IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 October 2023, accepted 19 October 2023, date of publication 23 October 2023, date of current version 2 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3326995

SURVEY

A Comprehensive Survey of Detection and Prevention Approaches for Online Radicalization: Identifying Gaps and Future Directions

OMRAN BERJAWI[1,2], GIUSEPPE FENZA[2], (Member, IEEE), AND VINCENZO LOIA[2], (Senior Member, IEEE)

[1]IMT School for Advanced Studies Lucca, 55100 Lucca, Italy
[2]Department of Management and Innovation Systems, University of Salerno, 84084 Fisciano, Italy

Corresponding author: Giuseppe Fenza (gfenza@unisa.it)

This work was supported in part by Project SERICS under the Ministero dell'Università e della Ricerca (MUR) National Recovery and Resilience Plan funded by the European Union—NextGenerationEU under Grant PE00000014.

ABSTRACT In the digital era, online radicalization has emerged as a significant concern for governments, social media platforms, and researchers. Detecting and preventing online radicalization have become key priorities, leading to extensive research efforts. This study presents a comprehensive survey of existing works in this field, covering various techniques and methodologies. An extensive assessment of 68 publications from databases such as IEEE, SCOPUS, and Web of Science (WoS) was conducted to analyze recent literature on detecting and preventing online radicalization. This research provides an overview of the definition of online radicalization and its relationship with social media. It explores different types and sources of datasets used in studying online radicalization. Additionally, it categorizes approaches and techniques, including Machine Learning (ML), Deep Learning (DL), and Graph algorithms, for detecting and preventing online radicalization. The survey identifies limitations and challenges in the field, highlighting existing gaps and suggesting potential directions for further study. To the best of the authors' knowledge, this work is the first of its kind to undertake such a holistic investigation that consolidates these methodologies presenting them in an accessible manner. The findings contribute as a valuable resource for academics, decision-makers, and professionals working in the field of counter-radicalization and provide insights into existing countermeasures against this expanding threat.

INDEX TERMS Radicalization, extremist, machine learning, deep learning, social network, survey.

I. INTRODUCTION

The digital revolution in the last decade has dramatically transformed the behaviors and way of life of billions of people throughout the world. It has become an indispensable pillar in people daily lives, whether from a commercial or personal perspective. The emergence of social network platforms like Twitter, Facebook, YouTube, and Instagram is considered one aspect of the digital revolution. Their usage has grown enormously recently. As seen in Figure 1, the number increased to 4.80 billion users in April 2022, representing a growth rate of 3.2% per year. This increase represents 59.9% of the population of the entire universe and 93% of internet users [1]. These statistics demonstrate the enormous impact of these platforms due to their substantial benefits in revolutionizing the way interaction occurs between users. This has prompted people to consider them as potent communication mediums that break down information barriers, seamlessly share political information, and open the door for interaction between users from various backgrounds. Additionally, these platforms help individuals identify information disorders and fake news. Furthermore,

The associate editor coordinating the review of this manuscript and approving it for publication was Adamu Murtala Zungeru.

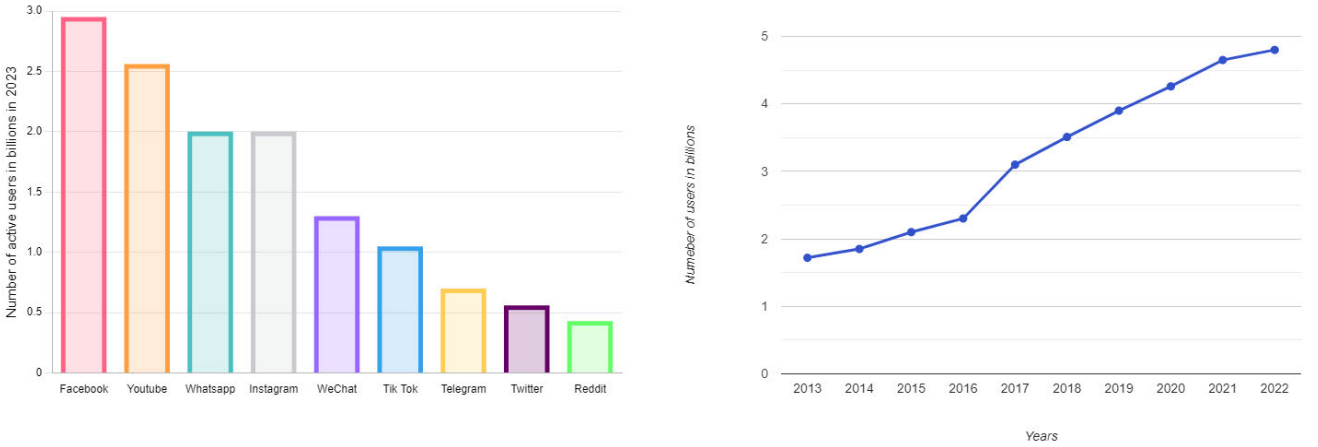


FIGURE 1. The raising of social media users.

social networking platforms have been utilized as suitable venues for many debates, including economic, political, and ideological disputes.

Despite the fact that the opportunities provided by social networks have a preponderance over traditional media, which is limited to a few consumers, they also have a dark side in some cases. Social networks have been accused of playing a role in increasing social polarization and contributing to it through their recommendation systems. This is because social networks are algorithm-based platforms that personalize content for users based on their interests, behaviors, and interactions. This personalized approach leads to the emergence of filter bubbles and echo chambers, which, in turn, widen ideological gaps and amplify extremist content. Furthermore, interacting only with like-minded individuals ultimately leads to radicalization [2]. Several previous studies have investigated the effects of social network algorithms on polarization and group clustering. Sharma et al. [3] demonstrated the effect of recommendation algorithms in increasing polarization by analyzing the Twitter graph. Daly et al. [4] found that network clustering increased as a result of friend recommendation systems. The mechanism of social network algorithms leads radicalized ideology makers and extremist communities to benefit from the characteristics of social media platforms, resulting in an increase in societal clustering and polarization. The activities of these radicalized communities on social media platforms highlight the importance of these platforms as a sharp sword for promoting extremist ideologies, gaining support, and recruiting individuals due to easy access to a large number of users worldwide, regardless of their location. Additionally, the anonymity provided by social media platforms allows these organizations to create an environment conducive to their activities, establishing echo chambers that polarize vulnerable individuals [5]. The use of radicalization organizations combined with the characteristics of social networks caused this phenomenon to have a substantial financial impact on these businesses and the governments. They are compelled to increase their efforts in a number of areas to minimize it, which calls for increased budgets for things like costs associated with content moderation, legal and compliance expenses, research and development, user education and awareness, and data storage and analysis. According to [100], the industry's cost of moderating digital material will be approximately 8.8 billion in 2024 and will rise to 13.6 billion by 2027. Consequently, they have collaborated to implement regulations, policies, and detection mechanisms to detect and prevent the rise of terrorist activities online. The following regulations outline the efforts made by these platforms to reduce the phenomenon of online radicalization.

Reference [6] presents that 278 ISIS accounts posted videos on YouTube in 2018. Despite the fact that these videos included content related to radicalization, only about 40% of these accounts were deleted by YouTube after the videos were removed by YouTube. According to [7], Facebook discovered and removed 2.5 million extremist posts in the first three months of 2018. Additionally, Twitter terminated around 1 million accounts that supported violence between 2015 and 2017. YouTube demonstrated its attempt at intervention by deleting more than 14k videos in just three months in 2018. Unfortunately, despite these best efforts and the significance of these restrictions, they have been unable to completely eradicate this phenomenon, resulting in the persistence of this process and its continued spread over time. Consequently, monitoring, detecting, and preventing the process of online radicalization becoming a hot topic for governments and researchers to study. This necessitates the development of new approaches and strategies to contain and restrict this process.

In addition to the efforts made by governments and social media platforms, cybersecurity researchers have also contributed to understanding and studying this phenomenon from various aspects, thereby aiding in the efforts of detecting and preventing it. They have utilized various approaches

TABLE 1. List of acronyms.

Artificial Intelligence	AI
Artificial Neural Networks	ANN
Bag-of-Words	BoW
Best First Search	BFS
Bidirectional Encoder Representations from Transformers	BERT
Convolutional Neural Network	CNN
DecisionTree	DT
Deep Learning	DL
Global Vectors for Word Representation	GloVe
Gradient Boosting	GB
Islamic State of Iraq and Syria	ISIS
Jay and ARTIS Transnational Terrorism Database	JATT
K_Nearest Neighbor	KNN
Latent Dirichlet Allocation	LDA
Long Short Term Memory	LSTM
Linear Regression	LE
Linguistic Inquiry and Word Count	LIWC
Machine Learning	ML
Naive Bayes	NB
Named Entity Recognition	NER
National Consortium for the Study of Terrorism and Responses to Terrorism	START
Radicalization Awareness Network	RAN
Random Forest	RF
Reinforcement Learning	RL
Right Wing	RW
Robustly Optimized BERT	RoBERTa
Shark Search Algorithm	SSA
Social Network Analysis	SNA
Stochastic Approach for Link-Structure Analysis	SALSA
Support Vector Machine	SVM
Term Frequency_Inverse Document Frequency	TF_IDF

such as artificial intelligence (AI) and social networks analysis (SNA), which have previously shown effectiveness in addressing issues in various fields, including detecting fake news [8], [9], examining public opinion changes towards COVID-19 vaccination in social networks [10], [11], [12], community detection [13], [14], business [15], [16], security, and intelligence [17], [18]. In this context, numerous studies have been conducted to minimize the impact of this process. They range between manual and automated approaches, such as graph analysis as a network-based approach to study the role of recommendation systems [19], [20], while others have used ML and DL techniques as a content-based approach to detect extremist and hate speech content [21], [22], [23], and others have utilized sentiment analysis to examine extremist emotions [24], [25]. Despite these efforts, there are still a number of important gaps in the understanding of the phenomenon and how to address it in order to alleviate and reduce its impact, and if feasible, fully prevent it [26], [27], [28].

A. PROBLEM AND GAPS

Understanding the gaps in the existing research on online radicalization is crucial for making a valuable impact in this field. The existing overviews that dealt with online radicalization address the detection and prevention as separate problems; they focused on the detection mechanism as an effective strategy to contain it and ignored the prevention mechanism, which is regarded as a complementary strategy to the detection mechanism.

The previous detection studies have focused on certain approaches while ignoring others. some surveys only analyzed Twitter and ignored other platforms, other reviews have focused solely on NLP techniques for detection without taking into account alternative techniques, while others have focused on hate speech as a frame of radicalism rather than common radicalization. The remaining review aligns more closely with this methodology but still highlights important differences. They only focus on ML, DL, and graph methods, addressing the detection mechanism but not demonstrating the prevention strategy.

There are no surveys that provide a thorough analysis of the methods and strategies used for prevention mechanisms; instead, studies that attempted to do so suffered from serious limitations because they concentrated only on certain techniques, such as educational strategies and counter-narrative approaches, while ignoring others like the technical approaches.

The literature reveals that the existing surveys that addressed the extremist dataset concentrated on its types while ignoring a crucial feature, namely how the studies mitigate the considerable influence of data imbalances. The aim of this work is to address the limitations of existing literature highlighting the gap in the field of online radicalization.

B. CONTRIBUTIONS

This research consolidates and analyzes findings from publications pertaining to extremist content within textual data to gain a comprehensive understanding of the studies focused on detecting and preventing radicalization. Additionally, it meticulously addresses various theoretical and practical aspects, bridging anticipated knowledge gaps in this domain by addressing the research questions outlined in Table 2. Indeed, this study aims to define the phenomenon, provide an overview, and delve into the common methods and strategies associated with the detection and prevention processes.

The main contributions of this survey are:

- A thorough analysis of the main concepts, context, and the impact of the phenomena of online radicalization based on a review of 68 publications.
- A thorough examination of the methodologies and techniques used to detect online radicalization.
- A detailed discussion of the popular techniques used in prevention of the online radicalization.
- A summary table listing the various datasets, their sources, categories, characteristics, and methods for addressing data imbalance.
- A thorough analysis of the results, key challenges, and recommendations for future research in the area of online radicalization.

TABLE 2. Research questions.

Research Question	Discussion
RQ1: What are the various datasets that have been utilized in the literature?	Study the usage of datasets by comparing their types, sources, characteristics, collection approaches, and the approaches used to mitigate the imbalances in data.
RQ2: What are the various methodological approaches in online radicalization detection and prevention?	Categorize the methodologies employed in the detection and prevention processes
RQ3: What are the effective techniques used for detecting online radicalization?	Analysis of various detection techniques used in each detection mechanism
RQ4: What are the existing strategies and interventions for the prevention process?	Analysis of the effective prevention techniques used in each prevention mechanism

An in-depth analysis of published papers, as well as various survey papers, on prevention and detection techniques for online radicalization in recent years was conducted to address aforementioned questions in Table 2. The aim was to provide a comprehensive understanding of this phenomenon, including its definition, detection methods, and prevention strategies. To achieve this, the datasets used in previous studies were examined, comparing their types, sources, characteristics, and public accessibility to address RQ1. Additionally, RQ2 was used to explore the different methodologies employed in the detection and prevention processes, categorizing each process into distinct approaches. Through the literature review, a detailed analysis of the detection and prevention techniques was conducted, as well as the identification of the frequently used feature extraction techniques in each approach, to answer RQ3 and RQ4.

This survey is structured as follows: In Section II, the research methodology will be outlined. Section III will provide background information on what radicalization is and how it works. In Section IV, relevant surveys conducted on this subject will be presented. The findings of the literature review on datasets, approaches, and methods for detection and prevention will be covered in Section V of this article. The application of online radicalization was presented in Section VI. The research questions and challenges encountered will be discussed in Section VII, the limitations are illustrated in Section VIII. Conclusion and tips for future directions are given in Section IX and Section X, respectively.

II. RESEARCH METHODOLOGY

This section outlines the search strategy to find publications that address the study questions established for the thorough literature evaluation.

A. SEARCH STRATEGY

In order to fulfill the research aim, studies from reputable journals and conferences were extracted as part of this search strategy using the scientific databases IEEE Explore, Web of Science, and Scopus. The adopeted search technique uses a variety of carefully chosen keywords to point to research regarding the issue of online radicalization, as well as ways for detection, prevention, and other relevant ideas. The search technique employed a variety of carefully chosen keywords to identify research related to online radicalization, detection, prevention, and other relevant ideas. Boolean operators (AND, OR) and truncation symbols (*) were used to refine the search procedure. The following keywords were used in the search query: ''Online extremist,'' ''online radicalization,'' ''radicalization detection and prevention,'' ''extremist detection and prevention,'' ''extremist prevention,'' and ''propaganda.''

Only the research works addressing the detection and prevention of online radicalization published between 2017 and 2023 were included in the selection process. Publications outside of this time period were excluded. The titles and abstracts of gathered study publications were reviewed to identify relevant studies. Initially, titles and abstracts were used to filter the search results. Subsequently, the study questions and findings of each full-text paper were carefully examined to determine its eligibility for inclusion in the survey paper. These procedures ensured that the selected research addressed the research objective, focusing on the detection and prevention of online radicalization. A total of 68 publications were obtained through this technique, which examined multilingualism datasets in languages such as English, Arabic, and Kazakh. Figure 2 shows the distribution of publications collected by year of publication.

TABLE 3. Filtering criteria.

Identifier	Description
FC1	Is the study objective sufficiently stated and aligned with the goal of this research?
FC2	Do the papers discuss various methods utilized in the process of detection and prevention?
FC3	Do the articles present the outcomes of the techniques employed?
FC4	Are the datasets and the methods of collection described in the articles?

B. FILTERING CRITERIA

The filtering criteria in Table 3 were used to select the most relevant and useful papers for the focus of this survey. After the filtering activity, 68 articles were used in the survey.

III. BACKGROUND

This section will provide a background of online radicalization. The definition, circumstances, and working mechanism

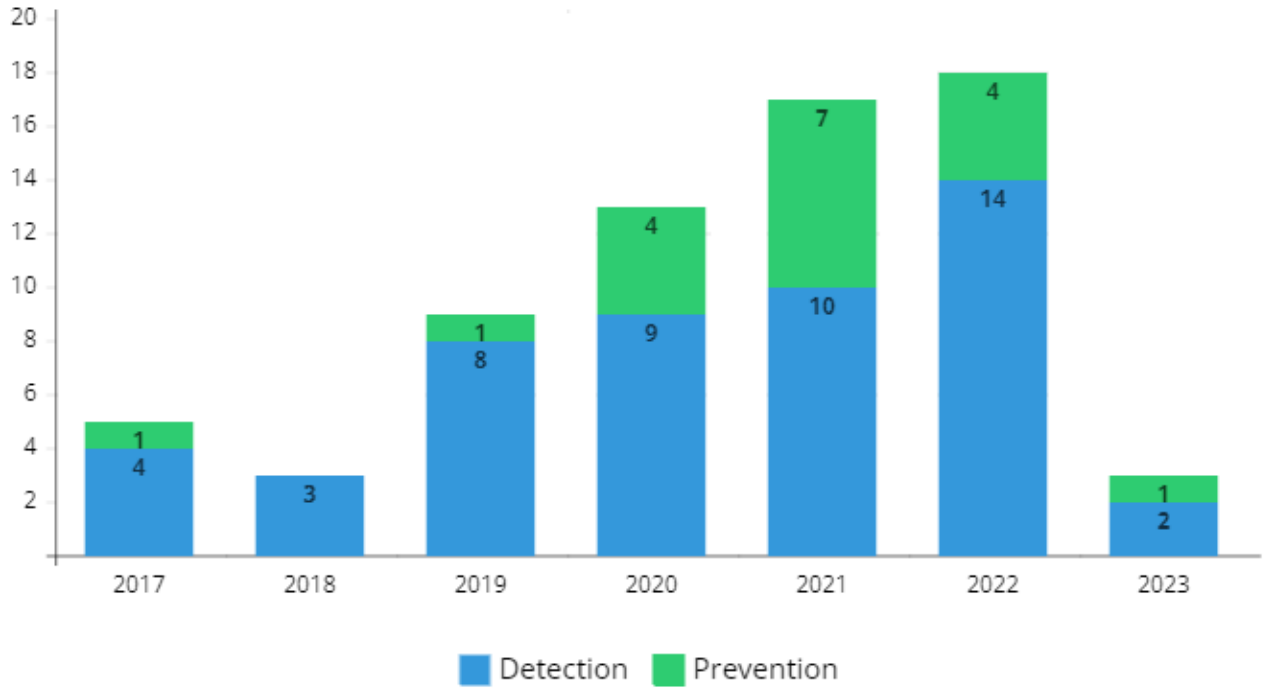


FIGURE 2. Published articles per year.

of this phenomenon will be demonstrated in the upcoming sections.

