocessed to make it suitable for modelling and testing.

D. SEED DATA COLLECTION

Seed data is collected based on political ideologies. The seed dataset's primary purpose is to contain text on propaganda, radicalization, and recruitment. For data collection, we collected various research articles, newspapers, websites identifying extremists, and blogs identifying influential propagandists, radicals, and recruiters.

1) RESEARCH ARTICLES AND NEWSPAPERS

The seed data was made from the research articles, and newspapers expressly identified the extremist text as propaganda, radicalization, or recruitment. The research articles and newspapers contained relevant tweets for this experiment thus, we collected the tweets from the research papers and news websites or downloaded the excel files from the websites. The collection of text was confined from January 2016 to December 2021. A total of 30 articles were gathered for this study.

2) WEBSITES AND BLOGS

The majority of seed samples chosen are from blogs and websites. Users were labeled as propagandists and recruiters on some websites. Such users' tweets or posts are regarded as propaganda or recruitment. For this experiment, 90 web blogs and websites were reviewed, of which 45 were deemed suitable for the study.

In some sources, only a few of the tweets were used, while in other sources, all the available tweets were utilized. Table 6 contains examples of tweets and its corresponding source.

E. SEED DATA FEATURES

The characteristics of seed data include Source Type, Text, and Label.

1. ***Source Type*** - Indicates whether the sample collected is from a research article, website, or newspaper article.

TABLE 5. Tweets scraped from Twitter and labelled.

Sr. No	Datetime	Tweet Id	Text	Label
1.	31-12-2020 03:23:00	1344401177417570 000	The Battle for the White House: Join us at 5 &amp; 7 PM/ET as Sen. Josh Hawley vows to stand up for @realDonaldTrump and object to fraudulent electoral votes on 6th January. @TomFitton @RepVernonJones @jaeson_jones @robertjeffress #MAGA #AmericaFirst #Dobbs https://t.co/aXkQsIhpaE	Recruitment
2.	30-12-2020 20:18:00	1344294193838120 000	Stand for freedom. Stand for America. And stand with President Trump in Washington DC on 6th January. #StandWithTrump	Radicalism
3.	2021-01-05 23:56:00+00:00	1346606404967210 000	Who the hell does @realDonaldTrump think he is? Threatening people who won't help him steal the election!You are a pathetic loser! #RemoveTrumpNow "Trump warns â€˜ineffective RINOâ€™ lawmakers his voters will revolt if they donâ€™t help him steal the election": https://t.co/x0VboBsJTT	Propaganda

2. ***Text*** - Contains extremist text or tweet provided by the source.
3. ***Label*** - As per the selected article, the label indicates the class to which the text belongs, such as propaganda, radicalization, or recruitment.

F. SEED DATA ANNOTATION

Each tweet in the seed dataset is classified as radicalism, propaganda, and recruitment. No manual annotator was used. Instead, those tweets were put in categories based on the content on the websites, research papers, news articles, and blogs mentioning radicalism, propaganda or recruitment.

IV. DATA PREPROCESSING

The collected Twitter data is studied through exploratory data analysis, and the text data is processed to continue with the models' development. Data preprocessing helps in the transformation of collected data into a proper format. The Preprocessing stage involves cleaning and removal of unwanted data as well as formatting the raw data into an understandable structure for machines to interpret text [42]. For preprocessing the dataset, the tweets were first converted into lowercase, and then noisy data was removed from the tweets to make the data suitable for further analysis. Eliminating noisy data involved removing URL links, placeholders, HTML references, non-letter characters (punctuation and special characters), Twitter handles and hashtags from text data.

The text data was further processed by removing stopwords, and tokenization was performed to divide the text into meaningful tokens [43]. The tokenized data was further lemmatized to get the base form of the word [43], which helps in performing sentiment analysis on the text data.

A. STOPWORDS

These are unwanted and meaningless words that are removed from the text to reduce noise, as their removal does not impact the performance of the models.

B. TOKENIZATION

This reduces the sentence into tokens by splitting text into words, which helps analyze the word's meaning.

C. LEMMATIZATION

This is used to normalize the words into their root form (dictionary-based approach) having the same meaning.

V. DATA LABELING

The data extracted from Twitter is unlabeled. To annotate the curated Twitter dataset, Pseudo-labeling is implemented. Pseudo-labeling is the technique of predicting labels for unlabeled data using a labelled data model. A seed dataset was created, consisting of 1000 samples gathered from multiple sources with careful annotations of labels such as propaganda, recruitment, and radicalization.

The labelled data is used to train an SVM model, which is used to forecast unlabeled tweets and to check the confidence level. Instead of using labels to help identify the confident guesses, we used the predicted probability, which signifies the class probability. The confidence level starts when the predicted probability is higher than 0.35 to 0.86. Then we add

TABLE 6. Sources of seed dataset.

Text example	Source
RT @realDonaldTrump: WE WILL WIN!	[34]
I can't stand back & watch this happen to a great American City, Minneapolis. A total lack of leadership. Either the very weak Radical Left Mayor, Jacob Frey, get his act together and bring the city under control, or I will send in the National Guard & get the job done right....	[35]
Watch: Hundreds of Activists Gather for #Stop the Steal: Rally in Georgia https://t.co/vUG1bqG9yg via BreitbartNews Big Rallies all over the Country. The proof pouring in is undeniable. Many more votes than needed. This was a LANDSLIDE!	[36]
"Once the #MuellerInvestigation is over, if there's nothing on @POTUS @realDonaldTrump then what are they going to do next? #DeepState coup? #FalseFlag operation? They want Trump impeached or assassinated. #FakeNews is stoking this narrative. @RealAlexJones @DRUDGE @seanhannity" @Jesus_Mohammad 2/7/18	[37]
Data... truly... doesn't... lie. Enjoy the show!	[38]
The DNC Data Breach Download Speed Points to an Internal Leak, Not an External Hack https://t.co/BwRVtcTajy #Qanon #WeAreTheNewsNow #FactsMatter #WWG1WGA #WakeUpAmerica #UnitedNotDivided #SaveAmerica #GreatAwakening	
These thugs are dishorning the memory of George Floyd, and I won't let that happen. Just spoke to Governer Tim Walz and told him that the Military is with him all the way. Any difficulty and we will assume control but, when the looting starts, the shooting starts. Thank you!	[39]
Peter Navarro releases 36-page report alleging electon fraud 'more than sufficient' to swing victory to Trump. A great report by Peter. Statistically impossible to have lost the 2020 Election. Big protest in D.C. on 6th January. Be there, will be wild!	[40]
#riotcleanup at Camden 11 am, Chalk farm 10 am, Roman Rd Hackney 9 am, Clapham 9 am, Peckham 10 am, Westbourne Grove 9 am	[41]
#NotMyPresident Anti-Trump rally planned for Downtown. Indianapolis on Saturday	[23]

these predictions to the labelled data and retrain the model using both labelled and unlabeled data. Basically, the labels are predicted in the actual dataset and then retrained, the model with seed and pseudo label dataset.

Later it is observed that if there was any improvement with a different threshold value with an accuracy of 94%. The final observation helped predict the final labels, i.e. propaganda, recruitment, and radicalization.

For the SVM training, the dataset was split into two subsets: the train set, which counts for 90% of the existing dataset, and the test set, which is 10% of the existing dataset. Next, the test set was labelled in order to check the confidence level. In the 90% train set, the subset was split into two sets, which are 20% labelled data and 80% unlabeled data. The first model was built with the help of the labelled train set and then classified into the unlabelled training datasets.

After this, measured confidence levels were according to the test sets that were already labelled. Measuring the confidence levels according to the test sets that were already labelled before was essential. Once this was done, we concatenated the labelled train data with predicted unlabeled data. As a result, when the predicted probability is higher than 0.35 till 0.86, this new data is called pseudo-labelled, which is similar to the actual label. This data is again retrained again from new train datasets.

