IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 2, 2021, accepted March 19, 2021, date of publication March 23, 2021, date of current version April 2, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3068313

Online Extremism Detection: A Systematic Literature Review With Emphasis on Datasets, Classification Techniques, Validation Methods, and Tools

MAYUR GAIKWAD[1], SWATI AHIRRAO[1], SHRADDHA PHANSALKAR[2], AND KETAN KOTECHA[3]

[1]Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune 412115, India
[2]MIT Art, Design and Technology University, Pune 412201, India
[3]Symbiosis Centre for Applied Artificial Intelligence, Symbiosis International (Deemed University), Pune 412115, India

Corresponding author: Swati Ahirrao (swatia@sitpune.edu.in)

This work was supported by Research Support Fund of Symbiosis International (Deemed University).

ABSTRACT Social media platforms are popular for expressing personal views, emotions and beliefs. Social media platforms are influential for propagating extremist ideologies for group-building, fund-raising, and recruitment. To monitor and control the outreach of extremists on social media, detection of extremism in social media is necessary. The existing extremism detection literature on social media is limited by specific ideology, subjective validation methods, and binary or tertiary classification. A comprehensive and comparative survey of datasets, classification techniques, validation methods with online extremism detection tool is essential. The systematic literature review methodology (PRISMA) was used. Sixty-four studies on extremism research were collected, including 31 from SCOPUS, Web of Science (WoS), ACM, IEEE, and 33 thesis, technical and analytical reports using Snowballing technique. The survey highlights the role of social media in propagating online radicalization and the need for extremism detection on social media platforms. The review concludes lack of publicly available, class-balanced, and unbiased datasets for better detection and classification of social-media extremism. Lack of validation techniques to evaluate correctness and quality of custom data sets without human interventions, was found. The information retrieval unveiled that contemporary research work is prejudiced towards ISIS ideology. We investigated that deep learning based automated extremism detection techniques outperform other techniques. The review opens the research opportunities for developing an online, publicly available automated tool for extremism data collection and detection. The survey results in conceptualization of architecture for construction of multi-ideology extremism text dataset with robust data validation techniques for multiclass classification of extremism text.

INDEX TERMS Extremism, machine learning, propaganda, radicalization, systematic literature review, terrorism.

I. INTRODUCTION

Social media platforms allow sharing their views, opinions, emotions, judgment, and beliefs. Social media platforms such as Twitter, Facebook, WhatsApp, Instagram are popular choices and flooded with messages, posts, and tweets. As per the recent survey, every minute, 4,74,000 tweets are uploaded on Twitter, and 2,93,000 statuses are uploaded on Facebook [1]. With billions of registered users, social media platforms offer wide outreach. Hence it is convenient media for the extremist group to propagate their harmful ideology. These extremist groups share violent content, hateful messages to make their agenda widespread in radicalization, recruitment, and propaganda [2]. The extremist organizations such as the Islamic State of Iraq and Syria (ISIS) and Al Qaeda now use social media platforms for propaganda, radicalization, and recruitment of the susceptible youth. These three concepts are elaborated in Section I(C).

As online extremism became a mainstay for extremist organizations, some researchers investigated the spread of

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

extremism through social media. Farwell [3] in 2014 counted an average of 44,000 tweets per day from ISIS supporters when ISIS attacked Mosul, a city in Iraq. Milton [4] measured 9,000 visual media like videos, images, and picture reports on Twitter during January 2015 - August 2016 related to ISIS.

Bill S-894 [5] presented before the US Congress in March 2019, from September 12, 2001, the 73 percent of the domestic violent incidents in the USA are connected to the far right-wing extremist groups and the remaining 27 percent are related to the radical Islamists. Many researchers have presented statistics about the presence and growth of right-wing extremism on the Internet. Berger [6] in Alt-Right Twitter Census 2018 poised that the number of Twitter users supporting the Alt-Right movement is nearly 1,00,000. The influential members of the Alt-Right movement have followers count up to 40,000. Christchurch Mosque Attack in New Zealand was live-streamed attack on Facebook by the extremists. Facebook claimed that only 200 people watched it live. Facebook had to stop nearly 1.2 million reuploads of the video, whereas 3,00,000 reuploads were undetected by Facebook [7].

Islamic State of Iraq and Syria (ISIS) [8] is a terrorist organization in the Middle East. ISIS fighters, ISIS recruiters, and ISIS supporters have a large number of accounts on Twitter. During ISIS's march to Mosul city, Social Media Platforms were extensively used, and 44,000 tweets from ISIS supporters were recorded [3]. The online magazines such as Rumiyah and Dabiq by ISIS and Inspire [9] and Al-Shamika [10] were used to reach the target audiences.

White supremacist is the racist ideology that also used these social media platforms to spread their Right-wing ideology. White supremacist propaganda blogs and websites like Stormfront [11] and Iron March [12] spread the hatred among the people. The recent attack on Christchurch mosque in New Zealand by an extremist was live-streamed over social media platforms [13]. Buckingham and Alali [13] claim that New Zealand doesn't show the active extremist movements. The online extremism which disseminated through Europe has led to this lone-wolf attack. The target audiences are radicalized, recruited, and even influenced to carry out violent acts [14].

The social media platforms like Facebook [15] and Twitter [16] have their own rules and regulations to combat the offensive and extremist text. This includes tagging some of the users as extremists. Some of them track community building amongst the extremist network. However, Social media platforms take action against offensive or extremist posts only after a user reports them. It is found that the extremists exploit the lacuna in content moderation by social media platforms. This increasing hatred on the social media has actively persuaded many countries to monitor and actively promote research in the extremism detection domain. The increasing use of social media platforms necessitates the development of the online extremism detection on social media platforms to help the regulatory agencies, automatically pinpointing extremist views. These efforts will help the government agencies in controlling the spread of extremism. Thus, online extremism detection on the social media platform is the significant research area.

A. SIGNIFICANCE

Social media platforms, websites, and blogs enhance outreach for the extremists [17]. Thus there is a rise in opportunities for extremists, radicalizing the youth [18]. The extremists present themselves with a positive narrative, enticing the youths with the hope of a better world. These narratives reach people due to an easy access to social media. Thus false report combined with the rapid spread of information is enough to convince gullible youths to commit the violent acts [19]. The researchers observe that the uninterrupted discussions between like-minded extremist individuals accelerate self-radicalization [17].

In addition to propaganda, radicalization, and recruitment, hate manifestos are also shared on the internet. Manifestos of high profile mass shooters have influenced other mass shooters [20]. In another study, Berger [20] states that all the recent activities of the violent White Supremacists can be can be directly traced to the manifesto of Oslo attacker. Thus, social media and forums have become a propaganda tool for extremists [21]. In her thesis Turner [22] observed that 6 out of 12 studied individual were recruited using the online strategies by ISIS. Thus, it is imperative to study and detect extremism spread via social media to restrict the spread of propaganda, radicalization, and eliminate extremist recruitment.

B. MOTIVATION

The extremist detection has few peer-reviewed research. Besides, there is no comprehensive survey [6], [18], [23] about extremism detection with an emphasis on datasets, classification techniques, validation methods, and online extremism detection tools. Literature shows a lack of exhaustive survey on the extremism datasets of standard and custom-built nature. The datasets from the multiple sources with different class labels need to be studied for identifying the research gaps in the extremism datasets.

The existing literature on extremism detection is focused on manual and automated trends, different classification techniques using network or graph approach, Machine Learning (ML) approach and Deep Learning (DL) approach [24]. The comparative analysis of these techniques based on various evaluation metrics is essential. Previous work shows a lack of sensitization towards the importance of data validation methods. This survey attempts to throw light on multiple datasets, algorithms, validation, and classification methods.

C. TERMS AND TERMINOLOGIES

Following are the few terms that are frequently used in extremism detection research:

Ideology is defined as the manner or content of thinking of a person, group, or culture [25]. Ideology is usually applied to political and religious thinking [26]. In recent times, the term ''Ideology'' is used to condemn the ideas of the person or group [26].

Extremism is defined as 'supporting beliefs that are extreme' [27]. Extremism usually refers to the ideology that may be religious or political, that are unacceptable to the general perception of the society. Extremism is not a recent phenomenon. The extremist thoughts are spread to gain religious and political gains for ages [28]. Extremism in religion is studied extensively and has led to associate it with a particular religion [29].

Propaganda is 'information, usually of biased nature, used to justify political cause or point of view' [30]. The spread of misinformation for political gains is also termed 'propaganda.' Online propaganda has a significant influence on the masses to garner support for conflict on social media and other online platforms [31]. Furthermore, researchers have classified propaganda as 'black' or 'white' [32]. Propaganda has served as an efficient tool in war times like encouraging people, which sometimes is termed as 'white propaganda.' Propaganda with half-truth and lies is known as 'black propaganda.' This 'black' propaganda is linked with Nazism in Germany and the former Soviet Union spreading their ideologies.

Radicalization is termed as 'changes in belief, feelings, and behaviour towards the extremity, calling for violence and sacrifice' [33]. Online radicalization is different from online propaganda. Online propaganda uses misinformation, while online radicalization misleads people by using their beliefs, either political or religious [34].

Recruitment *in the context of terrorism* is 'the enticement of youth to join and sacrifice for terrorist cause' [35]. Pre-social media times, extremist organizations recruited youth through a sympathetic family member, even duping, kidnapping, or force recruiting. Extremist organizations changed recruitment tactics after the arrival of social media [36]. These tactics include uploading texts, images, and videos with increasing extremist ideas. Even rap songs delivering extremist thoughts are used as a recruitment tactic. Thus, the extremist organization uses every tool and technique available, to spread their beliefs on social media. So, in this study, propaganda, radicalization and recruitment are considered as the focus areas in online extremism.

Inter-rater Agreement can be defined as a measure of agreement among experts, or annotators [37]. The score of inter-rater agreement denotes how much consensus exists in annotating or labelling given by various experts. Different coefficients statistically represent inter-rater agreement; two of them are Cohen's Kappa and Fleiss' Kappa. In extremism research, the inter-rater agreement is used to label extremist and non-extremist data and validate the data labelling of the author. Inter-rater agreement and its significance in extremism research is described in detail in Section IV(E).

D. EVOLUTION OF ONLINE EXTREMISM DETECTION TECHNIQUES

Early research handles the analysis of extremism on the Internet. Earlier work on online extremism with the emphasis on recruiting young people was observed in 2001 [38]. These early researches relied on manual identification method for extremism detection. Automated detection of extremism was initiated in 2007 [39]. A study of extremism spread on social media shows the prominent use of Social Network Analysis in an earlier trend. The ML approach is extensively used in online extremism detection research, since 2013. The important reason for using the ML approach is the classification and prediction capabilities of ML algorithms. Logistic Regression, SVM, Adaboost, and Random Forest, different algorithms are used for online extremism detection. As social media became prominent, researchers collected a huge amount of extremist data from Twitter, YouTube, and Facebook.

Deep Learning techniques are employed for processing a massive amount of data. With the advent of Deep Learning technologies, the researcher started using DL techniques for classification and prediction. Due to Deep Learning algorithms' ability to remember the context and long-term dependencies, researchers now prefer DL methods for online extremism detection. Algorithms like Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN) and Bidirectional Encoder Representation for Transformers (BERT) are used for online extremism detection.

E. PRIOR RESEARCH

On the topic of online extremism detection concerning computer science, a minimal number of Systematic Literature Reviews (SLRs) are published to the best of our knowledge.

Fernandez *et al.* [40], in 2020, categorized online extremism research into three types: *Analysis*, *Detection,* and *Prediction*. This survey concerns itself with communication and the process of radicalization. The authors also look into details of automatic detection of extremists and prediction of content adoption. This review examines the lack of validated data, lack of cooperation between researchers, extremist language evolution, and lack of ethical perspectives in online extremism detection research.

Research in extremism detection is evolved from hate speech research. So, it was worthwhile to consider SLRs in hate speech detection in this survey. Fortuna *et al.* [41] in 2018 elaborated the need to study the automatic detection of hate speech. The authors devised a new hate speech definition by analyzing the dimensions of hate speech. Studies are categorized into the topics like racism, sexism, prejudice towards refugees, and homophobia. The author finds very few publicly available datasets. The authors claim that due to the use of different metrics, and datasets by researchers, it is difficult to conclude which classification technique is better. Fortuna *et al.* provides an essential comparison between datasets, approaches, and metrics for researchers to follow in the online extremism research.

In their survey, Al-Hassan *et al.* [42] address questions like *What is Hate Speech, What constitutes Hate Speech,* and other hate speech detection approaches, but they do not discuss the dataset or validation of datasets. Al-Hassan *et al.* analyses

TABLE 1. Research goals.

Research Questions	Discussion
RQ1: What are the various datasets available for online extremism detection and classification?	Study the datasets by comparing various parameters like data collection methods, data sources, data size, classification labels, and class imbalance.
RQ2: What are the various methodological trends in online extremism detection and classification?	A study of extremism detection methods like Manual and Automated detection is carried out.
RQ3: What are different classification techniques used for online extremism detection and their comparative evaluation?	Comparative analysis of classification techniques with various evaluation metrics is performed.
RQ4: What validation techniques are used for data validation in online extremism detection and classification?	Different validation techniques for checking the quality of data are discussed.
RQ5: What are popular tools available for online extremism detection?	Various Online Extremism detection tools, projects, and prototypes with the comparison of performance standards are discussed.
RQ6: Is there empirical evidence available that the current literature is biased towards a specific ideology?	To find if there is literature bias about specific ideology in online extremism detection

the literature based on different hate speech detection dimensions like cyberbullying, radicalization, abusive language, religious, and racial hate speech. This survey focused on the Arabic social sphere for hate speech detection and put forth challenges in hate speech detection in Arabic.

Existing surveys [40]–[42] lack detailed analysis of datasets, technical approaches, evaluation metrics, and data validation. A lack of research contribution in extremism detection concerning datasets is found, methodological trends, classification techniques, validation approaches, and online extremism detection. This SLR opens opportunities for future research work in this area. This work attempts an exhaustive survey on various datasets, methodological trends, detection and classification methods, validation methods and evaluation metrics, and publicly available online extremism detection tools.

F. RESEARCH GOALS

This research aims to analyze the existing studies, their findings and leverage comparative analysis of existing online extremism detection techniques. The research questions are proposed in Table 1 to get a detailed survey of extremism detection:

G. CONTRIBUTION OF THE WORK

- Twenty-seven contributions towards applying custom built or standard online extremism datasets and analyzing them concerning various evaluation metrics are identified.
- Emerging trends in online extremism detection by analyzing 11 works in manual extremism detection techniques and 52 works in automated online extremism detection techniques are outlined.
- A comparative analysis of the popular techniques used for online extremism detection and classification is presented.
- Inference from the survey and establishment of the need for better validation techniques for quality assessment of the data as future work.
- The existing tools available for online extremism detection and classification and emphasize the need for automated, publicly available, and multiple classes to classify propaganda, radicalization, and recruitment are presented.
- It is established experimentally the existing research is biased towards limited ideologies.
- The work proposes the architecture to construct the ideology independent, extremism text dataset for multiclass classification with robust data validation techniques.

II. RESEARCH METHODOLOGY

A systematic literature review of the existing literature was conducted to address the research questions identified. PRISMA guidelines published by Kitchenham and Charters [43] were adapted for carrying out the detailed systematic literature review.

A. SELECTION CRITERIA

SCOPUS database was used to retrieve articles for extremism detection. A specialized query was formulated to retrieve the research articles from the SCOPUS database. Keywords used to create query are *('extremism,''radicalization,''extremists' and 'detection').*

The multiple database search approach was adopted, as shown in the Table 2.

B. INCLUSION/EXCULSION CRITERIA

The trends in online extremism detection and classification based on the trends that are included in the study are analyzed. Works with extremism detection using techniques like Network or Graph-based, Machine Learning-based, or Deep Learning-based are selected.

C. SELECTION RESULTS

Two hundred and forty-eight studies have been found with the initial query result. SCOPUS identified 69 works, Web of Science showed 17 papers, ACM displayed 162 studies, while IEEE returned 18 research work related to online extremism detection. Thirty-one studies were further shortlisted through inclusion/exclusion criteria. A total of 33 studies were obtained using Snowballing technique. Thus 64 works were obtained by applying inclusion/exclusion criteria. The studies from conferences, journals, thesis, and reports published from 2015 to 2020 were included. The document type per year of selected studies is presented in Fig. 3.

TABLE 2. Literature database and query.

Database	Query Executed
ACM	*[[All: "extremism"] OR [All: "radicalisation"] OR [All: "radicalization"] OR [All: "extremist propaganda"] OR [All: "extremists"]] AND [All: "detection" (or) "classification"] AND [[All: "network analysis"] OR [All: "graph"] OR [All: "graph-based"] OR [All: "machine learning"] OR [All: "deep learning"]] AND [Publication Date: (01/01/2015 TO 12/31/2020)]*
SCOPUS	TITLE-ABS KEY ((*"extremism" OR "radicalisation" OR "radicalization" OR "extremist propaganda" OR "extremists") AND ("detection" OR "classification") AND ("network analysis" OR "graph" OR "graph-based" OR "machine learning" OR "deep learning"*))
Web of Science	TOPIC: ((*"extremism" OR "radicalisation" OR "extremist propaganda" OR "radicalization" OR "extremists") AND ("detection" OR "classification") AND ("network analysis" OR "graph" OR "graph-based" OR "machine learning" OR "deep learning"*))
IEEE	*("extremism" OR "radicalization" OR "radicalisation" OR "extremist propaganda" OR "extremists") AND ("detection" OR "classification") AND ("network analysis" OR "graph" OR "graph-based" OR "machine learning" OR "deep learning")*

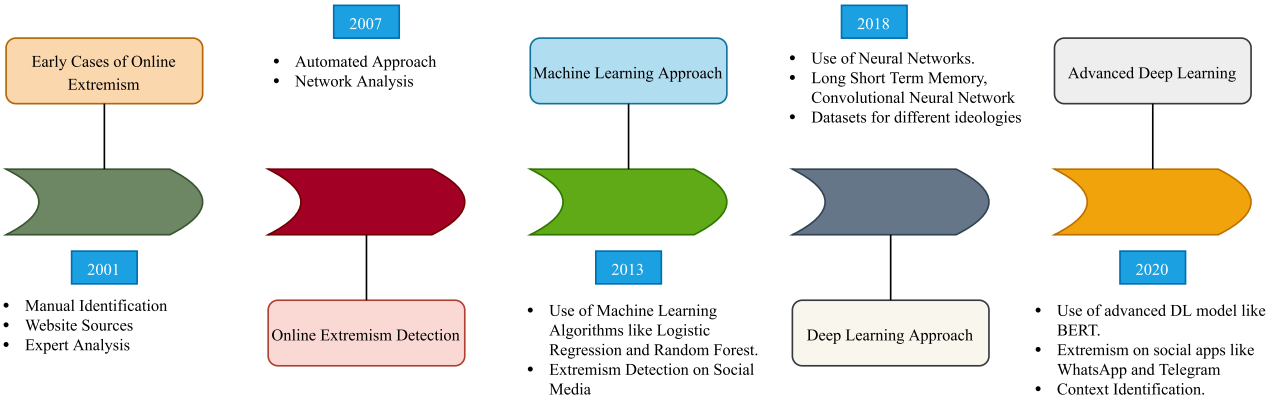


FIGURE 1. Evolution of online extremism detection.

D. QUALITY ASSESSMENT

For quality assessment, the following steps were used:

- *Extremism*–Research must be focused on either detection, classification, or datasets about extremism.
- *Types of extremism* (Online Radicalization or Online Recruitment) –Research must focus on radicalization or recruitment done through the internet sources like social media, websites, blogs, or messaging apps.
- *Detection*–Research must focus on the detection of extremism in digital media and different methodologies.
- *Datasets and Classification Algorithms*–The research work emphasizes dataset collection, building, or available datasets for classification using different techniques.
- *Data Validation*–The paper should use different validation techniques used for checking the quality of data.

III. REVIEW OF EXTREMISM DETECTION METHODS

A. DATA EXTRACTION FROM EXTREMISM DETECTION LITERATURE

A thematic diagram was created by studying title, abstract, and full-text of selected work to analyze existing extremism detection literature. Fig. 4 shows every online extremism study, follow themes: Sources, Extremist datasets, Methodological Trends, and Techniques. These themes are designed by the first author and reviewed by the second, third, and fourth author. Following the theme extraction, information was extracted from the selected literature based on research questions. In Table 4, the data extracted from a few papers is presented.

For RQ1, different datasets used, the size of those datasets, and the authors' label were found out. For RQ2, literary works on online extremism were classified as manual or automatic approaches. This extraction was performed by reading the full text of the literature. Techniques and Algorithms, features and evaluation metrics were extracted for answering RQ3. Data validation methods like inter-rater agreement [37] and validation metrics like Cohen's Kappa [44] were studied to answer RQ4. Any other data validation techniques used by studies were also looked for. For RQ5, tools or frameworks that are available for extremism detection were searched. To answer RQ6, title, abstract, and author keywords were extracted from the selected studies.

Detailed discussions of themes, analysis of extracted data, and their importance in extremism detection research are summarized in the following sections. The existing detection techniques are divided into Network or Graph-based, Machine Learning-based, and Deep Learning-based, as shown in Fig. 4.

B. NETWORK OR GRAPH-BASED APPROACH

Network or Graph-based techniques are used in various applications like Social Network Analysis, Pattern Identification,

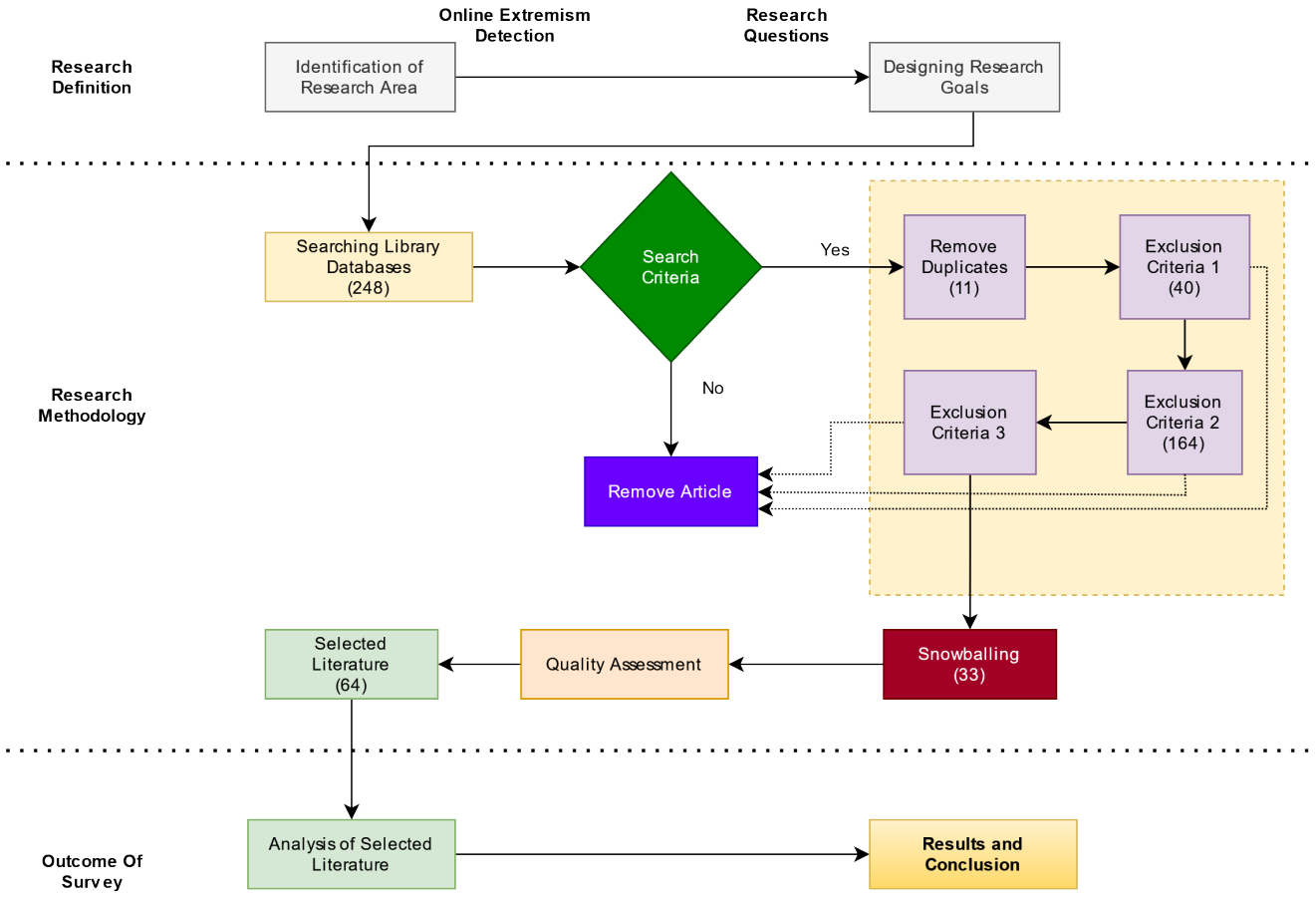
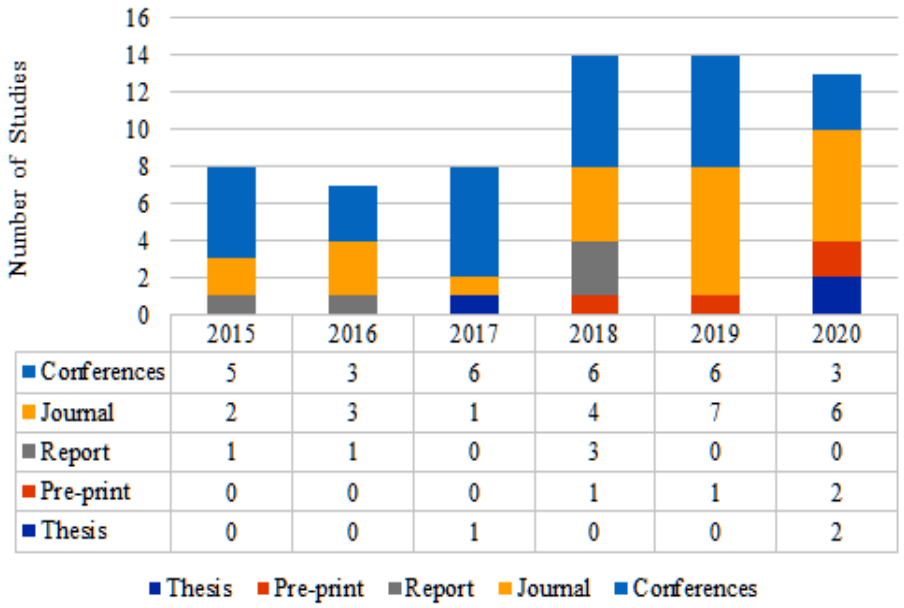


FIGURE 2. Systematic literature review process.