A. ONLINE RADICALIZATION AND ITS CIRCUMSTANCES

There are several contested meanings of radicalization terms that vary depending on context. Some agencies relate it with extremist beliefs [29], [30], while others define it as a terrorist term that refers to violent acts [31], [32], while others went even further and associate it with a particular religion [33]. In the context of this research work, radicalization underlies the process that drives individual behaviors and beliefs toward becoming more extreme toward a certain ideology, as well as the potential to translate this behavior into violent acts to achieve a certain purpose [34].

The radicalization phenomenon has various manifestations and methods; it can exist in a variety of daily life areas such as political, economic, and social aspects involving race, nationalism, and religion. It can occur both offline and online. Offline participation is the conventional way for individuals to participate in radicalized organizations, which represents face-to-face interaction in groups. This strategy is constrained by the geographical location of the groups and vulnerable people, which is considered a strong point for governments and policymakers in diminishing and combating the severity of this phenomenon by isolating it in a certain geographical region. In contrast to offline, online participation occurs through the use of communication technology such as social network platforms, which may cause people to get polarized and join like-minded groups as well as participate remotely without needing to contact them in person. This strategy is far more hazardous than the offline one due to its capacity to easily spread vast amounts of extremist ideology and significantly increases the ability of radicalized organizations to intensify their efforts in order to further their goals [35].

The significant influence of the online radicalization strategy led to the investigation of the causes that feed into this phenomenon. Its emergence was not a coincidence; rather, it is seen to be the consequence of a confluence of several factors, including personal circumstances, economic issues,political, and religious grievances [36].

In addition to the circumstances of the online radicalization phenomenon, radicalized communities have benefited from the characteristics of social networks to progressively change individuals' beliefs in order to attract them to join their organizations through a series of well-planned procedures. Therefore, there is no defined or consistent procedure for the online radicalization mechanism, although in most cases it goes through an entry phase as illustrated in Figure 4. These steps are as follows:

1) Pre-radicalization: This represents the initial stage, as individuals in this stage are exposed to a huge amount of extremist content that is published by terrorist groups through online platforms such as social media that can be compatible with their personal beliefs and behaviors. This leads to opening the door easily for individuals who want to get acquainted with new ideologies or deepen their knowledge, and in some cases, they can agree and accept them due to the individual being repeatedly exposed to this content.

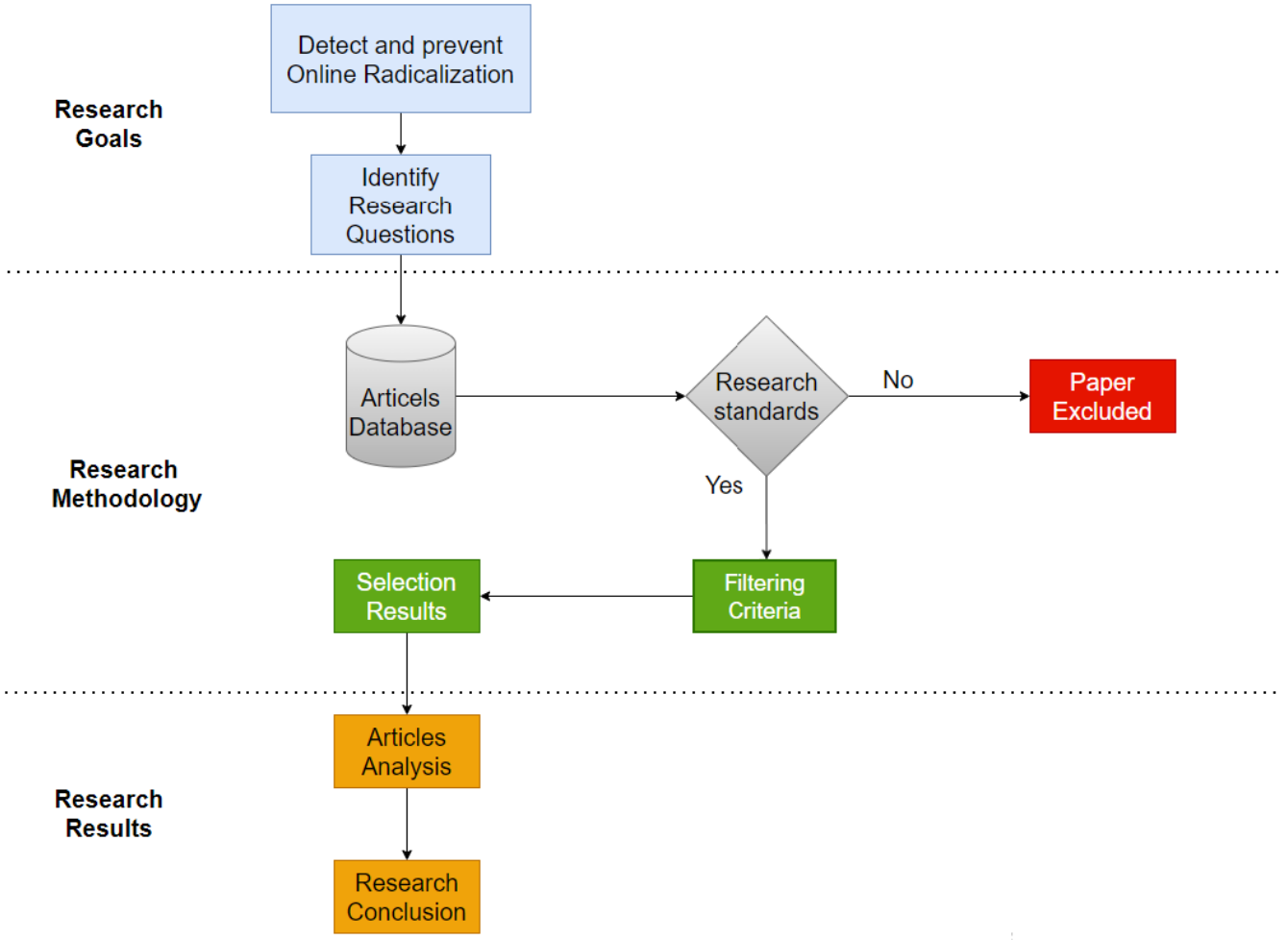


FIGURE 3. Review process.

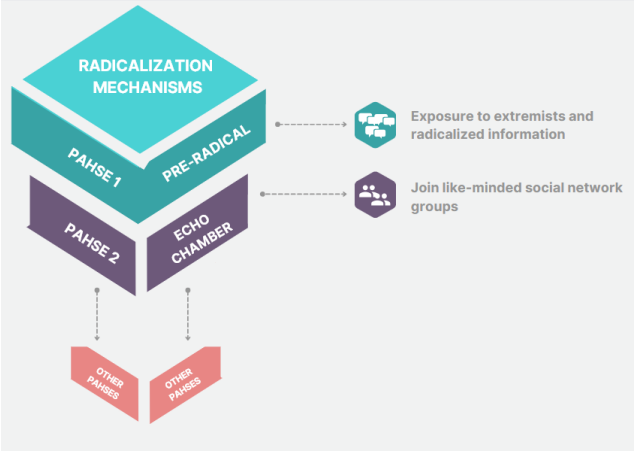


FIGURE 4. Radicalization phases.

2) Echo chamber zones: Personal curiosity and the recommendation algorithms of social networks play a key role in this stage of the process. As individuals use internet platforms to learn more about some of the beliefs they were exposed to during the pre-radicalization stage, recommendation systems occasionally might lead some individuals to be polarized into groups that share their beliefs and operate as echo chambers, making them more vulnerable to extremist organizations and driving them farther down the extreme path.

Based on the stages above, the first phase is considered the entry point in the radicalization process of changing individual beliefs that may push vulnerable individuals to access and interact with this terrorist content and join them later. Therefore, due to its importance and its significant impact on the radicalization pathway, this research work highlights the efforts of these organizations to use social networks such as Twitter, Facebook, and YouTube to spread their ideological content to as many users as possible [37].

The Islamic State of Iraq and Syria (ISIS) is one of the radicalized organizations that spread extremist content on

social media. The authors in [38] demonstrate that ISIS suggested a YouTube video ''one billion campaign'' to enlist support for their cause. Reference [6] shows that ISIS created a hashtag on Twitter in 2014 to show support for ISIS; this hashtag received 20k mentions in a single day. Reference [39] presents that during the ISIS invasion of Iraq in 2014, more than 5k people downloaded an ISIS-related Twitter application, and more than 40k tweets about ISIS accounts were sent out on the same day. Over the course of two months in 2018, 207 ISIS accounts posted 1,348 videos, garnering over 163,391 views.

The right-wing (RW) is another political movement that incites hatred towards a certain group of society. The report in [40] indicates that more than 100k Twitter users are regarded as active alt-right users in 2018. According to [41], the RW just needed a few months to garner more than 90k likes on posts critical of Islam (such as those on street rallies). In 2018, [42] hypothesizes that there are 169,071 members in Facebook groups who post racist content, and there were 5k tweets about violence against refugees in November 2015 and 6k tweets encouraging refugees to commit acts of violence.

IV. RELATED WORK

In this section, the research relevant to the present work was examined. The focus was on surveys that covered the detection and prevention techniques of online radicalization. It was found that surveys typically addressed the detection and prevention separately, necessitating the division of the work into two parts. The first subsection presents surveys related to the detection techniques, while the subsequent subsection focuses on surveys that specifically explore studies involving the prevention process.

Detecting surveys: Gaikwad et al. [43] conducted a survey following the systematic review approach. They examined the conclusions of 64 studies between 2015 and 2020 concerning the datasets used and techniques for detecting online radicalization. The authors divided the detection approaches into manual and automated categories and go deeper into automatic detection methods such as ML and DL. In addition to detecting techniques, the authors evaluated the datasets utilized and categorized them as public or Private. According to this work, the majority of studies used a Private dataset due to a lack of available datasets.

Adek et al. [44] presented a summary of studies that performed a detailed analysis of the applications utilized in detecting online radicalization. The author examined 36 studies that investigated the approaches employed in online detection applications, such as data mining and ML approaches, features, datasets, and their sources.

Torregrosa et al. [45] investigate the findings of three studies related to online extremism. They presented an overview of the extremes definition, its elements, and the approach used to detect its content. This study focused on papers that employed natural language processing (NLP) techniques and feature extraction approaches to detect extremist content, in addition to the analysis of the dataset used. The authors divided NLP techniques into two categories and compared their performance: classification techniques and descriptive techniques. This work shows that more than half of the software tools are used for extremism classification purposes, and the remaining tools are utilized for descriptive techniques such as sentiment analysis.

The survey in Trabelsi et al. [46], focused on extremist content on the Twitter platform by evaluating 68 studies. They provided a detailed overview of the prior article's results on content analysis approaches for detecting radical processes on Twitter as a source for disseminating extremist ideas. This work goes thoroughly into the utilized sentiment classification strategies, in addition to offering an outline of the sentiment methodologies used for feature extraction and the ML classification algorithm. The researchers determined that Twitter is the most common medium for academics to uncover radicals and ML approaches are extensively used in categorization methods.

Hate speech as one of the most extreme aspects was investigated by Chhabra et al. [47]. The authors provided a detailed analysis of the strategies used to contain this problem. They presented the phases of hate speech identification, such as feature extraction using text mining and classification approaches using ML and DL, as well as analyzing a different hate speech dataset used in the literature studies. According to the authors, the majority of researchers utilized Facebook and Twitter data as the primary platforms for their investigation.

Alghamdi et al. investigated 40 studies that analyze online extremism on the dark web in their survey [48]. The authors examined the existing extremist dataset and the machine learning techniques used to detect extremist content on the dark web. They found that content analysis is a suitable approach for the web detection method, despite the lack of research focusing on the Arabic language.

Prevention surveys: AKRAM et al. [49], in their survey, discussed prevention strategies for extremist content and its spread based on an educational approach. They conducted an in-depth analysis of over 80 articles that employed mitigation strategies. The authors focused on the psychological solutions presented in prior articles that lead people to hold de-radicalized views, such as spreading literacy and non-violent content.

Windisch et al. [50] studied the stages of frameworks that used the counter-narratives approach for hate speech prevention. In this survey, the authors conducted a systematic review of 22 counter-narrative strategies and performed statistical analyses to determine the effectiveness of mitigation strategies and interventions used to reduce hate speech.

The survey conducted by Iannello et al. [51] in 2023 focused on eight studies that employed educational software as mitigation strategies to moderate online radicalization content. They categorized the investigations into two groups. Four of the studies were evaluated using quantitative analysis, while the remaining studies used qualitative analysis. The authors emphasize that there is no

TABLE 4. Related works.

Article Title	Study,Year	Article Focus	Article Type
A literature survey on multimodal and multilingual automatic hate speech identification	[47], 2023	Present automatic classification methods and datasets used in hate speech detection:	Detection
Radicalization in Correctional Systems: A Scoping Review of the Literature Evaluating the Effectiveness of Preventing and Countering Interventions	[51], 2023	quantitative analysis and qualitative analysis to evaluates the effectiveness of the preventing programs	Prevention
Systematic Review of Radicalization through Social Media	[49], 2023	Focused on psychological solutions to hold de-radicalized views	Prevention
A survey of extremism online content analysis and prediction techniques in twitter based on sentiment analysis	[46], 2022	Focus o content analysis approaches to detect online extremists on Twitter	Detection
A survey on extremism analysis using natural language processing: definitions, literature review, trends, and challenges	[45], 2022	Present the frequent NLP approaches and datasets sources in online radicalization detection	Detection
Online interventions for reducing hate speech and cyberhate: A systematic review	[50],2022	Evaluate the effects of existing counter-narratives strategies to reduce online hate speech	Prevention
Systematic Review on the Outcomes of Primary and Secondary Prevention Programs in the Field of Violent Radicalization	[53], 2022	Focused on studies that evaluate the prevention programs in the field of violent radicalization	Prevention
Techniques to detect terrorists/ extremists on the dark web: a review	[48], 2021	present the detection techniques, feature selection, and dataset to detect radicalization in the dark web	Detection
Systematics Review on the Application of Social Media Analytics for Detecting Radical and Extremist Group	[44],2021	Describe detection techniques, data sources, features, machine learning methods	Detection
Online Extremism Detection: A Systematic Literature Review With Emphasis on Datasets, Classification Techniques, Validation Methods, and Tools	[43],2021	Focus on the datasets, manual and automatic classification techniques, validation methods of radiclaziotn detection	Detection
Cyberhate: A review and content analysis of intervention strategies	[52],2019	Present interventions approaches based on law, technology, and education approches	Prevention

evidence to determine the most effective mitigation technique and demonstrate that various approaches had an impact on individual-level mitigation rather than group-level mitigation.

In [52] 2019, Blaya et al. divided the mitigation process into three approaches: legal, technological, and educational. In the legal approach, the authors present laws and regulations to control hateful content, while in the technological approach, they analyze seven studies that used classification techniques for mitigation. Finally, the educational level is considered a soft approach. The author only considers classification methodologies and analyzes eight research for technical prescriptive.

More than 30 studies evaluating prevention programs based on different approaches, such as educational, technical monitoring, and psychological approaches, were discussed in [53]. The authors analyzed the outcomes of these studies and reported that the studies were focused on specific religious groups, as most research exclusively considers Islamic radicalization.

As observed in Table 4 which summarizes the literature studies mentioned above, these studies have made significant efforts and addressed various aspects of online radicalization.

V. LITERATURE REVIEW

In this section, recent literary analyses on the detection and prevention of the radicalization phenomenon are discussed. The state-of-the-art studies indicate that most researchers utilized a similar framework, as depicted in Figure 5. This framework involves dataset collection from various websites, online magazines, and social media platforms. Subsequently, the data undergoes a pre-processing phase that includes feature extraction techniques to prepare it for modeling. Finally, researchers apply detection or prevention models. The techniques employed in each phase of this framework will be discussed in the subsequent sections below.

A. DATASET AND ITS SOURCES

The dataset is considered one of the most important parts of the detection and prevention framework. The datasets used in existing studies are collected from various sources, ranging from social media platforms to websites and magazines. These datasets can be categorized into two types: public datasets and Private datasets. The public dataset refers to data that is publicly accessible. On the other hand, Private datasets are collected and designed by researchers based on their specific goals, and they are not publicly accessible. Figure 6 illustrates the sources of datasets commonly used by researchers. For both types of datasets, and in order to collect tweets, posts, videos, and comments, most researchers have used various words related to radical groups such as #Kofar, #DASH, #Terrorist, #white supremacy, #Jihad, #extremism, # Bomb, #far-right etc. They have labeled them into various categories, such as radicalized, non-radicalized, terrorist, and non-terrorist, or classified them into different radicalization categories, such as propaganda, recruitment, and ideology.