The selected pseudo-labelled ones from the unlabeled datasets, and then we retrained the model to predict the remaining unlabelled data. At this stage, we had to iterate the same step of combining labelled train data with the prediction of unlabeled data until there is no probability of predicted pseudo-labelled higher than 0.35 to 0.86. The training model was evaluated using automatic metrics such as accuracy,

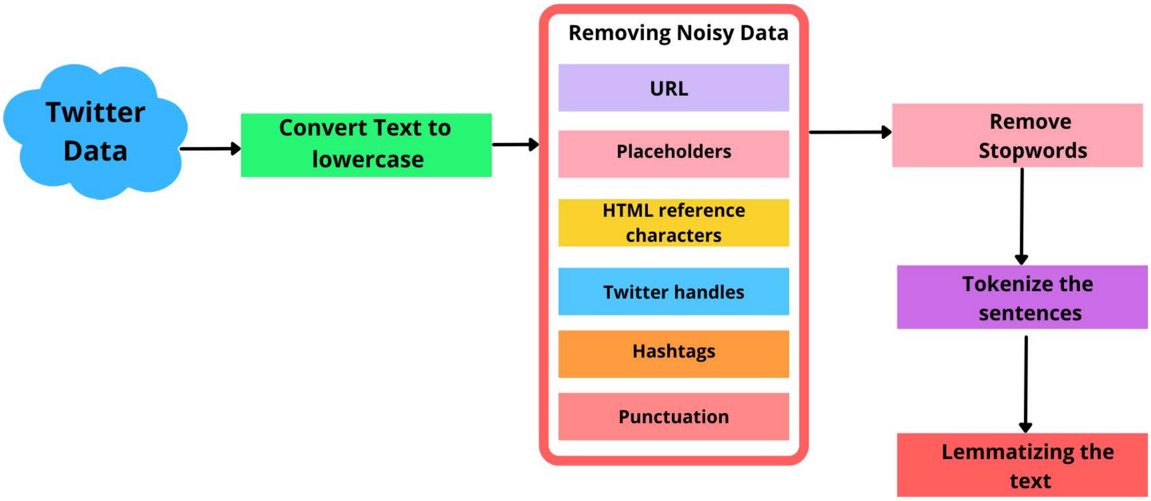

FIGURE 3. Data preprocessing steps.

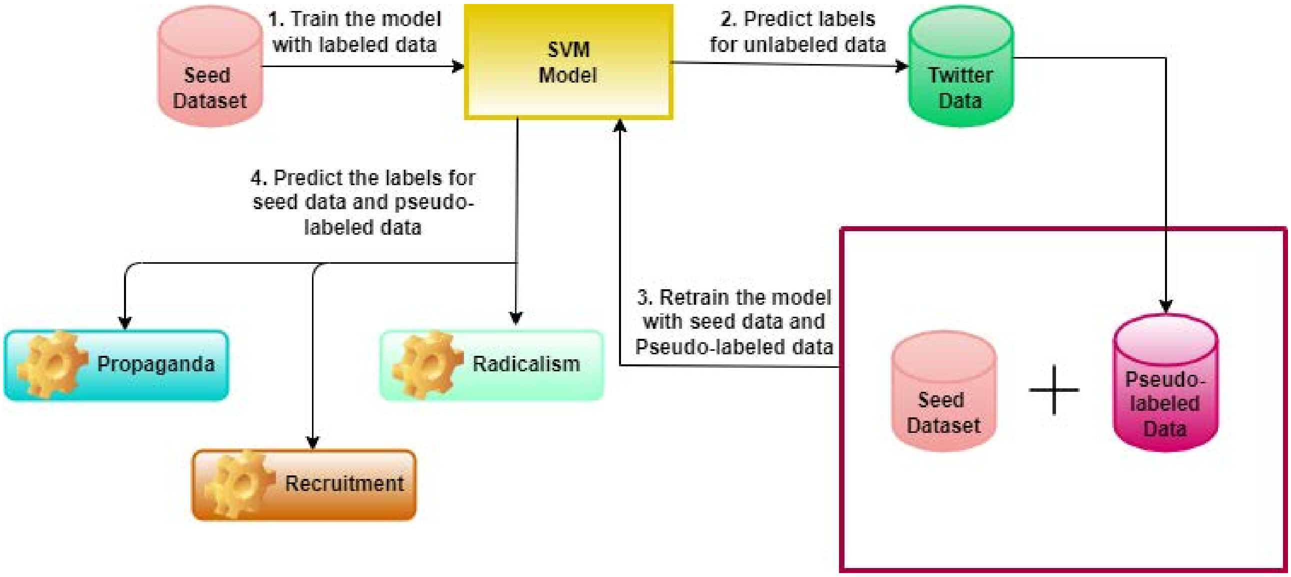
Seed Dataset
1. Train the model with labeled data
SVM Model
2. Predict labels for unlabeled data
Twitter Data
4. Predict the labels for seed data and pseudo-labeled data
3. Retrain the model with seed data and Pseudo-labeled data
Propaganda
Radicalism
Recruitment
Seed Dataset
+
Pseudo-labeled Data

FIGURE 4. Pseudo labeling architecture.

precision, recall, and F1-score to assess the performance of the pseudo-labelling model.

Following are the features of the seed dataset

1. ***Text*:** Contains an extremist tweet posted by the user.
2. ***Label*:** It identifies different categories of extremism in the tweet as propaganda, radicalization or recruitment.

VI. ANALYSIS OF TRENDING TWEETS THROUGH DATA VISUALIZATION

The entire dataset was analyzed using visualization techniques to understand the trends followed during the U.S. Capitol riot in 2021 and the U.S. presidential election in 2020. This analysis helped identify the most used hashtags, understand the public reaction, and check the distribution of labelled tweets after text classification.

A. WORD CLOUD

One of the most popular techniques to find the top keywords in the dataset is word cloud which indicates the frequency of the word according to their size. Figure 5 highlights the hashtags most used on Twitter during the U.S. Capitol riot.

B. TOP HASHTAGS USED DURING U.S. CAPITOL RIOT (BEFORE 6TH JANUARY)

The bar chart in Figure 6 presents the count of trending keywords according to the word cloud. The most used hashtags

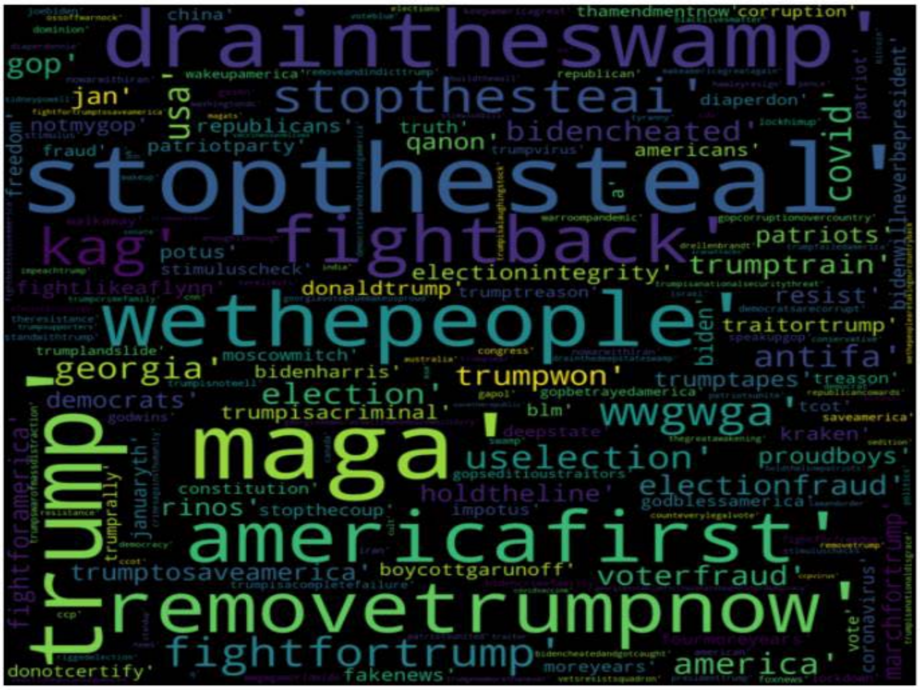

FIGURE 5. Word cloud of hashtags collected from tweet.