	2015	2016	2017	2018	2019	2020
Conferences	5	3	6	6	6	3
Journal	2	3	1	4	7	6
Report	1	1	0	3	0	0
Pre-print	0	0	0	1	1	2
Thesis	0	0	1	0	0	2

FIGURE 3. Types of studies and year.

Sentiment Analysis, and Text Classification. Graph algorithm shows the interconnection between different entities. Researchers leverage this property of graph techniques to identify the extremists and their interconnections on the social network [50]. Fig. 5 provides the details of the Network or Graph-based approach in extremism detection research.

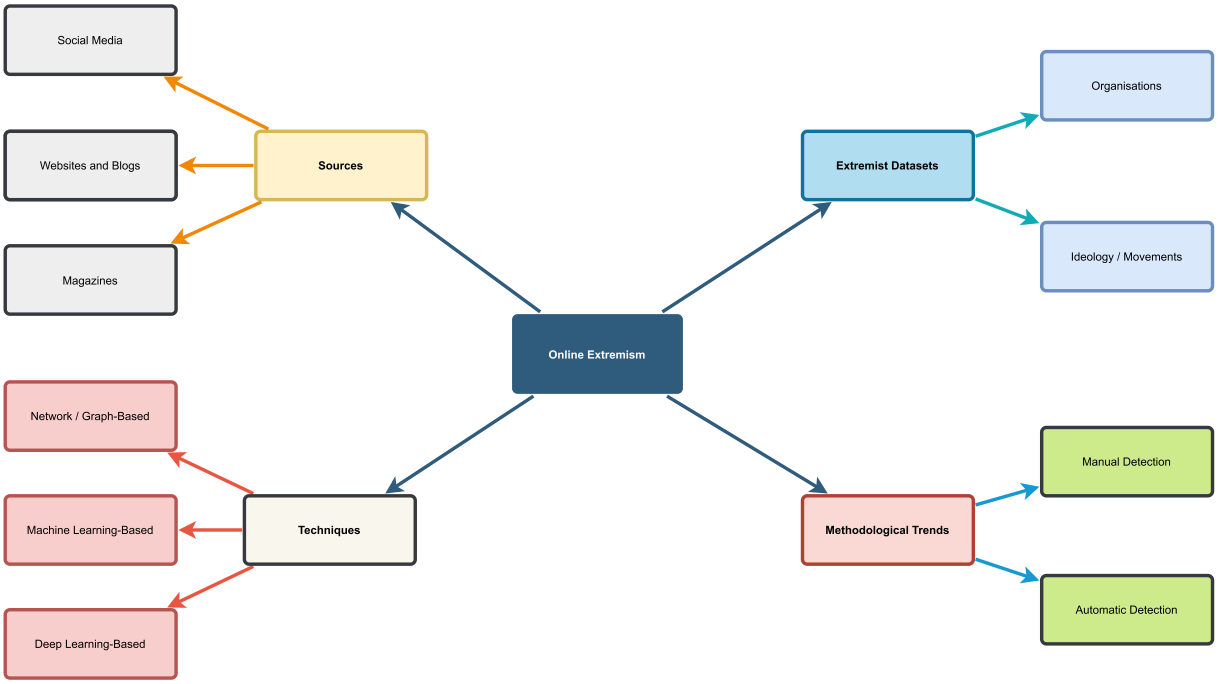


FIGURE 4. Thematic diagram of online extremism.

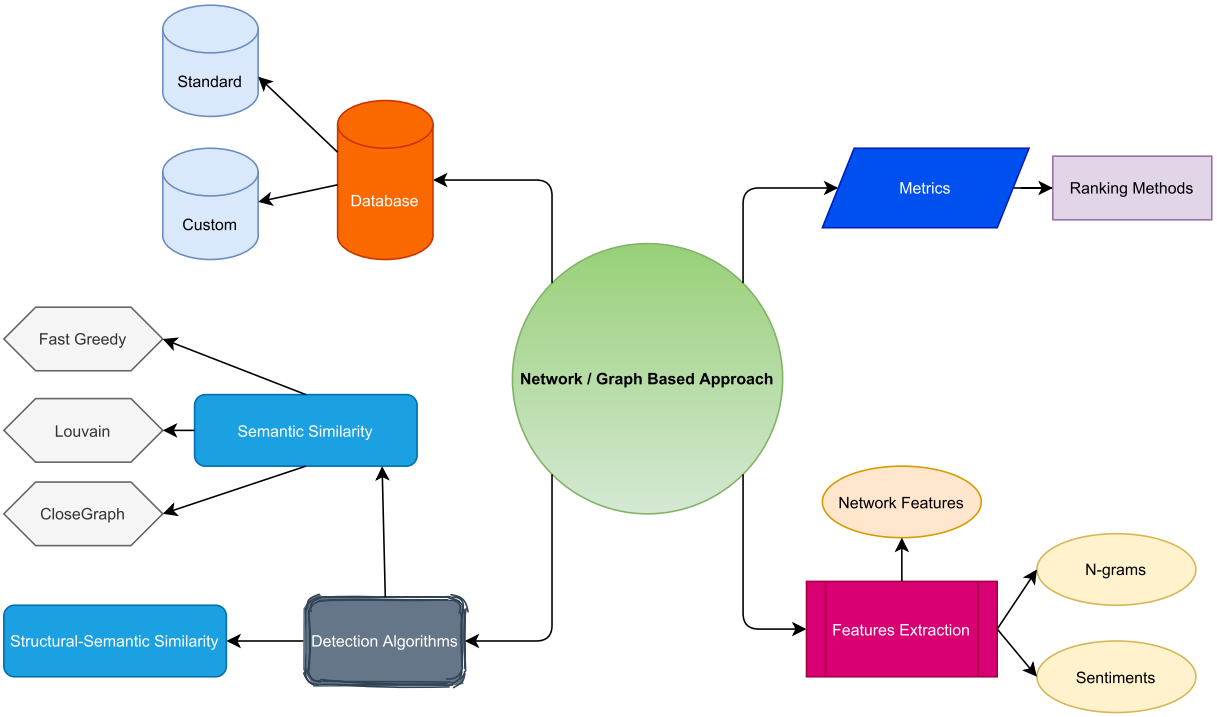


FIGURE 5. Network/ graph-based approach in online extremism detection.

A graph-based approach is used in extremism detection on standard datasets like Lucky Troll Club dataset [51] or custom datasets collected by studies [52]. Graph-based techniques rely on similarities between subgraphs for an accurate detection and prediction of data. Semantic similarity and structural similarity are frequently used network or graph-based techniques in online extremism detection research. As shown in Fig. 5, in semantic similarity, a node represents entities and edges represent concept [53]. In extremism detection using semantic similarity, nodes are social media users,

TABLE 3. Inclusion and exclusion criteria.

Criteria No.	Topic	Inclusion Criterion	Exclusion Criterion
1	Extremism Datasets	Creation of own custom dataset or use of standard datasets related to the extremist organization. Creation or use of datasets mentioning ideology.	Studies having Dark Web as data sources. Papers show lesser emphasis on information about datasets like labels and annotation methods.
2	Detection and Classification Techniques	Studies on trends in online extremism detection included. Works on AI-based techniques included.	Projects or initiatives emphasizing the social and economic aspects of extremism are excluded.
3	Extremism Detection Tools	Tools emphasizing extremism data collection and detection are included.	-

TABLE 4. Data extracted from selected studies.

Authors	Study	RQ1			RQ2	RQ3			RQ4		RQ5	RQ6
Extracted Data		D	S	L	M	TA	F	EM	VM	DVM	T	TAA
Araque *et al.*	[45]	Y	Y	Y	Y	Y	Y	Y	N	N	N	Y
Deb *et al.*	[46]	Y	N	N	Y	Y	Y	Y	N	N	N	Y
Heidarysafa *et al.*	[47]	Y	Y	N	N	Y	N	Y	N	N	N	Y
Ahmad *et al.*	[48]	Y	Y	Y	Y	Y	Y	Y	N	N	N	Y
Kaur *et al.*	[49]	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y

Y= Yes; N= Not Available; D = Datasets; S = Size; L = Labels; M = Methods; TA = Techniques and Algorithms; F = Features; EM= Evaluation Metrics; VM = Validation Method; DVM = Data Validation Metric; T = Tools; TAA = Extracted Title, Abstract and Author Keywords;

and edges represent communication between the users [54]. Nodes with similar subgraphs or partitions are considered as structurally similar [55]. Different algorithms like Fast Greedy, Louvain, and Close Graph are also used in extremism research. Table 5 gives a comparison between graph algorithms. The similarities of the graph are compared with different ranking methods as *betweenness centrality, proximity prestige, in-degree centrality (Refer Glossary).*

The similarities obtained by comparing subgraphs are used to find extremists and their communities [50]. Some researchers use graph techniques to extract semantic and structural similarities from text datasets for extremism. These are used as input features to ML algorithms for extremist detection. Some researchers also use different features in addition to the features obtained by the graph. They are as follows:

- *Sentiment features*–Sentiment is conveyed by words. They can be positive, negative, and neutral.
- *N-gram*–Sequence of n number of words. They can be Uni-gram (1 word), Bi-gram (2 words), and Tri-gram (3 words) depending on the value of n.
- *Network features*–They are different from features obtained by the graph. They are obtained from an interconnection in social media. These features are number of followers, number of following, number of hashtags, number of mentions, a profile description, and geographic location.

Advantages of Network/Graph-Based Approach:

- Graphs can have weighted nodes and edges, and their comparison ensures accurate text classification [59].
- Graphs represent structural relations, which are sometimes better than vector representations [60].

Disadvantages of Network/Graph-Based Approach

- Graphs rely on connections between nodes, but classification suffers, if there are no connections [60].
- Graphs lack an understanding of the meaning of textual representations that is, they consider the whole sentence as a unit for graph representation instead of individual words [60].

The researchers employed ML techniques to eliminate the drawbacks of network/graph-based approaches.

C. FEATURES

Features can be either extracted or reduced. Most of the time, ML or DL prefer feature extraction using Bag of Words, Term Frequency – Inverse Document Frequency (TF-IDF), and Word2Vec.

- *Bag-of-Words (BoW)*: BoW is a document representation in which the frequency-based feature vector is generated from tokens in documents, regardless of their position in it [61]. It is used to get quick results in an ML-based approach.

TABLE 5. Graph based techniques in extremism detection.

Techniques	Study	Working	Advantages	Disadvantages
Fast Greedy	[56]	Find Shortest or Optimal Paths between Nodes	The low time complexity for smaller graphs.	Inefficient for a large number of nodes. Fails for disconnected nodes.
Louvain	[57]	Two-step operation, that is, Optimal Path Finding and aggregation of nodes on the optimal path. Determines which nodes have a dense connection with each other.	Efficient in finding small communities.	Inefficient for a large number of nodes or communities. Fails for disconnected nodes. Can give a large number of frequent subgraphs.
CloseGraph	[58]	Use right-most extension and early termination of graph traversal. The researcher selects early termination criteria.	Early termination results in small number of important frequent subgraphs.	In some situations, early termination cannot be applied.

- *Term Frequency–Inverse Document Frequency* (TF-IDF): TF-IDF is a statistical measure to determine, how word is relevant to document in the corpus [62]. This works in two steps: first, counting occurrences of the word in the document, and second calculating the inverse of the number of documents in which the word appears. TF-IDF is frequently used in ML or DL-based approach as it determines how relevant the word is in a particular document.
- *Word2Vec*: Word2Vec represents words or tokens in distributed vector representation, preserving syntactic, and semantic word relationships [63]. ML and DL based methods use Word2Vec model, as it stores relationships between words.
- *Global Vectors* (GloVe): GloVe is an unsupervised method for word embedding, where vector representations are obtained by the semantic similarity between words [64]. GloVe finds probabilistic co-occurrences of words within a corpus. Concerning other word embeddings, GloVe helps in parallel implementation of the model. Thus, more data can be trained in less time.

Feature reduction is used to reduce the number of features or variables to take less computation time. Feature reduction also ensures only features that have a significant impact on outputs, are selected. In extremist detection research, two features reduction algorithms are used:

- *Principal Component Analysis* (PCA): PCA is a technique for dimensionality reduction for large datasets, such that there is minimum information loss [65]. Thus, removing uncorrelated data. Outliers present in large datasets can affect PCA dimensionality reduction.
- T-Distributed Stochastic Neighbour Embedding *(t-SNE)*: t-SNE is a non-linear dimensionality reduction technique that used probabilistic approaches to reduce high dimensionality data to low dimensionality data [66]. This is achieved by calculating conditional probabilities between points and their neighbours. Due to the probabilistic approach, outliers do not affect t-SNE as much as they affect PCA's outcome.

D. MACHINE LEARNING -BASED APPROACH

Machine Learning is a subset of Artificial Intelligence that learns automatically from provided data and improve the assessment, without being explicitly programmed. The research on online extremism text analysis, an ML-based approach, is used to detect extremist content.

Fig. 6 shows the steps involved in using the ML-Based approach in extremism detection research. Features are extracted or reduced using methods, as described in Section III(C). Machine Learning techniques allow selecting different features like semantic, lexicon, and emotions [45]. In semantic features selection, words are related to each other is found. Lexicon features are the vocabulary words associated with Twitter lexicons such as hashtags, retweets, and mentions. Emotions or sentiment features involve identifying the emotion conveyed by the words, that is, positive, negative, or neutral.

Logistic Regression, SVM, Adaboost, Random Forest, and XGBoost are used to classify the extremists. Logistic Regression is used in studies presenting a binary classification of extremism [67]. Algorithms like SVM, Adaboost, Random Forest, and XGBoost are used in multiclass classification of extremism [68], [69]. Most studies make use of supervised learning, that is, use pre-labelled dataset [67], [70]. Researchers use both standard datasets like ISIS Kaggle data or collect data to create the custom datasets [45], [71]. Researchers also use unsupervised learning to identify and classify the extremists using clustering algorithms [72]. For evaluation of these algorithms, different metrics are employed, some metrics require confusion matrix and terms like True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN). Most used evaluation metrics are *ROC-AUC Curve* [73], [74], *Precision* [75], [76], *Recall* [77], *Accuracy* [76], and *F1-measure / F1-Score* [76] (Refer Glossary).

Disadvantages of Machine Learning Approaches for online extremism detection:

- Heavy reliance on feature extraction, reduction, and selection to obtain good performance [78].

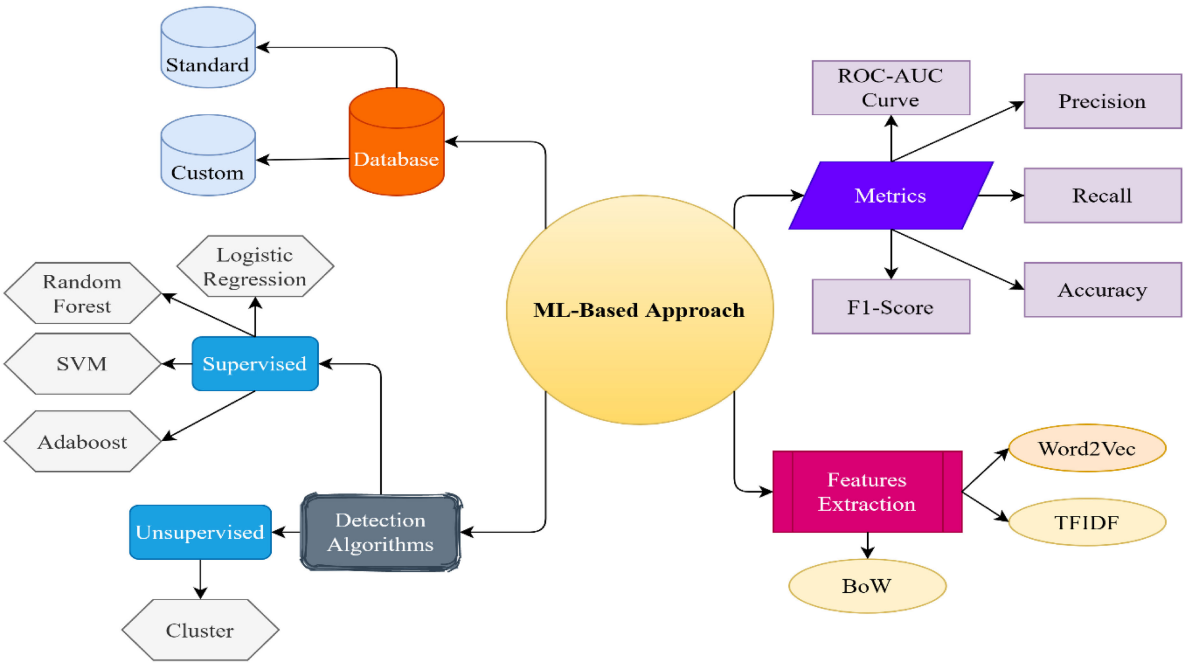


FIGURE 6. Machine learning-based approach in online extremism detection.

- Cannot take advantage of extensive data due to pre-defined features [78].
- Context analysis is a challenge using Machine Learning algorithms [79].

These problems of Machine Learning techniques are overcome by using Deep Learning-based (DL-based) techniques.

E. DEEP LEARNING -BASED APPROACH

Deep Learning techniques are a subset of Machine Learning which uses Artificial Neural Networks for computational tasks. Deep Learning relies on layer upon layer of data training to identify successfully the complex patterns [80]. In Fig. 7, the details of the DL-based approach as presented with respect to extremism research. The DL approach is similar to the ML approach. Here the difference lies in detection algorithms and models used. Deep Learning consists of different models like Feed Forward Network (FFN), CNN, LSTM, and Transformers. LSTMs are complex neural networks that take the sequence of inputs and outputs sequence, while considering contextual information [81]. LSTMs are primarily used in machine translation, speech recognition, etc. CNN takes input data that has a grid pattern like images [82]. So, CNN has been successfully applied in areas of face recognition, medical image analysis, etc. In recent years, 1-D CNN is used in text classification [83], sentiment analysis etc. [84]. The DL approach used two types of models, first the regular model, that is, models trained from scratch using training data, and second pre-trained models, i.e., models trained on data or features extracted from the same domain. In regular models, techniques like LSTM, that store long-term dependencies are used in extremism research [49]. CNN is also used in the extremism detection and classification [85]. CNN is also used in combination with LSTM for emotion-based extremism detection [48].

Works using the pre-trained model in extremism detection extract discriminatory features like hashtags used by extremists, frequently occurring words in the extremist corpus, etc. [86]. Then BERT is pre-trained on this discriminatory feature corpus. This pre-trained model is used for the classification of data collected by the researchers. Some studies using the DL approach collect their custom dataset for extremism detection [49], [48], [85]. Standard datasets like the Stormfront dataset are used with custom data for an accurate extremism detection [86]. DL-based approach uses a similar evaluation metric to that of the ML approach for the classification problem. Most DL libraries provide an in-built embedding layer for feature extraction [87], [88].

A comparison of all three techniques is provided in Section IV. For the comparison, criteria like features, algorithms, classification, metrics, performance, and validation are considered.

IV. OUTCOME OF SURVEY

Based on the methodology described above, the study of the literature is categorized into the following focus areas.

A. METHODOLOGICAL TRENDS IN ONLINE EXTREMISM DETECTION

Online extremism detection has been evolved with the advent of emerging areas like Artificial Intelligence. However,

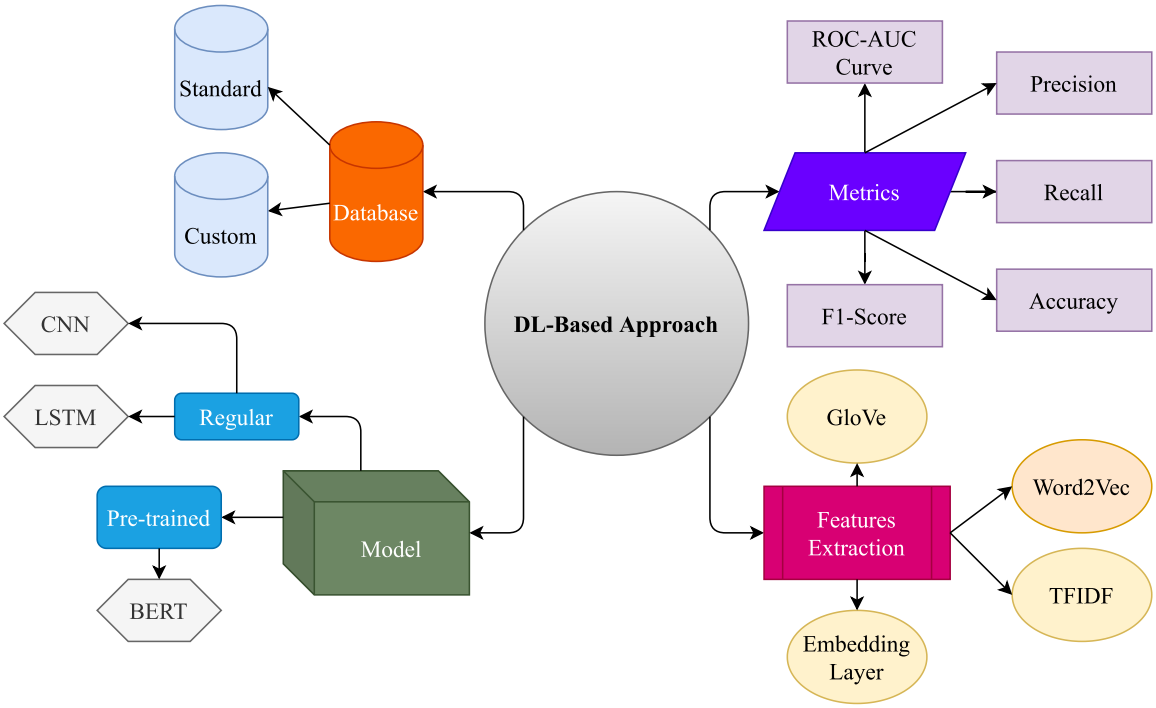


FIGURE 7. Deep learning-based approach in online extremism detection.

the initial works on extremism detection depend on perception of the expert and require manual identification

1) MANUAL DETECTION

Few studies like [18] rely on the manual detection of online extremism. Researchers and experts identify the spread of extremist views, by manually analyzing the social media accounts. The researchers manually identified online extremism raised by different organizations [18] and also within different ideologies [6]. This is time-consuming and inefficient as experts have to annotate each extremist text manually.