Social Media Datasets: For the Private and public datasets, different social media platforms such as Twitter, Facebook, Telegram, and WhatsApp were used as sources for collecting extremist data. Twitter is used widely in collecting extremist datasets as shown in Figure. 6, and it is used by various

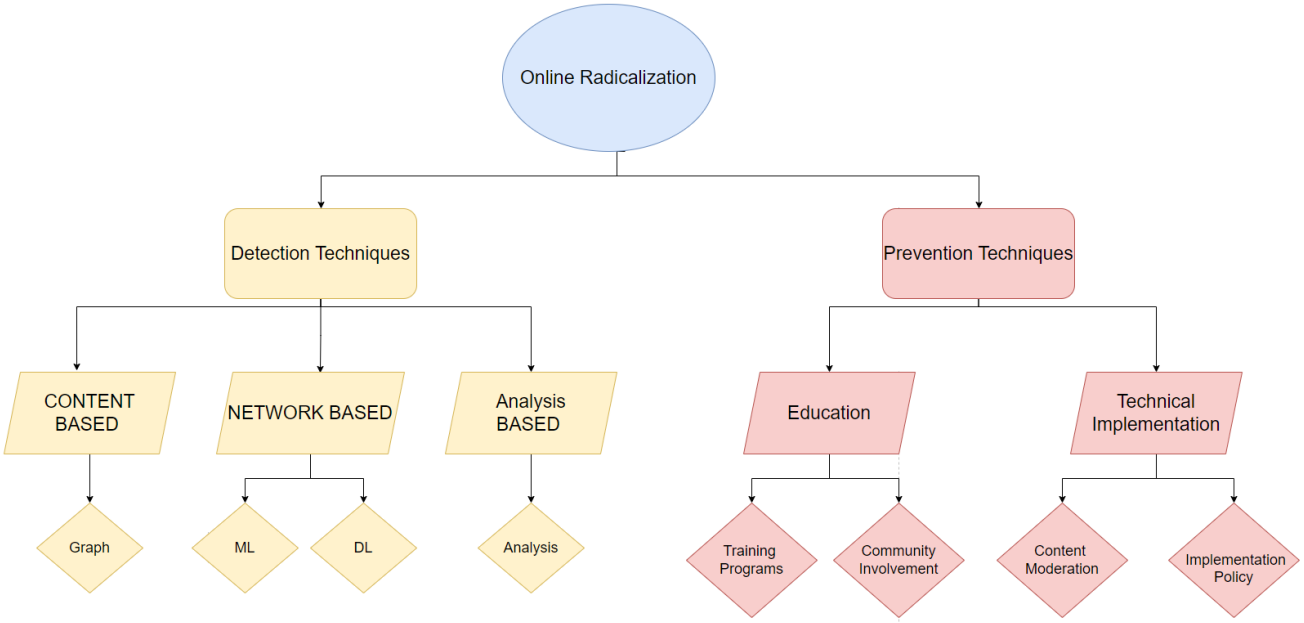


FIGURE 5. Detection and prevention framework.

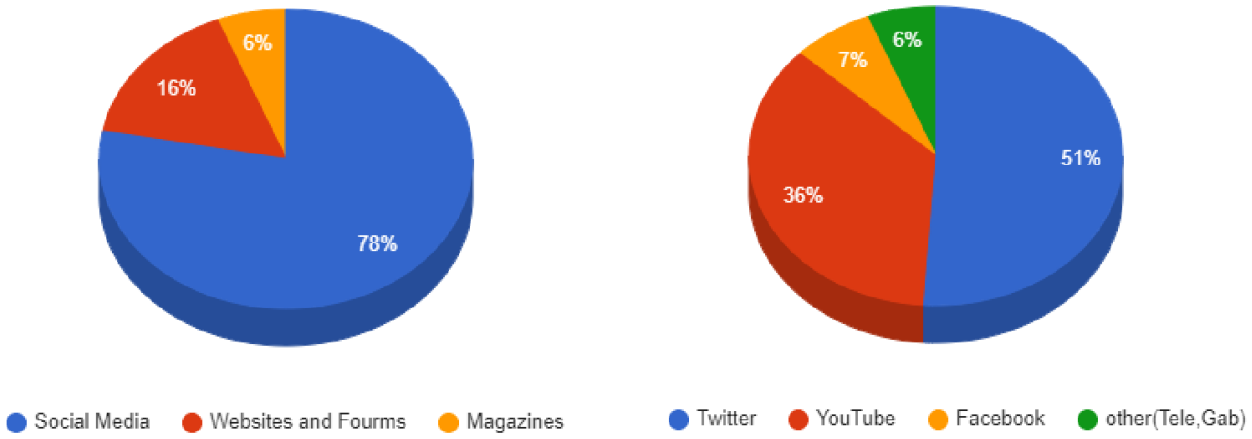


FIGURE 6. Dataset sources and most social media platforms used in previous articles.

studies for detection and prevention, [21], [54], [55] collected English tweets. Some researchers collected tweets that contain multilingualism. [22], [23], [56], [57] collected Arabic tweets and make them publicly accessible, [58] collect multilingualism tweets including using Urdu, and Roman in addition to the English language.

Aldera et al. [22] collected more than 89k Arabic tweets using extremist hashtags and manually labeled them as extremist or non-extremist. The dataset is available upon request. Mursi et al. [23] collected around 100k tweets and labeled them as hateful or non-hateful based on sentiment analysis results. Alharbi et al. [57] created ASAD, a dataset containing over 95k Arabic tweets from various countries, which were classified as favorable, negative, or neutral using sentiment annotation. Similarly, Alomari et al. [56] introduced the Arabic Jordanian General Tweets (AJGT) dataset, consisting of more than 1700 tweets collected using extremist Arab hashtags. They used sentiment analysis to categorize the tweets as positive or negative. Sharif et al. [59] compiled a dataset with over 7k tweets representing the Taliban movement in the Afghanistan region. The tweets were labeled as neutral or extremist, with the extremist category further divided into pro-Afghan and pro-Taliban organizations. The dataset was publicly published.

Reddit,as a social network platform, hosts various communities from different backgrounds. Theisen et al. [60] used keywords like #echochambers, and #radicalization to collect thousands of user comments from Reddit threads belonging to extremist communities. Reference [61] Proposed a SPINOS dataset to detect opinion shifts, they collected

threads from both extremist and moderate communities from the Reddit platform using keywords like #gun control, #abortion, #veganism, and #politics in order to detect opinion shifts. Kennedy et al. [62] proposed a dataset of hate speech from the Gab platform, collected over 27k posts. Three experts were involved in annotating the posts based on hate rhetoric, and the dataset, known as Gab Hate Corpus (GHC), is publicly available.

Facebook was used by Asif et al. [58] for data collection. They used words such as #Bomb, #terrorist attack, #fight, #ISIS and their translation to collect posts and comments that contained different languages such as Arabic, English, and Urdu. The authors performed sentiment analysis on these posts and comments to label them into four categories: one as neutral and the remaining three representing levels of extremism. Reference [63] used the Kazakh language and collected more than 15k posts from the Vkontakte platform, labeling them as extremist and non-extremist posts.

WhatsApp and Telegram messaging platforms were also used in different studies. WhatsApp, widely popular and used by individuals, was employed by [64] to detect and analyze extremist messages. On the other hand, Telegram, preferred by groups and organizations for spreading extremist content, was utilized by [65] and [66]. In [65], Schulze et al. collected messages related to far-right organizations to perform a longitudinal analysis.

A few researchers turned to video streaming platforms to collect metadata on extremist videos and study the influence of video streaming platforms and their recommendation systems. YouTube, being the most commonly used video streaming platform by researchers, was used by Albadi et al. in [28] to collect metadata for approximately 350k Arabic videos belonging to various religions. The videos were labeled as hateful or non-hateful. Ribeiro et al. [67] collected a dataset containing metadata for 300k videos from 349 channels and their recommendations, divided into two classes: neutral and radical. The radical class consisted of three categories: alt-right, alt-lite, and intellectual dark web. This dataset served as a primary dataset for Ravid et al. in [26] and as a secondary dataset for Fabbri et al. [68]. Sock puppets were utilized by Haroon et al. [69] to imitate social media accounts which are virtual accounts that can be used for misleading activities like watching videos. They collected 100k YouTube sock puppets divided into five classes, one of which was neutral, and the other four had distinct ideologies. The puppets were trained by watching videos from different categories to analyze how YouTube recommends videos to viewers. Faddoul et al. [27] collected 1146 seed data channels related to conspiracy theories and generated their up-next video recommendations. Kathuria et al. [70] used YouTube, Vimeo, and 4chan platforms to collect comments, as well as the number of likes and views for videos from left-wing channels like Antifa and right-wing groups like Proud Boys.

Some studies utilized public datasets in their work. The Kaggle website is considered one of the most reliable repositories for public datasets [71], [72], [73], [74], [75], [76]. In [77] and [78], the authors used the public datasets [71] and [72] that contain ISIS radical tweets for detection on Twitter. In [79], the author employed the dataset [73], while [27] used the YouTube conspiracy dataset [80] to study the impact of recommendation systems on the spread of conspiratorial material. Kursuncu et al. [81] used the Lucky Troll Club public dataset [82], which contains tweets posted by user accounts related to the ISIS organization.

Blog and Forums: In addition to social media platform datasets, existing studies on detecting and preventing online radicalization have collected data from websites, magazines, and blogs belonging to extremist organizations. Stormfrom is a website related to white racist organizations that spread extremist and hateful content towards non-white nationalists. Various studies have used this website to collect data for their studies. De Gibert et al. [83] collected data consisting of sentences from Stormfrom posts related to white supremacists and labeled them as hate or non-hate, making them publicly accessible.

The Ansar dataset is a public dataset that contains over 28k terrorist posts related to Western jihadist movements. It includes Arabic and English posts from the AlJihad Network, available only in English and Arabic [84]. Theodosiadou et al. in [85] used it to analyze the English posts of jihadist movements. Petrovskiy et al. in [86] used the KavkazChat dataset [87], which contains over 600k posts related to Islamic jihadists in the Caucasus, in multiple languages (Arabic, English, Russian, etc.). The posts in this dataset are divided into more than 16k topics and were collected from different jihadist forums by the AI lab team at the University of Arizona, making them publicly available.

Dabiq and Rumiyah are two online magazines published in different languages, including Arabic, English, and French. They are officially used by ISIS terrorist organizations to publish and spread ISIS content to receive endorsement, support, and recruitment from those wishing to join. These magazines were used by some researchers. Araque et al. in [54] used both magazines to collect more than 500 articles. Nouh et al. in [77] used Dabiq to collect English articles related to propaganda. [70] and [88] used both magazines to perform classifications of articles.

Some websites publish publicly accessible databases and datasets. The National Consortium for the Study of Terrorism and Responses to Terrorism (START) publishes a dataset named PIRUS, which contains profiles covering personal and demographic information of individuals in the United States related to four terrorist organizations [89]. Al-Zewairi et al. in [90] used [89] in their work to perform multi-classification for extremists in the United States. Tundis et al. in [91]] used Jay and ARTIS Transnational Terrorism Database (JATT) [92] as the primary dataset to detect the criminal profiles of users by extracting demographic and behavioral features from Facebook profiles. Fabbri et al. in [68] used the NELA-GT dataset [93] as a

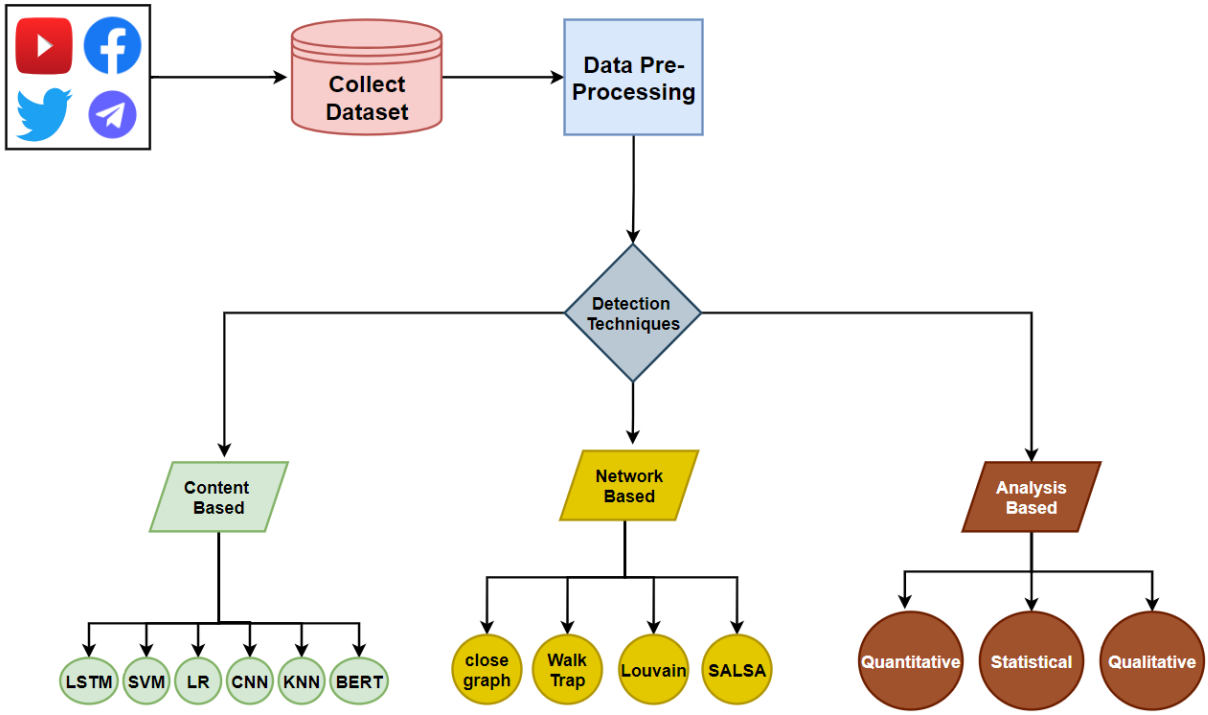


FIGURE 7. Radicalization detection framework.

secondary dataset in their work. This dataset was collected by Jeppe et al. to detect misinformation in articles. They collected more than 700k articles from over 190 new media sources.

B. DETECTION AND PREVENTION METHODS

This section presents a detailed analysis of the most commonly used detection and prevention approaches in existing studies for online radicalization. The discussion is developed into two sub-sections as follows:

1) DETECTION METHODS

Radicalization detection refers to the concept of detecting people who show indicators of disseminating radical ideas and joining organizations that promote extreme content. Using cutting-edge technologies like ML, DL, and SNA techniques, this procedure comprises a thorough study of digital sources such as social network content and online activity. The detection process may fall into three categories, as shown in Figure 7: network-based, content-based, and psychological analysis.

a: NETWORK BASED APPROACH

The network or graph approach technique depicts the mathematical structure of complex networks. It is used to explore and model the relationships between items in order to comprehend and understand how they are connected and mapped. It is based on graph theory and consists of a set of vertices and edges, where each vertex represents a single entity and each edge shows the connections among those entities. The weight of the edges determines the strength of the connection. There are two types of graphs: weighted graphs, in which the links are assigned varying strengths, and unweighted graphs, in which the edges are represented as binary representations. Due to its effective outcome in dealing with complicated networks, this method has been used in several disciplines, including economics, medicine, and intelligence security [94].

In the cyber intelligence field, this technique has been employed by several researchers and scientists for a variety of tasks, including community detection in social networks, the detection of malicious information, and the detection of radicalized user patterns. The researchers focused on the community detection approach, which was widely employed to better understand and identify the phenomena of the online radicalization process in social networks. Using hierarchical and partitioning techniques, which are regarded as the two most often used strategies in community discovery methodologies, it was possible to group similar items together into a single cluster based on the traits and properties of each object. The most frequently used graph analysis techniques for radicalization detection in existing research are Louvain, infomap, Walktrap, Fast Greedy, label propagation, Edge Betweenness, Multi-step Greedy, and Kernighan-Lin algorithms. In addition to analyzing techniques like Jaccard and modularity [95].

In addition to the community detection approach, some researchers have used graph-based techniques to study the

TABLE 5. Radiclazied dataset.

Study	Type	Platfrom	Language	Size	Labels
[21]	Private	Twitter	English	33,000 Tweets	Radical, Non_radical
[22]	Private	Twitter	Arabic	89,000 Tweets	Extremists, Non_extremists
[23]	Public	Twitter	Arabic	100,000 Tweets	Hatful, Non-hateful
[27]	Private	YouTube	English	1146 Videos	Conspiracy, Non_conspiracy
[28]	Private	YouTube	English, Arabic	350,000 Videos	Hatful, Non-hateful
[54]	Private	Twitter	English	200,000 Tweets	Radical, Non_radical
[56]	Public	Twitter	Arabic	1700 Tweets	Positive, Negative
[57]	Public	Twitter	Arabic	95000 Tweets	Positive, Negative, Neutral
[58]	Private	Facebook	English, Urdu	20,000 Posts	Low, High, Moderate
[59]	Private	Twitter	English	7500 Tweets	Pro-Afghanistan, Pro-Taliban, Neutral
[60]	Private	Reddit	English	50,000 Posts	Radical, Non_radical
[61]	Private	Reddit	English	3500 Posts	Strongly_against, Neutral,Weakly_against, Weakly_favor, Strongly_favor
[62]	Public	Gab	English	27,000 Posts	HT,VO, CV
[63]	Private	Vkontakte	Kazakh	7000 Posts	Extremists, Non_extremists
[67]	Private	YouTube	English	300,000 videos	Radical, Non_radical
[69]	Private	YouTube	English	100,000 Videos	Neutral, Right, Left.
[70]	Private	YouTube,Vimeo,4chan	English	31,000 Videos	Right_leaning, Left-leaning
[71]	Public	Twitter	English	17,000 Tweets	NA
[72]	Public	Twitter	English	122,000 Tweets	NA
[73]	Public	Magazines	English	2600 Posts	NA
[74]	Public	Twitter	English	2000 Tweets	Explicit, Implicit, Neutral, white supremacism
[75]	Public	Twitter	English	25,000 Tweets	Hate_speech, Neither, Offensive_language
[76]	Private	Twitter	English	40,000 Tweets	Radical, Non_radical
[80]	Private	YouTube	English	8 Million videos	Conspiracy, Non_conspiracy
[82]	Public	Twitter	English	1.9 Million Tweets	NA
[83]	Private	Forums	English	10,568 Sentences	Hate, Non-hate
[84]	Public	Forums	English, Arabic	28,000 Posts	Rerrorist, Non_terrorist
[87]	Public	Forums	English, Arabic,Russian	600,000 Posts	NA
[89]	Public	Forums	English	1,473 Profiles	NA
[92]	Public	Facebook	English	2157 Profiles	NA
[93]	Public	Forums	English	713,000 Posts	NA

recommendation system and its impact on the phenomenon of radicalization. They have done this by observing and studying how users interacted with radicalized items like watching videos and how recommendations produced high-interest items that push users further into the radicalization pathway. Graph-based techniques have also been employed by other researchers as a feature extraction technique to extract the semantics and similarities from the dataset to feed another detection model.

Studies Using Network Approach: Dhiraj in [96] utilized the Gephi software [97] to study the role and impact of the YouTube recommendation system in leading people down the path of radicalization. The authors collected radicalized videos belonging to the ISIS organization as seed data. Using this seed data, the authors constructed three levels of recommendation (recommendations of seed videos, recommendations of first-level videos, and recommendations of second-level videos) via the Gephi software. This resulted in a graph with over 15k nodes representing videos and about 180k edges representing the links between the videos, allowing them to study the relationship between extremist videos and their recommendations. Additionally, the authors performed qualitative analysis to study the video features that are used as indicators to influence the recommendation systems.