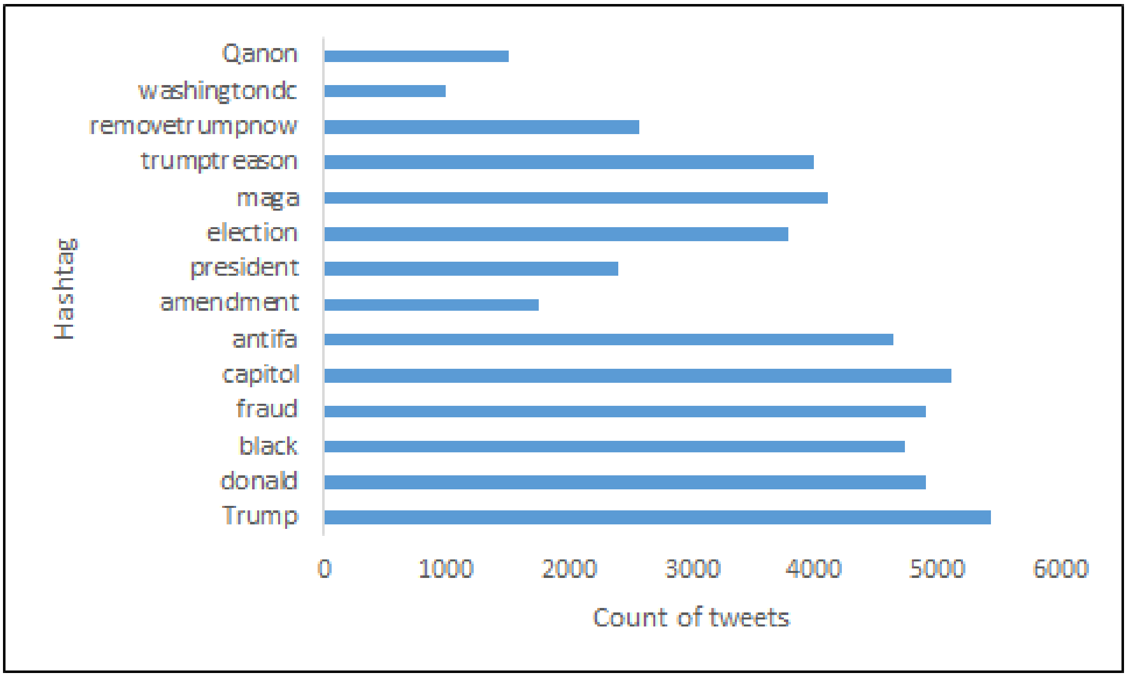

FIGURE 6. Trending hashtags before capitol riot.

before the U.S. Capitol riot in 2021 are #Trump, #capitol, #donald, #fraud, and #antifa.

C. TOP HASHTAGS USED DURING U.S. CAPITOL RIOT (AFTER 6TH JANUARY)

The bar chart in Figure 7 presents the count of trending keywords according to the word cloud. The most used hashtags after the U.S. Capitol riot in 2021 are #Antifa, #Kag, #Stopthesteal, #Trump2020, and #Voterfraud.

From Figure 6 and Figure 7, it is clear that there was a change in the trends before and after the Capitol Riot took place. Before 6th January 2021, tweets in support of Trump were trending. However, the trend changed after the riot resulting in trending hashtags gaining momentum.

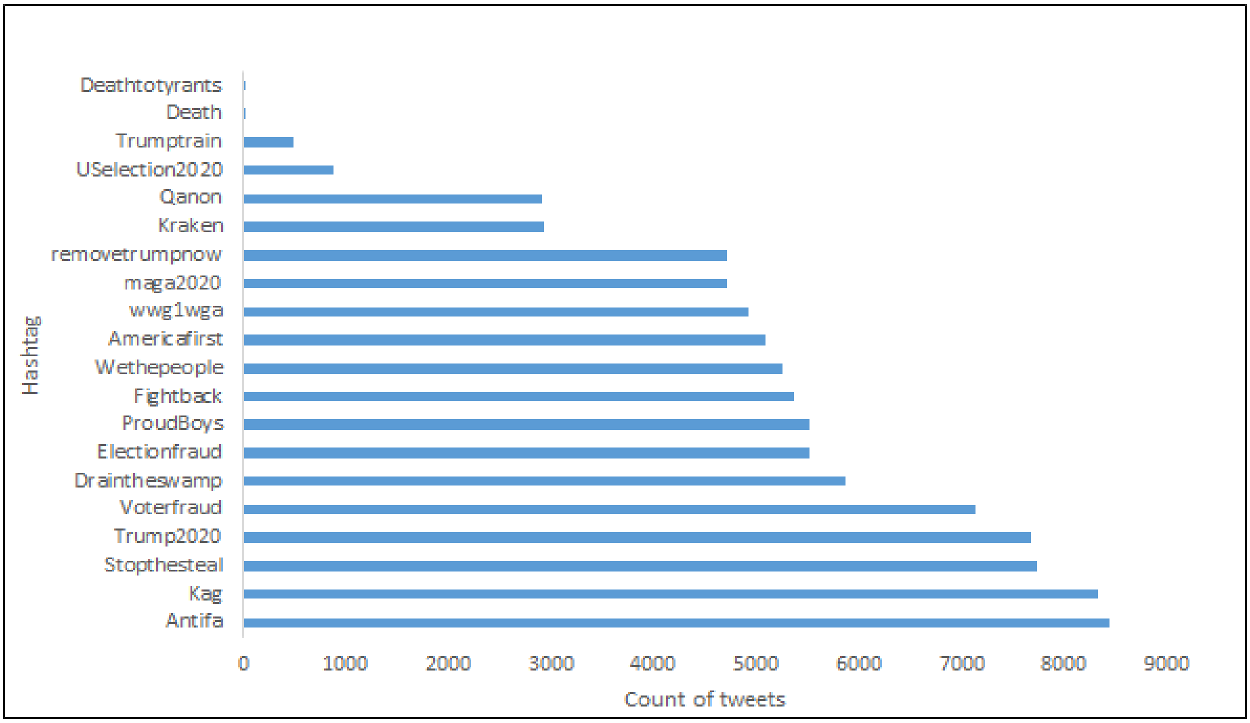

FIGURE 7. Trending hashtags after capitol riot.

D. EXTREMISM CLASSIFICATION OF LABELED DATA

The extracted data was labelled propaganda, recruitment, radicalization, and non-extremism. The pie chart shows the percentage of each type of labelled tweet. Almost 98% of tweets belong to at least one of the categories of Extremism. This information is significant because the labels will play a dominant role in the training of the models.

VII. DATASET EVALUATION

A. EXPERIMENTAL SETUP

For this research study, the following hardware and software requirements are mentioned in Table 7. The programming was done on a computer using the Anaconda Jupyter Notebook. The entire program was written in python language using python libraries that are utilized for deep learning.

B. DEEP LEARNING MODELS

The four deep learning models implemented in this experiment are Bi-Directional Long Short-Term Memory (Bi-LSTM), Bidirectional Encoder Representations from Transformers (BERT), RoBERTa, and DistillBERT.

C. Bi-LSTM

Bi-Directional Long Short-Term Memory is a Recurrent Neural Network and a Sequence Processing Model, which comprises two LSTM layers. The first layer takes input in one direction, and the second layer processes input in the opposite direction. The first recurrent layer is repeated in the network, resulting in two parallel layers. The input sequence is passed to the first layer as an input, and a reversed version of the same input is given to the second layer for processing.

The working of the Bi-LSTM can be understood in six phases as explained below:

D. WORD EMBEDDING

Words with the same meaning have an equal representation in a learned text representation. To extract semantic information from a tweet, it is first represented as a sequence of word embeddings.

E. BI-LSTM LAYER

This layer records the sequences that appear in the data.

F. DENSE LAYER

The output is passed to the dense layer, which has a sigmoid activation function and uses dropout between the two dense layers to avoid overfitting.

G. INPUT LAYER

It brings the initial data into the system to be processed further.

H. HIDDEN LAYER

The inputs entering the network are transformed nonlinearly by the hidden layers with the help of an activation function.

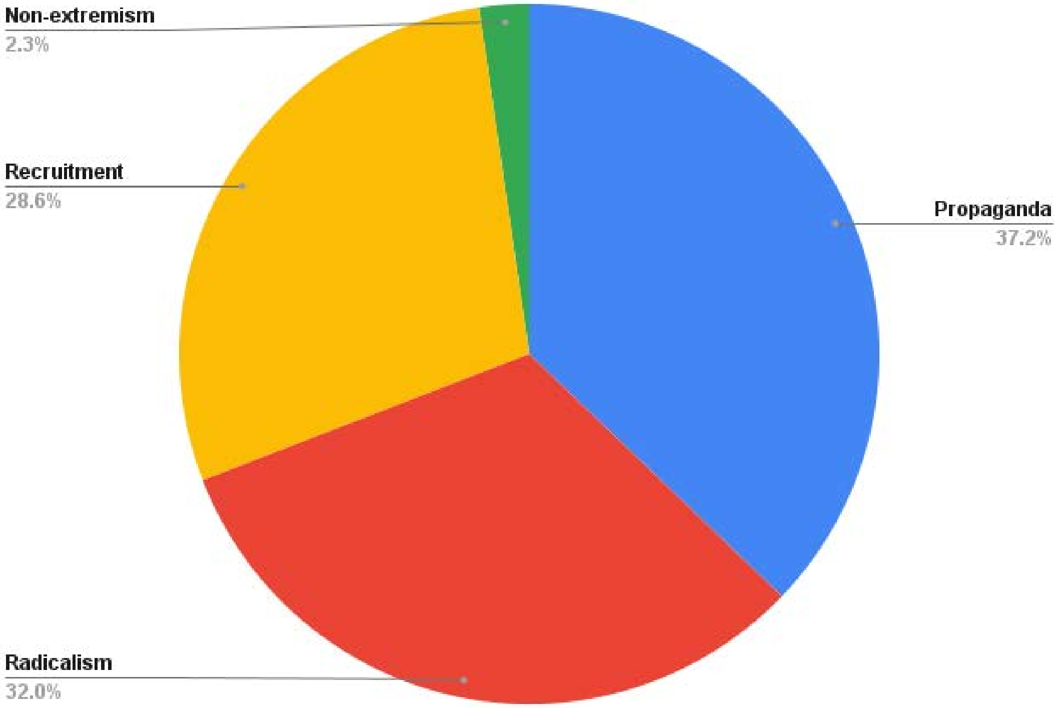

FIGURE 8. Percentage of tweets classified as one of the corresponding labels.