Chatfield *et al.* [89] analyzed and classified extremist communication into propaganda, radicalization and recruitment on Twitter. The authors referred to different works and articles that identify *@shamiwitness* as a Twitter user and ISIS supporter. The authors classify these tweets manually as propaganda, radicalization, and recruitment. To analyze the communication network of @shamiwitness, the authors use multi-sided network graphs (interactions between *@shamiwitness* and other Twitter users). Thus, multiple subgraphs of *@shamiwitness* communications with other users are created. The authors, then manually classified tweets based on keywords, phrases, hashtags, and religious references into propaganda, radicalization, and recruitment.

Another report by Berger [6] titled *'Alt-Right Twitter* Census' analyzes Alt-Right a right-wing online organization, and its spread on Twitter. The authors manually identified 27,895 Twitter user accounts supporting the far-right extremist group. The authors conduct manual coding of tweets by analyzing keywords, phrases, hashtags, and linguistics. This report identifies accounts based on user content like pro-Trump content, White nationalist content, general far-right content, anti-immigrant and anti-Muslim content, trolling and conspiracy, and fake news after manual coding. The report considers tweets, retweets, and mentions of users to assign the influential tag to users.

Berger identifies and compares neo-Nazis and ISIS based on the ability to spread online extremism in the report *Nazis vs ISIS* [90]. Berger identified 25,406 user accounts active on Twitter, which relate to white nationalism or Nazi supporter. These accounts are manually labelled as a white nationalist or Nazi sympathizer based on specific keywords. For comparison, 4,000 ISIS supporting accounts are identified in the report. Online extremism of both ideologies is compared on metrics like friends and followers counts, account suspensions, and average tweets per day. Top hashtags, tweet type (tweet, retweet, etc.), and suspensions are also considered.

2) AUTOMATED DETECTION

Manual identification suffers from multiple shortcomings as listed here:

- Its time consuming.
- The user accounts detected or analyzed, which are extremists, are smaller in numbers [89].
- No mechanism to detect new extremist accounts if the identified ones are suspended [6], [18].
- Considers only handpicked features [90].

Automated detection is a novel way to detect the online extremism. Automated detection overcomes mentioned

shortcomings of manual detection. Automated detection relies on the training of previously collected datasets, so time consumption is reduced. These trained models can be reused on unknown data to identify similar extremist accounts. Automated detection methods use the different features collected from large datasets. These features are usually divided into Data-dependent and Data independent features. Data-dependent features include entities like hashtags, urls, emoji, n-grams, and phrases. These features change as domain changes, like hashtags for extremism, are different from hashtags for movie-reviews. Data independent features include time-features, emotion words, and stylistic features. Data independent features combined with data-dependent features give better accuracy.

Automated detection covers different algorithms ranging from social network analysis to Deep Learning. For automatic identification, few pre-identified user accounts are selected as 'seed' accounts. More details about 'seed' data collection are given in Section IV(B). Some studies focus on a network or graph-based approach for automatic extremism detection.

Some research like [91] and [92] focus is on statistical analysis of radicalization, performed on multiple ISIS datasets. The authors extract using automated natural language processing. The authors use TF-IDF to identify important words within ISIS corpus. From these words, multiple topics like Jihadism, condemning Western civilization, Negativity and Swearing are identified. The authors use statistical analysis like data correlation, and Wilcoxon Signed Rank Test to provide details about the type of behaviour, emotions and different topics in ISIS datasets.

A study by De Bruyn [93] determine extremism based on behaviour profiles like ultra-peripheral, peripheral, satellite connector, kinless, provincial, connector and global. These behavioural profiles are collected using simple graphs. The author uses Multinomial Naïve Bayes to classify collected extremist data into these behaviour profiles

Gialampoukidis *et al.*[50] use five Arabic keywords to collect the data from Twitter, which gives a total of 4400 user accounts and 9,528 Twitter posts. The authors use a network/graph-based approach to automatically identify extremist communities. The authors use different graph-based algorithms like Fast Greedy and Louvain to detect the extremist community size.

Benigni [94] works on ISIS ideology. The author employs a ML-based approach of ensemble clustering algorithms and network or graph-based method with heterogeneous dense subgraph for automated extremism detection. Al-Saggaf [95] used hashtags related to Spam Daweish Army Campaign, which trended on Saudi Arabian Twitter in April 2016. The authors collected these exposed ISIS supporter' tweets and other metadata. The authors used network graphs and topic modeling to analyze interaction between the extremists.

Moussaoui *et al.* [55] collect data based on terms and keywords associated with ISIS. Authors use unknown ISIS supporters and gather the data about them and their followers. Researchers use a network/graph-based approach for an automated extremism detection. The authors extract features like semantic similarity, structural similarity, and possibilistic similarity using possibilistic graph-based approach. The possibilistic approach automatically clusters user accounts into terrorists, terrorist sympathizers, and non-supporting types. The authors state that possibilistic graph-based approach automatically detects extremist communities better than *GRAMI* (Graph Mining) and *gSpan* graph algorithms. More network or graph-based studies are discussed in Section IV(C) and Section IV(D).

Most research uses Machine Learning as the primary technique for automatic detection of online extremism as no explicit programming is needed for pattern identification. A study by De Bruyn [93] determine extremism based on behaviour profiles like ultra-peripheral, peripheral, satellite connector, kinless, provincial, connector and global. The author uses Multinomial Naïve Bayes to classify collected extremist data into these behaviour profiles.

Xie *et al.* [96] use the seed method to collect additional data for extremism detection. Hashtags related to a particular ideology are used to collect data from Twitter. A total of 4,820 unique ISIS supporting user accounts were collected. Tokens from hashtags, expected hitting, and harmonic closeness was used as features for extremism detection. The study used a graph-based method to extract features like expected hitting time and harmonic closeness. For automatic extremism detection, ML-based approach is used by this research work. Researchers use Adaboost algorithm for automated detection and classification of the candidates into ISIS-supporting and ISIS non-supporting social media users.

Ashcroft *et al.* [69] use hashtags to collect and identify ISIS supporting user accounts on Twitter. The study identified a total of 6,729 user accounts. Different features like stylometric, time-based, and sentiment-based features are considered. Stylometric features are words that are most frequent in dataset. Time-based features include hour of day, period of day, week-day and type of day. Sentiment-based features are classified as very negative, negative, neutral, positive, very positive. This study used ML-based approach for automated extremism detection. Researchers use SVM, Naïve Bayes, and Adaboost algorithms for extremism detection.

Kaati *et al.* [97] used hashtags related to ISIS for data collection. Tweets were collected from a list of known ISIS sympathizers, which were about 6,729 users. Authors extract data-dependent and data-independent features. Data-dependent features are like the most common hashtags, most common word bigrams, most common letter bigrams, and most frequent words. Data independent features consist of stylistic features, time features, and emotion words. Researchers leverage ML-based approach to detect extremism automatically using Adaboost algorithm.

In recent years, Deep Learning Approach is being adopted for automatic extremism detection. Kaur *et al.* [49] used Deep Learning Approach for automatic extremism detection. The data collected by authors is classified as radical, non-radical, and irrelevant type by using expert annotators. Word2Vec is

employed for the extraction of word embedding. Researchers consider LSTM to detect extremism and classified data as radical, non-radical, and irrelevant. Similarly Johnston and Marku [98] used LSTM models to identify extremism from different groups like Sunni Islamic, Antifascist Groups, White Supremacist and Sovereign Citizens. Nizzoli *et al.* [99] use different scenarios of balanced and imbalanced ISIS extremist datasets with Recurrent Convolutional Neural Network, and character-based Convolutional Neural Network.

More details about studies using Automated Identification are described in Section IV(C) and Section IV(D).

B. DATASETS

Most researchers try to collect data as well as use standard available data for extremism detection and analysis. These datasets come from various sources. Following studies detect extremism specific to sources:

1) SOURCES OF DATA COLLECTION: SOCIAL MEDIA

Social media platforms are sources for gathering data on extremism detection. Twitter is widely used social media platform for the collection of data [6], [18], [45], [69], [90]. Multiple ideologies use Twitter as a primary tool for propaganda, radicalization, and recruitment. Datasets include multilingual data (different languages like Arabic, English, and German) which are used by researchers to detect extremism in a particular language [23], [97].

Messaging applications like Telegram [100] and WhatsApp [46] are also monitored for extremism data collection and analysis. WhatsApp is a popular and widely used messaging platform, while Telegram is a similar messaging application but has a small user-base. Both messaging apps are preferred platforms for the extremists as they employ end-to-end encryption, thus securing the communication [101].

ISIS Kaggle data is a collection of tweets of ISIS supporters and is publicly available [102]. In [52], [70], and [103] studies, researchers use ISIS Kaggle Dataset as the primary dataset, and in this study [45] researchers use ISIS Kaggle Data as a seed dataset for collecting the custom data.

Seed data collection researchers select few extremist social media accounts from a standard dataset or identify extremist social media accounts from newspaper articles or previous studies. These samples or extremist user accounts are searched on social media like Twitter. Researchers then collect data like followers, following, friends, posts, tweets, retweets, and mentions [104]. Researchers build datasets by adding followers and friends to extremist accounts and also collect their metadata as well. Researchers also analyze keywords used by extremists in the seed dataset. These keywords are searched on social media. Accounts posting or tweeting with similar extremist keywords are then classified as extremists. Smedt *et al.* [105] use Facebook pages like La Dernière Heure and Le Figaro to collect extremist text. La Dernière Heure and Le Figaro are French newspapers. This dataset of Facebook pages contains about 60,000 public comments, out of which 10,000 were identified as racist content. Tundis *et al.* [106] also uses Facebook for data collection of individuals with features such as age, family, intimate relationships, associations, prison, religion and occupation. These features are then used to predict criminal profiles on Facebook. Mouhssine and Khalid [107] extracts sentiments and other features like leakage (intent to harm), fixation (deep interest in group) and identification (comparison of oneself to extremists) from Facebook posts.

2) SOURCES OF DATA COLLECTION: WEBSITES, BLOGS AND MAGAZINES

The current research on extremism detection and analysis used data from websites, forums, and magazines for data collection. Magazines like Dabiq and Rumiyah are used for detecting ISIS ideology [45], [70]. Dabiq and Rumiyah were published in various languages like Arabic, English, German and French. Forums like Stormfront are used to collect data about the white supremacist movement [85]. Stormfront is a forum dedicated to white nationalists and a popular hate-spreading website. Various accusations like the promotion of hatred and inciting violence are levelled against Stormfront.

The participation of women in the extremist organizations is showing a rising trend [108]. Various groups include women as sympathizers, participants, and perpetrators of violence. A specific blog is used to analyze the psychology of women extremists by Leah Windsor [109]. This manually selected and analyzed blog belonged to Aqsa Mahmood, who later joined ISIS. The author states that this blog is the perfect example of gradual descent of the individual into violent radicalization. Heidarysafa *et al.* gather women-specific articles from Dabiq and Rumiyah [47]. These articles contain topics, keywords, and themes that target women. The main aim of the authors is to compare the language of ISIS, used to radicalize women to language, regularly used by religious entities. Therefore, authors use articles from catholicwomensforum.org for the comparison.

Kapitanov *et al.* [110] collect data from various websites, blogs, forums, chats, and social networking. The collected data contains information from Russian sites referring to extremism content related to some regions in Russia. These sites discuss various political events that took place in Russia, Caucasus, and the Middle East.

Table 7. gives details about different sources from which data was collected. Datasets sometimes have designated names, for example, ISIS Kaggle dataset, Lucky Troll Club dataset. Datasets were classified into two types 1) Standard Dataset and 2) Custom Dataset. The dataset are considered as the standard if they are publicly available and used in multiple studies. The extremist dataset gathered by researchers that is not publicly available is termed as Custom Dataset. Standard extremist datasets are limited to Twitter and Stormfront forum. Researchers have used Twitter, Facebook, YouTube, different websites, forums, and magazines to create custom extremism datasets. The size of datasets depends upon studies. Ferrara *et al.* used over 33,95,901 tweets for automated

TABLE 6. Manual detection in online extremism.

Study	Data Sources	Data Size	Approach	Ideology	Technique
[89]	Twitter	Single Twitter User	Keywords, Phrases, Hashtags	Jihadist / ISIS	Mulit-sided Network Graphs
[18]	Twitter	20,000 Twitter Users	Keywords, Phrases, Hashtags. Multiple Users	Jihadist / ISIS	SNA
[6]	Twitter	27,895 Twitter Users	Keywords, Phrases, Hashtags. Multiple Users	White Supremacist	-
[90]	Twitter	25,406 Nazi Supporters, 4000 ISIS Supporters	Multiple Users	Jihadist / ISIS, White Supremacist / Nazi	-
[51]	Twitter	Deprecated	Crowd Sourcing	Jihadist / ISIS	-

classification, while Chatfield *et al.* only used 3039 tweets for analysis after manual detection of the extremism.

Comparison of datasets are made on three criteria that are size of dataset, classes balanced or imbalance and ideology. Some studies have used graph or clustering techniques for the detection of extremist communities. Therefore, their datasets do not contain any labels. Standard datasets like ISIS Kaggle Dataset and Lucky Troll Club also does not have any classification labels as all the tweets are considered extremists. Analytical studies only analyze extremist data; therefore, researchers did not give them any labels. Most of the standard datasets have only positive occurrences of extremism text, which leads to class-imbalance. The negative views are also collected to balance the class.

There are few problems in standard datasets. Standard datasets are older, and many of the user accounts present in standard datasets are suspended. Thus, accessing user accounts that are suspended is impossible. By creating custom datasets, researchers can collect recent extremist user accounts. The language, related to extremism change over time, but most of the social networks identify usual extremist keywords and phrases to suspend the users. So, creating a custom extremism dataset ensures that the dataset contains the latest extremist keywords, phrases, and recent events. These standard datasets or custom datasets are collected from different sources. Most researchers focus on Twitter as a primary source for extremist data collection. Forty-three studies use Twitter as the primary source from the current surveyed literature, 2 studies use YouTube as the primary source. Three research uses Facebook as the primary source, and 5 studies used magazines as the extremism text source.

Table 7, Table 8 and Fig. 10 give details about sources and number of studies using these sources. It can be observed that some studies use multiple sources to collect extremism data. Studies using social media applications like WhatsApp are counted as social media. Few research works mention Stormfront as forum and while some studies mention it as a website. For the sake of uniformity, Stormfront is considered as a website. Standard or custom data is collected based on organization or ideologies. Most research works collect extremism data related to ISIS organization. There are only four standard publicly datasets, with two for ISIS [51], [102] and one for Right-Wing White Supremacist ideology [85]. A total of three investigations collect the data from Twitter for Right-Wing or White Supremacist ideology [23] [86], [118]. Standard datasets or custom datasets have challenges, as shown in Fig. 11.

The challenges in extremism datasets are as follows:

- No availability or verification of data if social media remove user accounts or posts.
- Manual validation with the help of experts is tedious and time-consuming.
- The size of the positive sample, that is, extremism data, is far less than the negative sample, that is, neutral data. This leads to data imbalance and further leads to errors in classification.
- Data is collected based on specific extremist organizations, ideologies, movements or events.
- Few experts validate datasets with the low inter-rater agreement.

C. STRATEGIES, FEATURES AND ALGORITHMS

Existing literature on extremism is classified into Detection or Analysis. The content of this literature is subdivided into strategies, feature extraction/feature reduction, feature selection methods, and algorithms. Features extraction, reduction, and selection form an integral part of extremism detection and analysis research. This section aims to identify the surveys based on techniques used to detect or analyze extremist research. The comparison of techniques is one of the goals of this section.

Network or Graph-based is used to address the interconnections and behaviour of extremist in social media. As mentioned in Section III(B) and Section III(D), ML approach is preferred by researchers for detection and classification of extremism texts. As seen in Table 9, most of the surveyed studies focus on detection strategies. Most studies focus on the use of ML-based techniques for extremism detection. Different algorithms like Logistic Regression, SVM, Random Forest, Adaboost and XGBoost are used to classify data. ML algorithms depend on features for pattern identification, therefore diverse feature selection, feature extraction and feature reduction methods are used. The most popular feature extraction methods are TF-IDF, Word2Vec, and GloVe. Many studies employ features selection criteria as hashtags, keywords, sentiment or emotional features, lexical features and semantic features. As reasons stated in Section III(D) and

TABLE 7. Extremism datasets and their sources.

Dataset / Author	Study	Dataset Type	Source	Positive Labels	Negative Labels	Size / Total Labels	Ideology
ISIS Kaggle Dataset	[102]	Standard	Twitter	No Label	No Label	17000 tweets	Jihadist
Agarwal *et al.*	[111]	Custom Dataset	YouTube	Only Extremist Label	Only Extremist Label	35 YouTube Channels	Jihadist
Lucky Troll Club	[112]	Standard (Deprecated)	Twitter	No Labels	No Labels	Variable Size	Jihadist
Demographics Dataset	[18]	Custom Dataset	Twitter	Only Extremist Label	Only Extremist Label	20,000 Twitter Accounts	Jihadist
Demographics Dataset	[90]	Custom Dataset	Twitter	Only Extremist Label	Only Extremist Label	8000 Twitter Accounts	Jihadist, Nazis
Alt-Right Dataset	[6]	Custom Dataset	Twitter	Only Extremist Label	Only Extremist Label	27,895 Twitter Accounts	White Supremism
Benigni *et al.*	[72]	Custom Dataset	Twitter	No Labels Available	No Labels Available	119,156 Twitter Accounts	Jihadist
Chatfield *et al.*	[89]	Custom Dataset	Twitter	No Labels available	No Labels available	3039 Tweets of @shamiwitness	Jihadist
Gialampoukidis *et al.*	[50]	Custom Dataset	Twitter	No Labels Available	No Labels Available	9,528 Tweets	Jihadist
Heidarysafa *et al.*	[47]	Custom Dataset	Magazines and Website	No Labels Available	No Labels Available	20 articles from Dabiq and Rumiyah, 132 articles from catholicwomensforum.org	Jihadist
Stormfront Dataset	[85]	Standard	Forum	1,119 Hate Label	8,537 non-Hate	10,568 messages	White Supremism
TW-PRO, TW-RAND, TW-CON	[69]	Custom Dataset	Twitter	ISIS supporting. No label count given.	Random label, ISIS opposing. No label count given	7500 Tweets	Jihadist
Fernandez *et al.*	[113]	Custom Dataset	Twitter	17350 pro-ISIS [102]	197,743 non-ISIS.	2,150,93 Tweets	Jihadist
Rowe *et al.*	[114]	Custom Dataset	Twitter	602,511 pro-ISIS	1,368,827 non-ISIS	1,971,338 Tweets	Jihadist
Araque *et al.*	[45]	Custom Dataset	Twitter and Magazine	[113], [114] , 316 Radical Articles.	[113], [114] , 152 Neutral Articles	[113], [114] , 161 articles from Dabiq and 155 articles from Rumiyah , 129 articles from CNN, 23 articles from The New York Times	Jihadist
TW-PRO-E, TW-RAND-E, TW-PRO-A, TW-RAND-A	[97]	Custom Datasets	Twitter	27753 English Pro-ISIS, 16000 Arabic Pro-ISIS,	60000 English Non-ISIS, 45,013 Arabic Non-ISIS	87,753 English Tweets, 61,013 Arabic Tweets	Jihadist
Abrar *et al.*	[68]	Custom Datasets	Twitter	13,369 Terrorism Supporting.	16,506 Terrorism non-supporting, 38,617 random	55,123 Tweets	Jihadist
Ahmad *et al.*	[48]	Custom Dataset	Twitter	12,754 Extremist	8432 Non-extremists	21,186 Tweets	Jihadist
Asif *et al.*	[115]	Custom Dataset	Facebook	5279 Moderate, 6912 Highly Extreme, 2991 Low Extreme.	4315 Neutral	19,497 Facebook Comments	Jihadist
Charles	[116]	Custom Dataset	YouTube	41 White Supremacist Supporting	39 Non-Supporting	80 YouTube Channels	White Supremism
Ferrara *et al.*	[117]	Custom Dataset	Twitter	25,538 as positive label	25,000 as negative label.	i) 3,395,901 tweets with 25,538 ISIS accounts ii) 29,193,267 tweets with 25,000 selected as users exposed to ISIS	Jihadist

TABLE 7. *(Continued)* **Extremism datasets and their sources.**

Dataset / Author	Study	Dataset Type	Source	Positive Labels	Negative Labels	Size / Total Labels	Ideology
Hartung *et al.*	[118]	Custom Dataset	Twitter	15,911 Right-Wing	29,836 Non-Right Wing	45,747 Tweets	White Supremism
Jaki *et al.*	[23]	Custom Dataset	Twitter	50,000 Right Wing Extremist	50,000 Safe Tweets	100,000 Tweets	White Supremism
Fraiwan *et al.*	[119]	Custom Dataset	Twitter	10,793 Tweets as violence advocating	10,397 benign, 2,680 unrelated	23,870 Tweets	Jihadist
Alatawi *et al.*	[86]	Custom Dataset	Twitter, Forum	2294 White Supremacist	2294 non- White Supremacist.	[85], 1,299 Tweets	White Supremism

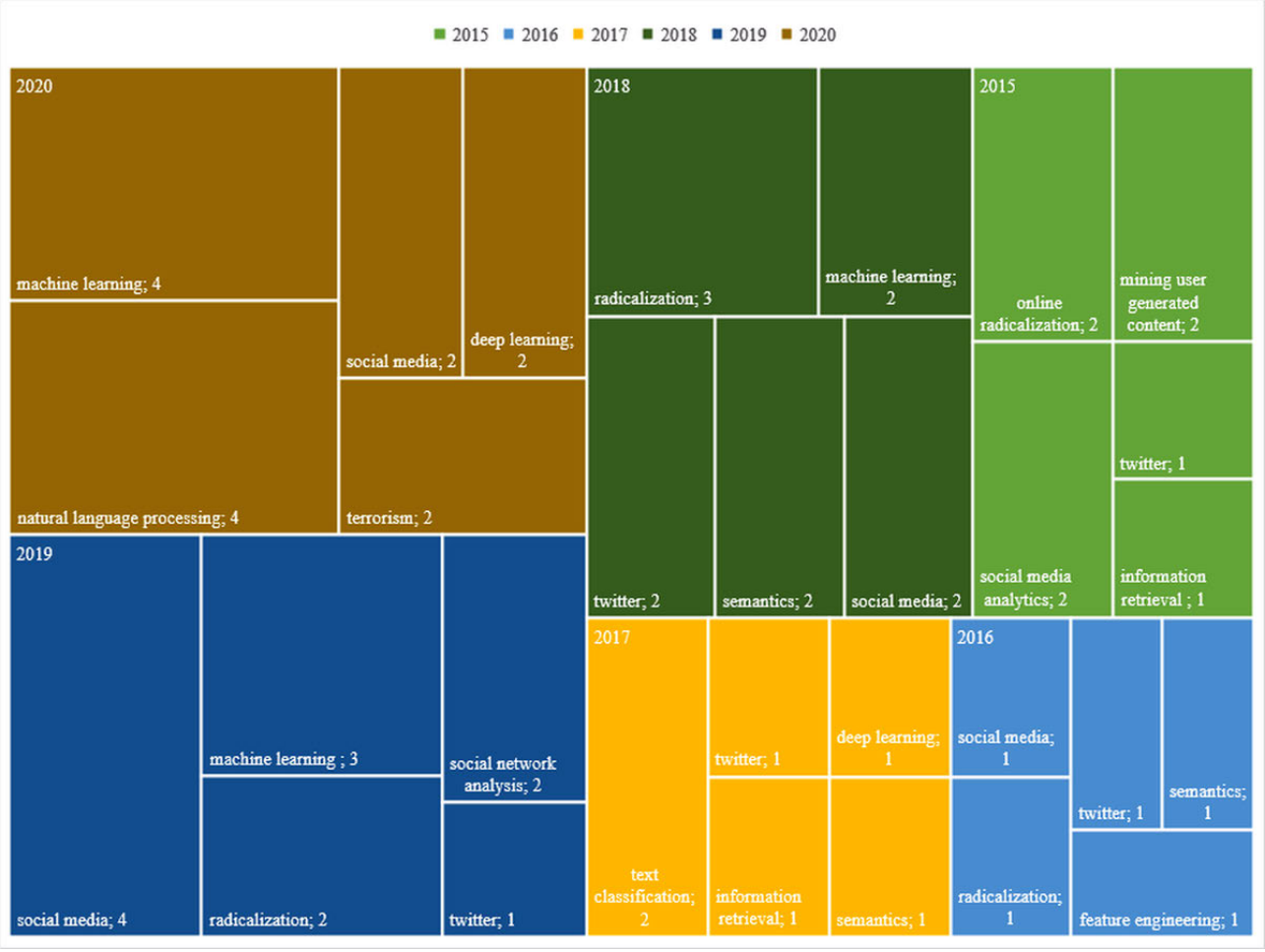


FIGURE 8. **Top keywords related to extremism detection per year.**

Section III(E), Deep Learning approach is gaining traction among researchers for extremism detection. Features considered for classification are usually obtained by GloVe [86] and Word2Vec [86]. The Embedding layer provided in deep learning libraries like Keras, are used by researchers. The Embedding layer converts the text input data to 2D vector and it is given as input to Deep Learning layers [86]. LSTM is the most preferred Deep Learning model in text classification and hence LSTM model is used for extremism detection. BERT is used for extremism detection over LSTM as BERT is better at identifying contextual representations and can be pre-trained on large corpus like Wikipedia [86].