Also in the same context as analyzing the impact of recommendation systems on polarization. Cinus et al. [98] provided a Monte Carlo framework to analyze the influence of recommenders on the evolution of user opinion. The suggested framework integrated two graph inputs: the first is a social opinion graph generated using a Random walk, and the second is a link recommender system using the Stochastic Approach for Link-Structure Analysis (SALSA) algorithm in order to construct a graph expressing the evolution of opinion. The recommendation influence was evaluated using Neighbor Correlation Index (NCI) and Random Walk Controversy score (RWC) metrics.

Regarding community detection, Agarwal et al. [19] employed a graph-based technique to detect radicalized communities in the Indian area that include malicious content and to identify influential people. In this work, the authors used crawling techniques such as best-first search (BFS) and shark search (SSA) to perform community detection by constructing a network graph using YouTube video data. They classified the users into positive, representing videos containing malicious content, and negative, representing moderate content.

Papadamou et al. [99] focused on the Incel community on YouTube by investigating the development of these groups over time and the role of a recommendation system in directing viewers toward Incel material. They collected over 12k videos regarding Incel and other videos, along with roughly 10 recommendations collected for each one. Using a Lexion dictionary created by specialists that contained 200 terms associated with the Incel group, they annotated if each video was an Incel or not. The authors divided their

work into two phases: first, they studied the evaluation of Incel communities by examining the number of Incel videos published per month and the comments on those videos, and how they increased over time. Second, they studied the role of the recommendation system by constructing recommendation graphs for Incel and random videos, and showing how that recommendation algorithm can distinguish an Incel video from a random video after five tiers using random walk techniques.

Some researchers have used the graph approach as a feature extraction tool to feed classification models. Petrovskiy et al. [86] suggested a methodology for detecting radical harmful people on social networks using their characteristics. They worked with two datasets: the first one collected from Kaggle related to the ISIS organization [71], [73], and the second one is the KavkazChat dataset [87]. The authors employed graph techniques to categorize features as dangerous, safe, or unknown. In this work, the authors constructed user graphs based on user interactions, which were then analyzed using various metrics techniques such as Betweenness Centrality and In-Degree Centrality to select and determine the features to feed ML classifiers. Saif et al. [100] used the close graph approaches to study the sentiment of features and the users' interactions, such as the number of likes, follows, and retweets. The authors in this research worked with datasets that contain classified tweets as pro-ISIS and anti-ISIS. They used the Close Graph algorithms to extract features as subgraphs to be used as features to feed ML algorithms.

Arruda et al. [20] focused on detecting the polarization of opinions in the United States. They worked with a dataset containing political tweets collected using hashtags like #Obamacare and #guns. They implemented a graph network technique to model the relationship between users and postings and demonstrated the influence of social posts on opinion polarization. The authors analyzed two phases: post transmission when the user posts, and post-distribution when people interact with the user's post through actions such as retweeting, following, and liking.

To analyze online hate speech, Nguyen et al. [101] combined graph theory and sentiment analysis for over 1,900k threads with more than 30 comments each. The authors used sentiment analysis to categorize each user's polarity as negative, neutral, or positive based on the average score of their comments. This score was then used as a node attribute in the graph to study how each user's polarity spread throughout the network.

b: CONTENT BASED APPROACH

The term ''content-based approach'' refers to the study and analysis of the content or attributes of texts, such as magazine articles, social media posts, comments, and messages. This approach aims to derive insightful information that aids in comprehending the sentiments and opinions expressed in the posts on particular subjects [102]. It has been used in various fields due to its success in dealing with complex networks. For example, in marketing, it is used to study consumer opinions towards products, while in social networks, it is employed to categorize posts and blogs. In the field of cyber intelligence security, researchers have utilized this approach in various aspects, including the detection of online extremist behavior patterns and the classification of radicalization and terrorist users. It allows academics to effectively identify extremists by studying the level of violence and extremist materials in posted comments and messages.

Recently, researchers and academics have employed a variety of techniques to apply the content-based approach. Some researchers have focused on studying the opinions and attitudes expressed in posts that support extremist content, using techniques such as Named Entity Recognition (NER), Sentiment Analysis, and Topic Modeling. Others have utilized machine learning (ML) and deep learning (DL) algorithms, such as Naive Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), Artificial Neural Networks (ANN), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), K-Nearest Neighbor (KNN), Gradient Boosting (GB), Logistic Regression (LR), Bidirectional Encoder Representations from Transformers (BERT), Decision Tree (DT), and Robustly Optimized BERT (RoBERTa) to classify text and posts as extremist or non-extremist. Furthermore, some research has gone deeper by categorizing extremist posts into several groups, which can aid in recognizing different types of extremists [103].

Feature Extraction Techniques: Features considered as inputs to any detection system, as well as how to deal with these elements, can significantly influence the overall results of the system. Therefore, the process of extracting these features plays a crucial role in systems that are based on content classification. This process involves transforming raw data into understandable patterns and appropriate representations that have an impact on system performance. It has been employed in various aspects, including extracting extremist and hateful characteristics from posts, comments, video materials, titles, and more. The following techniques are the most used in the literature.

Word embedding: Word embedding is an NLP technique that transforms text into a vector representation. It is frequently used for analyzing extremist texts in the literature. This technique computes the relationships and similarities between words by assigning the same representation in low dimensions to words with similar meanings.

Term Frequency-Inverse Document Frequency (TF_IDF): TF-IDF is another NLP technique considered a statistical method for determining the importance of a word in a text. It involves two steps: first, determining the word's frequency in the text, and second, determining the inverse of the number of documents in which the word appeared. A higher TF-IDF score indicates greater relevance of the word.

Bag of Words (BOW): BoW is a technique used to extract features from documents, sentences, websites, and more. It converts text into a word vector representation based on the frequency of words in the corpus of text.

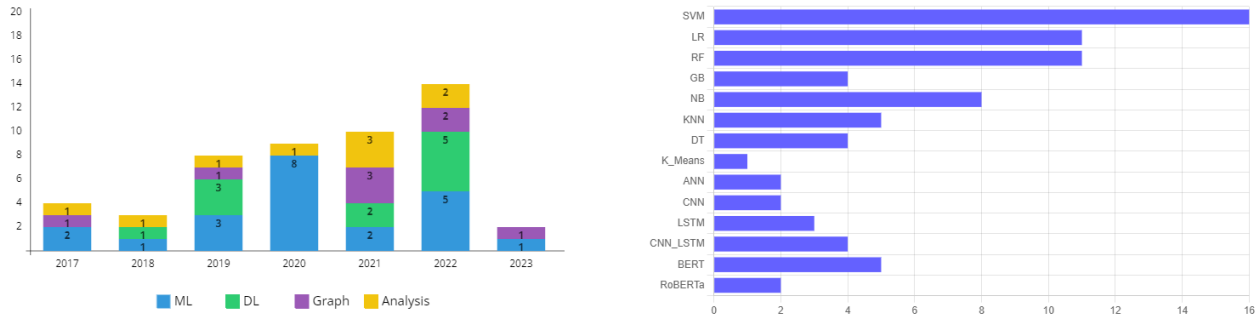


FIGURE 8. Detection techniques used by articles.

Topic Modeling: Topic modeling involves extracting and identifying topics from documents and texts by analyzing word relationships and extracting important information.

N-grams: N-grams are a method for analyzing and extracting patterns from sequential data, such as text. The "N" refers to the number of items in the text used to extract meaning units. N-grams can be 1-gram (uni-gram), 2-gram (bi-gram), 3-gram (tri-gram), or more.

Lexicon Based: Lexicon-based methods involve identifying aspects of text using predefined dictionaries. These dictionaries or lexicons are manually created by specialists and are available from different sources, such as sentiment lexicons or domain-specific dictionaries. They consist of a group of words or phrases that have been assigned a particular sentiment value.

Studies using Machine learning Techniques: ML is widely utilized in the study of extremism detection in social networks. These detection techniques are carried out on several platforms. Some studies incorporate social media sites like Twitter and Facebook as data repositories, while others utilize sites that stream videos, like YouTube, as effective tools for radicalized spreading. Furthermore, some studies examine discussion and blogging platforms like Reddit, while others examine chatting applications like WhatsApp, Telegram, and Discord. Researchers can efficiently examine extremist behavior and classification across these various platforms.

Batra et al. [21] implemented an ML model to classify tweets as radical or non-radical. The authors collected more than 30k tweets using hashtags related to Islamic organizations like #ISIS, #Taliban, and #jihad and labeled them as positive (radical) or negative (non-radical) using sentiment analysis. The authors performed binary classification based on sentiment results using the LR technique. Human intervention was necessary based on the classification result to prohibit the tweet. In conclusion, the authors suggested a method for automatically categorizing tweets using ML, followed by human interaction to decide whether or not to block them.

Araque et al. [54] studied the influence of emotional characteristics such as sadness, anger, and happiness in detecting online extremist content. They carried out various binary classification experiments to understand the effect of these features on classification performance and to determine which emotional features play a significant role. The authors collected a dataset containing more than 500 radicalized and neutral newspaper articles from Dabiq and Rumiyah magazines and approximately 200k radical and neutral tweets. They used NLP approaches, including lexicon-based techniques, FastText, and Word2Vec, to extract emotional features and feed them into LR and SVM algorithms for classification.

A few studies take into account the psychological perspective in extremism detection. Nouh et al. [77] investigated psychological and behavioral signs for the identification of online radical content and the influence of radical users. In this study, the authors worked with two public radicalized datasets collected from Kaggle [71], [72]. They extracted radical psychological and behavioral features using TF-IDF and Word2Vec approaches and utilized the LIWC dictionary to produce scores for each feature. RF, NN, SVM, and KNN algorithms were then used to perform content classification. In the second phase, the authors created an interaction graph among radicalized users to examine which users have the greatest impact. This graph shows various types of relationships, such as likes, followers, and retweets.

The effectiveness of religious and ideological features in identifying online radicalization on Twitter was evaluated by Kursuncu et al. [81]. They used the Lucky Troll Club public dataset [82] and the Kaggle dataset containing ISIS tweets [71], which were annotated by experts as radical or non-radical. The authors used Word2Vec to extract religion, ideology, and hate features and evaluated their relationship and effectiveness on the detection task. They fed SVM, LR, RF, and KNN classifiers with each feature category separately and then with combined features.

How religious material content detects online extremists was investigated by Rehman et al. [55]. They performed binary classification to investigate the relationship between the language of extremist organizations and the language of religion. In this study, the authors worked with five different datasets: two datasets containing religious texts and three

datasets containing radical and neutral tweets [71], [72], [73]. They used TF-IDF as a feature extraction tool to extract radical and religious features, which were then fed into RF and SVM algorithms. They evaluated the impact of religion by conducting two classification experiments, one based only on radical features and the other based on both radical and religious features.

Mussiraliyeva et al. [24] proposed a two-stage framework to detect online radicalization on the Twitter platform. The authors collected more than 12k posts from Twitter using hashtags like #islamophobia, #whitesupremacy, and #radicalizedwords, and labeled them as either radical or neutral. In the first stage, the radicalization score of users' posts was determined by performing sentiment analysis to determine whether the post was positive or negative. In the second stage, an ML algorithm was used to categorize the tweets as radicalized or not.

Berhoum et al. [78] evaluated 16 ML algorithms for binary and multi-classification to study their performance in extremist detection. They performed two experiments: the first classified tweets into extremist and non-extremist, and the second experiment classified extremism into three categories: religious, political, and intellectual. The authors merged two datasets, one publicly available from Kaggle containing more than 16k radical tweets [71], and another dataset they collected from Twitter using radical hashtags like #ISIS, #Jihad, and #Kofar. They labeled the dataset as extremist or non-extremist using sentiment analysis and then used the TF-IDF approach to extract features for feeding machine-learning algorithms.

Some researchers focused on the Arabic language in detecting online radicalization on the Twitter platform. Mursi et al. [23] proposed a methodology to detect hateful content in Arabic tweets. They collected more than 3k Arabic tweets using Arabic hashtags and had them annotated by experts as hateful or non-hateful. TF-IDF and count vectorizing techniques were utilized to extract hateful features from the tweets, which were then used to train SVM and MLP classifiers. They evaluated their classifier using public data containing more than 100k Arabic tweets. Similarly, Aldera et al. [22] performed binary classification on more than 89k Arabic tweets, labeling them as either extremist or non-extremist. In addition to the EDA technique, they employed TF-IDF, n-gram (Bigrams, Trigrams), and Word2Vec techniques to determine the interconnections of these characteristics and extract extremist features. LR, SVM, NB, RF, and BERT algorithms were used for the classification task.

Several works focused on geo-location to detect online radicalization on Twitter. Sharif et al. [59] focused on extremist activity in the Afghanistan region. They collected more than 7k tweets and classified them as neutral or extremist, further dividing the extremist tweets into pro-Afghan and pro-Taliban organizations, while labeling other tweets as neutral or irrelevant. The authors used four ML classifiers (SVM, NB, D.T, and KNN) for classification. Initially, they used TF-IDF and n-grams as feature extraction techniques, in addition to the PCA technique for dimensionality reduction. In the same context, Mussiraliyeva et al. [63] focused on Russia and surrounding regions to identify extremist behaviors. They focused on extremist features in the Kazakh language to detect extremist content. They performed feature extraction using Word2Vec and TF-IDF methods and used NB, LR, GB, and RF as classifiers for binary classification.

A few researchers used the Facebook platform as a source for their studies in detecting terrorist activities. Tundis et al. [91] studied extremist crimes on Facebook, focusing on demographic and behavioral features that represent the personality information of an individual. These features were used to detect criminal profiles of users. The authors used a public dataset from the John Jay & ARTIS database [92], consisting of 2157 rows. The demographic and behavioral features feed DT and SVM to perform binary classification.

Asif et al. [58] also used Facebook posts and comments in multiple languages, including Urdu, English, and Roman in order to do a multi-classification to detect online extremism. They collected a dataset of more than 20k posts and comments from various pages including ARY News, Express News, and others. In their study, to identify the degree of extremism between $+5$ and -5, they created lexicon dictionaries based on sentiment analysis for each language, classifying it into Neutral, Moderate, and Low, and then using the Tf_idf as feature selection to feed the NB and SVM algorithms in order to perform multi-classification.

Theisen et al. [60] collected more than 50k posts and comments from various radical and non-radical Reddit communities using general radical keywords in order to detect extremist content. The authors used the Linguistic Inquiry and Word Count (LIWC) approach to extract features from about 1 million comments. These features were then fed into the SVM, RBF, NB, RF, DT, and LR algorithms to investigate their performance in terms of content classification.

In their work, Al-Zewairi et al. [90] used the PIRUS dataset [89], which contains data about 1,473 people associated with extremist activity, to focus on personal and demographic variables for online extremist detection in the United States. The authors used the attributes provided by this dataset as features to feed the NB, GB, and RF models for binary classification (Islamist or not) as well as multi-classification (far right, far left, or single issue).

The changing of opinions over time on social networks was investigated by Sakketou et al. [61]. The authors collected a dataset containing posts expressing the stance from different Reddit communities. They proposed a methodology consisting of two steps, first the stability of the opinions was investigated using the user's entropy as an indicator of the evolution of opinion over time, and the language of users with the highest entropy was compared with low entropy users using the LIWC method, In the second task, a binary and multi-classification job was performed using the n-gram approach with NB and LR classifiers to detect various stances.

Several studies focused on the YouTube platform to identify extremist content and investigate the effect of recommendation algorithms on driving people toward radicalization paths by examining video information such as comments, titles, number of likes, and views. Ravid et al. [26] employed ML algorithms to conduct binary content classification based on the history of user interactions, such as comments, in order to study how users transition from more liberal forums to more extreme communities. They studied how the vocabulary used in user comments on particular subjects evolves over time. The authors made use of a dataset [67] that included more than 300k videos and 1.6 million comments from various communities, such as the right and Alt-right. An expert generated a lexicon including seed terms, and based on this lexicon, extreme opinions were extracted from comments to feed the ML classifiers.

In the same context, Kathuria et al. [70] analyzed the metadata of YouTube videos statistically and performed classification using ML algorithms to detect the stance of the videos. The author made use of a dataset that included the metadata for more than 31k videos from left- and RW (right-wing) groups on several video-streaming websites, including YouTube, Vimeo, and 4chan. In the first step, the author investigated the popularity of these communities by examining how the number of likes and views varied over time. In the second step, they performed feature extraction using NLP techniques like LIWC, NER, and BOW approaches. These features were then fed to the SVM and RF classifiers to study the stance detection of each community.

Faddoul et al. [27] looked into how a recommendation system affected the spread of conspiratorial material. The authors collected about 1k seed data channels and generated their up-next video recommendations. They proposed a methodology based on two sub-models: first, using fastText and TF-IDF as feature extraction to score the video transcript, snippet, and comment and label them as conspiratorial or non-conspiratorial, and then inputting the data into LR methods. The suggested model was tested using the recommendation videos after being trained on seed data.

Kennedy et al. [62] examined hate speech on the Gab platform, conducting multi-classification on the GHC dataset using SVMs and BERT algorithms. To feed the classifiers, the authors used TF-IDF and LIWC as feature extraction techniques.

Alomari et al. [56] compared the effectiveness of two machine-learning systems for classifying extremist Arabic tweets using an AJGT dataset. As feature extraction techniques, they used Unigrams, Bigrams, Trigrams, and TF_IDF to feed two SVM and NB classifiers, which then performed binary classification. Alharbi et al. [57] used BOW and Tf_Idf to extract extremist features from the ASAD dataset and fed them to LR for multi-classification.

Studies Using Deep Learning Techniques:

Online radicalization detection research typically uses DL, a branch of machine learning that focuses on neural networks. DL employed a variety of techniques to complete its objective, including the usage of graph neural networks and their varieties, such as GraphCNN for social network analysis, CNN for computer vision applications, and LSTM for natural language processing.. Different DL algorithms were used in the literature. Some authors employed CNN for the extraction of extremist features, while others employed mixed models like LSTM-CNN in addition to a variety of algorithms, including MLP, RoBERTa, and BERT, for the classification of extremist content. Few researchers incorporated ML and DL in their studies.

Batra et al. [21] built the CNN model in addition to the LR algorithm to classify Twitter posts as extremist or not. Aldera et al. [22] developed a pre-trained BERT model to classify Arabic postings, in addition to ML classifiers. Al-Zewairi et al. [90] used MLP to identify hateful material.