TABLE 7. Environment setup.

Environment	Configuration	
Hardware	CPU	Intel(R) Core(TM) i5-8250 CPU @1.60GHz
	GPU	12GB NVIDIA Tesla K80
	Memory	8 G.B.
	Operating System	Windows 10, 64bit
Software	Programming Environment	Google Colab, Anaconda (Jupyter)
	Python Libraries	PyTorch, Tensorflow, Keras, Scikit-Learn, nltk

I. OUTPUT LAYER

This layer extracts the desired named entities.The Bidirectional feature of the Bi-LSTM model serves as an enhancement to RNN, making it possible for the neural networks to memorize the current and previous information. Therefore, Bi-LSTM has been implemented in this study to understand the neural network's performance in the classification of text data and to note the difference in the performance of Bi-LSTM with other advanced models.

J. BERT

Bidirectional Encoder Representations from Transformers is a transformer model that uses a masked attention mechanism to assign weight to each input and output element. It first chooses a pre-trained BERT model according to the language need and later modifies the architecture as per the need of the task, as shown in Figure 10. Lastly, the training data was prepared after fine-tuning the modified model on the dataset.

The BERT encoder anticipates a token sequence [CLS], a unique token that appears at the beginning of the first sentence. Each sentence has [SEP] at the end of it. To distinguish between the sentences, a segment 'A' or 'B' is added to the embeddings. BERT takes a sequence of inputs and moves them up the stack. Before being sent into a feed-forward neural network, each block passes through a Self-Attention layer. The data is subsequently passed on to the next encoder. Each point generates a concealed size vector (768 in BERT Base).

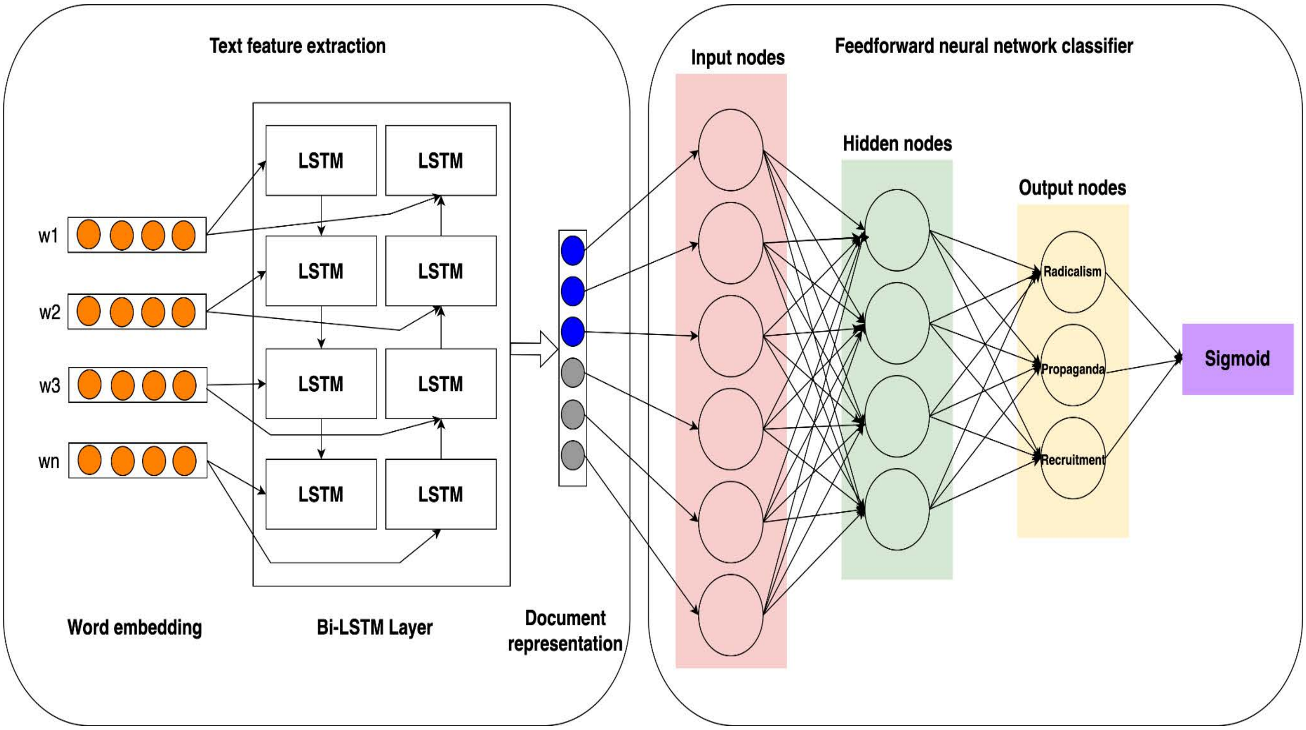

FIGURE 9. BI-LSTM model architecture.

TABLE 8. Dataset evaluation results using BI-LSTM.

Model	Labels	Precision	Recall	F1-Score
BiLSTM	Propaganda	0.84	0.90	0.86
	Radicalization	0.81	0.83	0.84
	Recruitment	0.80	0.84	0.83

The dataset is trained further on the pre-trained model, and the result is fed to a sigmoid layer. Any error gets back-propagated through the entire architecture, and the model's pre-trained weights are adjusted depending on the updated dataset.

This model performs better with massive data with the advantage of a masking feature that helps in better identification of keywords. Therefore, BERT is exclusively used for text classification, whose performance has been tested in this study to achieve better classification results.

K. RoBERTa

The Robustly Optimized BERT Pre-training Approach optimizes BERT architecture, which reduces the time taken during the pretraining of the model. The architecture is very much similar to BERT. Roberta extends BERT's language masking approach by changing key hyperparameters, such as removing BERT's next-sentence pretraining goal and training with much larger mini-batches and learning rates. RoBERTa was trained on data with an order of magnitude more than BERT for an extended period.

This is an advanced version of the BERT model that gives better performance than BERT itself. RoBERTa has been used in this study to achieve good results in extremist text classification.

L. DistilBERT

DistilBERT is a Transformer model trained on a BERT base and is small, quick, cheap, and light. To mimic Google's BERT, it uses a process called distillation, which involves substituting a more extensive neural network with a smaller one. After a vast neural network has been trained, the whole output distributions of the network can be approximated using a smaller network. The data is tokenized using the DistilBERT tokenizer and then turned into a series of tokens, converted to tensors and supplied to the model, as shown in Figure 12. The DistillBERTClass is used to create a neural network. This network will use the DistilBERT Language model to obtain the final outputs, a dropout, and a Linear layer.