D. CLASSIFICATION, METRICS AND PERFORMANCE

The existing literature on extremism detection used different classification algorithms for online extremism detection. Most extremist detection research classifies extremism texts into binary or tertiary classes. Extremist detection literature

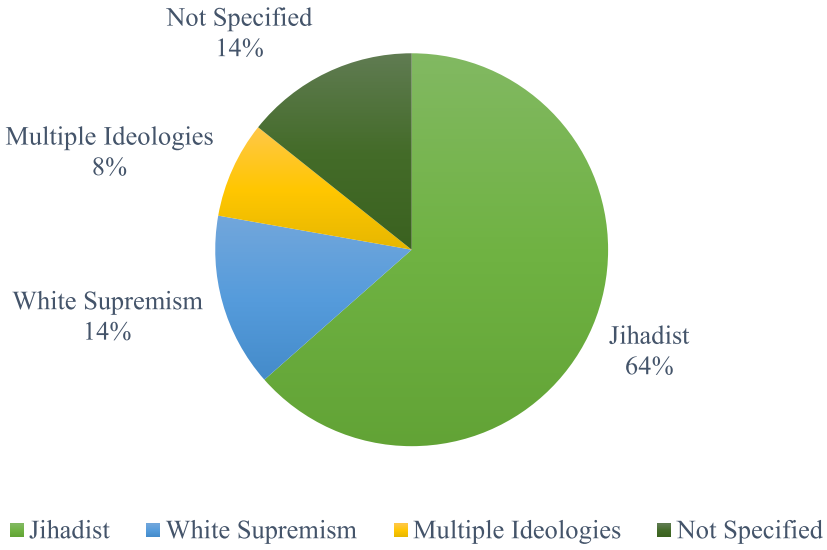


FIGURE 9. Percentage of online extremism research works in particular ideology.

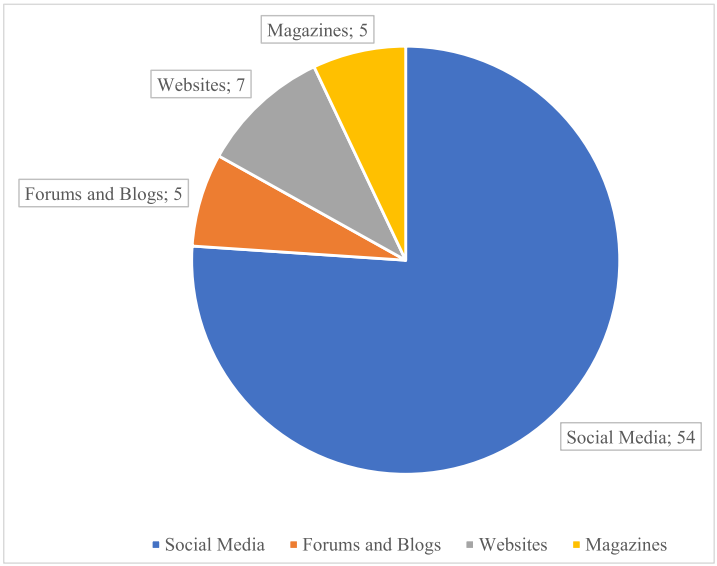


FIGURE 10. Frequency of sources used by articles in extremism detection.

TABLE 8. Sources for extremism data.

Sources of Extremism Data	Number of articles
Social Media	54
Forums and Blogs	5
Websites	7
Magazines	5

defines classes like extremist–non-extremist, radical–non-radical, terrorism supporting–non-supporting etc. The classification algorithms are further evaluated with various performance metrics.

The Section is further subdivided into 3 techniques of extremism detection in the first part followed by a tabulated view of performance metrics of classification algorithms shown in Table 10. The most popular techniques in extremism detection are shown by Fig. 12.

1) STUDIES USING NETWORK OR GRAPH TECHNIQUES

In network or graph-based techniques, some works used simple graph techniques like Breadth-First Search (BFS) [111]

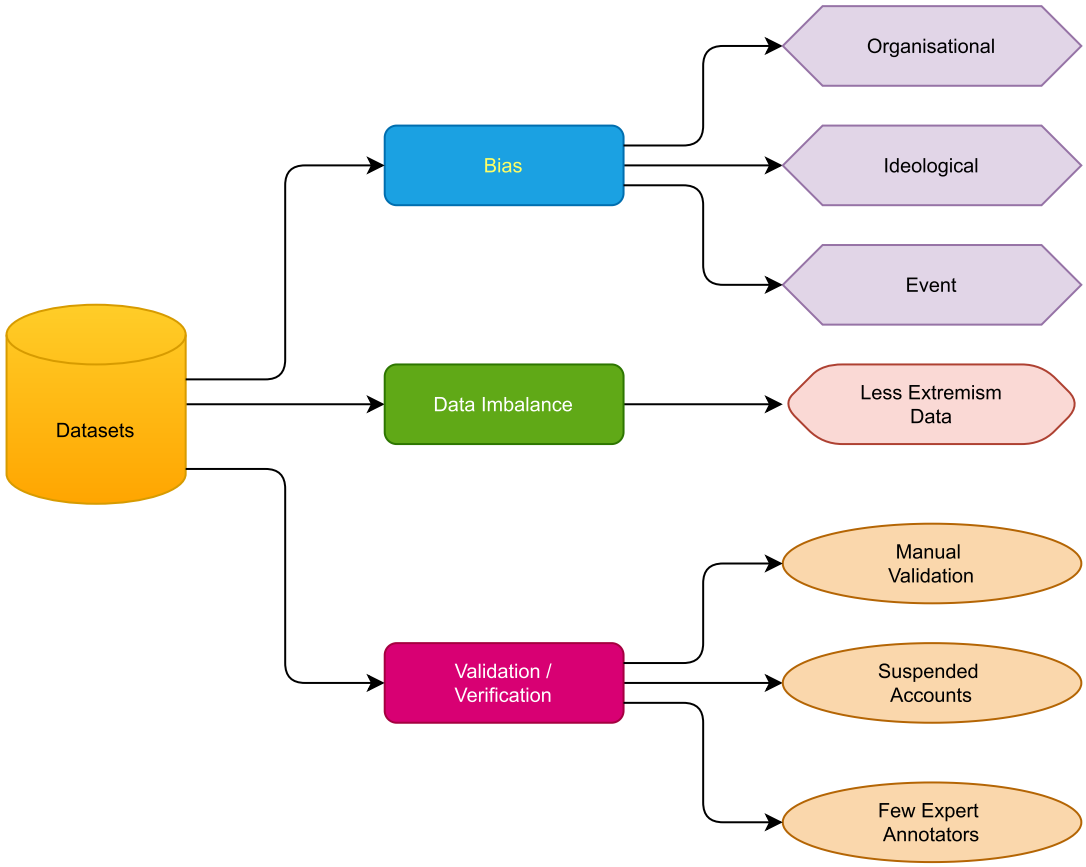


FIGURE 11. Challenges in online extremism datasets.

and Network graph [95], while some use the sophisticated graph algorithms like possibilistic graph [53].

Agarwal *et al.* [111] used simple Best-First Search (BFS) and Shark Search Algorithms (SSA) to find extremist communities and influential users, using data collected from YouTube. The authors classified YouTube users as relevant and irrelevant. The relevant class comprises of extremist content, while the irrelevant class may have different content.

Sophisticated graph algorithms used in extremism detection include semantic graphs, probabilistic graph, Fast Greedy, Louvain, and Close Graph. Saif *et al.* [53] used the semantic graph to extract semantic features and user relations to detect extremism. The article also uses other features like the number of followers, retweets, unigram features, sentiment features and topic features. The author applies Close Graph to extract subgraphs from collected data. These subgraphs are used as features for the classification. Saif *et al.* categorize the data into pro-ISIS and anti-ISIS. The authors use different algorithms for classification, like Naïve Bayes, Maximum Entropy, and SVM. Authors use SVM with different features like unigrams, sentiments and semantic features for the comparison of the result.

Another complex graph algorithm is proposed by Moussaoui *et al.* [55] called Probabilistic Similarity Graph for extremist community detection. In this graph technique, a probabilistic approach is employed to assign, whether the user belongs to the extremist community. The features extracted using probabilistic similarity graph and hybrid structural similarity are used for classification. The authors categorize the data into classes as Leaders of ISIS, Pro-ISIS and Anti-ISIS by using Possibilistic Graph Approach. They use different algorithms like Naïve Bayes, Multinomial Naïve Bayes and Stochastic Gradient Descent classifier (SGD) for classification. SGD classifier uses decision boundary methods efficiently with convex loss function. The performance of Possibilistic Similarity is compared with standard graph algorithms like gSpan and GRAMI. The number of subgraphs obtained by Possibilistic Similarity exceeds gSpan and GRAMI. Some studies explicitly use a network or graph approach to detect online extremism, while some as feature extraction techniques.

Gialampoukidis *et al.* [50] used *degree centrality between nodes* (number of links), *betweenness centrality* (node lies on shortest path of other nodes), *closeness centrality* (average distance between nodes), *eigen vector centrality* (measure of influence of node), *Page Rank*, *mapping entropy* and mapping *entropy betweenness* to identify influence of the extremist user and community surrounding the extremist influencer.

TABLE 9. Strategies in online extremism research: feature engineering and classification techniques.

Year	Studies	Strategies	Technical Approach	Feature Extraction / Reduction Methods	Features Selection	Techniques
2020	[45]	Detection	ML	Word2Vec, FastText, GloVE.	Affect Lexicon	Linear SVM and Logistic Regression
2020	[86]	Detection	DL	GloVe, Word2Vec	NA	LSTM, BERT
2020	[120]	Detection	ML	Word2Vec, TF-IDF	NA	Gradient Boosting, Random Forest
2020	[47]	Analysis	ML	TF-IDF	Emotional Features	Latent Dirichlet Allocation (LDA), Depechemood
2019	[68]	Detection	ML	TF-IDF	NA	SVM and Logistic Regression
2019	[48]	Detection	DL	Embedding Layer	NA	LSTM + CNN
2019	[103]	Detection	Network / Graph	NA	Sentiment	Degree of Centrality, Fuzzy Clustering,
2019	[121]	Analysis	NA	NA	Hashtags, LIWC, Effect Size	NA
2019	[122]	Analysis	NA	NA	Physical Characteristics, Type of Content	Helfstein's Model of Self-Radicalization, Galtung's Model of Peace Journalism
2019	[123]	Analysis	NA	NA	Keywords, Linguistic Features	NA
2019	[55]	Detection	Network / Graph	TF-IDF, Word2vec	Semantic and Structural Similarity	Classic Naïve Bayes, Multinomial Naïve Bayes, Stochastic Gradient Descent Classifier, Possibilistic Clustering, Probabilistic Labelling
2019	[124]	Analysis	NA	NA	Keywords, Linguistic Features, Word Co-occurrences, Particular Word Occurrences	NA
2019	[125]	Detection	ML	TF-IDF, PCA	NA	Naïve Bayes, SVM, Decision Tree, Random Forest.
2019	[70]	Detection	ML	TF-IDF, Word2Vec	Textual Features, Psychological Features, Behavioural Features,	KNN, SVM, RF, and Neural Network
2019	[52]	Detection	ML	Word2Vec	Religious Features, Ideological Features, Hate Features	LDA, t-SNE, Random Forest.
2019	[54]	Detection	Network / Graph	BoW	Node Page Rank, Hub and Authority Measure, Betweenness Centrality, In-Degree Centrality, Sociability	XGBoost
2018	[85]	Detection	DL	NA	NA	SVM, CNN, LSTM
2018	[126]	Detection	ML	NA	Various features depending upon type of classification	Logistic Regression
2018	[6]	Analysis	NA	NA	Hashtags, Followers, Keywords, Locations	NA
2018	[71]	Detection	ML	NA	Trigrams	LibSVM
2018	[127]	Analysis	NA	NA	Keywords, Linguistic Feature, Topic	Directed Content Analysis
2018	[110]	Detection	ML	NA	N-grams, Tone Analysis	Naïve Bayes
2017	[128]	Detection	ML	NA	SentiWordNet	Naïve Bayes
2017	[53]	Detection	ML	NA	Semantic Based Features	Naïve Bayes, Maximum Entropy and SVM.
2017	[72]	Detection	Network / Graph	NA	Hashtags, Followers, Following, Degree of Centrality	Iterative Vertex Clustering and Classification
2016	[117]	Detection	ML	NA	User metadata and activity features, Timing features, Network Statistics	Logistic Regression and Random Forest
2016	[114]	Detection	Network / Graph	NA	Lexical, Sharing and Interactions	Relative Entropy, Adoption Probability
2016	[96]	Detection	ML	NA	Hashtags, Harmonic Closeness, Random Walk	Random Forest, Adaboost.
2015	[69]	Detection	ML	Custom Feature Vector	Stylometric features, time-based features, sentiment Based features	SVM, Naïve Bayes, Adaboost
2015	[97]	Detection	ML	NA	Data Independent Features, Data Dependent Features	Adaboost
2015	[111]	Detection	Network / Graph	NA	NA	Breadth-First Search, Shark Search Algorithm

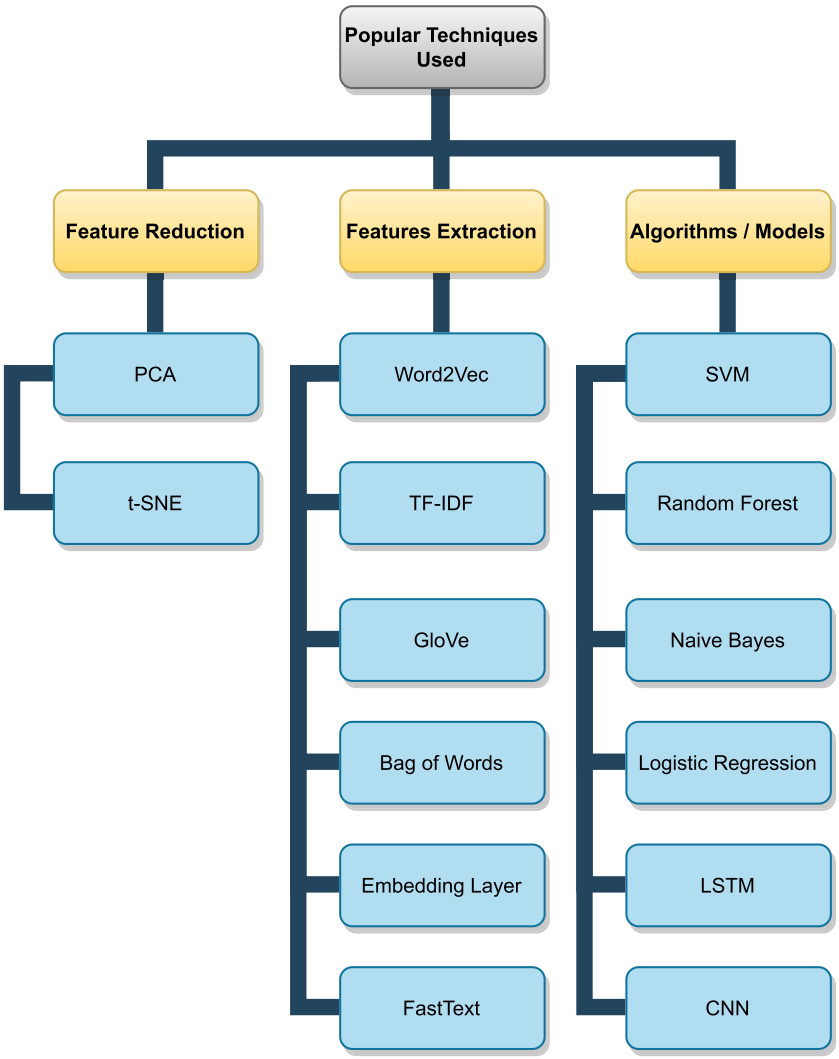


FIGURE 12. Popular techniques in online extremism detection: Feature reduction, feature extraction and classification techniques.

The authors use different graph algorithms like Fast Greedy, WalkTrap, Louvain, InfoMap and DBSCAN to identify the extremist communities. Fast Greedy and Louvain identify 58 extremist communities, using above mentioned graph evaluation methods.

Lopez-Sanchez *et al.* [133] proposes case-based reasoning framework with human expert for validation of identified extremist profiles. The authors use interaction case-base to detect adjacent extremist twitter profiles, with following, tweets, and mentions. The authors use different image descriptors to identify extremist images. The profiles termed extremist by case-based reasoning framework are handed to human expert. These profiles if flagged as extremist by human experts are then selected for monitoring. The authors monitor three extremist organisations as Golden Dawn, ISIS and Hogar Social Madrid.

Similarly, Petrovskiy *et al.* [54] used graph evaluation metrics like Node Page Rank, Hub and Authority Measure, Betweenness Centrality, Proximity Prestige (importance of a particular node in the given domain), In-degree Centrality, and Sociability (the number of incoming edges divided by all edges for vertex) as the features for classification. The authors classified texts as dangerous, safe, and unknown. The authors also use different features like user connections, number of users in the nearest neighbours, and minimal distance to user with and without weights of edges. All these features are input data to the different ML algorithms used.

2) STUDIES USING MACHINE LEARNING TECHNIQUES

Network or Graph methods have different limitations, as mentioned in Section III(B), like the inability to access disconnected data and learning from the semantics of trained data. These requirements are addressed by using Machine Learning approaches with appropriate features selection methods.

TABLE 10. Summary of classification techniques in online extremism detection: Metrics and performance.

Year	Study	Approach	Classification	Metrics Used	Algorithms Used	Algorithm with The Best Performance
2020	[45]	ML	Positive (radical), Negative (non-radical)	F1-Score.	Linear SVM, Logistic Regression	Linear SVM + EmoFeat + SIMON+ Pro-Neu, F1-Score = 94.02 Linear SVM + EmoFeat + SIMON + Pro-Anti, F1-Score = 87.90 Linear SVM + EmoFeat+ SIMON+ Magazines, F1-Score = 94.02
2020	[86]	DL	Hate, Non-Hate, explicit white supremacist, implicit white supremacist, other hate speech and neutral	Precision, Recall, F1-Score, AUC.	Logistic Regression, LSTM , BERT	BERT + Word2Vec – F1-score = 0.79, Precision = 0.80 LSTM + Word2Vec, F1-score = 0.74, Precision = 0.75
2020	[120]	ML	Extremist, Non-extremist	Precision, Recall, F1-Score, AUC.	SVC, Random Forest, Multinomial Naïve Bayes, Gradient Boosting	Gradient Boost + Word2Vec – F1-score = 86 Random Forest + Word2Vec – F1-score = 85
2020	[115]	ML	Neutral, Moderate, Low Extreme and High Extreme	ROC-AUC Curve, Accuracy	Multinomial Naïve Bayes, KNN, Support Vector Classifier (SVC)	SVC, Accuracy = 82
2020	[129]	ML	Extremist, Non-Extremist	Precision, Recall, Accuracy, F1-score	SVM, Random Forest, Naïve Bayes	SVM + All Features, F1-Score = 0.87 Random Forest + All Features, F1-score = 0.84
2019	[68]	ML	Terrorism Supporting, Terrorism non-supporting, Random	Precision, Recall, Accuracy.	Logistic Regression, SVM	SVM – Accuracy = 95, Precision = 87, Recall = 98
2019	[48]	DL	Extremist, Non-Extremist, Anger, Joy, Fear, Sadness, Analytical	Precision, Recall, Accuracy, F1-measure.	CNN, LSTM, Gated Recurrent Unit (GRU)	LSTM+CNN , Accuracy = 92.66, Precision = 0.9, Recall = 0.88, F1-Score =0.88, Loss Score = 0.96
2019	[103]	ML	NA	Number of Clusters, Fuzzy Partition Coefficient (FPC)	Fuzzy Clustering	Fuzzy Clustering – Clusters = 2, FPC = 0.85
2019	[55]	Network / Graph	Leaders of ISIS, Pro-ISIS, Anti-ISIS	Accuracy, Precision, Sensitivity, Specificity	Naïve Bayes, Multinomial Naïve Bayes and Stochastic Gradient Descent classifier	Stochastic Gradient Descent + Naïve Approach, Precision = 0.81, Accuracy = 0.86
2019	[125]	ML	Pro-Afghan, Pro-Taliban, Neutral, Irrelevant	Accuracy, Precision, Recall, F-score	Random Forest, SVM, Naïve Bayes	SVM + TF-IDF Bigram, Precision = 84.71 Random Forest + PCA, Accuracy = 0.76 Naïve Bayes + unigram, Accuracy = 83.16
2019	[70]	ML	Known bad, Random good, Non-radical	Precision, Recall, F-measure, Accuracy, ROC curve.	Random Forest, SVM, KNN, Neural Network	Random Forest + Psychological Features, Accuracy = 100 Random Forest + Radical Language, Accuracy = 80 Random Forest + Behavioural Features, Accuracy = 91
2019	[52]	ML	Extremist Users, Non-Extremist Users.	Precision, Recall, F1-Score, ROC	Random Forest	Random Forest + Imputation, F1-score = 0.93
2019	[49]	DL	Radical, Non-Radical, Irrelevant	Precision, Recall, F-score	SVM, Max Entropy, and Random Forest	LSTM+Word2Vec, Precision = 85.96
2019	[130]	ML	Pro-ISIS, non-pro-ISIS	Precision, Recall, F-measure	Naïve Bayes, SVM, J48	SVM + Semantic Context Features, F1-score = 0.85, Precision = 0.85, Recall = 0.84
2019	[131]	ML	non-dangerous, dangerous and unknown	AUC Curve	Decision Tree, Logistic Regression, Random Forest, LightGBM, XGBoost	XGBoost + All Features, Test AUC = 0.943

TABLE 10. *(Continued)* **Summary of classification techniques in online extremism detection: Metrics and performance.**