Kaur et al. [104] present an LSTM algorithm for performing binary classification for data collected from various online sources such as news and blog sites. They focused on posts related to the India region, collected over 60k posts, and annotated them as radical or non-radical based on predefined features. Cohen's Kappa was used as a label measurement for different experts. The authors employed the Word2vec model as a feature extraction method to feed the LSTM network, which then performed the classification task. They evaluated it with ML techniques to compare their performance with the LSTM classifier. Similarly, Johnston et al. [105] performed multi-classification of online propaganda of terrorist organizations. The authors collected articles from websites and online forums that contain radicalized and neutral articles. They developed LSTM models to categorize the articles into several categories such as Islamic, Sovereign, and White Supremacism.

Some researchers used combined DL algorithms for the detection process. De Gibert et al. [83] employed a combination of CNN and LSTM models to conduct a binary classification of content related to hateful actions associated with White Supremacy. They worked with a dataset containing sentences collected from posts on white supremacist websites like Stormfront and other forums. The dataset was labeled as hate or non-hate based on sentence content if it was a deliberate attack toward some group based on their religious or identity, and as non-hate if it was a normal sentence without an extremist meaning. In a similar vein, Sofat et al. [106] employed a CNN_LSTM as a combined model in their assessment of the radicalization score of users in social networks. They collected articles and posts from news articles and online blogs. The authors suggested a framework in this study to detect radical content and measure users' radicalness score based on three features, with the total score representing the sum of the scores for each feature. The first feature is retrieved using the cosine similarity approach between the content and a domain glossary comprising radical terms. The second feature is sentiment analysis using the TextBlob method, and the final feature is a classification flag based on the CNN_LSTM model output.

Ahmad et al. [25] employed sentiment analysis to detect extremist tweets using the combined model. In this study, the authors collected over 30k tweets and posts from Twitter and the dark web using keywords such as #bomb and #ISIS. They then conducted a sentiment analysis on the post text to extract emotions such as anger, sorrow, happiness, and fear as input, in addition to the tweet text. The authors employed TF_IDF, n-grams, and BoW as feature extraction to transform words into vector representation to feed the CNN_LSTM classifier and perform binary classification. They compared the performance of each one and its influences on the classification model.

The influence of recommendation algorithms on YouTube platforms was investigated by Albadi et al. [28] to study the effect of personalization as viewing history on the suggested video. The author collected metadata for over 350k Arabic videos as well as their recommended videos. Word2vec and TF_IDF techniques were utilized to extract features from the video's title, description, and tags to be used as input for the CNN_LSTM classifier. Two experiments were carried out in this study: first, the CNN_LSTM was trained on seed data, and then the CNN_LSTM was trained on non-personalized seed data and verified on a recommendation video.

CNN algorithm was employed by Theodosiadou et al. [85] to study the dynamics of opinions. They used the Ansar dataset [84] in their work, which is a public dataset with over 29k Arabic and English terrorist posts. The authors only used the English dataset for their work. The authors suggested a change point detection framework that consists of two stages to detect the shifting point of users into extreme viewpoints based on time series data. In the first stage, CNN was employed to perform binary classification as terrorist or not, and they used the embedding layer as feature extraction. In the second stage, they employed the E-Divisive approach to detect the change point based on time series that classified as terrorist text.

In certain studies, the researchers employed pre-trained DL models for extremism detection purposes. Gaikwad et al. [107] classified tweets using the BERT and RoBERTa models into three categories: propaganda, radicalization, and recruiting. The authors' work collected over 61k tweets and posts using extremist hashtags from Twitter, articles, and news websites, and labeled them into three categories using Latent Dirichlet Allocation (LDA) and Word2Vec approaches.

In the same context, Rajendran et al. [108] studied the US Capitol riot during Trump's presidency using the pre-trained models BERT, RoBERTa, and DistilBERT. The authors utilized newspaper keywords and hashtags such as Stopthesteal, Donald, and Trump2020 to collect a dataset including around 90k tweets in order to perform a multi-classification of the tweets. Instead of manually labeling the tweets, they utilized the SVM model with around 1k seed data to categorize the dataset into three extremism forms or non-extremism. They then employed the TF_IDF, GloVe, FastText, and Word2Vec approaches as feature extraction to feed the pre-trained models.

The pre-trained Bert model and Bi_LSTM model were employed by Alatawi et al. [79] to detect extremist content on Twitter related to white supremacists. In this work, the dataset was collected from combined platforms (Twitter and Stormfront) using hashtags related to white supremacist organizations. The collected dataset was labeled as white supremacy or neutral based on Cohen's kappa measurement [109]. The author proposed two approaches for the detection process: the first one based on word embedding by extracting words from the dataset using Word2Vec and GloVe methods to feed the Bi_LSTM classifier, and in the second one, they deployed a pre-trained Bert model as a classifier.

Barachi et al. [110] proposed a system for detecting online extremist activity utilizing NLP and data mining tools. The authors of this study concentrated on far-right activity in the United States during Trump's presidency, collecting around 250k tweets including neutral and extremist statements. They used mood and emotion analysis to extract extremist information from tweets and then built a K-means algorithm to partition the dataset into groups based on retrieved attributes.

c: ANALYSIS BASED APPROACH

The statistical approach entails examining and studying people's social characteristics and behaviors to understand the factors that make them susceptible to extremism, ultimately leading to terrorist acts. While there are common behaviors that characterize all individuals, this procedure relies on the unique characteristics of each individual [111].

Recently, researchers have shifted from a technical perspective to a more psychological one. This approach involves various techniques, such as conducting interviews with vulnerable individuals and studying their relationships, to gain a better understanding of how their personalities and behaviors influence the phenomenon of radicalization. Some researchers also use statistical models to study the transition phase from moderate to extremist pathways in individuals within social networks. Additionally, quantitative and qualitative analyses of social media users from a psychological perspective are used to determine the intensity of their extremist views. These methods, among others, are employed to gain a better understanding of the complex phenomenon of radicalization.

Studies using analysis techniques: In addition to AI and graph-based techniques, several researchers have employed a variety of analytic techniques to detect and counter the phenomenon of online radicalization in social networks. Ottoni et al. [112] employed statistical analysis to investigate the extremist content of RW organizations on YouTube. They collected comments for over 17 million of 7k videos on various extremist channels. The authors used three approaches to analyze user comments: lexical analysis (investigating the semantics of vocabulary used by users), topic analysis (extracting the latent presence in texts), and

TABLE 6. Summary of detection methods: Techniques and datasets.

Year	Study	Strategies	Type	Appraoch	Techniques	Features Extraction	Dataset	Objective
2023	[79]	Classification	Binary	ML	SVM, LR, RF, GB,KNN	Sentiment_Based,TF_IDF	Private,Public (English)	Detecting Radical /Extremism Content
2023	[101]	Analysis	NA	Graph	NA	NA	Private (English)	Detect Hate Speech Content
2022	[21]	Classification	Binary	ML, DL	LR, CNN	Countvectorizer	Private (English)	Detecting Radical /Extremism Content
2022	[23]	Classification	Binary	ML	SVM	Word2Vec, TF-IDF	Private,Public (Arabic)	Detect Islamic Radicalism / Extremism
2022	[22]	Classification	Binary	ML, DL	LR,SVM,NB,RF,BERT	Word2Vec, TF_IDF, N_gram	Private (Arabic)	Detect Islamic Radicalism / Extremism
2022	[61]	Classification	Binary, Multi	ML	NB,LR	NA	Private (English)	Detect Opinion Shift
2022	[26]	Classification	Binary	ML	RF,GB,LR	Detecting Radical / Extremism Videos	Private (English)	Detecting Radical / Extremism Videos
2022	[106]	Classification	Binary	DL	CNN_LSTM	Sentiment_Based,TextBlob,	Private (English)	Detect Radicalism / Extremism Users
2022	[28]	Classification	Binary	DL	CNN_LSTM	TF_IDF,word2vec	Private (English& Arabic)	Detect Islamic Radicalism / Extremism
2022	[107]	Classification	Multi	DL	BERT, RoBERTa	NA	Private(English)	Detecting Radical /Extremism Types
2022	[108]	Classification	Multi	DL	BERT, RoBERTa	GloVe, FastText, Word2Vec, TF-IDF	Private (English)	Detect Radicalism / Extremism Users
2022	[110]	Clustering	Binary, Multi	Clustering	K_Means	NA	Private (English)	Detect Radical /Extremism Signals
2022	[114]	Analysis	NA	Analysis	Statistical	NA	Private (English)	Detecting Radical / Extremism Videos
2022	[65]	Analysis	NA	Analysis	Longitudinal , Quantitative	NA	Private (English)	Detect Radical / Extremism Communities
2022	[98]	Analysis	NA	Graph	SALSA	NA	Private (English)	Detect Radical / Extremism Communities
2022	[19]	Analysis	NA	Graph	BFS,SSA	NA	Private (English)	Detecting Radical / Extremism Videos
2021	[60]	Classification	Binary	ML	SVM,RBF,NB, RF,DT, LR	LIWC	Private(English)	Detecting Radical /Extremism Content
2021	[70]	Classification	Binary	ML	SVM, FR	BOW, NER, LIWC	Private (English)	Detect Radicalism / Extremism Users
2021	[85]	Classification	Binary	DL	CNN	Embedding Layer	Public (English)	Detect Radicalism / Extremism Users
2021	[79]	Classification	Binary	DL	Bert Bi_LSTM	Word2Vecv, GloVe	Private (English)	Detect Hate Speech Content
2021	[115]	Analysis	NA	Analysis	Representative Panel	NA	Private (English)	Detecting Radical / Extremism Videos
2021	[67]	Analysis	NA	Analysis	Statistical	NA	Private (English)	Detecting Radical / Extremism Videos
2021	[118]	Analysis	NA	Analysis	Statistical	NA	NA	Study the Radical / Extremism Polarization
2021	[96]	Analysis	NA	Graph	Gephi software	NA	Private (English)	Detecting Radical / Extremism Videos
2021	[20]	Analysis	NA	Graph	NetworkX Library	NA	Private (English)	Detect Opinion Polarization
2021	[99]	Analysis	NA	Graph	NA	NA	Private (English)	Detect Radical / Extremism Communities
2020	[54]	Classification	Binary	ML	LR, SVM	Lexicon-Based , FastText, Word2Vec	Private (English)	Detect Radical /Extremism Signals
2020	[55]	Classification	Binary	ML	RF, SVM	TF_IDF	Public (English)	Detecting Radical /Extremism Content
2020	[24]	Classification	Binary	ML	KNN, SVM	Sentiment_Based	Private (English)	Detecting Radical /Extremism Content
2020	[91]	Classification	Binary	ML	DT, SVM	NA	Public (English)	Detect Radicalism / Extremism Users
2020	[58]	Classification	Multi	ML	NB, SVM	TF-IDF	Private (English& Urdu)	Detecting Radical /Extremism Content
2020	[81]	Classification	Binary	ML	SVM, LR,RF,KNN	Word2Vec	Public (English)	Detect Islamic Radicalism / Extremism
2020	[27]	Classification	Binary	ML	LR	FastText,TF_IDF	Private(English)	Detecting Radical / Extremism Videos
2020	[116]	Analysis	NA	Analysis	NA	NA	Private (English)	Detect Radical / Extremism Communities
2020	[57]	Classification	Multi	ML	LR	Bow,TF_IDF	Private (Arabic)	Detect Islamic Radicalism / Extremism
2019	[59]	Classification	Binary	ML	SVM, NB KNN, DT, RF	TF_IDF	Private (English)	Detect Radical /Extremism Signals
2019	[63]	Classification	Binary	ML	NB, LR, GB, RF	Word2vec ,TF_IDF	Private (Kazakh)	Detecting Radical /Extremism Content
2019	[77]	Classification	Binary	ML	RF, ANN, SVM, KNN	TF-IDF, Word2Vec, LIWC	Public (English)	Detect Radical /Extremism Signals
2019	[104]	Classification	Binary	DL	LSTM	Word2vec	Private (English)	Detecting Radical /Extremism Content
2019	[105]	Classification	Multi	DL	LSTM	Word2vec	Private (English)	Detecting Radical /Extremism Content
2019	[25]	Classification	Binary	DL	CNN_LSTM	N_grams, BoW, TF_IDF, FastText	Private (English)	Detecting Radical /Extremism Content
2019	[113]	Analysis	NA	Analysis	Statistical	NA	Private (English)	Detecting Radical / Extremism Videos
2019	[86]	Analysis	NA	Graph	NetworkX Library	Betweenness, In-Degree Centrality	Public (English& Arabic)	Detect Radical / Extremism Communities
2018	[83]	Classification	Binary	DL	CNN_LSTM	NA	Private (English)	Detect Hate Speech Content
2018	[112]	Analysis	NA	Analysis	Topic , Lexical Analysis	NA	Private (English)	Detecting Radical / Extremism Videos
2018	[62]	Classification	Multi	ML, DL	SVM, BERT	TF–IDF, LIWC	Private (English)	Detect Hate Speech Content
2017	[90]	Classification	Binary, Multi	ML, DL	NB, GB, RF, ANN	NA	Private (English)	Detect Islamic Radicalism / Extremism
2017	[117]	Analysis	NA	Interview	Interview	NA	NA	Detecting Radical /Extremism Content
2017	[100]	Analysis	NA	Graph	Close Graph	NA	Private (English)	Detect Islamic Radicalism / Extremism
2017	[56]	Classification	Binary	ML	SVM, NB	Unigrams, Bigrams, Trigrams, TF-IDF	Private (Arabic)	Detect Islamic Radicalism / Extremism

implicit bias analysis (studying the presence of extremist content and comparing it to the moderate channel). In the same context of extremist detection on YouTube, Ledwich et al. [113] conducted a statistical investigation of extremist detection on YouTube and investigated the influence of recommendation systems. They gathered over 800 channels representing various ideologies and their recommendations and classified them into 18 groups. The authors examined the impression of recommendation channels by estimating the number of suggestions and views for each seed channel, and comparing the number of views and likes in each ideology type.

The influence of recommendation algorithms in moving users from moderate videos toward radicalized videos was investigated by Ribeiro et al. [67]. They conducted a study of a dataset containing over 300k videos belonging to three groups: one extremist and the other two moderate. The authors used user activity, such as comments and views, to determine the elevation of users' interaction in various communities and the intersections between them, such as user remarks that shift over time from a moderate channel to an extreme channel.

The relationship between the video producer and video consumption of extremist material was studied in Munger et al. [114]'s work in order to understand why YouTube is seen as a different venue for extremist rights content. The authors collected a dataset containing more than 70 million videos from about 6k political channels. They provided a Supply and Demand approach based on quantitative analysis of video information such as views, likes, and uploaded videos from right-wing sources. The number of views and uploads are the two main measures considered for supply and demand, respectively.

Hosseinmardi et al. [115] used a representative panel to study far-right information consumption by analyzing user browsing behaviors. The authors collected more than 300k user behaviors that had at least one pageview. The proposed framework examined the user's session by estimating the users' view duration throughout a month to label the users based on their ideological leanings, and then studied changes in the consumption of users to understand their dynamics.

Schulze et al. [65] used longitudinal analysis and quantitative analysis to study the dynamics of far-right organizations in messaging applications. They collected a dataset containing more than 4k messages from nine extremist channels on the Telegram platform. Based on the analysis results, it shows that the extremist indicator in message content was increasing over time.

Rekik et al. [116] proposed a recursive methodology to detect extremist communities in social networks. They collected extremist users' profiles from Twitter and YouTube platforms and annotated them using experts in this domain as extremist or not. The authors measured the degree of danger profiles after developing a radical vocabulary and comparing the user content with it. Users with high-danger profiles were considered radicalized users.

In Daniel's study [117], the author employed the Grounded Theory technique to investigate the function of the internet in driving individuals into extremist beliefs in order to discover online radicalization. The author conducted interviews with eight people who are members of RW extremist organizations, four of whom were members of extremist organizations before the Internet and four of whom were members of extremist organizations after the Internet.

Axelrod et al. [118] research examines the evolution of ideological polarization. To determine how the positions of these individuals were altered, the author conducted a statistical analysis of the individual data generated using an agent-based model, dividing the data into clusters and assigning each individual a unique ideological variance. They then measured the ideological variance of each individual after exposure to various factors, such as tolerance and extremist exposure. High polarization is represented by high variance, and vice versa.

2) PREVENTION MECHANISM

Prevention mechanisms refer to building a solid foundation in communities to assist individuals in dealing with extremist and propaganda-like digital information. The use of prevention mechanisms, as an integrated strategy with the detection approach, has been employed to achieve the primary objective of reducing the online radicalization pathway and the spread of extremist ideologies on social networks. Despite the importance of the detection mechanism in understanding the phenomenon of extremism, the last decade has witnessed a surge in interest from governments and researchers in the prevention of online radicalization and polarization [119]. This has necessitated the implementation and development of novel strategies and methods to keep users away from the path of radicalization. The prevention mechanism cannot be approached solely from a technical perspective; it is also considered complementary to soft approaches that aim to address and prevent this phenomenon from a psychological standpoint in the long term. Examples of such soft approaches include training programs for youth individuals in society [120]. Therefore, the prevention mechanism is divided into two approaches:

The hard approaches refer to the technical interventions implemented by public and private companies in collaboration with governments to forcefully prevent manifestations of extremism. These approaches include technological techniques such as content moderation, policy controls, and link recommendation [121]. Content moderation involves monitoring user content to ensure compliance with legislation and the public. Governments and companies have widely employed this technique to remove extremist content and ban extremist accounts [122]. Link recommendation is another technique used in the prevention mechanism. It is an AI technique that involves suggesting or recommending new connections to users by rewiring previous connections or suggesting new ones. The aim is to reduce user polarization,

TABLE 7. Summary of detection methods: Techniques and performance.