The data is incorporated into the DistilBERT Language model. It builds a single, dense output layer with a sigmoid

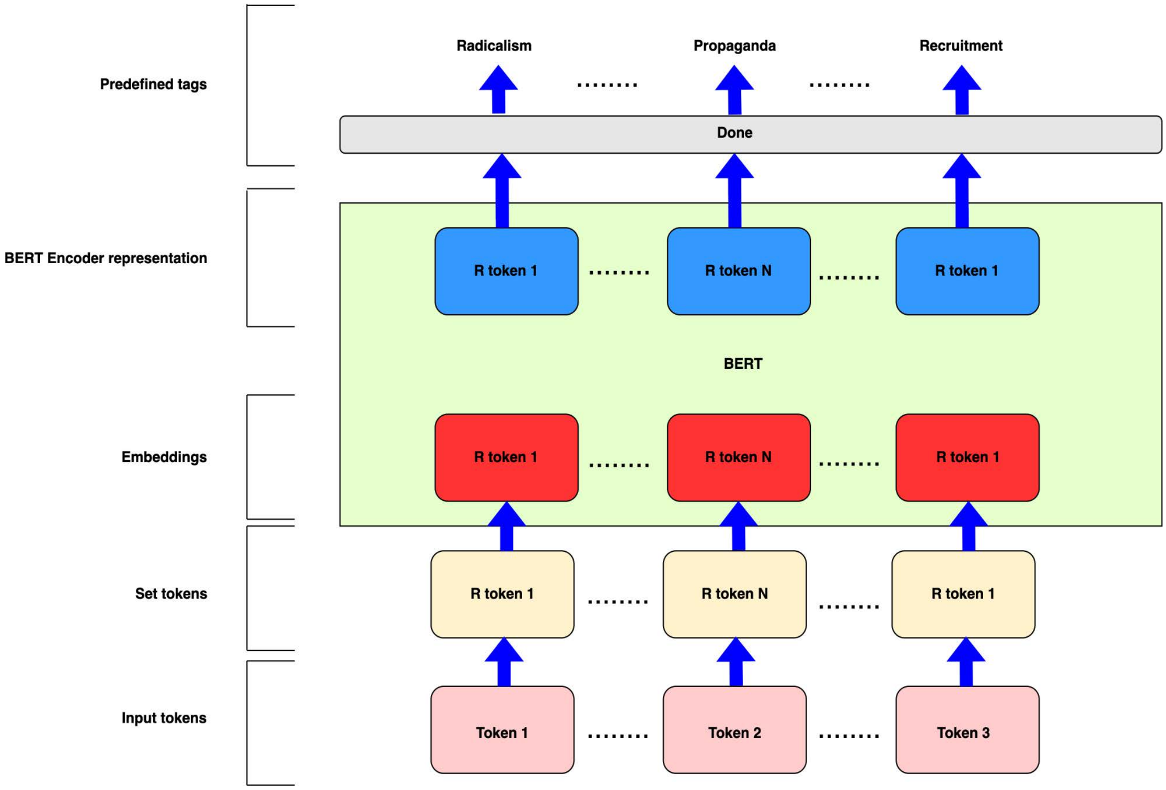

FIGURE 10. BERT model architecture.

TABLE 9. Dataset evaluation results using BERT.

Model	Labels	Precision	Recall	F1-Score
BERT	Propaganda	0.89	0.90	0.85
	Radicalization	0.85	0.87	0.91
	Recruitment	0.92	0.89	0.93

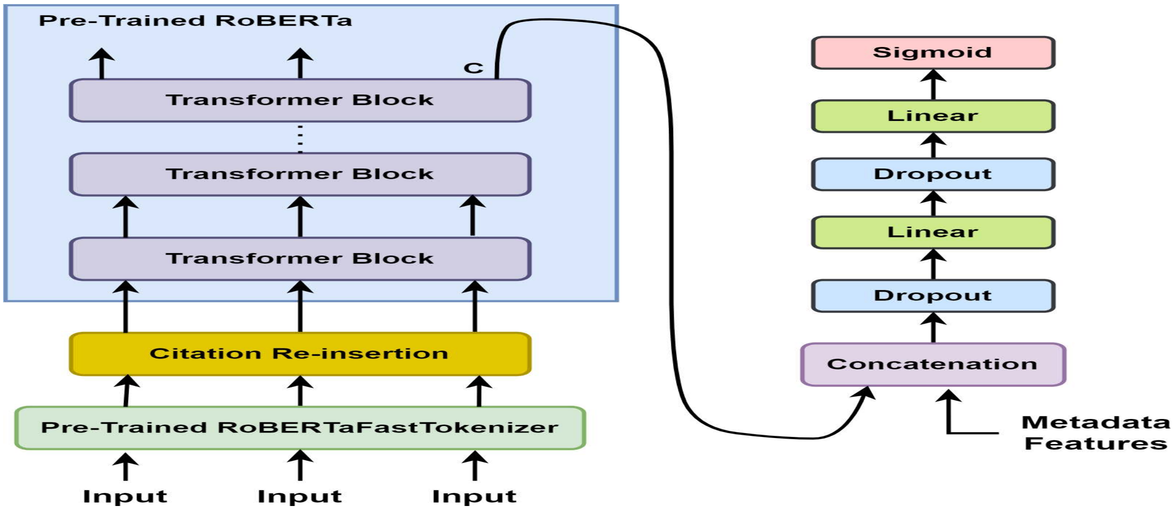

FIGURE 11. RoBERTa model architecture.

activation function on top of the [CLS] token's sentence-level embedding to acquire a baseline for the model's performance. Begin training the classification layer's randomly initialized weights until the model converges in performance. After

TABLE 10. Dataset evaluation results using RoBERTa.

Model	Labels	Precision	Recall	F1-Score
RoBERTa	Propaganda	0.93	0.91	0.93
	Radicalization	0.91	0.85	0.92
	Recruitment	0.89	0.93	0.90

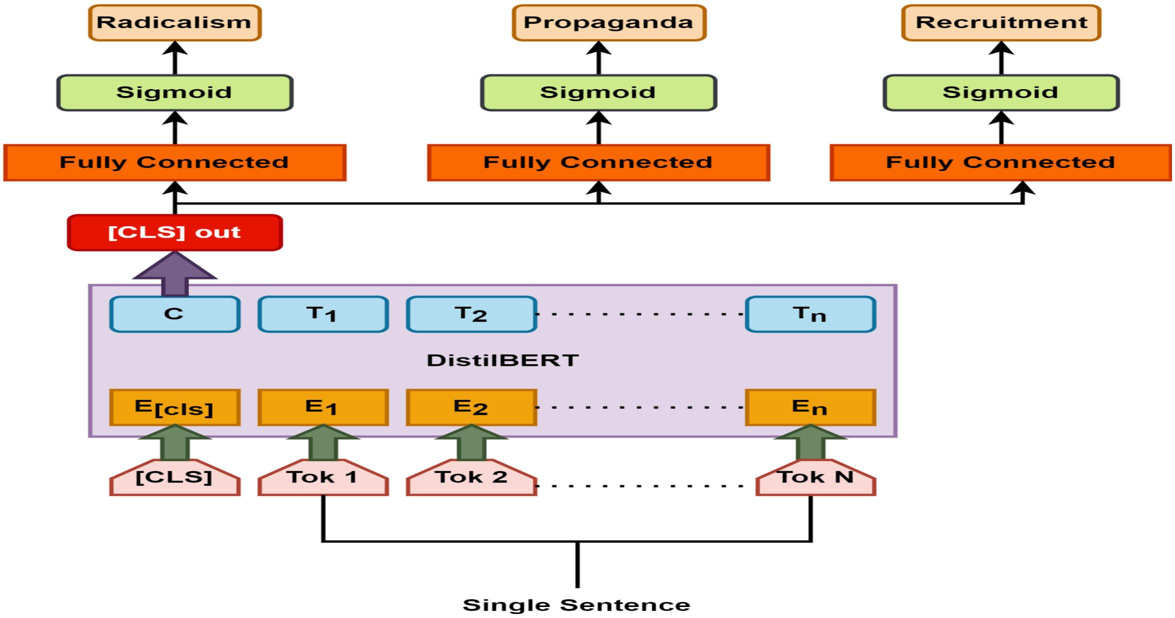

FIGURE 12. DistilBert model architecture.

TABLE 11. Dataset evaluation results using DistilBERT.

Model	Labels	Precision	Recall	F1-Score
DistilBert	Propaganda	0.92	0.89	0.93
	Radicalization	0.88	0.90	0.93
	Recruitment	0.87	0.92	0.90

training the additional classification layers, unfreeze DistillBERT's embedding layer and fine-tune all weights with a reduced learning rate to extract even more performance out of the model.

DistilBert is another version of BERT that offers good performance; therefore, it was selected to analyze its performance with other models.

M. RESULTS AND DISCUSSIONS

The dataset is evaluated with different deep learning classifiers, as shown in Tables 9, 10, and 11. The dataset is divided into training and testing in the ratio of 80:20. For all of the deep learning models examined, Table 11 shows the reported accuracy, precision, recall, and F1-Score.