2019	[54]	Network / Graph	Dangerous, Safe, Unknown	Node Page Rank, Hub and Authority Measure, Betweenness Centrality, Proximity Prestige, ROC-AUC Curve	Simple Graph, Logistic Regression, Decision Tree, Random Forest, XGBoost and LightGM	XGBoost, ROC-AUC Curve Train = 0.95, Test = 0.94
2018	[126]	ML	Extremist (Suspended), Non-Extremist (Not Suspended), Same User, Different Users	ROC-AUC curve.	Logistic Regression	Logistic Regression, AUC = 0.93
2018	[71]	ML	Hate, Safe	Precision, Recall, Accuracy	libSVM	libSVM + Arabic, F1-score = 84 libSVM + English, F1-Score = 79 libSVM + French, F1-Score = 80
2018	[110]	ML	Positive, Negative	Precision, Recall, F1-Score	Naïve Bayes	Naïve Bayes + 4 grams, Precision = 88.94
2018	[105]	ML	Hate, Safe, Left, Right	Precision, Recall.	SVM, Single Layer Perceptron, Decision Trees	SVM + English, Precision = 82 SVM + Arabic, Precision = 82 SVM + Hate-Safe, Precision = 84 SVM + Left-Right , Precision = 82
2018	[113]	ML	Micro, Meso, Macro	Precision, Recall, F1-Score	Collaborative Filtering, Naïve bayes	Naïve Bayes + Pro-ISIS Macro, Precision = 90 Naïve Bayes + Pro-ISIS Micro, Precision = 79 Naïve Bayes + Pro-ISIS Meso, Precision = 69
2018	[23]	ML	Hate, non-Hate	Precision, Recall, Accuracy	Single Layer Perceptron	Single Layer Perceptron + 5 features, Precision = 84.21
2018	[85]	DL	Hate, non-Hate, Relation, Skip	Accuracy	LSTM, CNN, SVM	LSTM + Hate+ non-Hate, Accuracy = 0.78
2017	[128]	ML	Positive, Negative and Neutral	Accuracy		NA
2017	[53]	Network / Graph	Pro-ISIS, Anti-ISIS	Precision, Recall, F1-Score	Semantic Graph, SVM	SVM, Precision = 0.923 Recall = 0.923 and F1-score = 0.923 for pro-ISIS and anti-ISIS
2017	[72]	Network / Graph	ISIS OEC Member, non-member, official account, non-official accounts.	Accuracy, F1-Score	KMeans, Newman Method and Louvain grouping, IVCC	IVCC, Accuracy =0.96, F1-score = 0.93
2017	[132]	ML	Islamist, Non-Islamist, Far Right, Far Left	Accuracy, F1-Score, Precision, Recall	H2O DL, Distributed Random Forest, Naïve Bayes, Gradient Boosting Machines	H20 DL + Binary, Accuracy = 0.97 H20 + Multiclass, Accuracy = 0.72
2017	[118]	ML	Right-wing Extremist, Non-extremist	Precision, Recall, F1-Score.	NA	BoW, F1-score = 0.95
2016	[117]	ML	Positive, Negative	Precision, Recall, F1-score, ROC-AUC Curve.	Logistic Regression, Random Forest	Random Forest, AUC = 0.923
2016	[114]	Network / Graph	Pro-ISIS, Anti-ISIS	Micro-ROC, Macro-ROC	Custom Algorithms	Per User Adoption Probability + Sharing Features, Micro ROC-AUC = 0.602 Per User Adoption Probability + Lexical Features, Micro ROC-AUC = 0.476 Per User Adoption Probability + Interaction Features, Micro ROC-AUC = 0.55

Some studies focus on the social media like Twitter and Facebook. Agrawal *et al.* [67] propose ML-based approach for extremism detection on Twitter. The authors classified data using one-class labelled as hate-supporting. If data

TABLE 10. *(Continued)* **Summary of classification techniques in online extremism detection: Metrics and performance.**

Year	Study	Approach	Classification	Metrics Used	Algorithms Used	Algorithm with The Best Performance
2016	[96]	ML	Positive, Negative	Precision, Sensitivity, Specificity, Negative Predictive Value (NPV)	Adaboost	Adaboost + All Features, Specificity = 0.9937, Precision = 0.8850 Sensitivity = 0.7977 NPV = 0.9857
2015	[69]	ML	Radical, Non-Radical	Accuracy	SVM, Naïve Bayes, Adaboost	Adaboost + All Features, Accuracy = 99.5
2015	[97]	ML	Jihadist, Random	Precision, Accuracy, Recall.		Adaboost + All Features + English Tweeps, Accuracy = 1.0
2015	[67]	ML	Positive. Unknown	Precision, Recall, True Positive Rate, Negative Predictive Value, Accuracy.	KNN, libSVM	libSVM, Accuracy = 0.97, Precision = 0.78, Recall = 0.83
2015	[111]	Network / Graph	Relevant, Irrelevant	True Positive Rate, False Positive Rate, Positive Predictive Value, Negative Predictive Value.	BFS, SSA	Shark Search, Accuracy = 0.74, F1-score = 0.85

does not fit the hate-supporting class, then it is labelled as unknown. This research work collects tweets of extremist and hate speech by identifying particular keywords. The study used seed hashtags like #Terrorism, #Islamophobia, #Extremist, and #Islam. Any tweets having these hashtags are manually checked and labelled as hate speech. The authors make use of UDI-TwitterCrawl-Aug2012 and ATM-TwitterCrawl-Aug2013 dataset to collect more hate and extremism promoting tweets. The paper extract discriminatory features like religious, offensive, slang, negative emotions, punctuations, and war-related terms, and use Term Frequency for vectorization. For the training purposes, 10,486 labelled tweets are considered.

Asif *et al.* [115] use the multi-class classification of texts to detect the extremism on Facebook. This work collects data from Facebook pages of ARY news [134], PTV news [135], Dawn [136], The News [137], Samaa [138], Express news [139], Dunya [140] and Geo [141]. These pages contain posts and comments in Urdu, English, and Roman Urdu. Lexicon based sentiment analysis is considered for the classification of texts. A total of 20,000 posts are considered. Sentiment Lexicons are used for assessing the sentiment score for labelling. Intuition followed by the authors is that more the negative sentiment, more extreme the text. Texts are classified into Neutral, Moderate, Low Extreme and High Extreme. Due to the expanse of the social media, the detection of the extremist communities is also a primary concern [142]. Few works used the clustering-based extremist community detection. Benigni *et al.* [72] propose Iterative Vertex Clustering and Classification (IVCC) for identifying ISIS supporting communities. IVCC consists of two phases 1) Vertex Clustering/Community Optimisation and 2) Multiplex Vertex Classification. Phase 1 involves identification of the extremist communities using algorithms like KMeans, Newman Method and Louvain grouping. Phase 2 involves identification of vertices in community, using KMeans and other algorithms. The authors collect data of ISIS supporters from Twitter, using criteria like followers, tweet mentions and hashtags. A total of 1,19,156 Twitter accounts are considered in this study. Using the Louvain algorithm, data is divided into 10 clusters, with the data classified into member, non-member, and suspended. The authors conclude from the obtained clusters that there are three types of users in the obtained data which are ISIS community member, non-member, official accounts (which are News accounts).

Some works dwell into sentiments of extremist texts available on social media [143]. Araque *et.al* [45] propose similarity-based and emotion-based detection of ISIS extremist content. The authors classified text as radical, non-radical, neutral. The authors use NRC Hashtag Emotion Lexicon [144] tool to identify Emotional Lexicon in the collected data. NRC Hashtag Emotion Lexicon contains nearly 16,000 words with the emotional annotations like sad, anger, fear etc. Collected Emotion Lexicon is given as input to Emotion Features (EmoFeat) extraction algorithm. While only words cannot mean that content is radicalized, the frequency of the appearance of radical words (FreqSelect) is used as a radical lexicon for Similarity-based feature extraction (SIMON). For EmoFeat, the number of emotions and the number of statistical measures are used. Recursive Feature Elimination (RFE) is used to select all the possible combinations of emotions and statistical measures. The authors use Linear SVM to classify the texts into positive, negative.

Ashcroft *et al.* [69] use 579 stylometric, 36 time-based, and 4 sentiment-based features to classify tweets into pro-ISIS, anti-ISIS, and random. In stylometric features, the authors consider punctuations, letter bi-grams, word bi-grams, and the most frequently used hashtags. Time-based features include the time of tweets divided into hour of day, period of day (morning, noon etc), day (Sunday, Monday etc), type of

day (weekend, weekday). Sentiment extraction is performed using Stanford CoreNLP Toolkit. The sentiments of words are labelled as very negative, negative, neutral, positive and very positive. The author detect extremism on Twitter by classifying tweets into pro-ISIS, anti-ISIS and random. The authors collect tweets related to *Jihadist related hashtags* for extremist data collection. This work collects random tweets for non-radical class. For the classification, study used tweets in the range 4000-7000.

Some works use the features like content sharing, interaction, and influence of extremists for their detection. Rowe *et al.* [114] use lexical, sharing and interaction features to assess the extremist ideology adoption by Twitter users. Experts are used to classify tweets into pro-ISIS and anti-ISIS. The study creates their own algorithm to determine per user adoption probabilities. The authors identified crucial milestones in user radicalization such as activation points (time at which user exhibited radicalized behaviour), pathways to activation and behaviour point to activation. The activation points are further classified into pre-control (words used before radicalization), activation (words used when radicalized), and post-control (words used after radicalization). For features angular cosine similarity for finding similarity in language, with sharing of content and communications between the users are used.

Fernandez *et al.* [113] applied influence of the extremist users as a feature to detect the roots of radicalization in the users. The authors collect their custom extremist data as well as neutral data and manually verify them. The study classifies users into Micro (Individual) influence, Meso (Group) influence, and Macro (Global) influence. The micro influence is captured by the posts of individual, meso influence is denoted by shared posts, and macro influence by links or urls from different websites. Similarity function is created to compute influence whether it is micro or meso or macro.

Few studies discuss context like religion, behaviour, event, war zone and language in extremism detection research. Fernandez *et al.* in another work [130] use the contextual dimensions like categories, topics, entities, and entity types for radicalization detection. To extract contextual dimensions, the authors use TextRazor Semantic Annotator [145]. The keywords in contextual dimensions are determined for both radical and non-radical texts based on confidence scores. In confidence scores, each keyword represents the appearance of specific keyword in the annotated category. The cosine function is then used to calculate the similarity between the contextual dimensions.

Mussiraliyeva *et al.* [120] focus on extremism in Kazakh language. The authors use Vkontakte [146] social network, which is popular in countries bordering Russia. For data collection, the authors created a parser with Vkontakte API. The authors collected different extremist posts with total of 3,000 words. Similarly, the authors collected non-extremist posts with total of 15,000 words. For classification into extremist and non-extremist, the authors use Linear Support Vector Classifier (SVC), Multinomial Naïve Bayes, Logistic Regression, Classification Trees, Gradient Boosting and Random Forest.

Smedt *et.al* [71] consider terrorist event-based context for the extremism detection. The extremism data was collected for the specific terrorist events like the Charlie Hebdo shootings [147], and Sinai attack [148]. The authors classified tweets into hate and non-hate. Text profiling is done to identify the author's age, gender, education, or personality. The author uses character trigrams as a feature extraction method. Character trigrams are preferred, as they efficiently model word endings, word functions, emoticons, and spelling variations.

Ul Rehman *et al.* [129] incorporates religious features, and radical features with violent and bad words to detect extremism. The authors use five different dataset ISIS Kaggle Dataset [102] as Radical Corpus, ISIS-related Dataset [149] as Neutral Corpus, ISIS Kaggle Religious Text [150] Dataset with text from Rumiyah, and Dabiq as Religious Corpus and new dataset with both extremist and non-extremist tweets. The extremist tweets in new dataset were identified by Ctrl-Sec group which is non-profit organization reporting ISIS activities on Twitter. The authors perform exploratory analysis on these datasets for radical and religious terminologies. The authors extract radical and religious features from Radical Corpus and Religious Corpus respectively, using TF-IDF. The authors classify texts into radical and neutral. For classification, Naïve Bayes, SVM and Random Forest algorithms are used.

Sharif *et al.* [125] identified extremism within Twitter communities in the context of the Afghan war zone. The authors classified tweets into extremists and non-extremists. Extremist tweets are sub labelled as pro-Afghan and pro-Taliban tweets. The authors collected data with geo-location of Afghanistan and are related to the Kunduz Madrassa attack [151]. They have manually labelled collected tweets as pro-Taliban, pro-Afghan, Neutral, and Irrelevant. Uni-grams, bi-grams and TF-IDF are used for the feature extraction. PCA is used for reducing the dimensionality.

Mashechkin *et al.* [131] use topic analysis and time series analysis to identify extremist communities, extremist users, and extremist message flow characteristics. The authors follow similar methods used by Petrovskiy *et al.* [54], in addition, they use Latent Sematic Analysis (LSA) for topic identification. The authors use the same KavakazChat dataset mentioned in Petrovskiy *et al.* [54]. The authors consider web links, hashtags and author of document to extract primary keywords or summaries. To extract topics, the authors use LSA based on orthogonal non-negative matrix factorization. In orthogonal non-negative matrix factorization, text data is represented as matrix of topics and matrix of fragments in topics. The authors extracted relevance of topics with help of these matrices. The authors analyze time intervals of messages and their topic weights to gather the details about message flow. To make analysis relevant, the authors classify extremism data into non-dangerous, dangerous and unknown. The authors extract all topics, relevance of topics and time

series information from the data of defined classes. The authors create detailed graph using extracted data to identify extremist communities. From these graphs, the authors select specific features as mentioned by Petrovskiy *et al.*

Kursuncu *et al.* [52] used the contexts like religion and ideology to detect the extremism on Twitter. The authors separate tweets based on religion, ideology, and hate. The authors used identified extremist data from Lucky Troll Club and obtained 47,376 tweets. The authors also collected 13,000 non-extremist tweets. Tweets are classified into extremist and non-extremist. This work used Word2Vec for word embedding. This work also analyzed tweets on different topics and investigated the similarities among the users. Manual data validation with experts led to the inter-rater agreement of 0.82.

Nouh *et al.* [70] used radical language, psychological signals, and behavioural features as the context for the extremism detection. The authors used ISIS Kaggle Extremist Dataset. Extremist textual properties, topics, and linguistic cues were used for radical language features using TF-IDF and Word2Vec. LIWC dictionary is used to calculate scores of psychological, personal, and emotional categories. To collect behavioural signals, the authors create an interactional graph between the extremist users that monitor retweets, following, followers, and mentions. To measure degree of influence of the extremist users, the authors use the degree of centrality and betweenness centrality.

Some studies use ideologies as well as a different part of hate speech for the extremist detection. Another study by Smedt *et al.* [105] involves collecting and classifying data into different hate speech categories like *jihadism*, *right-wing extremism, sexism*, and *racism*. The study gathers jihadism, right-wing extremism, and racism data from Twitter using keywords related to the specific categories. Racism data was obtained from Facebook while sexism data was obtained from Incels.me now incels.co [152] website. In this work, extremism detection is classified into Hate or Safe and Left or Right. Hate class contains jihadist, right wing extremism, racist and sexist texts. The right class contains tweets supporting hate texts while Left class refers to opposing hate speech. The authors extract features like trigrams, word frequencies and sentiment analysis, using LIWC. This work used various classification models like SVM, single-layer perceptron and decision trees.

Jaki *et al.* [23] detected hate and extremism in right-wing German Twitter users. The authors identified various dehumanizing keywords that were used by right-wing extremists. For automatic detection, the study classified and collected tweets as hate and non-hate. To correctly train the model, the authors used both German and English tweets. The study used character trigram as a feature extraction method. The different features like emojis, uni-grams, bigrams, punctuation marks etc are also considered. The authors also test their model on different unknown samples, some of which were manually labeled by the experts. These unknown samples were retrieved from various sources like the German Far Right-wing conspiracy website and some random articles from German Wikipedia pages.

3) STUDIES USING DEEP LEARNING TECHNIQUES

Although, ML methods are popular among researchers in the extremism detection domain, they have few limitations as mentioned in Section III(D). This includes the need of feature extraction and selection procedure for better performance metrics. Deep Learning models do not require explicit feature selection and can take advantage of a large dataset to identify features.

Most extremism detection in DL-approach uses LSTM models. Kaur *et al.* [49] use the LSTM model to identify the extremist texts in context with India. The collected data is classified as Radical, Non-radical, and Irrelevant. The authors use expert annotators for labelling the data. The labelling of texts is based on the features predefined by the authors like abuse of Indian service personnel, anti-national speech, support of terrorism and terrorists, and inciting other people to support terrorism. Word2Vec is used to generate word embeddings. The authors use different ML algorithms like SVM, Max Entropy, and Random Forest. Ahmad *et al.* [48] use sentiment analysis for the detection with LSTM and CNN models. The researchers classified tweets into extremists and non-extremists. The authors used Twitter API to collect their custom extremist data based on keywords like *ISIS*, *suicide*, and *bomb*. These tweets are matched with seed words present in BiSAL (Bilingual Sentiment Analysis Lexicon) [153]. Sentiments considered are anger, joy, fear, sadness, confident, analytical, and tentative. The authors also use n-grams, TF-IDF, and BoW as feature extraction to compare accuracy with word embedding layer of a neural network model. The authors use the different combinations of CNN, LSTM, FastText with word embedding, and Gated Recurrent Unit (GRU). But LSTM with the CNN model outperforms other models.

Few researchers have a focus on the extremism detection related to white supremacist ideology. De Gibert *et al.* [84] aimed to create a standard dataset for White Supremacy detection. The authors collected data from the StormFront website about white supremacists. The authors created rules for annotation of hate and extremism data. These rules dictate that text must be hate if a sentence indicates a deliberate attack on a specific individual or group. Another rule says the text depicts hatred, if the text includes words or context supporting the extremist group. The authors labelled data as Hate, NoHate, and Skip. NoHate contains normal texts, while Skip contains generic text like a different language and different taglines. Algorithms like SVM, CNN, and LSTM are used to classify messages into hate and non-hate. StormFront Dataset is compared with the standard Hatebase dataset. Only 9.2% of the vocabulary is similar in both datasets. Thus, confirming the StormFront dataset is sufficiently unique.

Al-Zewairi and Naymat [132] use H2O platform [154] which provides multilayer feedforward artificial neural network to identify radical islamists. The authors use PIRIUS

TABLE 11. Comparison of techniques in online extremism detection.

Comparison	Network / Graph Approach	Machine Learning	Deep Learning
Contextual Applications	Finding Interconnections between Extremists	Classification of Extremist Content	Heterogenous automated feature extraction of Extremist Content
Most Commonly Used Algorithms	Louvain Grouping	Logistic Regression, Naïve Bayes, SVM, Random Forest	LSTM
Most Commonly Used Performance Metrics	Degree of Centrality, In-Degree Centrality	Precision, Recall, Accuracy, F1-Score and ROC	Precision, Recall, F1-Score and Accuracy
Use	Used in conjunction with ML techniques recently.	Standalone most of the times.	Standalone most of the times.
Technical Purpose	To get contextual data about communications of the extremists.	To classify and detect extremists or extremism.	To extract features automatically, classify and detect extremists or extremism.

dataset containing information of 1,473 individuals in USA who were associated with extremist activities. The dataset includes information like personal, demographics, socio-economic, radical group, radicalization ideology and extremist activities. The authors consider binary and multiclass classification. For binary classes, the authors consider Islamist and non-Islamist. For multiclass classification, the authors use classes like Islamist, far-right and far-left.

Alatawi *et al.* [86] use BERT to identify white supremacist content from Twitter. Two different datasets are used that is Stormfront Dataset and custom Twitter dataset. The custom dataset was developed by using known white supremacist hashtags like #white_privilege, and #its_ok_to_be_white. The custom dataset was labelled as explicit white supremacist, implicit white supremacist, other hate speech, and neutral. The authors employ three annotators using Amazon Mechanical Turk [155]. GloVe and Word2Vec are used for feature extraction. The authors used pre-trained word vectors like Google-News word vectors, GloVe trained on Wikipedia, GloVe trained on Twitter, and White Supremacist Word2Vec (WSW2V) trained on Twitter. Pre-training is performed to obtain the initial parameters in unlabelled data. The Authors employed LSTM and BERT models for classification. BERT with WSW2V outperformed LSTM with WSW2V.

4) COMPARISON

As it can be observed in the trends shown in Table 10 and Table 11, ML and DL methods are preferred over Network-based techniques in the recent years. This can be attributed to an increase in computational speed for ML and DL techniques. Network/Graph-based approaches are still used to identify extremist networks over social media, whereas ML and DL are used to classify extremist content. Comparison of Network/Graph, ML and DL techniques is present in Table 11.

E. VALIDATION

The researchers use their intuitions based on hashtags and keywords to check if the collected data is extremist or not. Hashtags and keywords used by the researchers are provided by Law Enforcement Agencies as well as the existing researches. So, the researchers assume if collected data contains specific hashtags and keywords, it is extremist text. For data validation, the researchers used Expert agreement, also known as Inter-rater Agreement. The studies collect custom data or use publicly available extremist dataset, annotate the data using field experts [49], [113], [114] [119], [125]. Annotation with labels as extremist–non-extremist, radical–non-radical, ISIS supporting–ISIS non- supporting, hate–non-hate were performed by experts. Multiple experts were employed for the annotation to prevent bias. After annotation, methods like Cohen's Kappa coefficient [49] and Fleiss' Kappa coefficient [114] were used to calculate inter-rater agreement. Cohen's Kappa coefficient measures the agreement between two annotators, who classified data into different categories [156]. Coefficient of 1 implies experts or annotators are in complete agreement, while coefficient $= 0$ means they are in complete disagreement. Fleiss' Kappa measures the agreement for more experts or annotators, while categorizing the data [157].

Kaur *et al.* [49] used two experts to annotate the collected texts into radical, non-radical, and irrelevant classes. Cohen's Kappa for agreement between two experts was 0.76, which is a substantial agreement. De Gibert *et al.* [85] used three experts and two data batches for the annotation into hate and non-hate categories. For the first batch, Cohen's Kappa is 0.61 while Fleiss' Kappa is 0.60. For the second batch, Cohen's and Fleiss' Kappa are 0.62 and 0.63, respectively. Fernandez *et al.* [113] identified accounts as non-ISIS supporting, by labelling them through two annotators with Cohen's Kappa of 1. Rowe *et al.* [114] used two experts to label tweets as pro-ISIS, anti-ISIS, or neutral. Fleiss' Kappa for inter-rater agreement for classes was 0.418 and 0.504 for English and Arabic pro-ISIS tweets. For English and Arabic anti-ISIS tweets, inter-rater agreement was 0.43 and 0.52, respectively. The difference between inter-rater agreement coefficients of two languages was due to a larger number of hate keywords in ISIS Arabic tweets over ISIS English tweets. Alatawi *et al.* [86] used three annotators to label

TABLE 12. Summary of validation metrics for online extremism detection in existing literature.

Authors	Study	Number of Experts	Cohen's Kappa	Fleiss' Kappa	Others
Kaur et al.	[49]	2	0.76	NA	NA
De Gibert *et al.*	[85]	3	0.61, 0.62	0.60, 0.63	NA
Fernandez *et al.*	[113]	2	1.0	NA	NA
Rowe *et al.*	[114]	2	NA	ISIS: Arabic – 0.5, English – 0.4 Anti-ISIS: Arabic – 0.5, English – 0.4	NA
Alatawi *et al.*	[86]	3	Original Labels - 0.076, Modified labels – 0.10	NA	NA
Asif *et al.*	[115]	109	NA	NA	Label Matching Percentage – 88% and K-fold cross validation
Advantages		-	Easy to calculate.	Works with multiple experts.	-
Disadvantages		-	Works with just 2 judges or experts.	More the experts or labels, lower the agreement	No statistical significance.

their custom hate and extremism dataset into explicit white supremacist, implicit white supremacist, neutral and irrelevant. The inter-rater agreement of these annotators shows Cohen's Kappa of 0.076, which is quite low. The authors conclude that low agreement scores were due to neutral and implicit white supremacist labels. Even though when the authors created new binary labels like white supremacy and non-white supremacy, inter-rater agreement was as low as 0.1047.