Year	*Study*	*Approach*	*Techniques*	*Performance*
2023	[79]	ML	SVM, LR, RF, GB, KNN	F1= 0.80
2023	[101]	Graph	Na	F1=0.87
2022	[21]	ML,DL	LR, CNN	Accuracy= 0.97, Precision=0.95, Recall= 0.95
2022	[23]	ML	SVM	Precision= 0.95, Recall= 0.89, F1=0.92, Accuracy= 0.92
2022	[22]	ML,DL	LR, SVM, NB, RF, BERT	F1= 0.97, AUC= 0.99
2022	[61]	ML	NB, LR	F1= 0.66
2022	[26]	ML	RF, GB, LR	F1= 0.61, AUC = 0.71 Precision= 0.59, Recall= 0.63
2022	[106]	DL	CNN, LSTM	Accuracy= 0.93
2022	[28]	DL	CNN, LSTM	F1=0.69, Accuracy= 0.76, Precision= 0.67, Recall= 0.71, AUC = 0.79
2022	[107]	DL	BERT, RoBERTa	F1=0.72
2022	[108]	DL	BERT, RoBERTa	F1= 0.95, Accuracy= 0.95, Precision= 0.95, Recall= 0.95
2022	[110]	Clustering	K-Means	F1= 0.93, Accuracy= 0.94, Precision= 0.92, Recall= 0.93
2022	[114]	Analysis	Statistical	Na
2022	[65]	Analysis	Longitudinal, Quantitative	Na
2022	[98]	Graph	SALSA	Na
2022	[19]	Graph	BFS, SSA	F1=0.71, Precision=0.74, Recall=0.69
2021	[60]	ML	SVM, RBF, NB, RF, DT, LR	F1= 0.90, Accuracy= 0.91, Precision= 0.97, Recall= 0.84
2021	[70]	ML	SVM, FR	F1= 0.92, Precision= 0.87, Recall= 0.98
2021	[85]	DL	CNN	F1= 0.93, Accuracy= 0.93, AUC= 0.99
2021	[79]	DL	Bert BiLSTM	F1=0.80
2021	[115]	Analysis	Representative Panel	Na
2021	[67]	Analysis	Statistical	Na
2021	[118]	Analysis	Statistical	Na
2021	[96]	Graph	Gephi software	Na
2021	[20]	Graph	NetworkX Library	Na
2021	[99]	Graph	Na	F1= 0.79, Accuracy= 0.83 Precision= 0.79, Recall= 0.83
2020	[54]	ML	LR, SVM	F1=0.97
2020	[55]	ML	RF, SVM	F1= 0.93, Accuracy= 0.93 Precision= 0.93, Recall= 0.93
2020	[24]	ML	KNN, SVM	F1=0.72
2020	[91]	ML	DT, SVM	F1= 0.79, Accuracy= 0.77 Precision= 0.77, Recall= 0.87
2020	[58]	ML	NB, SVM	F1=0.82
2020	[81]	ML	SVM, LR, RF, KNN	F1= 0.82, Accuracy= 0.93 Precision= 0.83, Recall= 0.82
2020	[27]	ML	LR	F1= 0.85 Precision= 0.78, Recall= 0.86
2020	[116]	Analysis	Na	F1= 0.96, Precision= 0.95, Recall= 0.97
2020	[57]	ML	LR	F1= 0.80, Accuracy= 0.81
2019	[59]	ML	SVM, NB KNN, DT, RF	F1= 0.84, Accuracy= 0.84 Precision= 0.83, Recall= 0.85
2019	[63]	ML	NB, LR, GB, RF	F1= 0.95, Accuracy= 0.96 Precision= 0.94, Recall= 0.95
2019	[77]	ML	RF, ANN, SVM, KNN	F1= 0.94, Accuracy= 0.94 Precision= 0.93, Recall= 0.95
2019	[104]	DL	LSTM	F1= 0.86, Accuracy= 0.93 Precision= 0.85, Recall= 0.86
2019	[105]	DL	LSTM	F1=0.91
2019	[25]	DL	CNN, LSTM	F1= 0.88, Precision= 0.90, Recall= 0.88
2019	[113]	Analysis	Statistical	Na
2019	[86]	Graph	NetworkX Library	F1= 0.93, Accuracy= 0.93 Precision= 0.91, Recall= 0.95
2018	[83]	DL	CNN, LSTM	Accuracy= 0.76
2018	[112]	Analysis	Topic, Lexical Analysis	Na
2018	[62]	ML,DL	SVM, BERT	F1=0.64
2017	[90]	ML,DL	NB, GB, RF, ANN	F1= 0.98, Accuracy= 0.97, Precision= 0.97, Recall= 0.98
2017	[117]	Interview	Interview	Na
2017	[100]	Graph	Close Graph	F1= 0.93, Accuracy= 0.99 Precision= 0.87, Recall= 0.62, AUC= 0.97
2017	[56]	ML	SVM, NB	F1= 0.84, Accuracy= 0.83

TABLE 8. Summary of imbalances mitigation techniques.

Year	Study	Approach	Dataset Type	Mitigation Techniques
2023	[79]	ML	Imbalanced	Na
2023	[101]	Graph	Imbalanced	Na
2022	[21]	ML,DL	Imbalanced	SMOTE
2022	[23]	ML	Balanced	Na
2022	[22]	ML,DL	Balanced	Na
2022	[61]	ML	Balanced	Na
2022	[26]	ML	Imbalanced	SMOTE
2022	[106]	DL	Imbalanced	Na
2022	[28]	DL	Imbalanced	weighted loss function
2022	[107]	DL	Balanced	Na
2022	[108]	DL	Imbalanced	oversampling, undersampling
2022	[110]	Clustering	Imbalanced	SMOTE
2022	[114]	Analysis	Balanced	Na
2022	[65]	Analysis	Balanced	Na
2022	[98]	Graph	Balanced	Na
2022	[19]	Graph	Imbalanced	Na
2021	[60]	ML	Imbalanced	undersampling
2021	[70]	ML	Imbalanced	Na
2021	[85]	DL	Imbalanced	oversampling, undersampling
2021	[79]	DL	Imbalanced	oversampling, undersampling
2021	[115]	Analysis	Imbalanced	weighted sampling
2021	[67]	Analysis	Imbalanced	Na
2021	[118]	Analysis	Na	Na
2021	[96]	Graph	Imbalanced	Na
2021	[20]	Graph	Balanced	Na
2021	[99]	Graph	Imbalanced	oversampling
2020	[54]	ML	Imbalanced	oversampling, undersampling
2020	[55]	ML	Imbalanced	oversampling, undersampling
2020	[24]	ML	Imbalanced	Na
2020	[91]	ML	Imbalanced	SMOTE
2020	[58]	ML	Imbalanced	SMOTE
2020	[81]	ML	Imbalanced	SMOTE
2020	[27]	ML	Balanced	Na
2020	[116]	Analysis	Imbalanced	SMOTE
2020	[57]	ML	Imbalanced	oversampling , undersampling
2019	[59]	ML	Imbalanced	oversampling , undersampling
2019	[63]	ML	Imbalanced	SMOTE
2019	[77]	ML	Imbalanced	SMOTE
2019	[104]	DL	Imbalanced	SMOTE
2019	[105]	DL	Imbalanced	SMOTE
2019	[25]	DL	Imbalanced	SMOTE, undersampling
2019	[113]	Analysis	Imbalanced	Na
2019	[86]	Graph	Imbalanced	oversampling, undersampling
2018	[83]	DL	Imbalanced	oversampling, undersampling
2018	[112]	Analysis	Imbalanced	Na
2018	[62]	ML,DL	Imbalanced	Na
2017	[90]	ML,DL	Imbalanced	SMOTE
2017	[117]	Interview	Na	Na
2017	[100]	Graph	Imbalanced	oversampling, undersampling
2017	[56]	ML	Balanced	Na

and this technique is particularly effective when dealing with social network recommendation systems.

On the other hand, soft approaches encompass psychological techniques implemented in the long term, such as education about the dangers of radicalization, community involvement, and training programs to raise awareness of polarization and radicalization on social media. Education is considered a cornerstone and fundamental preventive measure against the radicalization mechanism. Its effective strategies are employed by governments to reduce radicalization and polarization on social networks by spreading knowledge and equipping individuals with the skills to evaluate digital extremist information and resist its influence [123]. Another significant component contributing to the preventive mechanism is community participation. This strategy relies on civil associations, educational institutions, and local communities, which provide platforms for open expression of opinions and beliefs, facilitating forums for discussion and emotional support. As a result, a cultured environment is created that reduces vulnerability to extremist narratives and offers counterarguments to radicalization [124]. In conclusion, the prevention mechanism is multifaceted and incorporates both soft and hard measures. It encompasses the enforcement of legislation and regulations while also educating people through intensive programs and involving communities, thus reducing the radicalization process and avoiding societal divisions based on culture or religion.

Prevention studies: Several studies, in addition to governments, are focused on content moderation to prevent and reduce the phenomenon of radicalization and polarization. In 2019, the European Parliament passed two legislative laws: the first addressed the spread of extremist material online, and the second mandated companies to implement filtering systems for content uploads to control and prevent the upload of extremist content [125]. The Federal Bureau of Investigation (FBI) [126] offers some soft suggestions to help reduce online and offline radicalization, like being aware of one's surroundings. Avoid giving out too much personal information and keep an eye out for any changes in behavior that could indicate someone is about to use violence. [44] shows how social network platforms utilized content moderation. YouTube removed violent videos produced by 278 ISIS accounts, Facebook deleted over 2 million extremist comments, and Twitter banned more than 1 million accounts linked to extremist organizations.

In the same context, Borelli et al. [127] proposed an automated method based on hashing technology to remove extremist information from social network sites. Similar to copyright systems, this algorithm purges all variants of content that have been previously identified as extreme by tracking the fingerprints of that content. This technique was also used in [128], [129], and [130]. Ganesh et al. [131] suggested that social network platforms not only remove extremist content but also consider deleted content as illegal and implement algorithms to prevent the re-upload of deleted content. The removal and suspension of extremist accounts were proposed by [130], [131], and [129], who also suggested implementing policies to force users to verify their identities and reduce anonymity. Bilazarian et al. [132], in their prevention framework, suggested implementing the redirect technique, which involves redirecting online advertising for users searching for radicalized content to moderate content.

Few studies have utilized AI techniques to prevent radicalization and polarization. Fabbri et al. [68] focused on reducing the risk of the YouTube recommendation system in increasing the online radicalization pathway by rewiring the recommendation graph. The authors worked with a dataset of over 30,000 videos and their recommendations. They first built a network to represent the recommended videos and then used a greedy approach to determine the ideal number of rewiring edges to reduce the issue of segregation.

TABLE 9. Summary of prevention techniques.

Year	Study	Strategy	Techniques
2023	[127]	Hard	Content Moderation
2022	[68]	Hard	Link prediction
2022	[69]	Hard	Reinforcement learning
2022	[134]	Soft	Education, Spreading anti-radicalization narratives
2022	[137]	Soft	Spreading anti-radicalization narratives
2021	[44]	Hard	Content Moderation
2021	[124]	Hard	Content Moderation
2021	[125]	Hard, Soft	Content Moderation, Spreading anti-radicalization narratives
2021	[131]	Soft	Education
2021	[132]	Soft	Education
2021	[133]	Soft	Education, Spreading anti-radicalization narratives
2021	[135]	Soft	Education
2020	[126]	Hard	Content Moderation
2020	[128]	Hard	Content Moderation
2020	[129]	Hard, Soft	Redirecting technique, Spreading anti-radicalization narratives
2020	[136]	Soft	Education
2018	[123]	Soft	Education
2017	[130]	Hard	Decrease Controversial

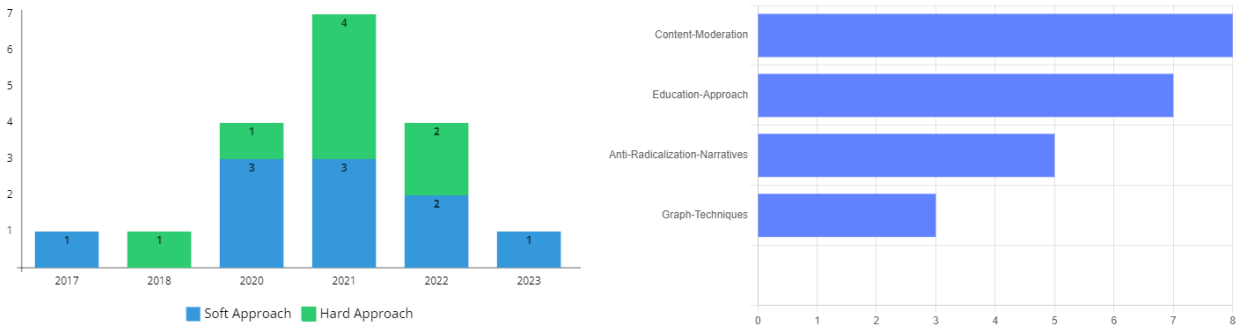


FIGURE 9. Prevention techniques used by articles.

In the same context of mitigating the impact of the recommendation system, Haroon et al. [69] developed an intervention tool based on Reinforcement Learning to reduce the extreme bias in recommendations. The authors worked with a dataset including 100k YouTube sock puppets divided into five groups, one of which was neutral and the other four had distinct ideologies. The puppets were trained by watching videos from different categories to analyze how YouTube recommends videos to viewers. The authors first conducted a statistical analysis of the recommendations to identify radicalized recommendations and then applied the Reinforcement Learning (RL) model as a mitigation strategy by viewing supplementary videos to balance the user diversity.

Garimella et al. [133] proposed a framework based on a graph-based approach to decrease controversial debates on social networks. In this study, they worked with various datasets containing contentious tweets collected using hashtags such as #ukra and #guncontrol. The authors developed an opinion graph to reflect user retweets and then utilized the random walk approach to quantify the level of debate in the user opinion graph on diverse topics. The controversy score was decreased using the ROV-AP algorithm, which is responsible for detecting connections between opposing viewpoints.

Regarding the education approach, several studies conducted by governments and researchers have suggested different training programs to educate and raise awareness among individuals about the impact of radicalization in order to reduce this phenomenon. In 2020, the Radicalization Awareness Network (RAN) [119] provided young people with a novel platform for instructional activities focused on the risks of radicalization online. References [129], [130], [134], and [135] utilized educational training as a tool to spread and raise individual awareness about this phenomenon on social media.

Regarding the education approach, several studies performed by governments and researchers suggested different training programs to educate and spread awareness for individuals about the impact of radicalization to reduce this phenomenon. In 2020, the Radicalization Awareness Network (RAN) [119] provided young people with a novel platform for instructional activities centered on the risks of radicalization online. References [129], [130], [134], and [135] used education training as a tool to spread and raise individual awareness about this phenomenon on social media.

Ebers et al. [136] examined the influence of spreading awareness films on the dangers of online radicalization in the prevention process. The authors studied the impact of this method by measuring the level of radicalization through a survey including various questions before and after individuals watched the film in the German region. This work showed that this type of prevention method had a better outcome for young individuals compared to adults. In the same context, [125] proposed minimizing the exposure to content from similar opinions and increasing exposure to opinions from different backgrounds.

Some researchers have focused on creating spaces for individuals to share information and opinions. Stray et al. [137] implemented a discussion platform as a space for dialogue to encourage users to accept opposing opinions by exposing them to diverse information from different backgrounds and topics. Schulten et al. [138] focused on evaluating training programs based on discussion platforms between individuals. The authors studied the psychological aspect to determine the impact of training programs on individual trainees and whether they become more polarized or not.

Some studies have utilized spreading anti-radicalization narratives as a tool in the prevention framework. Effendi et al. [120] used social media as a prevention method in the Indonesia region by publishing and propagating anti-radicalization narratives, including information about peace, acceptance of diverse opinions, and media literacy. This technique was also used by [128], [132], and in [135], they proposed contacting users who display extremism to warn them and offer positive narratives.

VI. ONLINE RADICALIZATION APPLICATIONS

According to the literature, a few applications have been developed for the detection and prevention of online radicalization.

[139] Proposed a system called INSiGHT, a technology to identify Islamic radicalization in online content. This system was developed using graph algorithms, which match patterns of behavior over time to find groupings connected to extremist organizations.

The TTDF system has been proposed by [140] as a real-time framework for analyzing and detecting terrorist content on the Twitter platform. The proposed framework is composed of the following phases: crawling, which gathers real-time tweets, pre-processing, which eliminates stop words and useless hashtags, training, which employs machine learning techniques, and classification, which determines whether a tweet is terrorist or not.

NewsGuard application [141] is a score system for identifying misinformation in news websites. Each website is graded according to nine criteria, ranging from 0 to 100, this application was accessible through a web browser extension and mobile application.

A multilingual web service called Perspective API [142] uses machine learning to score words and phrases in relation to the impact they have on the text. It aims to moderate content by reducing the harmful content on online platforms. It receives input in the form of sentences and outputs a score along one or more ‘‘attributes like Toxicity, DENTITY_ATTACK, INSULT, and PROFANITY.

VII. DISCUSSION AND CHALLENGES

A comprehensive understanding of the existing literature on online radicalization is provided by addressing the research questions selected for this survey. The outcome is a detailed picture of the literature studies available in this area that provides a thorough grasp of how to deal with this phenomenon, including works with the dataset, methods, and current detection and preventive tactics. However, the research process has presented certain challenges that need to be addressed. These challenges will be discussed in the upcoming section, shedding light on the complexities and limitations encountered during the study.

A. DISCUSSION

RQ1: What are the various datasets that have been utilized in the literature?

Studying the widely used aspects of the dataset and its sources in the investigation of the identification and prevention of online radicalization is necessary to respond to this question. As indicated in Table 5, the dataset used in the literature was compiled from various sources, including websites, magazines, and social media. The resulting analysis provides information about the types, variety, and accessibility of datasets utilized in earlier research, as depicted in Picture 5. Various sources were employed to gather the data; nevertheless, social media data is typically used in studies. Twitter and YouTube were employed as the key data sources in 51% and 36% of the studies, respectively. Given their popularity and ease of use by users, it is obvious that academics prefer to work with radicalization phenomena on these two platforms. Furthermore, websites and forums were employed as dataset sources in 16% of the studies, and they were used for semantic and psychological analysis. These datasets are divided into two types: Private and public datasets, which signify whether the dataset is publicly accessible or not. In addition to the sources and types of datasets used in the literature, this research demonstrates how researchers coped with the problem of data imbalance, which accounts for around 80 % of all datasets used. Essentially, three approaches were used by researchers to mitigate its significant influence. Table 8 shows that 50% of the researchers employed SMOTE, over-sampling, and under-sampling methods to mitigate data imbalances,which enables us to conclude that these techniques are considered the effective course to reduce the impact of unbalanced dataset. This study reveals that the majority of the datasets were Private datasets collected depending on the researcher's aim, which led to a paucity of public datasets.

RQ2: What are the various approaches in online radicalization detection and prevention?