As evident from Table 11, RoBERTa outperformed the other three models by almost 6% in accuracy, proving its supremacy over BERT and its improved models known as DistilBERT. RoBERTa's precision, recall, and F1-score show an increase of almost 0.06, 0.11, and 0.07 compared to Bi-LSTM. The experiment with neural network Bi-LSTM helped to understand the difference between the result of transformer models and neural networks. BERT models are trained for text data, so their performance is better than the traditional neural networks.

The dynamic masking of RoBERTa is an advantage over BERT, which contributes to the model performance and makes it robust. It outperforms BERT by 3% in terms of accuracy by achieving an accuracy of 95%. Although RoBERTa performs better than DistilBERT, the performance score of DistilBERT was also remarkable. It performed better than BERT and Bi-LSTM due to its architectural advantage, which makes it a good choice for text classification. Lastly, the

TABLE 12. Dataset evaluation results.

Models	Evaluation Metrics			
	Accuracy	Precision	Recall	F1-Score
Bi-LSTM	0.89	0.87	0.83	0.88
BERT	0.92	0.89	0.92	0.93
RoBERTa	0.95	0.93	0.94	0.95
DistilBERT	0.93	0.9	0.89	0.94

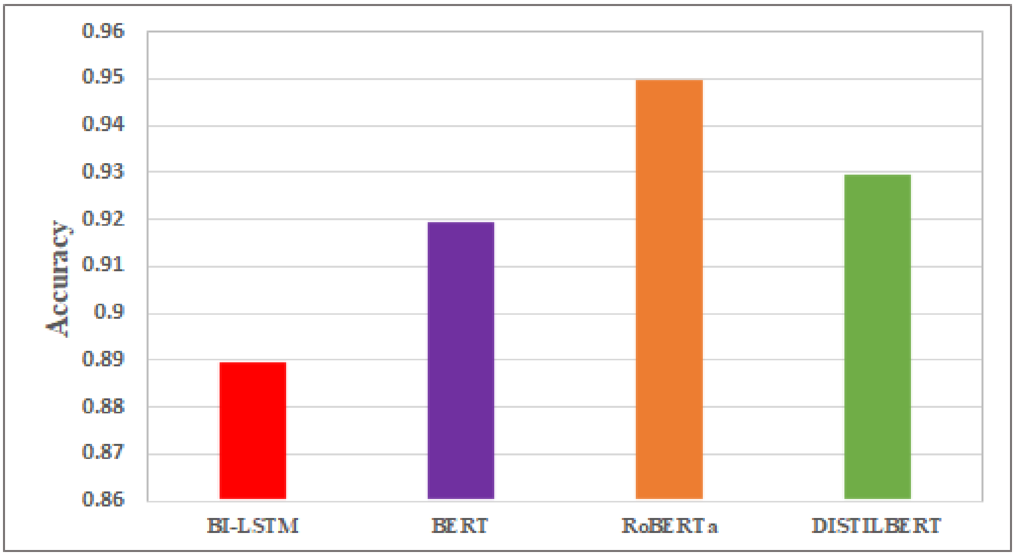

FIGURE 13. Accuracy of deep learning classifiers.

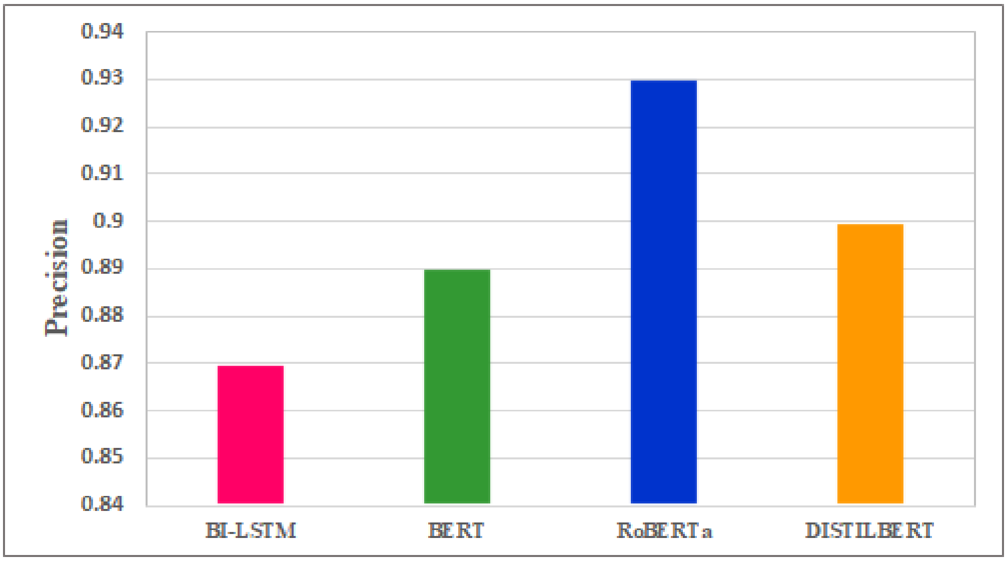

FIGURE 14. Precision values of deep learning classifiers.

training time of RoBERTa was low compared to BERT, DistilBERT, and Bi-LSTM, which makes it optimal for classifying text, as shown in this study.

N. PERFORMANCE EVALUATION

The performance of deep learning models was evaluated using an automatic verification dataset in which 33% of the

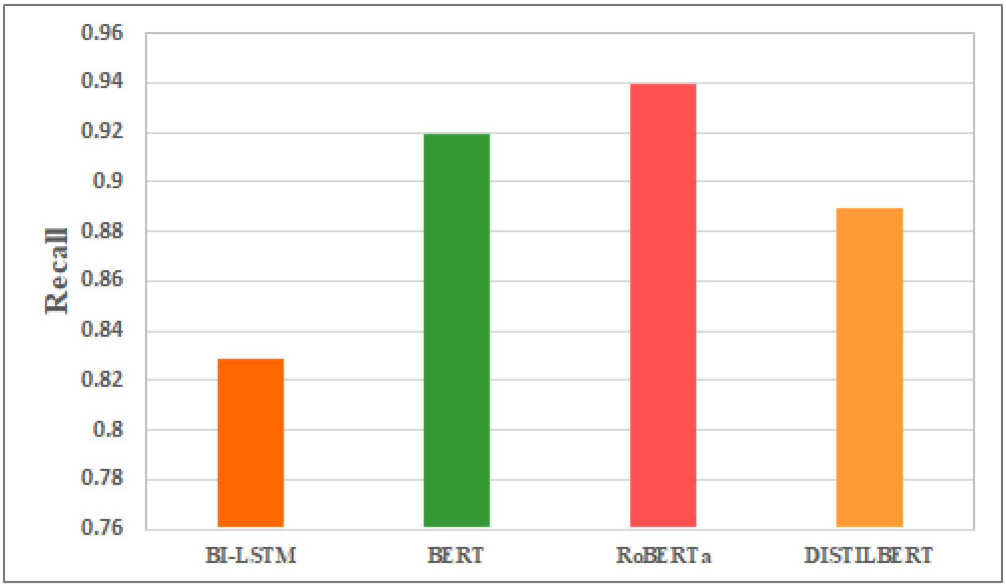

FIGURE 15. Recall values of deep learning classifiers.

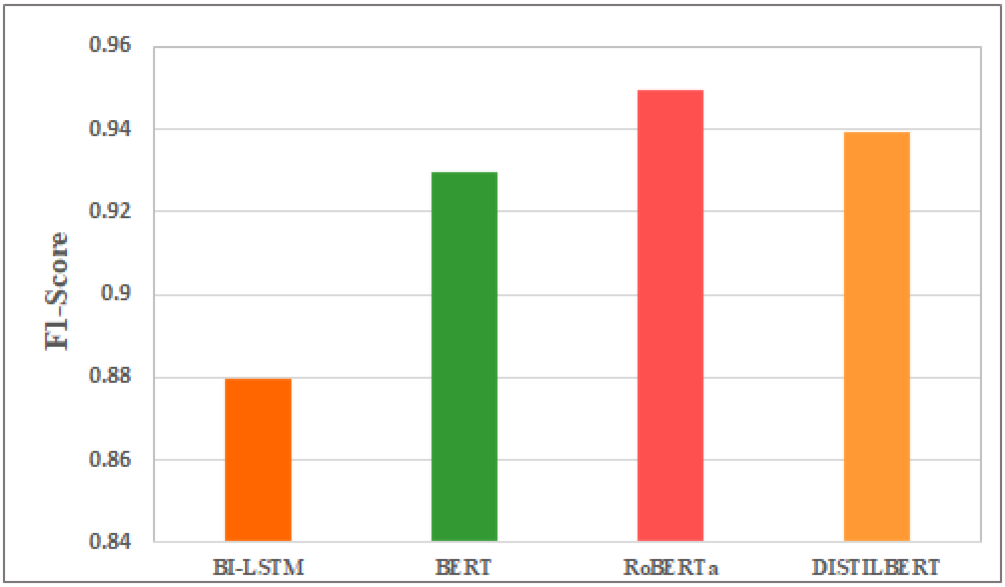

FIGURE 16. F1-Score values of deep learning classifiers.

data was held back for validation. On running the verbose output on each epoch it showed the loss and accuracy on both the training dataset and the validation dataset, which was used to evaluate their performances.