In article [115], a survey-based validation is used. The authors create a small dataset of 25 posts and comments at random, annotated by 109 different people. These people belonged to various domains like Computer Science, Literature, and Psychology. Then surveyed labels were compared with the author assigned labels. This work reported 88% matching between surveyed labels and author labels. Although the inter-rater agreement is the best way to annotate and validate data, it is challenging to determine the neutrality of the experts or the annotators.

Advantages of Inter-rater agreement:

- Uses statistical methods for measuring consensus.
- Easier to calculate.

Disadvantages of Inter-rater agreement:

- It is manual method, thus time-consuming.
- No information about neutrality of experts or annotators.

It gives optimal results, when there are fewer categories or labels.

F. TOOLS

The existing extremism detection literature shows that very few tools are developed for extremism detection. The availability of tools is an issue, as more comprehensive tools are restricted to private use. The details about the tools like methods, techniques, and metrics are kept confidential. Table 12 shows some tools, that are used in the extremism research. It is also observed that there are no commercial tools for extremism detection.

Investigative Search for Graph Trajectories (INSiGHT) [158] used graph pattern matching techniques to identify the violent extremism radicalization. The authors proposed using different datasets from Open Source Intelligence (OSINT) [159] and restricted criminal and terrorist datasets. OSINT is the data collected from the authorized civilian sources designated to monitor social media platforms. The authors developed Investigative Graph Search (IGS), which takes radicalization query patterns and extremist data graph as input. Radicalization query patterns may include keywords, sentiments, and behaviours identified by psychologists, criminologists, and political scientists. An extremist data graph is the representation of data obtained through OSINT and other sources. The most inherent factor in the determination of radicalized behaviour is the analyst. The analyst interprets context, identifies new patterns, and confirms the radicalized extremist behaviour. The authors classified the data into Query Focus (QF), Individually Innocuous but Related Activities (IIRA), Indicator (IND), and Red Flag Indicator (RFI). Data is labelled into these categories by IGS and verified by the analyst. The study used synthetic as well as real-world data to evaluate INSiGHT. The authors use similarity scores to identify behaviour of the extremist in current data. The similarity score of INSiGHT tool for Klausen's Radicalized behaviour dataset [160], ranges from 0.130 to 0.696.

Twitter Terrorism Detection Framework [68] (TTDF) is the tool proposed for automatic extremism detection on Twitter. This framework consists of five modules: Twitter Data Crawler module, Storage Module, Tweet Classification Module, Output Module and Training Module. Twitter Data Crawler module collects real-time tweets available at time of streaming. In the Storage module, various pre-processing steps are performed, like removing urls, hashtags etc. These tweets are classified into terrorism supporting, terrorism non-supporting, and random. In the classification module, SVM is used for the classification of tweets. The authors used n-grams as a feature extraction method. The authors used precision, recall, f1-score, and accuracy for measuring

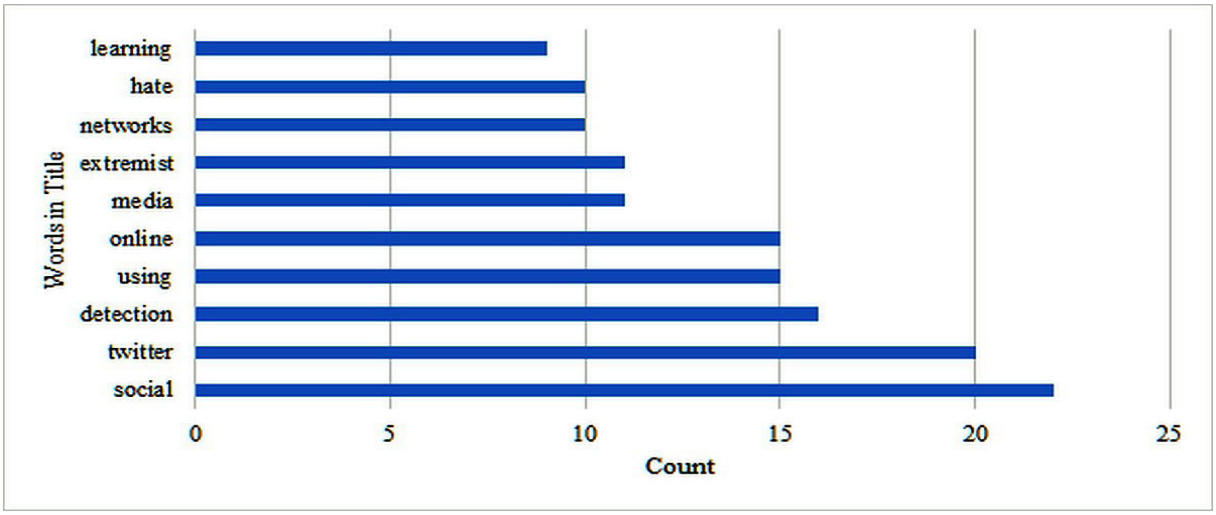


FIGURE 13. Top words by count in title.

classification performance. The framework results in an overall accuracy of 95% for the detection of terrorism supporting tweets. The real-time framework provides 72 % accuracy for the detection of real-time supporting tweets.

Whole Post Integrity Embedding (WPIE) [161] is the tool used to detect Facebook content violations. WPIE is a pre-trained model to analyze the content for violations. WPIE works on multimodal data, with different violation types, and even time of events. WPIE used different models based on CNN for image and video and RNN+CNN for different text types. Using WPIE, Facebook has automatically detected and removed 98.5% of extremist data on Al-Qaeda and ISIS. Facebook has not released the complete architecture of the tool, so information about the extracted features and used evaluation metrics are not available. Thus, it is observed that few tools are available for extremism detection and their public availability is also an issue.

V. EXPERIMENTS

A survey of the literature shows the inclination of the literature towards specific ideologies, datasets, and classification labels. As a result, this observation is investigated. The goal of this experiment is to figure out frequently used words, phrases, terms in online extremism detection research and to prove that current literature is less diverse with respect to ideologies. As title, abstract, and author keywords mirror the research work highlights, search is limited to frequently used words in literature. The researches use word clouds to determine the importance of words in a particular corpus [162]. TF-IDF with uni-gram is used for the feature extraction. The investigations are conducted in the following steps:

A. DATASET

To create a dataset, 64 studies from the literature survey were considered. For dataset creation, three attributes are selected that is Title, Abstract, and Author Keywords.

B. TEXT PRE-PROCESSING

Text pre-processing is the task of cleaning and preparing text data. It helps to remove unwanted parts of text data and to make text data suitable for further analysis.

The regular expression is used to select any word characters. All the text present in the dataset is converted to lower case. This helps in creating a uniform model. NLTK library is used to remove stopwords. Word tokenization is performed after removing stopwords.

C. FEATURE EXTRACTION

For feature extraction, Term Frequency–Inverse Document Frequency (TF-IDF) is used. TF-IDF is used to get importance of the word in the corpus. TF-IDF score of the word depends on frequency of appearance of the word in the corpus. The uni-gram model is selected for feature extraction to get important words.

D. FINDINGS

Fig. 14, 15, and 16 show, top 10 words by count for 'Title', 'Abstract' and 'Author Keywords', respectively. Word count is different from TF-IDF, as TF-IDF considers the numbers of documents in which word appears. Words like 'social', 'media', 'radicalization' and 'extremism' frequently appear in 'Title', 'Abstract' and 'Author Keywords'. Words like 'isis' and 'twitter' frequently appear within the corpus of surveyed literature. By observing word count, the researchers focus on 'social media' for extremism detection. Word frequency gives, which words are repeated but does not refer to words relevant in the corpus. Therefore, TF-IDF is used to extract relevant and important words from Title, Abstract, and Author Keywords. Fig 17 shows, word cloud for Titles of studies using TF-IDF. It is observed from word cloud that words like social', 'media', 'extremist', 'detection', and 'twitter' have more weightage in a given corpus. This can be attributed to detection research focused on extremism detection on Twitter. Word 'isis' denotes that ISIS organization

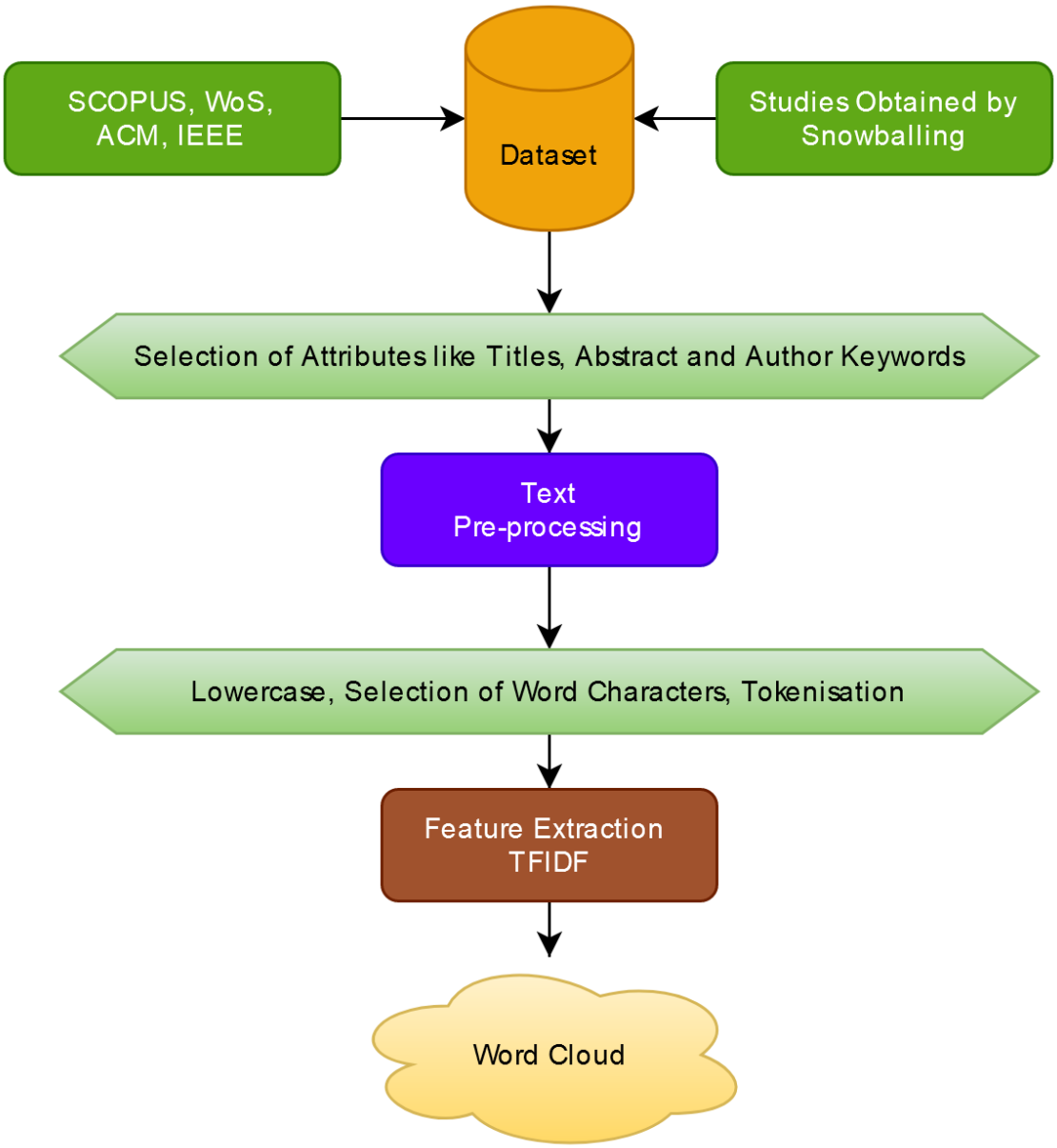


FIGURE 14. Text analysis of research articles on online extremism detection.

dominates in frequency in current literature. Fig. 18, shows word cloud for the abstract of studies using TF-IDF. The words similar to the title appear in the abstract. Words like 'social', 'media', 'isis', 'twitter' and 'extremist' appear in the Abstract. Thus, both Title and Abstract confirm research is bias towards ISIS organization on Twitter. The author keywords are used for indexing and identifying the focus areas in the study. Fig. 19 shows 'regression', 'theory, 'forensic' has higher weightage in author keywords. Thus, it is safe to conclude the keywords usually represent the techniques and methods used in the studies. The words like 'radicalization', 'islamic' and 'islamist' also appear in word cloud, denoting the ideological perspective. The words like sentiments', 'semantics', 'possibilistic', 'clustering' and 'classification' indicate the techniques which are used for analysis and detection.

Findings from Word Cloud:

- Research works are limited to ISIS organization for extremism detection.
- Research works investigated the application of the social media as a primary tool for extremist organizations' propaganda and recruitment.
- The researchers primarily use Twitter as the source for the extremism data collection.

VI. DISCUSSIONS

The survey helps us to formulate answers to our research questions as follows:

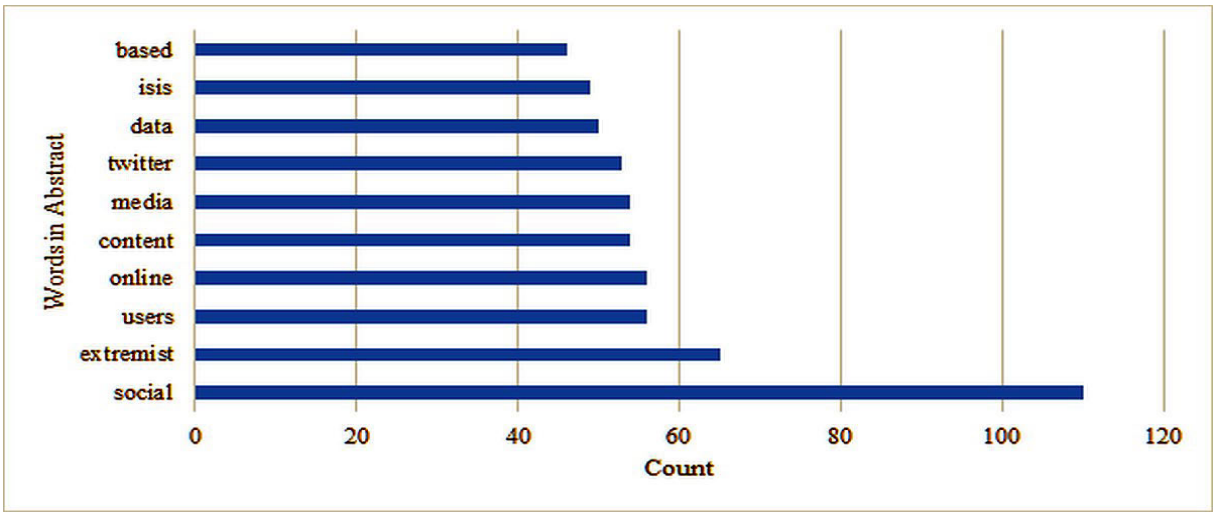


FIGURE 15. Top words by count in abstract.

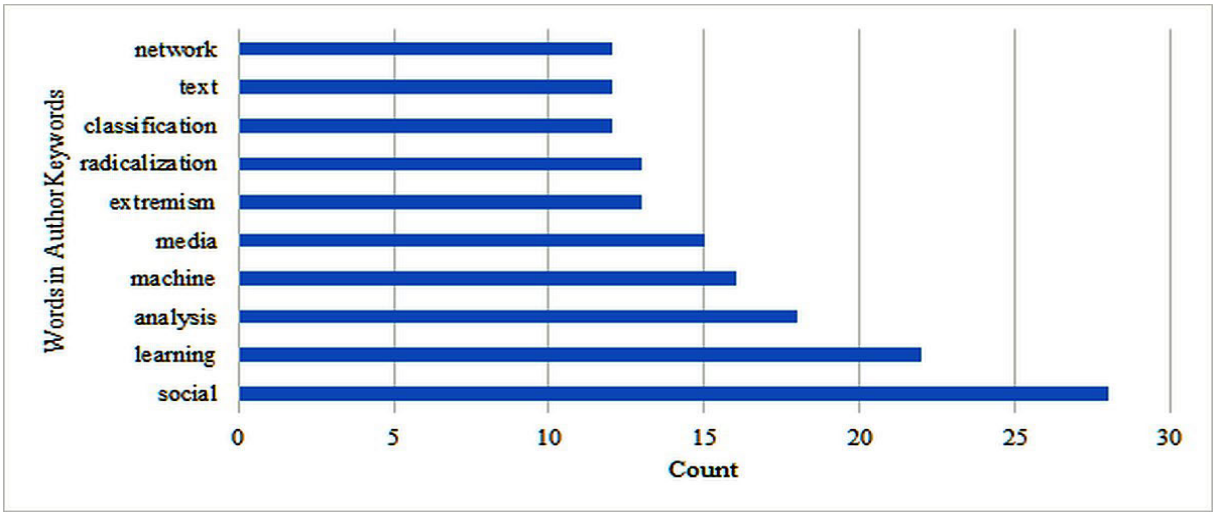


FIGURE 16. Top 10 words by count from author keywords.

- *RQ1: What are the various datasets available for online extremism detection and classification?*

Most of the literature surveyed, use Twitter as the primary source of extremist propaganda. Twitter is used as a primary source in 43 studied works of literature. This can be attributed to the popularity of Twitter, its ease of access to people and developers, and short message format. Text accessed from Twitter is either available as a public dataset or obtained by researchers using Twitter API. Most of the data available in the public dataset from Twitter comes from discontinued accounts. This is due to strict policy of Twitter to curb hate speech on its platform.

The researchers use magazines like Rumiyah and Dabiq which are studied for linguistic, semantic, and psychological analysis of ISIS organization. 5 studies make use of these ISIS extremist magazines. No studies were found, that use magazines that are related to other extremist organizations or ideologies. The data collection about far right-wing ideology or White Supremacist is limited due to the small number of researches on them. Twitter and Stormfront are used frequently by the researchers to study far right-wing ideology or right-wing movements. Out of the total surveyed literature, only nine are exclusively on right-wing extremist ideology, out of which two used Stormfront and two used Twitter as a source of the dataset.

The challenges with the dataset are visualized in Fig. 11. The most important problem in the dataset is a class imbalance. Class frequency of extremism positive elements are much less than non-extremist or non-supporting class. As most datasets are collected from social media, data availability is challenged by suspending extremist accounts and stringent social media policies. These problems lead to lesser availability of datasets in the public domain affecting the

FIGURE 17. Word cloud using TF-IDF of title.

reproducibility of results. Another challenge is manual data validation, which suffers from a threat to bias with few experts and a lesser sample size for expert labelling. This results in less accurate classification. Another problem is the collection of data about the specific organization, ideology, and specific events. Thus, classification algorithms get biased towards these organizations, ideologies, and events.

- *RQ2: What are the various methodological trends in online extremism detection and classification?*

Two important methodological trends were observed in existing online extremism detection literature: manual detection and automatic detection of extremism. The earlier studies depended upon manual detection of extremist accounts. few articles used manual identification of extremists' users. Manual detection of extremist user accounts relies on the intuition of authors. Thus, manual detection takes a lot of effort and time as the authors manually analyze every tweet or post.

Another method, that is popular among researchers is automated detection. In automated detection, two techniques are used extensively for extremist speech detection, that are Graph and Machine Learning. Less number of studies used the Graph approach for extremism detection. Graph methods used the social network to find associates and followers of known extremists. Graph techniques effectively find the radicalization sphere and reach of extremists. The graph method fails to categorize news services portraying terrorists and actual terrorists. Another problem with the graph approach is the identification of the new extremist user.

Machine Learning algorithms can overcome these problems of graph approaches. ML algorithms effectively find new extremist users as well as learn from evolving extremist content. ML algorithms classify news sources from the terrorist hate speech with an acceptable accuracy. State-of-art ML methods use a combination of datasets and different algorithms to detect online extremism. ML algorithms like Logistic Regression, Naïve Bayes, SVM and Random Forest are popular in surveyed literature. However, ML-based approaches have problems like over-dependence on features extraction or selection or reduction, and they are unable to identify context as mentioned in Section III(D).

In recent years, the DL approach has been evolving for an automatic extremism detection. In extremism detection research, 4 significant studies using DL approach are chosen for this survey. Extremism detection studies make use of Deep Learning models like LSTM, CNN and BERT. The Deep Learning method does not require the sophisticated feature extraction/ selection/ reduction processes and can work with the huge amounts of data. This makes the DL-based approach suitable for an automatic extremism detection research characterized by newer vocabulary in extremism.

- *RQ3: What are different classification techniques used for online extremism detection and their comparative evaluation?*

State-of-art methods use a combination of features and algorithms to detect online extremism. Feature extraction based on metadata like the number of retweets, number of comments, number of followers etc. are present in some articles. Semantic and Network features are also extracted. Table 9, gives information of features reduced or features extracted.

Graph techniques rely on simple graphs and standard graph algorithms like Fast Greedy, Louvain, and Close Graph. ML algorithms like Logistic Regression, Naïve Bayes, SVM, and Random Forest are prevalent in the surveyed literature. Recent Deep Learning research is focused on using LSTM and CNN-1D. Table 10, provides algorithms and their respective performance in detail.

Classification of text in extremism detection research is either binary or tertiary. Most research use binary classification like extremist–non-extremists, ISIS supporting–ISIS non-supporting, and White Supremacist supporting–White Supremacist non-supporting etc. In some research work, the neutral class is added, while in some studies irrelevant class is added for tertiary classification. The few articles use labels based on sentiments and behaviour. Fifty-eight

TABLE 13. Summary of online extremism detection tools.

Tool Name	Study	Techniques Used	Availability	Features	Evaluation Metric	Pros	Cons
INSiGHT	[158]	Graph	Prototype	Keywords, behaviour, sentiments	Similarity Score	Detection of extremist links via identification of risk factors.	Human intervention is needed.
TTDF	[68]	ML	Prototype	N-grams	Accuracy, Precision, Recall	Automated Tweet Collection.	Absence of Validation. Tertiary Classification
WPIE	[161]	DL	Private	NA	NA	Use of multimodal context identification.	Binary Classification. Still preference to the user reporting.

studies performed binary classification of extremist texts, and 5 used tertiary classification. All these binary or tertiary classifications does not provide the details about online extremism. No information is learned from binary or tertiary classification. So multi-class classification with the labels, that will provide analytical insights is needed.

The researchers have adopted different metrics to evaluate these classification methods. ROC-AUC curve is used in a few studies [114], [126]. Numerical metrics like precision, recall, and F1-score are frequently used, when the data is imbalanced. Numerical metrics are used mostly in the researches involving the use of ML or DL algorithms. As different evaluation metrics with different datasets are used in selected extremism literature, direct comparison for techniques is difficult.

Extremist text forms a small proportion of total posts or tweets on the social media and other sources. Thus, it is imperative to evaluate performance metrics, considering all possible scenarios like data imbalance, multi-class classification, micro and macro averages of precision, recall, and F1-score. Table 7, Table 9 and Table 10 provide the details about techniques, classification, and metrics of the surveyed studies. It can be concluded that the researchers prefer binary classification. The classification evaluation metrics like precision, recall, and F1-score are frequently used.

Context identification is one of the biggest challenges in the online extremism detection research. Most research in the context about extremism revolves around religion, psychology, and behaviour. These contexts are good enough to detect extremist content, but not enough for the detailed analysis. The spread of online extremism heavily depends on the political scenario in a particular location. But the absence of location context in the surveyed research was not found. Thus, there is a need for extremism detection using location as a context.

- *RQ4: What validation techniques are used for data validation in online extremism detection and classification?*

Researchers rely mostly on hashtags and keywords to determine data is extremists or not. This may help in early data collection, without worrying about data validation. As data validation is a precondition for the classification, the studies employ expert validation to validate data. In expert validation, multiple experts give their opinions on the particular data. These opinions are aggregated, and inter-rater agreement is calculated. Two interrater agreement methods Fleiss's Kappa coefficient and Cohen's Kappa coefficient, are most commonly used in extremism detection and analysis. Cohen's Kappa method is used, if only two experts or annotators are used. Fleiss' Kappa can be used when there are more than two experts.