It is crucial to note that there is no one right way to handle the phenomena of internet radicalization. This research identifies three strategies for the detection mechanism: the network-based strategy, the content-based strategy, and the analysis strategy. The majority of currently published publications classified their content using a content-based method. Researchers used the network approach to identify radicalized communities based on user interaction, but most studies that used the analysis approach sought to understand how radicalized communities are evaluated and how attitudes change. On the other hand, the researcher deals with two approaches for the prevention mechanism: the soft approach, which refers to raising awareness and providing information about the risk of radicalization, was widely used in the existing studies, and the hard approach, which refers to technical interventions by blocking and removing radicalized content, was used by fewer studies.

RQ3: What are the effective methods and techniques used for detecting online radicalization?

Numerous efficient methods and techniques were employed in the detection mechanism. Researchers employ two ways to do classification: ML and DL. As shown in Figure 8, machine learning (ML) is the most widely used algorithm for study in conjunction with feature extraction methods in order to train the model to recognize extremist and hateful content. Figure 8 demonstrates that SVM, LR, and RF were the most often employed algorithms in the literature; they were the main classification methods in more than 11 articles. The literature analyzed in this work reveals that 11 works used LSTM, CNN, BERT, and RoBERTa as their main detection algorithms. This is due to DL characteristics in dealing with the problem and its lack of requirement for feature extraction prior to training the model. The results also show that most studies used the following four metrics Accuracy, F1_score, Presecion, and Recall to measure the performance of the classification models as seen in Table 7. This research demonstrates that the studies that employed the analysis strategy concentrated on statistical analysis methods to examine the number of comments, likes, views, and subscriptions in order to evaluate radicalized communities. Techniques like quantitative and qualitative analysis were rarely used in investigations. Few studies used various graph detection techniques, including close graph, Random Walk, and SALSA, in the network-based approach. Some researchers employed graph algorithms as a means of feature extraction in their study. Table 6 shows the summary of detection strategies and methods used by previous studies.

RQ4: What are the existing strategies and interventions for the prevention process?

Regarding prevention strategies, this survey reveals that education programs and training to spread awareness and content moderation strategies are the most widely used methods by studies as seen in Figure 9. Some studies depend on anti-radicalization narratives as effective tools to spread and reduce radicalization. On the other hand, content moderation techniques like removing extremist content, and filtering uploads were the tools that were only used by social media, and there are only three studies that used the graph method as a prevention tool by studying the relation between users and proving and monitoring new connections in order to reduce radicalization. Table 9 shows the summary of prevention strategies and methods used by previous studies. The performance of the employed approaches is not explicitly discussed in this table due to the complexity of assessing the effectiveness of certain research techniques. This complexity arises from researchers frequently utilizing a combination of methods, making it particularly challenging to measure performance, especially in areas like education.

B. CHALLENGES

This research reveals some of the critical challenges in online radicalization studies described following.

Dataset misbalancing: The problem of misbalancing in the dataset arises when the data distribution of categories is severely skewed, which produces inaccurate results. The majority of the dataset used in the reference literature is unbalanced; the non-radicalized class is far more important than the radicalized class, which affects the detection of the radicalized content. The researchers in existing studies used a variety of techniques to minimize the impact of this issue, including oversampling, undersampling, and SMOTE techniques.

Another significant challenge in the context of this survey is the bias in the data. First, the data was biased as a result of the collection process from different sources, such as historical prejudices, sources of particular specialized ideologies, or sampling techniques. Second, there is a lack of the labeled data. Manual labeling based on expert opinion runs the danger of introducing bias into the data and leading to inaccurate results. This difficulty necessitates continual attempts to create new equines for bias prevention, as well as taking ethical standards into account when collecting data. In the literature, most of the studies employed Cohen's kappa metrics as bias mitigation to measure different experts during the data labeling process.

In terms of the detection and prevention methods, this research found that the graph method failed to accurately categorize new users as radicalized or not when the graph was formed. However, the issue with the ML method is that it depends on the effectiveness of feature extraction or selection. The main finding is that the training program for raising awareness may only be effective over a lengthy period of time, requiring a significant amount of resources and cooperation between social movements and the general public.

VIII. LIMITATION

Although the goal of this survey is to provide a vision of detecting detecting and preventing extremist ideologies, this study has certain limitations affecting the scope and depth of the investigation.

First, the literature evaluation is only comprised of 68 publications that were published between 2017 and 2023. While this time period enables us to capture recent breakthroughs in the area, it may unintentionally leave out important insights from earlier research.

Second, the main limitation of this work is its exclusive focus on classic terrorist organizations. This method helped us comprehend these organizations in great detail, but it left out other kinds of extreme organizations.

Lastly, the work was limited to the examination of textual data; it ignored other sorts of data, such as videos and photos. This restriction made it impossible to conduct a thorough analysis of the additional computer vision and image processing techniques that are employed in the detection of online radicalization.

IX. CONCLUSION

In this work, a comprehensive overview of the detection and prevention mechanisms of online radicalization is provided by analyzing a large number of articles based on the research questions. Different aspects such as datasets, methodological approaches, and techniques used in the detection and prevention mechanisms are covered. Initially, a thorough analysis of existing surveys related to the objective in this field was conducted. However, it was found that none of these surveys addressed both the detection and prevention mechanisms simultaneously. Consequently, a search methodology was developed to gather relevant articles for the study. Searches were conducted in article databases, and articles published between 2017 and 2023 that aligned with specific criteria were selected. As a result of this rigorous process, 68 papers fulfilling the study objectives were identified.

A deep analysis of the datasets used in the selected articles was conducted, identifying two main types: public datasets that are publicly available and private datasets collected by researchers. These datasets were sourced from a diverse range of platforms, with Twitter and YouTube emerging as primary sources due to their extensive user bases. Regarding the methodologies employed in both mechanisms, various approaches used by researchers to detect and prevent online radicalization were presented. For the detection mechanism, it was divided into three methodologies: the content-based approach, the network-based approach, and the analysis approach. On the other hand, the prevention mechanism was divided into two approaches: the education approach and the technical approach. In terms of techniques used in each approach, the content-based approach was observed as the most popular method employed by researchers. Machine learning (ML) and deep learning (DL) techniques in natural language processing (NLP), such as KNN, DT, SVM, CNN, and LSTM, were implemented to perform tasks like text classification and content analysis. Additionally, NLP techniques like GloVe, Word2Vec, and word embedding were used for sentiment analysis to study the emotions of users presented in discussion posts. Some articles utilized social network analysis (SNA) techniques like a close graph, Random Walk, and SALSA to analyze user interactions and detect radicalized communities within the social network.

For the prevention mechanism, the approaches were divided into two categories: the soft approach and the hard approach. The soft approach, which is the most commonly used, involves education and training programs to raise awareness about the dangers of online radicalization. Few studies focused on the hard approach, which involves technical interventions such as content moderation policies and content filtering. Additionally, some works employed graph models to address prevention strategies.

In conclusion, the work has provided a comprehensive analysis of the detection and prevention mechanisms used for online radicalization. The examination of dataset types, sources, methodological approaches, and techniques used highlights the complexity of studying this phenomenon. While significant progress has been made, several challenges are encountered. By leveraging the insights gained from this survey, researchers, policymakers, and stakeholders can collaborate to develop more advanced detection methods, deepen the understanding of radicalization processes, and establish effective prevention strategies. This collaborative effort will contribute to creating a safer and more resilient digital society, reducing the risks associated with online radicalization, and fostering a culture of tolerance, inclusivity, and democratic discourse.

X. FUTURE WORKS AND RESEARCH TRENDS

According to the findings of the survey, the number of detection studies increased from 4 in 2017 to 14 in 2022, as depicted in Figure 2. Prevention studies also witnessed growth, rising from 1 in 2017 to 7 in 2021 (note that studies conducted in 2023 were not considered due to data collection ending in May 2023). These findings illustrate a growing interest within the cybersecurity community in the field of online radicalization detection and prevention. There is a notable emphasis on leveraging advanced technologies such as machine learning, natural language processing, and social network analysis to identify and mitigate extremist content, indicating ongoing developments in research trends related to online radicalization detection.

Future research in this domain is expected to delve deeper into various aspects, and thus, the following areas are suggested for future study:

- In this research, the focus is on the strategies and impacts of social media in the development of spreading terrorism. In the future, it is necessary to broaden the emphasis beyond established terrorist groups. There are different radicalized groups that constitute an increasing threat, such as conspiracy theories, extreme gender ideologies, and anti-vaccination views.
- Future studies must focus on the role of social media algorithms and investigate the relationship between recommendation systems and the spread of extremist ideologies. The research activities will also the potential

methods to detect and mitigate the detrimental effects of recommendation systems as well as to understand the subtleties of how they contribute to the echo chambers of radicalization.

- Future directions will be also dedicated to helping researchers build balanced, multilingual, multi-sourced, unbiased datasets that will not only aid in the identification of new trends and threats but also make it easier to develop proactive counter-extremism strategies.
- Again, in the future of research on online radicalization, we suggest expanding the analytical approach beyond text analysis, by considering other media resources, like video, images, and so forth.
- This literature review highlights the predominant focus on the soft approach in the prevention mechanism, with limited attention given to the hard approach. Future investigations should explore the utilization of both the hard and soft approaches in conjunction to enhance their effectiveness.

By addressing these recommendations in future research, significant advancements can be made in understanding and combating online radicalization, leading to more effective detection and prevention strategies.

REFERENCES

[1] (2023). *Global Social Media Statistics*. [Online]. Available: https://datareportal.com/social-media-users
[2] J. J. Van Bavel, S. Rathje, E. Harris, C. Robertson, and A. Sternisko, "How social media shapes polarization," *Trends Cognit. Sci.*, vol. 25, no. 11, pp. 913–916, Nov. 2021.
[3] J. Su, A. Sharma, and S. Goel, "The effect of recommendations on network structure," in *Proc. 25th Int. Conf. World Wide Web*, Apr. 2016, pp. 1157–1167.
[4] E. M. Daly, W. Geyer, and D. R. Millen, "The network effects of recommending social connections," in *Proc. 4th ACM Conf. Recommender Syst.*, Sep. 2010, pp. 301–304.
[5] E. Ferrara, "Contagion dynamics of extremist propaganda in social networks," *Inf. Sci.*, vols. 418–419, pp. 1–12, Dec. 2017.
[6] J. P. Farwell, "The media strategy of ISIS," *Survival*, vol. 56, no. 6, pp. 49–55, Nov. 2014.
[7] B. Heller, "Combating terrorist-related content through AI and information sharing," pp. 1–8, Apr. 2020.
[8] S. Aphiwongsophon and P. Chongstitvatana, "Detecting fake news with machine learning method," in *Proc. 15th Int. Conf. Electr. Eng./Electron., Comput., Telecommun. Inf. Technol.*, Jul. 2018, pp. 528–531.
[9] C. K. Hiramath and G. C. Deshpande, "Fake news detection using deep learning techniques," in *Proc. 1st Int. Conf. Adv. Inf. Technol. (ICAIT)*, Jul. 2019, pp. 411–415.
[10] T. Zhou and P. Mengoni, "Reexamining the echo chamber effect of COVID-19 vaccine videos on YouTube," in *Proc. IEEE/WIC/ACM Int. Joint Conf. Web Intell. Intell. Agent Technol.*, Nov. 2022, pp. 415–422.
[11] A. Baby, "Computational modelling of world leaders' COVID-19 opinions: A sentiment analysis approach," in *Proc. Int. Conf. Comput., Commun., Secur. Intell. Syst.*, Jun. 2022, pp. 1–5.
[12] F. H. Calderón, L.-K. Cheng, M.-J. Lin, Y.-H. Huang, and Y.-S. Chen, "Content-based echo chamber detection on social media platforms," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Mar. 2019, pp. 597–600.
[13] J. Zhu, P. Ni, G. Tong, G. Wang, and J. Huang, "Influence maximization problem with echo chamber effect in social network," *IEEE Trans. Computat. Social Syst.*, vol. 8, no. 5, pp. 1163–1171, Oct. 2021.
[14] B. Rao, H. S. Maharana, and S. N. Mishra, "An approach to detect sub-community graph in n-community graphs using graph mining techniques," in *Proc. IEEE Int. Conf. Comput. Intell. Comput. Res. (ICCIC)*, Dec. 2016, pp. 1–6.
[15] R. Ahn, R. G. Junior, T. Hill, L. Chung, S. Supakkul, and L. Zhao, "Discovering business problems using problem hypotheses: A goal-oriented and machine learning-based approach," in *Proc. IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, Jan. 2021, pp. 137–140.
[16] M. Kamalzadeh and A. T. Haghighat, "Applying the approach based on several social network analysis metrics to identify influential users of a brand," in *Proc. 8th Int. Conf. Social Netw. Anal., Manage. Secur. (SNAMS)*, Dec. 2021, pp. 01–08.
[17] X. Liu, T. Sun, F. Bu, and H. Qin, "The analysis on the role of social network in the field of anti-terrorism take the 'East Turkistan' organization as an example," in *Proc. 5th Int. Conf. Mech., Control Comput. Eng. (ICMCCE)*, 2020, pp. 2282–2285.
[18] W. Liu, X. Zheng, T. Wang, and H. Wang, "Collaboration pattern and topic analysis on intelligence and security informatics research," *IEEE Intell. Syst.*, vol. 29, no. 3, pp. 39–46, May 2014.
[19] S. Agarwal and A. Sureka, "Topic-specific YouTube crawling to detect online radicalization," in *Proc. Int. Workshop Databases Networked Inf. Syst.* Cham, Switzerland: Springer, 2015, pp. 133–151.
[20] H. F. de Arruda, F. M. Cardoso, G. F. de Arruda, A. R. Hernández, L. da Fontoura Costa, and Y. Moreno, "Modelling how social network algorithms can influence opinion polarization," *Inf. Sci.*, vol. 588, pp. 265–278, Apr. 2022.
[21] V. Batra and S. Kumar, "A semi-automated hybrid approach to identify radicalization on social digital platform," *Indonesian J. Electr. Eng. Comput. Sci.*, vol. 27, no. 1, p. 563, Jul. 2022.
[22] S. Aldera, A. Emam, M. Al-Qurishi, M. Alrubaian, and A. Alothaim, "Exploratory data analysis and classification of a new Arabic online extremism dataset," *IEEE Access*, vol. 9, pp. 161613–161626, 2021.
[23] K. T. Mursi, M. D. Alahmadi, F. S. Alsubaei, and A. S. Alghamdi, "Detecting Islamic radicalism Arabic tweets using natural language processing," *IEEE Access*, vol. 10, pp. 72526–72534, 2022.
[24] S. Mussiraliyeva, M. Bolatbek, B. Omarov, Z. Medetbek, G. Baispay, and R. Ospanov, "On detecting online radicalization and extremism using natural language processing," in *Proc. 21st Int. Arab Conf. Inf. Technol. (ACIT)*, Nov. 2020, pp. 1–5.
[25] S. Ahmad, M. Z. Asghar, F. M. Alotaibi, and I. Awan, "Detection and classification of social media-based extremist affiliations using sentiment analysis techniques," *Hum.-Centric Comput. Inf. Sci.*, vol. 9, no. 1, pp. 1–23, Dec. 2019.
[26] E. Ravid, S. Solomon, A. Segal, and K. Gal, "Detecting radicalization on YouTube using computational models," in *Proc. Behav.-Cultural Model. Predict. Behav. Represent. Modeling Simulation*, 2022.
[27] M. Faddoul, G. Chaslot, and H. Farid, "A longitudinal analysis of YouTube's promotion of conspiracy videos," 2020, *arXiv:2003.03318*.
[28] N. Albadi, M. Kurdi, and S. Mishra, "Deradicalizing YouTube: Characterization, detection, and personalization of religiously intolerant Arabic videos," 2022, *arXiv:2207.00111*.
[29] S. Trip, C. H. Bora, M. Marian, A. Halmajan, and M. I. Drugas, "Psychological mechanisms involved in radicalization and extremism. A rational emotive behavioral conceptualization," *Frontiers Psychol.*, vol. 10, p. 437, Mar. 2019.
[30] R. Borum, "Radicalization into violent extremism I: A review of social science theories," *J. Strategic Secur.*, vol. 4, no. 4, pp. 7–36, Dec. 2011.
[31] J. Kenyon, J. F. Binder, and C. Baker-Beall, "Online radicalization: Profile and risk analysis of individuals convicted of extremist offences," *Legal Criminolog. Psychol.*, vol. 28, no. 1, pp. 74–90, Feb. 2023.
[32] A. W. Kruglanski, M. J. Gelfand, J. J. Bélanger, A. Sheveland, M. Hetiarachchi, and R. Gunaratna, "The psychology of radicalization and deradicalization: How significance quest impacts violent extremism," *Political Psychol.*, vol. 35, no. S1, pp. 69–93, Feb. 2014.
[33] J. Githens-Mazer, "The rhetoric and reality: Radicalization and political discourse," *Int. Political Sci. Rev.*, vol. 33, no. 5, pp. 556–567, Nov. 2012.
[34] B. Doosje, F. M. Moghaddam, A. W. Kruglanski, A. De Wolf, L. Mann, and A. R. Feddes, "Terrorism, radicalization and de-radicalization," *Current Opinion Psychol.*, vol. 11, pp. 79–84, Jan. 2016.
[35] J. Whittaker, "Rethinking online radicalization," *Perspect. Terrorism*, vol. 16, no. 4, pp. 27–40, 2022.
[36] 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.

[37] G. F. Hollewell and N. Longpré, "Radicalization in the social media era: Understanding the relationship between self-radicalization and the internet," *Int. J. Offender Therapy Comparative Criminology*, vol. 66, no. 8, pp. 896–913, Jun. 2022.

[38] I. Awan, "Cyber-extremism: Isis and the power of social media," *Society*, vol. 54, no. 2, pp. 138–149, Apr. 2017.

[39] (2018). *What Does Isis Post on YouTube?* [Online]. Available: https://www.hsdl.org/c/what-does-isis-post-on-youtube/

[40] A. S. Khawaja and A. H. Khan, "Media strategy of isis," *Strategic Stud.*, vol. 36, no. 2, pp. 104–121, 2016.

[41] C. Schwemmer, "The limited influence of right-wing movements on social media user engagement," *Social Media+ Soc.*, vol. 7, no. 3, 2021, Art. no. 20563051211041650.

[42] M. Wahlström and A. Törnberg, "Social media mechanisms for right-wing political violence in the 21st century: Discursive opportunities, group dynamics, and co-ordination," *Terrorism Political Violence*, vol. 33, no. 4, pp. 766–787, May 2021.