O. ERROR ANALYSIS

Even though the trained deep learning models have good performance scores, there was still some misclassification of tweets, especially those incorrectly identified as radicalism but were propaganda and vice versa. Some of the reasons behind the incorrect identification of these tweets are probably due to the improper labelling of the seed dataset and epoch sizes during training. It is possible that the tweets labelled as radicalism or propaganda in the seed dataset could have been mixed up as both seem to have similar contexts. Therefore, when the model is trained, it predicts more or fewer tweets as propaganda or radicalism because their context is somewhat similar. Moreover, epoch size is used in training the model, which could have affected how unlabelled data was misclassified. When the epoch size is larger, then more data is adequately trained in the model and a result, it returns with a low validation loss and high accuracy. Thus, in this experiment, a moderate epoch size might have trained some of the propaganda and radicalism tweets properly, which could have caused misclassification for some of those tweets.

VIII. LIMITATIONS

This work is limited to Twitter and can be expanded to other networking platforms, such as Facebook and Parler. Although this study explores the social media responses to the U.S. presidential election and the U.S. Capitol riot in English, the response in different languages is to be investigated. The collected dataset lacked emoticons, which can help understand the user's sentiment. Therefore, emoji detection can be used to obtain better results. The bot tweets present in them can significantly affect the data, resulting in incorrect classification. Hence, it is best to detect bots through advanced approaches and increase the model's efficiency in identifying bots from posts. Finally, a user interface can be designed to detect extremism in posts.

IX. CONCLUSION

This research contributes to the field by creating a high-quality diversified dataset on the U.S. Capitol riot gathered via Twitter. This study explains how semi-supervised learning is used to predict labels. This research also compares and evaluates several deep learning classifiers such as BERT, BI-LSTM, RoBERTa, and DistilBert. According to the experimental data, Roberta achieved the most competitive outcomes with 95% accuracy. The results of the models show that they can help identify extremist messages on social media, thereby preventing the tragic consequences of the spread of radical posts. Other platforms, such as Facebook and Parler, can be analyzed to gain a broader perspective of the riot and investigate social media's influence on the masses. A few advanced features can be developed or explored to suggest a more accurate model for detecting extremism.

ACKNOWLEDGMENT

The authors would like to thank Symbiosis International University and Symbiosis Institute of Technology for their support and help in this research, and would also like to thank anonymous reviewers who made this study better with their valuable input.

REFERENCES

[1] K. Smith. (Jan. 2, 2018). *60 Incredible and Interesting Twitter Stats and Statistics | Brandwatch*. Brandwatch. Accessed: Oct. 4, 2022. [Online]. Available: https://www.brandwatch.com/blog/twitter-stats-and-statistics/

[2] D. Sayce. (2022). *The Number of Tweets Per Day in 2022*. Accessed: Oct. 4, 2022. [Online]. Available: https://www.dsayce.com/social-media/tweets-day/

[3] ANI. (Nov. 14, 2020). *Twitter Flags Around 300,000 Tweets for Misleading Content Regarding U.S. Presidential Elections—The Economic Times*. Economic Times. Accessed: Oct. 4, 2022. [Online]. Available: https://economictimes.indiatimes.com/news/international/world-news/twitter-flags-around-300000-tweets-for-misleading-content-regarding-us-presidential-elections/articleshow/79219543.cms

[4] *The Evolution of Social Media: How Did It Begin and Where Could It Go Next*. Accessed: Oct. 4, 2022. [Online]. Available: https://online.maryville.edu/blog/evolution-social-media/

[5] M. Aliapoulios, E. Bevensee, J. Blackburn, B. Bradlyn, E. De Cristofaro, G. Stringhini, and S. Zannettou, "An early look at the parler online social network," 2021, *arXiv:2101.03820*.

[6] C. Yong. (Jan. 2021). *U.S. Capitol Riot: How Social Media Helped Enable Attack by Die-Hard Trump Fans | the Straits Times*. Straits Times. Accessed: Oct. 4, 2022. [Online]. Available: https://www.straitstimes.com/world/united-states/how-social-media-helped-enable-the-storming-of-the-us-capitol

[7] A. Papasavva, J. Blackburn, G. Stringhini, S. Zannettou, and E. D. Cristofaro, "'Is it a qoincidence?': An exploratory study of QAnon on voat," 2020, *arXiv:2009.04885*.

[8] D. Zaldivar and I. Alsmadi, "Capitol riot analysis," *SSRN Electron. J.*, Nov. 2021, doi: 10.2139/ssrn.3959184.

[9] A. Prabhu, D. Guhathakurta, J. Jain, M. Subramanian, M. Reddy, S. Sehgal, T. Karandikar, A. Gulati, U. Arora, R. Ratn Shah, and P. Kumaraguru, "Capitol (Pat)riots: A comparative study of Twitter and parler," 2021, *arXiv:2101.06914*.

[10] A. Sipka, A. Hannak, and A. Urman, "Comparing the language of QAnon-related content on parler, gab, and Twitter," 2021, *arXiv:2111.11118*.

[11] H. Alvari, S. Sarkar, and P. Shakarian, "Detection of violent extremists in social media," in *Proc. 2nd Int. Conf. Data Intell. Secur. (ICDIS)*, Jun. 2019, pp. 43–47, doi: 10.1109/ICDIS.2019.00014.

[12] 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, pp. 1–23, Dec. 2019, doi: 10.1186/s13673-019-0185-6.

[13] U. Yaqub, N. Sharma, R. Pabreja, S. A. Chun, V. Atluri, and J. Vaidya, "Location-based sentiment analyses and visualization of Twitter election data," *Digit. Government, Res. Pract.*, vol. 1, no. 2, pp. 1–19, Apr. 2020, doi: 10.1145/3339909.

[14] B. Bansal and S. Srivastava, "On predicting elections with hybrid topic based sentiment analysis of tweets," *Proc. Comput. Sci.*, vol. 135, pp. 346–353, Jan. 2018, doi: 10.1016/J.PROCS.2018.08.183.

[15] S. Zambezi. (2021). *Predicting Social Unrest Events in South Africa Using LSTM Neural Networks*. Accessed: Oct. 4, 2022. [Online]. Available: https://open.uct.ac.za/handle/11427/33986

[16] C. Saroufim, A. Almatarky, and M. Abdel Hady, "Language independent sentiment analysis with sentiment-specific word embeddings," in *Proc. 9th Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal.*, 2018, pp. 14–23, doi: 10.18653/V1/W18-6204.

[17] 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.

[18] S. Aldera, A. Emam, M. Al-Qurishi, M. Alrubaian, and A. Alothaim, "Online extremism detection in textual content: A systematic literature review," *IEEE Access*, vol. 9, pp. 42384–42396, 2021, doi: 10.1109/ACCESS.2021.3064178.

[19] C. Fuhriman, R. M. Medina, and S. Brewer, "Introducing a dataset of multi-scale geographies of ISIS ideology from ISIS sources," *Terrorism Political Violence*, vol. 34, no. 4, pp. 817–834, 2020, doi: 10.1080/09546553.2020.1742707.

[20] A. I. Abd-Elaal, A. Z. Badr, and H. M. Mahdi, "Detecting violent radical accounts on Twitter," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 8, pp. 1–7, 2020, doi: 10.14569/IJACSA.2020.0110865.

[21] S. Mussiraliyeva, B. Omarov, M. Bolatbek, K. Bagitova, and Z. Alimzhanova, "Bigram based deep neural network for extremism detection in online user generated contents in the kazakh language," *Commun. Comput. Inf. Sci.*, vol. 1463, pp. 559–570, Sep. 2021, doi: 10.1007/978-3-030-88113-9_45/COVER.

[22] S. Aldera, A. Emam, M. Al-Qurishi, M. Alrubaian, and A. Alothaim, "Exploratory data analysis and classification of a new Arabic online extremism dataset," *IEEE Access*, vol. 9, pp. 161613–161626, 2021, doi: 10.1109/ACCESS.2021.3132651.