The problem with these methods is that they indicate agreement. The inter-rater agreement coefficients of most studies settle around 0.6, which is low and degrades, if there are many labels. The annotation of labels also depends upon the perception of experts or annotators. Few research works show that the opinion of experts substantially disagrees with the opinion of laymen. Another problem is the neutrality of few experts, which is hard to determine. The number of experts also affects inter-rater coefficients. Less number of experts may have a good agreement but can lead to higher bias. In contrast, an increasing number of experts may reduce bias, but inter-rater agreement coefficients may be low [86]. Thus, even using inter-rater agreement will not guarantee an accurate annotation or validation of data. Therefore, it is a research challenge to choose better values of inter-rater agreement with lesser bias.

- *RQ5: What are popular tools available for online extremism detection?*

There are very few tools used for extremism detection, and no tool is available in the public domain to detect online

TABLE 14. TF-IDF scores from title, abstract and author keywords.

Title		Abstract		Author Keywords	
Terms	**TF-IDF Score**	**Terms**	**TF-IDF Score**	**Terms**	**TF-IDF Score**
twitter	4.713606	social	3.043853	theory	2.043773
social	4.522558	extremist	2.886977	regression	1.519368
online	4.029438	isis	2.558884	forensic	1.437135
detection	3.990293	radical	2.459978	forest	0.709300
using	3.487205	users	2.457826	radicalization	0.709016
extremist	2.995607	media	2.202699	possibilistic	0.698813
networks	2.835558	online	2.165110	risk	0.669774
media	2.828251	radicalization	2.139750	semantics	0.622783
hate	2.712497	twitter	2.090576	radicalisation	0.614911
learning	2.653241	content	2.051777	time	0.605166

FIGURE 18. Word cloud using TF-IDF of abstract.

FIGURE 19. Word cloud using TF-IDF of author keywords.

extremism. Two tools and one framework for extremism detection could be identified, as shown in Table 13. The available tools are restricted to Government agencies, specific independent institutes, and respective social media platforms. Tools and frameworks available also focus on binary or tertiary classification. Thus, no insights on kinds of extremism

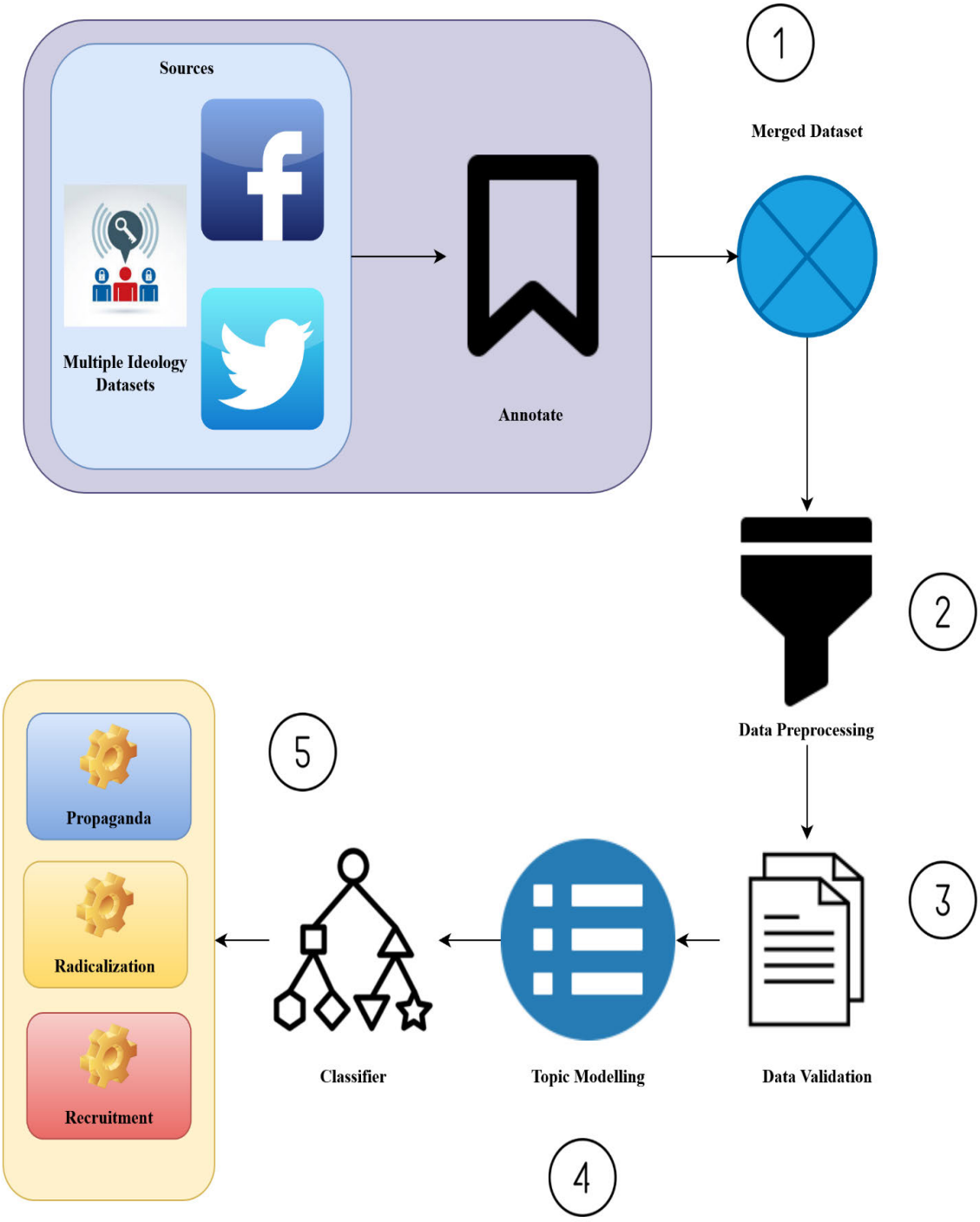


FIGURE 20. Architecture of proposed system for online extremism detection.

are gained. Therefore, there is a need for a comprehensive public tool for online extremism detection.

- *RQ6: Is there empirical evidence available that the current literature is biased towards a specific ideology?*

The word clouds were created for surveyed articles and reports. Using TF-IDF with uni-grams, word clouds for Title, Abstract, and Author Keywords for research articles and reports were generated. The words like 'jihad', 'jihadism', 'islamic', and 'isis' are observed in all the three Word clouds. By observing Fig. 10, Fig. 17, Fig. 18 and Fig. 19, it can be deduced that the researchers focus on fewer ideologies or organizations for extremism detection. The author J.M.Berger confirms this observation in the book 'Extremism', and states that research in jihadism outnumbers the study in white nationalism or supremism by a

3:1 ratio [163]. Thus, there is a need for research about extremism detection with the versatile ideologies.

Next section discusses the architecture of proposed system for online extremism detection, by addressing identifying research gaps in current section.

VII. ARCHITECTURE OF PROPOSED SYSTEM

The proposed architecture helps to identify extremism text and to further categorize it into propaganda, radicalization, and recruitment. The definition of these terms is provided in Section I(C). The outline of the architecture of the proposed system is elaborated in the following steps.

- Step 1: Dataset Construction

As shown in Fig. 20, the extremism text will be extracted from various sources like Twitter, Facebook, publicly available extremism standard datasets, and custom extremism datasets. This data will be collected using the standard Application Programming Interfaces (API) of Twitter and Facebook. After extraction of data, merging operation will be performed by stacking the datasets. Data augmentation techniques can be used to improve the class balance [45], [164]. Data annotation will be performed by labeling extremism texts using measures of similarity [113], [130].

- Step 2: Data Preprocessing:

In data pre-processing, different operations like removing stopwords, tokenization, and lemmatization [67] can be carried out. After these operations, cleaned data with tokens will be obtained.

- Step 3: Data Validation

Data validation techniques can be used to improve the quality of training data. The quality of data will be assessed using statistical methods to find the level of significance (p-value) of features [52]. Manual methods like crowd-sourcing using different tools like Amazon Mechanical Turk [86] can also be employed to validate the data.

- Step 4: Topic Modeling

Validated data is passed as input to topic modeling algorithms like Latent Dirichlet Allocation [52] and Hierarchical Dirichlet Allocation [165] which will form topics based on the group of tokens appearing together. We propose use of literary articles, tweets, and messages for seeding datasets of propaganda, radicalization, and recruitment extremism text. Topic modeling methods will help to cluster the extremism text into three peculiar types: propaganda, radicalization, and recruitment.

- Step 5: Classification

SVM, and Decision Trees classifiers are trained on dataset from Step 4 to classify the extremism text into three categories as propaganda, radicalization and recruitment. The classifiers can also evaluate the accuracy of topic modeling.

VIII. LIMITATIONS OF WORK

This systematic literature review is limited by keywords used, while selecting the literature. Extremism research is a specialized domain, and thus very few published investigations are available. Only 64 studies were selected for survey, including grey literature like thesis and technical and analytical reports from 2015 to 2020. Manual screening of studies was conducted which were obtained from libraries like SCOPUS, ACM, Web of Science, and IEEE. This review was limited to popular techniques like network/ graph, ML, and DL. The study was limited to a few ideologies or organizations like ISIS, Alt-Right, jihadism, and white nationalism retrieved by the query. Thus, survey suffers from the threat of bias due to the search query used. This paper shows conceptualization of the proposed system for online extremism detection as alternative to studied solutions in literature with respect to the variety of datasets, lack of validation and binary classification. The proposed architecture is under experimentation and evaluation is not stated in this paper.

IX. CONCLUSION

In this study, different SLR phases were planned, conducted, and executed on online extremism detection. The research questions were based on criteria like datasets, methodologies, classification methods, validation methods, and investigating bias in the literature.

In this study, different datasets used by extremism researchers were analyzed. Datasets were classified into two types: Standard dataset and Custom dataset. It was observed that datasets are collected from diverse sources ranging from magazines to social media. It can be concluded that there is a need for publicly available and verified standard datasets in online extremism research.

Existing literature on extremism research shows manual and automated detection trends. Manual methods had problems like time consumption, limited number of identified users and individual bias. It was observed that automated methods are more popular among researchers. ML approaches are more popular in automated extremism research. However, DL-based approaches are, however, evolving for extremism detection research with larger datasets and automated feature extraction. The proposed architecture aims to build an ideology independent, unbiased, balanced, verified extremism text dataset that will be publicly available to enhance future research in online extremism detection.

In RQ3, features, algorithms, and evaluation metrics used in online extremism detection research were identified. Most researchers have used binary or tertiary classification, so online extremism is evolving, making multiclass classification necessary. The proposed system aims to further enhance extremism detection research by categorizing extremism text into multiple classes like propaganda, radicalization, and recruitment. With a study on data validation techniques, it can be concluded that manual inter-rater agreement has limitations. There is a need for statistical validation techniques for verification and validation of extremism data. The accuracy of DL-based extremism detection methods can be improved with unbiased, validated extremism text dataset that will be constructed with proposed architecture. A hybrid

approach of data validation with statistical techniques and crowd-sourcing can be explored to reduce the expert bias.

In the study, tools available for extremism detection were studied. Most of the tools are proprietary and private in access. It can be concluded that there is a need for publicly available tools for online extremism detection. It can be observed that researches on online extremism detection are limited to social media and on a particular ideology. It can be concluded that there is a need for extremism detection research irrespective of ideology.

Conclusively the review opens the opportunities for research in online extremism detection, classification, validation methods with robust datasets, extremism detection tools that are not limited to fewer ideologies.

X. FUTURE WORK

This SLR will help researchers, intelligence analysts and government agencies to draw up and compare datasets, techniques, and methods to identify online extremism.

A. DATASETS

It is observed from the literature, that the creation of validated, verified, and publicly available datasets are prerequisite to accurate extremism detection research. Construction of multilabel dataset, with blend of extremism text from multiple ideologies, for enhanced extremism research is need of time.

B. METHODOLOGY, TECHNIQUE AND CONTEXT

With the abilities like context identification, remembering long-term dependencies, and handling large data need of employing pre-trained models and multiclass classification for more accurate extremism detection research.

C. GEOGRAPHICAL CONTEXT

Detecting extremism content is difficult due to the varied geographical and political scenarios. This concern can be handled by geographical or the location-based extremism detection research. Identifying geographical location or target from the extremist text can help security agencies to concentrate their efforts.

D. COMPREHENSIVE TOOL

Social media spreads information and misinformation at unprecedented speed, so there is a case for real-time extremism detection. There is a dire need for a publicly available tool, that the researchers and social media can use to detect online extremism.

ACKNOWLEDGMENT

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

GLOSSARY

- ISIS–Islamic State of Iraq and Syria
- ML–Machine Learning
- DL–Deep Learning
- SVM–Support Vector Machine
- CNN–Convolutional Neural Network
- LSTM–Long Short-Term Memory
- BERT–Bidirectional Encoder Representations from Transformers
- SLR–Systematic Literature Reviews
- TF-IDF–Term Frequency-Inverse Document Frequency
- PCA–Principal Component Analysis
- t-SNE–t-Distributed Stochastic Neighbour Embedding
- TP–True Positive, FP - False Positive, TN - True Negative and FN - False Negative
- *Betweenness centrality*–a subset of all minimal paths between users
- *Proximity prestige*–the normalized mean distance between each reachable node from the current vertex
- *In-degree centrality*–number of incoming edges.
- *ROC-AUC Curve*: Receiver Operating Characteristic (ROC) is a bidirectional graph used to evaluate and compare the performance of the classifiers [73]. Area Under Curve (AUC) is a scalar value, representing the overall performance of the binary classifier [74]. AUC is a reliable measure for classification performance, as it provides performance measures at all possible thresholds. ROC-AUC curve is recommended, when all the classification labels are balanced.
- *Precision*: Precision provides the number of predicted positives that were truly correct [75]. It can be given by:

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{1}$$

Precision is used when the dataset is unbalanced, and the number of false positives is high [76].

- *Recall*: Recall gives the number of actual positive classes that were predicted positive [77]. It can be provided by:

$$Recall = \frac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FN}} \tag{2}$$

The recall is also used, when the dataset is unbalanced, and the number of false negatives is high [76].

- *Accuracy*: Calculates the correctly predicted instances. The accuracy fails if classes are imbalanced. It can be given by:

$$Accuracy = \frac{\mathbf{TP} + \mathbf{TN}}{\mathbf{TP} + \mathbf{FP} + \mathbf{TN} + \mathbf{FN}} \tag{3}$$

Accuracy works correctly when classes are balanced [76].

- *F1-measure / F1-Score*: Combines both precision and recall and present their harmonic mean. It can be stated as:

$$F1 - score = \mathbf{2}*\frac{\mathbf{Precision}*\mathbf{Recall}}{\mathbf{Precision} + \mathbf{Recall}} \tag{4}$$

F1-measure or F1-score is used when data is unbalanced, and if the difference between Precision and Recall is significant.

- INSiGHT–Investigative Search for Graph Trajectories
- OSINT–Open Source Intelligence
- WPIE–Whole Post Integrity Embedding

REFERENCES

[1] J. Schultz. (2019). *How Much Data is Created on Internet Each Day? Micro Focus Blog*. Accessed: May 14, 2020. [Online]. Available: https://blog.microfocus.com/how-much-data-is-created-on-the-internet-each-day/

[2] 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 (ASONAM)*, Jul. 2009, pp. 231–236, doi: 10.1109/ASONAM.2009.31.

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

[4] D. Milton. (2016). Communication breakdown: Unraveling the islamic state's media efforts. New York, NY, USA. [Online]. Available: https://www.stratcomcoe.org/milton-d-communication-breakdown-unraveling-islamic-states-media-efforts

[5] Congress.Gov. (2019). *Domestic Terrorism Prevention Act of 2019*. Accessed: Oct. 10, 2020. [Online]. Available: https://www.congress.gov/bill/116th-congress/senate-bill/894

[6] J. M. Berger. (2018). *The Alt-Right Twitter Census: Defining and Describing the Audience for Alt-Right Content on Twitter*. [Online]. Available: https://www.voxpol.eu/download/vox-pol_publication/AltRightTwitterCensus.pdf

[7] C. Timberg, D. Harwell, E. Dwoskin, and T. Romm, "How social media's business model helped the New Zealand massacre go viral," Washington Post, Mar. 19, 2019.

[8] S. Baele. (2018). ISIS online propaganda. CREST Research. Accessed: Apr. 6, 2020. [Online]. Available: https://crestresearch.ac.uk/projects/isis-online-propaganda/

[9] J. Matusitz, A. Madrazo, and C. Udani, "Inspire, Dabiq, Rumiyah, and Gaidi Mtaani," in *Online Jihadist Magazines to Promote the Caliphate*. New York, NY, USA: Peter Lang, 2019, ch. 6.

[10] J. Cavendish, "Al-Qa'ida glossy advises women to cover up and marry a martyr," The Independent, 2011. Accessed: May 15, 2020. [Online]. Available: https://www.independent.co.uk/news/world/asia/al-qaida-glossy-advises-women-to-cover-up-and-marry-a-martyr-2240992.html

[11] *Stormfront*. Accessed: Aug. 20, 2020. [Online]. Available: https://www.stormfront.org

[12] J. Wilson, "Leak from neo-Nazi site could identify hundreds of extremists worldwide," The Guardian, 2019. Accessed: Aug. 20, 2020. [Online]. Available: https://www.theguardian.com/us-news/2019/nov/07/neo-nazi-site-iron-march-materials-leak

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

[14] A.-M. Bliuc, N. Faulkner, A. Jakubowicz, and C. McGarty, "Online networks of racial hate: A systematic review of 10 years of research on cyber-racism," *Comput. Hum. Behav.*, vol. 87, pp. 75–86, Oct. 2018, doi: 10.1016/j.chb.2018.05.026.

[15] Facebook. (2020). *Hate Speech*. Accessed: Aug. 20, 2020. [Online]. Available: https://www.facebook.com/communitystandards/recentupdates/hate_speech/

[16] Twitter Safety. (2020). *Updating Our Rules Against Hateful Conduct*. Accessed: Aug. 20, 2020. [Online]. Available: https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate.html

[17] I. Von Behr, A. Reding, C. Edwards, and L. Gribbon. (2013). *Radicalisation in the Digital Era*. [Online]. Available: https://www.rand.org/content/dam/rand/pubs/research_reports/RR400/RR453/RAND_RR453.pdf

[18] J. M. Berger and J. Morgan, "The ISIS Twitter census," Center Middle East Policy, Brookings Inst., Washington, DC, USA, Tech. Rep. 20, 2015.

[19] K. Cohen, F. Johansson, L. Kaati, and J. C. Mork, "Detecting linguistic markers for radical violence in social media," *Terrorism Political Violence*, vol. 26, no. 1, pp. 246–256, Jan. 2014, doi: 10.1080/09546553.2014.849948.

[20] J. M. Berger, "The dangerous spread of extremist manifestos," The Atlantic, 2019. Accessed: Oct. 10, 2020. [Online]. Available: https://www.theatlantic.com/ideas/archive/2019/02/christopher-hasson-was-inspired-breivik-manifesto/583567/

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

[22] L. Turner, "The path to terrorism: The islamic state and its recruitment strategies," Univ. Connecticut, Storrs, CT, USA, Tech. Rep. 585, 2018.

[23] S. Jaki and T. De Smedt, "Right-wing German hate speech on Twitter: Analysis and automatic detection," Oct. 2019, *arXiv:1910.07518*. [Online]. Available: http://arxiv.org/abs/1910.07518

[24] S. C. Dharmadhikari, M. Ingle, and P. Kulkarni, "Empirical studies on machine learning based text classification algorithms," *Adv. Comput. Int. J.*, vol. 2, no. 6, pp. 161–169, Nov. 2011, doi: 10.5121/acij.2011.2615.

[25] Merriam-Webster. *Ideology*. Accessed: Oct. 10, 2020. [Online]. Available: https://www.merriam-webster.com/dictionary/ideology

[26] T. A. van Dijk, "Politics, ideology, and discourse," in *Encyclopedia of Language & Linguistics*. Amsterdam, The Netherlands: Elsevier, 2006, pp. 728–740.

[27] The Free Dictionary. *Extremism Definition*. Accessed: Oct. 10, 2020. [Online]. Available: https://www.thefreedictionary.com/extremism

[28] S. M. Lipset, "Social stratification and 'right-wing extremism,'" *Brit. J. Sociol.*, vol. 10, no. 4, pp. 346–382, Oct. 1959, doi: 10.2307/587800.

[29] S. Wibisono, W. R. Louis, and J. Jetten, "A multidimensional analysis of religious extremism," *Frontiers Psychol.*, vol. 10, p. 2560, Nov. 2019, doi: 10.3389/fpsyg.2019.02560.

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

[31] S. S. M. Hamdani, "Techniques of online propaganda: A case study of western sahara conflict," *Int. J. Media, J. Mass Commun.*, vol. 3, no. 2, pp. 18–24, 2017, doi: 10.20431/2454-9479.0302003.

[32] M.-R. Ali. (2015). *Isis and Propaganda: How Isis Exploits Women*. [Online]. Available: https://reutersinstitute.politics.ox.ac.uk/our-research/isis-and-propaganda

[33] 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.

[34] US Department of Justice. (2018). *Awareness Brief?: Online Radicalization to Violent Extremism*. Accessed: Oct. 10, 2020. [Online]. Available: https://www.theiacp.org/sites/default/files/2018-07/RadicalizationtoViolentExtremismAwarenessBrief.pdf

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

[36] A. Dornbierer, "How Al-Qaeda recruits online," The Diplomat, 2011. Accessed: May 15, 2020. [Online]. Available: https://thediplomat.com/2011/09/how-al-qaeda-recruits-online/

[37] N. Gisev, J. S. Bell, and T. F. Chen, "Interrater agreement and interrater reliability: Key concepts, approaches, and applications," *Res. Social Administ. Pharmacy*, vol. 9, no. 3, pp. 330–338, May 2013, doi: 10.1016/j.sapharm.2012.04.004.

[38] 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.

[39] M. Chau and J. Xu, "Mining communities and their relationships in blogs: A study of online hate groups," *Int. J. Hum.-Comput. Stud.*, vol. 65, no. 1, pp. 57–70, Jan. 2007, doi: 10.1016/j.ijhcs.2006.08.009.

[40] M. Fernandez and H. Alani, "Artificial intelligence and online extremism: Challenges and opportunities," in *Predictive Policing and Artificial Intelligence*, J. McDaniel and K. Pease, Eds. New York, NY, USA: Taylor & Francis, 2020.

[41] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," *ACM Comput. Surv.*, vol. 51, no. 4, pp. 1–30, Sep. 2018, doi: 10.1145/3232676.

[42] A. Al-Hassan and H. Al-Dossari, "Detection of hate speech in social networks: A survey on multilingual corpus," in *Proc. 6th Int. Conf. Comput. Sci. Inf. Technol.*, 2019, pp. 1–18, doi: 10.5121/csit.2019.90208.

[43] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," Dept. Eng., Durham Univ., Keele Univ., Keele, U.K., Tech. Rep. EBSE-2007-01, 2007. [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.117.471&rep=rep1&type=pdf

[44] M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochem. Medica*, vol. 22, no. 3, pp. 276–282, 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23092060

[45] 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.

[46] K. Deb, S. Paul, and K. Das, "A framework for predicting and identifying radicalization and civil unrest oriented threats from Whatsapp group," in *Emerging Technology in Modelling and Graphics*. Singapore: Springer, 2020, pp. 595–606.

[47] M. Heidarysafa, K. Kowsari, T. Odukoya, P. Potter, L. E. Barnes, and D. E. Brown, "Women in ISIS propaganda: A natural language processing analysis of topics and emotions in a comparison with a mainstream religious group," in *Proc. Sci. Inf. Conf.*, 2020, pp. 610–624.

[48] 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.

[49] A. Kaur, J. K. Saini, and D. Bansal, "Detecting radical text over online media using deep learning," Jul. 2019, *arXiv:1907.12368*. [Online]. Available: http://arxiv.org/abs/1907.12368

[50] I. Gialampoukidis, G. Kalpakis, T. Tsikrika, S. Papadopoulos, S. Vrochidis, and I. Kompatsiaris, "Detection of terrorism-related Twitter communities using centrality scores," in *Proc. 2nd Int. Workshop Multimedia Forensics Secur. (ICMR)*, Jun. 2017, pp. 21–25, doi: 10.1145/3078897.3080534.

[51] LuckyTrollClub. (2015). *Lucky Troll Club Archive*. Accessed: Oct. 10, 2020. [Online]. Available: https://archive.is/9ZXeA

[52] 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," *Proc. ACM Hum.-Comput. Interact.*, vol. 3, pp. 1–22, Nov. 2019, doi: 10.1145/3359253.

[53] 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.

[54] 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.

[55] 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.

[56] Z. He, S. Deng, X. Xu, and J. Z. Huang, "A fast greedy algorithm for outlier mining," in *Advances in Knowledge Discovery and Data Mining*. Berlin, Germany: Springer, 2006, pp. 567–576.

[57] L. Rita. (2020). *Louvain Algorithm Towards Data Science*. Accessed: Oct. 10, 2020. [Online]. Available: https://towardsdatascience.com/louvain-algorithm-93fde589f58c

[58] X. Yan and J. Han, "CloseGraph: Mining closed frequent graph patterns," *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2003, pp. 286–295, doi: 10.1145/956750.956784.

[59] C. Jiang, F. Coenen, R. Sanderson, and M. Zito, "Text classification using graph mining-based feature extraction," *Knowl.-Based Syst.*, vol. 23, no. 4, pp. 302–308, May 2010, doi: 10.1016/j.knosys.2009.11.010.

[60] A. H. Osman and O. M. Barukub, "Graph-based text representation and matching: A review of the state of the art and future challenges," *IEEE Access*, vol. 8, pp. 87562–87583, 2020, doi: 10.1109/ACCESS.2020.2993191.

[61] G. Paltoglou and M. Thelwall, "More than bag-of-words: Sentence-based document representation for sentiment analysis," in *Proc. RANLP*, Hisarya, Bulgaria, 2013, pp. 546–552.

[62] J. Ramos, "Using TF-IDF to determine word relevance in document queries," in *Proc. 1st Inst. Conf. Mach. Learn.*, 2003, pp. 29–48.

[63] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," 2013, *arXiv:1310.4546*. [Online]. Available: https://arxiv.org/abs/1310.4546

[64] Sciforce. (2018). *Word Vectors in NLP*. Accessed: Oct. 10, 2020. [Online]. Available: https://medium.com/sciforce/word-vectors-in-natural-language-processing-global-vectors-glove-51339db89639

[65] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Phil. Trans. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 374, no. 2065, Apr. 2016, Art. no. 20150202, doi: 10.1098/rsta.2015.0202.

[66] L. Van Der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 1–27, 2008.

[67] S. Agarwal and A. Sureka, "Using KNN and SVM based one-class classifier for detecting online radicalization on Twitter," in *Proc. Int. Conf. Distrib. Comput. Internet Technol.*, 2015, pp. 431–442, doi: 10.1007/978-3-319-14977-6_47.

[68] M. F. Abrar, M. S. Arefin, and M. S. Hossain, "A framework for analyzing real-time tweets to detect terrorist activities," in *Proc. 2nd Int. Conf. Electr., Comput. Commun. Eng. (ECCE)*, Feb. 2019, pp. 1–6, doi: 10.1109/ECACE.2019.8679430.

[69] M. Ashcroft, A. Fisher, L. Kaati, E. Omer, and N. Prucha, "Detecting Jihadist messages on Twitter," in *Proc. Eur. Intell. Secur. Informat. Conf.*, Sep. 2015, pp. 161–164, doi: 10.1109/EISIC.2015.27.

[70] 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.

[71] T. De Smedt, G. De Pauw, and P. Van Ostaeyen, "Automatic detection of online Jihadist hate speech," Feb. 2018, *arXiv:1803.04596*. [Online]. Available: http://arxiv.org/abs/1803.04596

[72] M. C. Benigni, K. Joseph, and K. M. Carley, "Online extremism and the communities that sustain it: Detecting the ISIS supporting community on Twitter," *PLoS ONE*, vol. 12, no. 12, pp. 1–23, 2017, doi: 10.1371/journal.pone.0181405.

[73] F. Melo, "Receiver operating characteristic (ROC) curve," in *Encyclopedia of Systems Biology*. New York, NY, USA: Springer, 2013, pp. 1818–1823.

[74] F. Melo, "Area under the ROC curve," in *Encyclopedia of Systems Biology*. New York, NY, USA: Springer, 2013, pp. 38–39.

[75] H. Wang and H. Zheng, "Positive predictive value," in *Encyclopedia of Systems Biology*. New York, NY, USA: Springer, 2013, pp. 1723–1724.

[76] K. P. Shung. (2018). *Accuracy, Precision, Recall or F1? Towards Data Science*. Accessed: Oct. 10, 2020. [Online]. Available: https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

[77] H. Wang and H. Zheng, "True positive rate," in *Encyclopedia of Systems Biology*. New York, NY, USA: Springer, 2013, pp. 2302–2303.

[78] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning based text classification: A comprehensive review," Apr. 2020, *arXiv:2004.03705*. [Online]. Available: http://arxiv.org/abs/2004.03705

[79] D. Yan and S. Guo, "Leveraging contextual sentences for text classification by using a neural attention model," *Comput. Intell. Neurosci.*, vol. 2019, pp. 1–11, Aug. 2019, doi: 10.1155/2019/8320316.

[80] J. Brownlee. (2020). *What is Deep Learning? Machine Learning Mastery*. Accessed: Oct. 10, 2020. [Online]. Available: https://machinelearningmastery.com/what-is-deep-learning/

[81] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proc. INTERSPEECH*, Singapore, 2014, pp. 338–342.

[82] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: An overview and application in radiology," *Insights Imag.*, vol. 9, no. 4, pp. 611–629, Aug. 2018, doi: 10.1007/s13244-018-0639-9.

[83] Y. Kim, "Convolutional neural networks for sentence classification," Aug. 2014, *arXiv:1408.5882*. [Online]. Available: http://arxiv.org/abs/1408.5882

[84] K. Biney. (2018). *Sentiment Analysis Using 1D-CNN in Keras*. Accessed: Oct. 10, 2020. [Online]. Available: https://romannempyre.medium.com/sentiment-analysis-using-1d-convolutional-neural-networks-part-1-f8b6316489a2
[85] O. de Gibert, N. Perez, A. García-Pablos, and M. Cuadros, "Hate speech dataset from a white supremacy forum," Sep. 2018, *arXiv:1809.04444*. [Online]. Available: http://arxiv.org/abs/1809.04444
[86] 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," Oct. 2020, *arXiv:2010.00357*. [Online]. Available: http://arxiv.org/abs/2010.00357
[87] TensorFlow. (2020). *Embedding Layer Tensorflow*. Accessed: Oct. 10, 2020. [Online]. Available: https://www.tensorflow.org/api_docs/python/tf/keras/layers/Embedding
[88] PyTorch Documentation. *Embedding Layer PyTorch*. Accessed: Oct. 10, 2020. [Online]. Available: https://pytorch.org/docs/stable/generated/torch.nn.Embedding.html
[89] 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.
[90] 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/_les/zaxdzs2191/f/downloads/Nazisv.ISIS.pdf
[91] R. Lara-Cabrera, A. G. Pardo, K. Benouaret, N. Faci, D. Benslimane, and D. Camacho, "Measuring the radicalisation risk in social networks," *IEEE Access*, vol. 5, pp. 10892–10900, 2017, doi: 10.1109/ACCESS.2017.2706018.
[92] R. Lara-Cabrera, A. Gonzalez-Pardo, and D. Camacho, "Statistical analysis of risk assessment factors and metrics to evaluate radicalisation in Twitter," *Future Gener. Comput. Syst.*, vol. 93, pp. 971–978, Apr. 2019, doi: 10.1016/j.future.2017.10.046.
[93] P. C. De Bruyn, "Predicting behavioral profiles of online extremists through linguistic use of social roles," *Behav. Sci. Terrorism Political Aggression*, pp. 1–25, Jul. 2020, doi: 10.1080/19434472.2020.1775675.
[94] M. Benigni, "Detection and analysis of online extremist communities," Ph.D. dissertation, Inst. Softw. Res., Carnegie Mellon Univ., Pittsburgh, PA, USA, 2017.
[95] Y. Al-Saggaf, "Understanding online radicalisation using data science," *Int. J. Cyber Warfare Terrorism*, vol. 6, no. 4, pp. 13–27, Oct. 2016, doi: 10.4018/IJCWT.2016100102.
[96] U. Xie, J. Xu, and T.-C. Lu, "Automated classification of extremist Twitter accounts using content-based and network-based features," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2016, pp. 2545–2549, doi: 10.1109/BigData.2016.7840895.
[97] 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.
[98] A. Johnston and A. Marku, "Identifying extremism in text using deep learning," in *Development and Analysis of Deep Learning Architectures* (Studies in Computational Intelligence), vol. 867. Cham, Switzerland: Springer, 2020, pp. 267–289, doi: 10.1007/978-3-030-31764-5_10.
[99] 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. (WebSci)*, 2019, pp. 203–204, doi: 10.1145/3292522.3326050.
[100] M. Bloom, H. Tiflati, and J. Horgan, "Navigating ISIS's preferred platform: Telegram1," *Terrorism Political Violence*, vol. 31, no. 6, pp. 1242–1254, Nov. 2019, doi: 10.1080/09546553.2017.1339695.
[101] H. Agerholm, "ISIS using Whatsapp and Telegram to sell sex slaves," The Independent, Jul. 7, 2016.
[102] Fifth Tribe. (2015). How ISIS uses Twitter. Kaggle. [Online]. Available: https://www.kaggle.com/fifthtribe/how-isis-uses-twitter
[103] C. Sánchez-Rebollo, C. Puente, R. Palacios, C. Piriz, J. P. Fuentes, and J. Jarauta, "Detection of Jihadism in social networks using big data techniques supported by graphs and fuzzy clustering," *Complexity*, vol. 2019, pp. 1–13, Mar. 2019, doi: 10.1155/2019/1238780.
[104] D. Parekh, A. Amarasingam, L. Dawson, and D. Ruths, "Studying Jihadists on social media: A critique of data collection methodologies," *Perspect. Terrorism*, vol. 12, no. 3, pp. 3–21, 2018.
[105] T. De Smedt, S. Jaki, E. Kotzé, L. Saoud, M. Gwóźdź, G. De Pauw, and W. Daelemans, "Multilingual cross-domain perspectives on online hate speech," CLiPS Res. Center, Antwerp, Belgium, Tech. Rep. CTRS-008, Sep. 2018, vol. 8. [Online]. Available: http://arxiv.org/abs/1809.03944
[106] A. Tundis, L. Böck, V. Stanilescu, and M. Mühlhäuser, "Experiencing the detection of radicalized criminals on Facebook social network and data-related issues," *J. Cyber Secur. Mobility*, pp. 203–236, Jan. 2020, doi: 10.13052/JCSM2245-1439.922.
[107] E. Mouhssine and C. Khalid, "Social big data mining framework for extremist content detection in social networks," in *Proc. Int. Symp. Adv. Elect. Commun. Technol.*, 2019, pp. 1–5, doi: 10.1109/ISAECT.2018.8618726.
[108] (2016). *Women and Violent Radicalization in Jordan*. [Online]. Available: https://www.refworld.org/pdfid/5881d4e44.pdf
[109] L. Windsor, "The language of radicalization: Female Internet recruitment to participation in ISIS activities," *Terrorism Political Violence*, vol. 32, no. 3, pp. 506–538, 2020, doi: 10.1080/09546553.2017.1385457.
[110] A. I. Kapitanov, I. I. Kapitanova, V. M. Troyanovskiy, V. F. Shangin, and N. O. Krylikov, "Approach to automatic identification of terrorist and radical content in social networks messages," in *Proc. IEEE Conf. Russian Young Researchers Electr. Electron. Eng. (EIConRus)*, Jan. 2018, pp. 1517–1520, doi: 10.1109/EIConRus.2018.8317386.
[111] S. Agarwal and A. Sureka, "Topic-specific YouTube crawling to detect online radicalization," in *Proc. Int. Workshop Databases Netw. Inf. Syst.*, 2015, pp. 133–151.
[112] A. Badawy and E. Ferrara, "The rise of Jihadist propaganda on social networks," *J. Comput. Social Sci.*, vol. 1, no. 2, pp. 453–470, Sep. 2018, doi: 10.1007/s42001-018-0015-z.
[113] 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.
[114] M. Rowe and H. Saif. (2016). *Mining Pro-ISIS Radicalisation Signals From Social Media Users*. [Online]. Available: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13023/12752
[115] 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.
[116] C. Charles, "(Main)streaming hate: Analyzing white supremacist content and framing devices on YouTube," Ph.D. dissertation, Dept. Sociol., Univ. Central Florida, Orlando, FL, USA, 2020.
[117] E. Ferrara, W.-Q. Wang, O. Varol, A. Flammini, and A. Galstyan, "Predicting online extremism, content adopters, and interaction reciprocity," in *Proc. Int. Conf. Social Inform.*, 2016, pp. 22–39.
[118] 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.
[119] M. Fraiwan, "Identification of markers and artificial intelligence-based classification of radical Twitter data," *Appl. Comput. Inform.*, vol. 16, no. 1, pp. 1–7, 2020.
[120] S. Mussiraliyeva, M. Bolatbek, B. Omarov, and K. Bagitova, "Detection of extremist ideation on social media using machine learning techniques," in *Proc. Int. Conf. Comput. Collective Intell.*, 2020, pp. 743–752, doi: 10.1007/978-3-030-63007-2_58.
[121] J. Torregrosa, J. Thorburn, R. Lara-Cabrera, D. Camacho, and H. M. Trujillo, "Linguistic analysis of pro-ISIS users on Twitter," *Behav. Sci. Terrorism Political Aggression*, vol. 12, no. 3, pp. 171–185, Jul. 2020, doi: 10.1080/19434472.2019.1651751.
[122] M. Neelamalar and M. V. Vivakaran, "A critical analysis of the Jihadi discourse through online magazines with special reference to 'Wyeth' magazine," *India Quart., J. Int. Affairs*, vol. 75, no. 4, pp. 456–471, Dec. 2019, doi: 10.1177/0974928419874548.
[123] M. Lakomy, "Recruitment and incitement to violence in the islamic state's online propaganda: Comparative analysis of Dabiq and Rumiyah," *Stud. Conflict Terrorism*, pp. 1–16, Feb. 2019, doi: 10.1080/1057610X.2019.1568008.
[124] S. J. Baele, G. Bettiza, K. A. Boyd, and T. G. Coan, "ISIS's clash of civilizations: Constructing the 'west' in terrorist propaganda," *Stud. Conflict Terrorism*, pp. 1–33, Apr. 2019, doi: 10.1080/1057610X.2019.1599192.
[125] 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.
[126] J. Klausen, C. E. Marks, and T. Zaman, "Finding extremists in online social networks," *Oper. Res.*, vol. 66, no. 4, pp. 957–976, Aug. 2018, doi: 10.1287/opre.2018.1719.

[127] T. Welch, "Theology, heroism, justice, and fear: An analysis of ISIS propaganda magazines Dabiq and Rumiyah," *Dyn. Asymmetric Conflict*, vol. 11, no. 3, pp. 186–198, Sep. 2018, doi: 10.1080/17467586.2018.1517943.
[128] S. A. Azizan and I. A. Aziz, "Terrorism detection based on sentiment analysis using machine learning.pdf," *J. Eng. Appl. Sci.*, vol. 12, no. 3, pp. 691–698, 2017.
[129] 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.
[130] M. Fernandez and H. Alani. (2018). *Contextual Semantics for Radicalisation Detection on Twitter*. [Online]. Available: http://ceur-ws.org/Vol-2182/paper_4.pdf
[131] 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.
[132] M. Al-Zewairi and G. Naymat, "Spotting the islamist radical within: Religious extremists profiling in the United State," *Procedia Comput. Sci.*, vol. 113, pp. 162–169, Jan. 2017, doi: 10.1016/j.procs.2017.08.336.
[133] D. López-Sánchez, J. M. Corchado, and A. G. Arrieta, "Dynamic detection of radical profiles in social networks using image feature descriptors and a case-based reasoning methodology," in *Proc. Int. Conf. Case-Based Reasoning*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11156, 2018, pp. 219–232, doi: 10.1007/978-3-030-01081-2_15.
[134] *ARY News*. Accessed: Oct. 10, 2020. [Online]. Available: https://arynews.tv/en/
[135] *PTV News*. Accessed: Oct. 10, 2020. [Online]. Available: https://ptv.com.pk/
[136] *Dawn News*. Accessed: Oct. 10, 2020. [Online]. Available: https://www.dawn.com/
[137] *The News Pakistan*. Accessed: Oct. 10, 2020. [Online]. Available: https://www.thenews.com.pk
[138] *Samaa News*. Accessed: Oct. 10, 2020. [Online]. Available: https://www.samaa.tv/
[139] *Express News*. Accessed: Oct. 10, 2020. [Online]. Available: https://www.express.com.pk/
[140] *Dunya News*. Accessed: Oct. 10, 2020. [Online]. Available: https://dunyanews.tv/
[141] *Geo News*. Accessed: Oct. 10, 2020. [Online]. Available: https://www.geo.tv/
[142] M. Alfifi, P. Kaghazgaran, J. Caverlee, and F. Morstatter, "A large-scale study of ISIS social media strategy: Community size, collective influence, and behavioral impact," in *Proc. Int. AAAI Conf. Web Social Media*, 2019, pp. 58–67.
[143] M. Oussalah, F. Faroughian, and P. Kostakos, "On detecting online radicalization using natural language processing," in *Proc. Int. Conf. Intell. Data Eng. Automated Learn.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 11315, 2018, pp. 21–27, doi: 10.1007/978-3-030-03496-2_4.
[144] S. M. Mohammad and S. Kiritchenko, "Using hashtags to capture fine emotion categories from tweets," *Comput. Intell.*, vol. 31, no. 2, pp. 301–326, May 2015, doi: 10.1111/coin.12024.
[145] *TextRazor Semantic Annotator*. Accessed: Oct. 10, 2020. [Online]. Available: https://www.textrazor.com/
[146] Vkontakte. (2014). *Vkontakte Social Network*. Accessed: Oct. 10, 2020. [Online]. Available: https://vk.com/topic-78863260_30603285
[147] K. Willsher, "Charlie Hebdo attack," The Guardian, 2020. Accessed: Dec. 16, 2020. [Online]. Available: https://www.theguardian.com/world/2020/dec/16/charlie-hebdo-trial-french-court-convicts-14-over-2015-terror-attacks
[148] F. Farid, "Sinai attack," The Age, 2017. Accessed: Oct. 10, 2020. [Online]. Available: https://www.theage.com.au/world/egypt-launches-airstrikes-after-deadly-mosque-attack-20171125-gzstmg.html
[149] ActiveGalaxy, Kaggle. (2016). *ISIS Related Dataset*. Accessed: Oct. 10, 2020. [Online]. Available: https://www.kaggle.com/activegalaxy/isis-related-tweets
[150] FifthTribe, Kaggle. (2017). *ISIS Religious Text*. Accessed: Oct. 10, 2020. [Online]. Available: https://www.kaggle.com/fifthtribe/isis-religious-texts
[151] T. Ruttig. (2018). *Kunduz Madrassa Attack Al Jazeera*. Accessed: Oct. 10, 2020. [Online]. Available: https://www.aljazeera.com/opinions/2018/4/5/kunduz-madrassa-attack-losing-the-moral-high-ground
[152] *Incels.me*. Accessed: Oct. 10, 2020. [Online]. Available: https://incels.co/
[153] K. Al-Rowaily, M. Abulaish, N. A.-H. Haldar, and M. Al-Rubaian, "BiSAL—A bilingual sentiment analysis lexicon to analyze dark Web forums for cyber security," *Digit. Invest.*, vol. 14, pp. 53–62, Sep. 2015, doi: 10.1016/j.diin.2015.07.006.
[154] J. Kranjc, R. Orač, V. Podpečan, N. Lavrač, and M. Robnik-Šikonja, "ClowdFlows: Online workflows for distributed big data mining," *Future Gener. Comput. Syst.*, vol. 68, pp. 38–58, Mar. 2017, doi: 10.1016/j.future.2016.07.018.
[155] Amazon. *Amazon Mechnical Turk*. Accessed: Oct. 10, 2020. [Online]. Available: https://www.mturk.com/
[156] J. Cohen, "A coefficient of agreement for nominal scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, Apr. 1960, doi: 10.1177/001316446002000104.
[157] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychol. Bull.*, vol. 76, no. 5, pp. 378–382, 1971, doi: 10.1037/h0031619.
[158] B. W. K. Hung, A. P. Jayasumana, and V. W. Bandara, "INSiGHT: A system to detect violent extremist radicalization trajectories in dynamic graphs," *Data Knowl. Eng.*, vol. 118, pp. 52–70, Nov. 2018, doi: 10.1016/j.datak.2018.09.003.
[159] RecordedFutureTeam. (2019). *Open Source Intelligence*. Accessed: Oct. 10, 2020. [Online]. Available: https://www.recordedfuture.com/open-source-intelligence-definition/
[160] J. Klausen, S. Campion, N. Needle, G. Nguyen, and R. Libretti, "Research note: Toward a behavioral model of 'homegrown' radicalization trajectories," *Stud. Conflict Terrorism*, vol. 39, no. 1, pp. 67–83, 2016, doi: 10.1080/1057610X.2015.1099995.
[161] M. Schroepfer. (2019). Community standards report. Facebook. Accessed: Oct. 10, 2020. [Online]. Available: https://ai.facebook.com/blog/community-standards-report/
[162] P. Wignell, K. Chai, S. Tan, K. O'Halloran, and R. Lange, "Natural language understanding and multimodal discourse analysis for interpreting extremist communications and the re-use of these materials online," *Terrorism Political Violence*, vol. 33, no. 1, pp. 71–95, Nov. 2018, doi: 10.1080/09546553.2018.1520703.
[163] J. M. Berger, "Jihadism studies outnumber white nationalism," in *Extremism*. Cambridge, MA, USA: MIT Press, 2018, p. 22.
[164] S. Das Bhattacharjee, B. V. Balantrapu, W. Tolone, and A. Talukder, "Identifying extremism in social media with multi-view context-aware subset optimization," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2017, pp. 3638–3647, doi: 10.1109/BigData.2017.8258358.
[165] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *J. Amer. Stat. Assoc.*, vol. 101, no. 476, pp. 1566–1581, 2006, doi: 10.1198/016214506000000302.

MAYUR GAIKWAD 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.

SWATI AHIRRAO received the Ph.D. degree from the Department of Computer Science and Information Technology, Symbiosis Institute of Technology (SIT), Symbiosis International (Deemed University), Pune, India. She is currently an Associate Professor with SIT. She has published over 29 research papers in international journals and conferences. According to Google Scholar, her articles have 60 citations, with an H-index of three and an i10-index of two. Her research interests include big data analytics, machine learning, and deep learning.

SHRADDHA PHANSALKAR received the Ph.D. degree from the Department of Computer Science and Information Technology, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, India. She is currently a Professor with MIT ADT, Pune. She has published two book chapters and over 16 research articles and presented six research papers in international journals and conferences, respectively. According to Google Scholar, her articles have 39 citations, with an H-index of four and an i10- index of one. Her research interests include big data analytics, machine learning, and deep learning predictive analytics for big data application, blockchain in finance, and healthcare.

KETAN KOTECHA has worked as an Administrator with Parul University and Nirma University and has a number of achievements in these roles to his credit. He has expertise and experience of cutting-edge research and projects in AI and deep learning for more than last 25 years. He also has pioneered Education Technology. He is a Team Member for the nationwide initiative on AI and deep learning skilling and research named Leadingindia.ai initiative sponsored by the Royal Academy of Engineering, the UK under Newton Bhabha Fund. He currently heads the Symbiosis Centre for Applied Artificial Intelligence (SCAAI). He is considered a Foremost Expert in AI and aligned technologies. He is also with his vast and varied experience in administrative roles. He has published widely in a number of excellent peer-reviewed journals on various topics ranging from education policies and teaching learning practices and AI for all.