[23] M. Bahrami, Y. Findik, B. Bozkaya, and S. Balcisoy, "Twitter reveals: Using Twitter analytics to predict public protests," 2018, *arXiv:1805.00358*.

[24] L. J. Anastasopoulos and J. R. Williams, "A scalable machine learning approach for measuring violent and peaceful forms of political protest participation with social media data," *PLoS ONE*, vol. 14, no. 3, Mar. 2019, Art. no. e0212834, doi: 10.1371/JOURNAL.PONE.0212834.

[25] J. T. Chan and W. Zhong, "Predicting authoritarian crackdowns: A machine learning approach," *SSRN Electron. J.*, Mar. 2020, doi: 10.2139/SSRN.3545999.

[26] A. Chiorrini, C. Diamantini, A. Mircoli, and D. Potena, "Emotion and sentiment analysis of tweets using BERT," in *Proc. EDBT/ICDT Workshops*, 2021.

[27] 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," 2020, *arXiv:2010.00357*.

[28] B. Mathew, R. Dutt, P. Goyal, and A. Mukherjee, "Spread of hate speech in online social media," 2018, *arXiv:1812.01693*.

[29] K. T. Mursi, M. D. Alahmadi, F. S. Alsubaei, and A. S. Alghamdi, "Detecting Islamic radicalism Arabic tweets using natural language processing," *IEEE Access*, vol. 10, pp. 72526–72534, 2022, doi: 10.1109/ACCESS.2022.3188688.

[30] M. L. Jamil, S. Pais, J. Cordeiro, and G. Dias, "Detection of extreme sentiments on social networks with BERT," *Social Netw. Anal. Mining*, vol. 12, no. 1, pp. 1–16, Dec. 2022, doi: 10.1007/S13278-022-00882-Z/FIGURES/10.

[31] Z. Ul Rehman, S. Abbas, M. A. Khan, G. Mustafa, H. Fayyaz, M. Hanif, and M. Anwar 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.

[32] N. Alsaedi, P. Burnap, and O. Rana, "Can we predict a riot? Disruptive event detection using Twitter," *ACM Trans. Internet Technol.*, vol. 17, no. 2, pp. 1–26, Mar. 2017, doi: 10.1145/2996183.

[33] R. Rezapour, "Using linguistic cues for analyzing social movements," Aug. 2018, *arXiv:1808.01742*, doi: 10.48550/arXiv.1808.01742.

[34] N. Rattner. (Jan. 13, 2021). *Trump Tweets: Legacy of Lies, Misinformation, Distrust*. CNBC. Accessed: Oct. 8, 2022. [Online]. Available: https://www.cnbc.com/2021/01/13/trump-tweets-legacy-of-lies-misinformation-distrust.html

[35] D. Alba, K. Conger, and R. Zhong, *Twitter Places Warning on Trump Minneapolis Tweet, Saying it Glorified Violence—The New York Times*. New York, NY, USA: New York Times, Jun. 2020. Accessed: Oct. 8, 2022. [Online]. Available: https://www.nytimes.com/2020/05/29/technology/trump-twitter-minneapolis-george-floyd.html

[36] J. Holt. (Feb. 10, 2021). *#StopTheSteal: Timeline of Social Media and Extremist Activities Leading to 1/6 Insurrection*. The Atlantic Council's Digital Forensic Research Lab. Accessed: Oct. 8, 2022. [Online]. Available: https://www.justsecurity.org/74622/stopthesteal-timeline-of-social-media-and-extremist-activities-leading-to-1-6-insurrection/

[37] H. Thayer. *21st Century Propaganda: The Age of Twitter*. Pleasantville, NY, USA: Dyson College Arts Science Pace Univ., 2018. Accessed: Oct. 8, 2022. [Online]. Available: https://digitalcommons.pace.edu/honorscollege_theses

[38] L. Dilley, W. Welna, and F. Foster, "QAnon propaganda on Twitter as information warfare: Influencers, networks, and narratives," Jul. 2022, *arXiv:2207.05118*, doi: 10.48550/arXiv.2207.05118.

[39] I. A. Hamilton. (May 29, 2020). *Twitter Slapped a 'Glorifying Violence' Label on a Trump Tweet That Threatened George Floyd Protesters in Minneapolis With Getting Shot | Business Insider India*. Business Insider. Accessed: Oct. 8, 2022. [Online]. Available: https://www.businessinsider.in/tech/news/twitter-just-slapped-a-glorifying-violence-label-on-a-trump-tweet-that-threatened-george-floyd-protesters-in-minneapolis-with-getting-shot/articleshow/76087411.cms

[40] C. Goforth. (Mar. 26, 2021). *Donald Trump: Investigators Believe Tweet Incited Capitol Riot*. Daily Dot. Accessed: Oct. 8, 2022. [Online]. Available: https://www.dailydot.com/debug/donald-trump-tweet-incited-capitol-riot/

[41] R. Procter, F. Vis, and A. Voss, "Reading the riots on Twitter: Methodological innovation for the analysis of big data," *Int. J. Social Res. Methodol.*, vol. 16, no. 3, pp. 197–214, May 2013, doi: 10.1080/13645579.2013.774172.

[42] *What is Natural Language Processing? | IBM*. Accessed: Oct. 4, 2022. [Online]. Available: https://www.ibm.com/cloud/learn/natural-language-processing

[43] J. Samuel, G. G. M. N. Ali, M. M. Rahman, E. Esawi, and Y. Samuel, "COVID-19 public sentiment insights and machine learning for tweets classification," *Information*, vol. 11, no. 6, p. 314, Jun. 2020, doi: 10.3390/INFO11060314.

ARUNDARASI RAJENDRAN received the B.Tech. degree in computer science and engineering from the Symbiosis Institute of Technology, Pune, Maharashtra, India. She has completed her five month research internship at the Symbiosis Center for Applied Artificial Intelligence. Her research interests include data science, machine learning, deep learning, and natural language processing.

VATTIKUTI SREE SAHITHI received the B.Tech. degree in computer science and engineering from the Symbiosis Institute of Technology, Pune, Maharashtra, India. Her research interests include machine learning, deep learning, and natural language processing.

CHHAVI GUPTA received the B.Tech. degree in computer science and engineering from the Symbiosis Institute of Technology, Pune, Maharashtra, India. Her research interests include mathematics, machine learning, deep learning, and natural language processing.

MADHURI YADAV received the B.Tech. degree in computer science and engineering from the Symbiosis Institute of Technology, Pune, Maharashtra, India. She has completed her five month research internship at the Symbiosis Center for Applied Artificial Intelligence. She is interested in data science, machine learning, and web development.

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, Pune. Her research interests include big data analytics, machine learning, deep learning, natural language processing, and reinforcement learning. 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 3 and an i10-index of 2.

KETAN KOTECHA received the M.Tech. and Ph.D. degrees from (IIT Bombay). He is currently the 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 over the last 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 was a recipient of the two SPARC projects worth INR 166 lakhs from MHRD Government of India in AI in collaboration with Arizona State University, USA, and The University of Queensland, Australia, and 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 a Grant from Royal Academy of Engineering, the U.K. under Newton Bhabha Fund. He has published three patents and delivered keynote speeches at various national and international forums, including at the Machine Intelligence Laboratory, IIT Bombay, USA, under the World Bank project at the International Indian Science Festival organized by the Department of Science Technology, Government of India, and many more. He is also an Academic Editor of the *Peerj Computer Science* journal and an Associate Editor of IEEE ACCESS journal.

MAYUR GAIKWAD (Member, IEEE) received the master's degree in computer science and engineering from the Symbiosis Institute of Technology, Pune. 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.

AJITH ABRAHAM (Senior Member, IEEE) received the Master of Science 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, 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 U.S. $. Currently, he holds two university professorial appointments. He works as a Professor in artificial intelligence at Innopolis University, Russia, and the Yayasan Tun Ismail Mohamed Ali Professorial Chair in artificial intelligence at 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 (2011–2013). 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 ISI.

NADA AHMED received the B.E. degree from Omdurman Ahlia University, Sudan, in 2002, the master's degree from Jazira University, and the Ph.D. degree from Sudan University for Science and Technology. She is currently an Assistant Professor in computer science at the Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, Saudi Arabia. A strong theme of her work is in cloud computing, machine learning, and data mining.

SARAH M. ALHAMMAD received the Ph.D. degree in computer science from the University of Plymouth, U.K. She is currently an Assistant Professor in computer science at the Department of Computer Science, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, Saudi Arabia. A strong theme of her work is in programming, software engineering, and HCI.