[43] M. Gaikwad, S. Ahirrao, S. Phansalkar, and K. Kotecha, "Online extremism detection: A systematic literature review with emphasis on datasets, classification techniques, validation methods, and tools," *IEEE Access*, vol. 9, pp. 48364–48404, 2021.

[44] R. T. Adek and M. Ula, "Systematics review on the application of social media analytics for detecting radical and extremist group," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 1071, no. 1, Feb. 2021, Art. no. 012029.

[45] J. Torregrosa, G. Bello-Orgaz, E. Martínez-Cámara, J. D. Ser, and D. Camacho, "A survey on extremism analysis using natural language processing: Definitions, literature review, trends and challenges," *J. Ambient Intell. Humanized Comput.*, vol. 14, pp. 1–37, Jan. 2022.

[46] Z. Trabelsi, F. Saidi, E. Thangaraj, and T. Veni, "A survey of extremism online content analysis and prediction techniques in Twitter based on sentiment analysis," *Secur. J.*, vol. 36, pp. 1–28, Apr. 2022.

[47] A. Chhabra and D. K. Vishwakarma, "A literature survey on multimodal and multilingual automatic hate speech identification," *Multimedia Syst.*, vol. 29, pp. 1–28, Jan. 2023.

[48] H. Alghamdi and A. Selamat, "Techniques to detect terrorists/extremists on the dark web: A review," *Data Technol. Appl.*, vol. 56, no. 4, pp. 461–482, Aug. 2022.

[49] M. Akram and A. Nasar, "A bibliometric analysis of radicalization through social media," *Ege Academic Rev.*, vol. 23, no. 2, pp. 279–296, Mar. 2023.

[50] S. Windisch, S. Wiedlitzka, A. Olaghere, and E. Jenaway, "Online interventions for reducing hate speech and cyberhate: A systematic review," *Campbell Systematic Rev.*, vol. 18, no. 2, p. e1243, 2022.

[51] N. M. Iannello, M. G. L. Cricchio, P. Musso, I. Grattagliano, C. Inguglia, and A. L. Coco, "Radicalization in correctional systems: A scoping review of the literature evaluating the effectiveness of preventing and countering interventions," *J. Deradicalization*, vol. 10, no. 34, pp. 177–210, 2023.

[52] C. Blaya, "Cyberhate: A review and content analysis of intervention strategies," *Aggression Violent Behav.*, vol. 45, pp. 163–172, Mar. 2019.

[53] S. Brouillette-Alarie, G. Hassan, W. Varela, S. Ousman, D. Kilinc, É. L. Savard, P. Madriaza, S. Harris-Hogan, J. McCoy, and C. Rousseau, "Systematic review on the outcomes of primary and secondary prevention programs in the field of violent radicalization," *J. Deradicalization*, vol. 2022, no. 30, pp. 117–168, 2022.

[54] 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.

[55] 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.

[56] K. M. Alomari, H. M. ElSherif, and K. Shaalan, "Arabic tweets sentimental analysis using machine learning," in *Proc. Int. Conf. Ind., Eng. Appl. Appl. Intell. Syst.* Cham, Switzerland: Springer, 2017, pp. 602–610.

[57] B. Alharbi, H. Alamro, M. Alshehri, Z. Khayyat, M. Kalkatawi, I. Ibrahim Jaber, and X. Zhang, "ASAD: A Twitter-based benchmark Arabic sentiment analysis dataset," 2020, *arXiv:2011.00578*.

[58] 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.

[59] 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.

[60] E. Theisen, P. Bours, and N. Agarwal, "Us against the world: Detection of radical language in online platforms," in *Proc. Norwegian Inf. Secur. Conf. (NISK)*, Norway, 2021.

[61] F. Sakketou, A. Lahnala, L. Vogel, and L. Flek, "Investigating user radicalization: A novel dataset for identifying fine-grained temporal shifts in opinion," 2022, *arXiv:2204.10190*.

[62] B. Kennedy et al., "The gab hate corpus: A collection of 27k posts annotated for hate speech," *PsyArXiv*, vol. 18, Jul. 2018.

[63] 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.* Cham, Switzerland: Springer, 2020, pp. 743–752.

[64] 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*. Cham, Switzerland: Springer, 2020, pp. 595–606.

[65] H. Schulze, J. Hohner, S. Greipl, M. Girgnhuber, I. Desta, and D. Rieger, "Far-right conspiracy groups on fringe platforms: A longitudinal analysis of radicalization dynamics on telegram," *Converg., Int. J. Res. Media Technol.*, vol. 28, no. 4, pp. 1103–1126, Aug. 2022.

[66] 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.

[67] M. H. Ribeiro, R. Ottoni, R. West, V. A. F. Almeida, and W. Meira, "Auditing radicalization pathways on YouTube," in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2020, pp. 131–141.

[68] F. Fabbri, Y. Wang, F. Bonchi, C. Castillo, and M. Mathioudakis, "Rewiring what-to-watch-next recommendations to reduce radicalization pathways," in *Proc. ACM Web Conf.*, Apr. 2022, pp. 2719–2728.

[69] M. Haroon, A. Chhabra, X. Liu, P. Mohapatra, Z. Shafiq, and M. Wojcieszak, "YouTube, the great radicalizer? Auditing and mitigating ideological biases in YouTube recommendations," 2022, *arXiv:2203.10666*.

[70] L. Ai, A. Kathuria, S. Panda, A. Sahai, Y. Yu, S. I. Levitan, and J. Hirschberg, "Identifying the popularity and persuasiveness of right- and left-leaning group videos on social media," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2021, pp. 2454–2460.

[71] FifthTribe. *How ISIS Uses Twitter*. Accessed: Jun. 10, 2023. [Online]. Available: https://www.kaggle.com/datasets/fifthtribe/how-isis-uses-twitter

[72] ActiveGalaxy. *ISIS Related Dataset*. Accessed: Jun. 10, 2023. [Online]. Available: https://www.kaggle.com/activegalaxy/isis-related-tweets

[73] FifthTribe. *ISIS Religious Text*. Accessed: Jun. 10, 2023. [Online]. Available: https://www.kaggle.com/fifthtribe/isis-religious-texts

[74] H. S. Alatawi. *A Dataset for White Supremacist Tweets on Twitter*. [Online]. Available: https://github.com/Hind-Saleh-Alatawi/WhiteSupremacistDataset

[75] *Hate Speech and Offensive Language Dataset*. Accessed: Jun. 18, 2023. [Online]. Available: https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset

[76] P. Gupta, P. Varshney, and M. Bhatia, "Identifying radical social media posts using machine learning," GitHub, California, 2017.

[77] 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.

[78] A. Berhoum, M. C. E. Meftah, A. Laouid, and M. Hammoudeh, "An intelligent approach based on cleaning up of inutile contents for extremism detection and classification in social networks," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 22, no. 5, pp. 1–20, May 2023.

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

[80] *GitHub—YouTube-Dataset/Conspiracy: Open Dataset of Conspiracy Theories Recommended by YouTube's Watch Next Engin*. Accessed: Jun. 18, 2023. [Online]. Available: https://github.com/youtube-dataset/conspiracy

[81] 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: Religion, ideology, and hate," *Proc. ACM Hum.-Comput. Interact.*, vol. 3, pp. 1–22, Nov. 2019.

[82] Luckytrollclub. (2015). *Lucky Troll Club Archive*. Accessed: Jun. 19, 2023. [Online]. Available: https://archive.is/9ZXeA

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

[84] *Alsayra Forum April 5 2011-May 1, 2012*. Accessed: Jun. 19, 2023. [Online]. Available: http://azsecure-data.org/get

[85] O. Theodosiadou, K. Pantelidou, N. Bastas, D. Chatzakou, T. Tsikrika, S. Vrochidis, and I. Kompatsiaris, "Change point detection in terrorism-related online content using deep learning derived indicators," *Information*, vol. 12, no. 7, p. 274, Jul. 2021.

[86] 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.

[87] Y. Zhang, S. Zeng, L. Fan, Y. Dang, C. A. Larson, and H. Chen, "Dark web forums portal: Searching and analyzing Jihadist forums," in *Proc. IEEE Int. Conf. Intell. Secur. Informat.*, Jun. 2009, pp. 71–76.

[88] N. K. Trivedi and S. K. Singh, "A systematic survey on detection of extremism in social media," *Int. J. Res. Sci. Innov.*, ol. 4, no. 7, pp. 94–103, 2017.

[89] *National Consortium for the Study of Terrorism and Responses to Terrorism (Start). Profiles of Individual Radicalization in the United States*. Accessed: Jun. 19, 2023. [Online]. Available: https://www.start.umd.edu/pirus-download-full-dataset

[90] M. Al-Zewairi and G. Naymat, "Spotting the Islamist radical within: Religious extremists profiling in the United State," *Proc. Comput. Sci.*, vol. 113, pp. 162–169, Jan. 2017.

[91] 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.

[92] S. Atran, "John jay & artis transnational terrorism database," College Criminal Justice, 2009.

[93] J. Nørregaard, B. D. Horne, and S. Adalı, "NELA-GT-2018: A large multi-labelled news dataset for the study of misinformation in news articles," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 13, Jul. 2019, pp. 630–638.

[94] J. Wang, X. Zuo, and Y. He, "Graph-based network analysis of restingstate functional MRI," *Frontiers Syst. Neurosci.*, vol. 4, p. 16, Jun. 2010.

[95] Y. Lee, Y. Lee, J. Seong, A. Stanescu, and C. S. Hwang, "A comparison of network clustering algorithms in keyword network analysis: A case study with geography conference presentations," *Int. J. Geospatial Environ. Res.*, vol. 7, no. 3, p. 1, 2020.

[96] D. Murthy, "Evaluating platform accountability: Terrorist content on YouTube," *Amer. Behav. Scientist*, vol. 65, no. 6, pp. 800–824, May 2021.

[97] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: An open source software for exploring and manipulating networks," in *Proc. Int. AAAI Conf. Weblogs Social Media*, vol. 3, 2009, pp. 361–362.

[98] F. Cinus, M. Minici, C. Monti, and F. Bonchi, "The effect of people recommenders on echo chambers and polarization," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 16, 2022, pp. 90–101.

[99] K. Papadamou, S. Zannettou, J. Blackburn, E. De Cristofaro, G. Stringhini, and M. Sirivianos, "'How over is it?' Understanding the incel community on YouTube," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. 2, pp. 1–25, 2021.

[100] 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.* Cham, Switzerland: Springer, 2017, pp. 571–587.

[101] L. Nguyen, *A Graph-Based Approach to Studying the Spread of Radical Online Sentiment*. Rochester, NY, USA: Rochester Institute of Technology, 2023.

[102] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: A content-based approach to geo-locating Twitter users," in *Proc. 19th ACM Int. Conf. Inf. Knowl. Manage.*, 2010, pp. 759–768.

[103] R. H. Nidhi and B. Annappa, "Twitter-user recommender system using tweets: A content-based approach," in *Proc. Int. Conf. Comput. Intell. Data Sci. (ICCIDS)*, Jun. 2017, pp. 1–6.

[104] A. Kaur, J. Kaur Saini, and D. Bansal, "Detecting radical text over online media using deep learning," 2019, *arXiv:1907.12368*.

[105] A. Johnston and A. Marku, "Identifying extremism in text using deep learning," in *Development and Analysis of Deep Learning Architectures*. Switzerland: Springer, 2020, pp. 267–289.

[106] C. Sofat and D. Bansal, "RadScore: An automated technique to measure radicalness score of online social media users," *Cybern. Syst.*, vol. 54, no. 4, pp. 1–26, 2022.

[107] M. Gaikwad, S. Ahirrao, K. Kotecha, and A. Abraham, "Multi-ideology multi-class extremism classification using deep learning techniques," *IEEE Access*, vol. 10, pp. 104829–104843, 2022.

[108] A. Rajendran, V. S. Sahithi, C. Gupta, M. Yadav, S. Ahirrao, K. Kotecha, M. Gaikwad, A. Abraham, N. Ahmed, and S. M. Alhammad, "Detecting extremism on Twitter during U.S. capitol riot using deep learning techniques," *IEEE Access*, vol. 10, pp. 133052–133077, 2022.

[109] M. L. McHugh, "Interrater reliability: The Kappa statistic," *Biochemia Medica*, vol. 22, no. 3, pp. 276–282, 2012.

[110] M. El Barachi, S. S. Mathew, F. Oroumchian, I. Ajala, S. Lutfi, and R. Yasin, "Leveraging natural language processing to analyse the temporal behavior of extremists on social media," *J. Commun. Softw. Syst.*, vol. 18, no. 2, pp. 193–205, 2022.

[111] I. González, M. Moyano, R. M. Lobato, and H. M. Trujillo, "Evidence of psychological manipulation in the process of violent radicalization: An investigation of the 17-A cell," *Frontiers Psychiatry*, vol. 13, pp. 1–12, Feb. 2022.

[112] R. Ottoni, E. Cunha, G. Magno, P. Bernardina, W. Meira Jr., and V. Almeida, "Analyzing right-wing Youtube channels: Hate, violence and discrimination," in *Proc. 10th ACM Conf. Web Sci.*, May 2018, pp. 323–332.

[113] M. Ledwich and A. Zaitsev, "Algorithmic extremism: Examining YouTube's rabbit hole of radicalization," 2019, *arXiv:1912.11211*.

[114] K. Munger and J. Phillips, "Right-wing YouTube: A supply and demand perspective," *Int. J. Press/Politics*, vol. 27, no. 1, pp. 186–219, Jan. 2022.

[115] H. Hosseinmardi, A. Ghasemian, A. Clauset, M. Mobius, D. M. Rothschild, and D. J. Watts, "Examining the consumption of radical content on Youtube," *Proc. Nat. Acad. Sci. USA*, vol. 118, no. 32, Aug. 2021, Art. no. e2101967118.

[116] A. Rekik, S. Jamoussi, and A. B. Hamadou, "A recursive methodology for radical communities' detection on social networks," *Proc. Comput. Sci.*, vol. 176, pp. 2010–2019, Jan. 2020.

[117] D. Koehler, "The radical online: Individual radicalization processes and the role of the internet," *J. Deradicalization*, no. 1, pp. 116–134, May 2014.

[118] R. Axelrod, J. J. Daymude, and S. Forrest, "Preventing extreme polarization of political attitudes," *Proc. Nat. Acad. Sci. USA*, vol. 118, no. 50, Dec. 2021, Art. no. e2102139118.

[119] *Extremely EUnited: Prevent Radicalization Among Youth*. Accessed: May 5, 2023. [Online]. Available: https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/projectsdetails/31077817/866883/ISFP

[120] R. Effendi, V. Sukmayadi, and A. A. Unde, "Social media as a medium for preventing radicalization (a case study of an Indonesian youth community's counter-radicalization initiatives on Instagram)," *Plaridel*, vol. 19, no. 2, pp. 1–20, Oct. 2021.

[121] S. Zeiger and J. Gyte, *Prevention of Radicalization on Social Media and the Internet*. The Hague, Netherlands: Int. Centre Counter-Terrorism (ICCT), 2020.

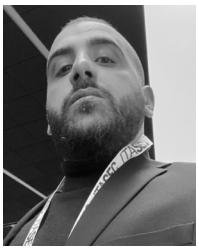
[122] S. Myers West, "Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms," *New Media Soc.*, vol. 20, no. 11, pp. 4366–4383, Nov. 2018.

[123] M. Sklad and E. Park, "Examining the potential role of education in the prevention of radicalization from the psychological perspective," *Peace Conflict, J. Peace Psychol.*, vol. 23, no. 4, pp. 432–437, Nov. 2017.

[124] A. Macaluso, "From countering to preventing radicalization through education: Limits and opportunities," *Hague Inst. Glob. Justice Work. Pap*, vol. 18, pp. 1–15, Oct. 2016.

[125] (2018). *European Commission, Proposal for a Regulation of the European Parliament and of the Council on Preventing the Dissemination of Terrorist Content*. Accessed: May 5, 2023. [Online]. Available: https://ec.europa.eu/commission/sites/beta-political/files/soteu2018-preventing-terrorist-content-online-regulation-640_en.pdf

[126] *Terrorism*. [Online]. Available: https://www.fbi.gov/investigate/terrorism
[127] M. Borelli, "Social media corporations as actors of counter-terrorism," *New Media Soc.*, vol. 25, no. 1, 2021, Art. no. 14614448211035121.
[128] S. Amit, L. Barua, and A.-A. Kafy, "Countering violent extremism using social media and preventing implementable strategies for Bangladesh," *Heliyon*, vol. 7, no. 5, May 2021, Art. no. e07121.
[129] C. Winter, P. Neumann, A. Meleagrou-Hitchens, M. Ranstorp, L. Vidino, and J. Fürst, "Online extremism: Research trends in internet activism, radicalization, and counter-strategies," *Int. J. Conflict Violence*, vol. 14, pp. 1–20, Dec. 2020.
[130] H. Wolbers, C. Dowling, T. Cubitt, and C. Kuhn, "Understanding and preventing internet facilitated radicalisation," *Trends Issues Crime Criminal Justice*, vol. 2023, p. 673, Jun. 2023.
[131] B. Ganesh and J. Bright, "Countering extremists on social media: Challenges for strategic communication and content moderation," *Policy Internet*, vol. 12, no. 1, pp. 6–19, Mar. 2020.
[132] T. Bilazarian, "Countering violent extremist narratives online: Lessons from offline countering violent extremism," *Policy Internet*, vol. 12, no. 1, pp. 46–65, Mar. 2020.
[133] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis, "Reducing controversy by connecting opposing views," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 81–90.
[134] É. Bourgeois-Guérin, D. Miconi, A. Rousseau-Rizzi, and C. Rousseau, "Evaluation of a training program on the prevention of violent radicalization for health and education professionals," *Transcultural Psychiatry*, vol. 58, no. 5, pp. 712–728, Oct. 2021.
[135] R. M. Andrian, "The role of young adults in preventing violent extremism in Indonesia," in *Tackling Terrorists Exploitation of Youth*. New York, NY, USA: Religions for Peace, 2021.
[136] A. Ebers and S. L. Thomsen, "Evaluating an interactive film on the prevention of political radicalization," *J. Deradicalization*, vol. 2022, no. 30, pp. 169–222, 2022.
[137] J. Stray, "Designing recommender systems to depolarize," 2021, *arXiv:2107.04953*.
[138] N. Schulten, F. F. Vermeulen, and B. Doosje, "Preventing polarization: An empirical evaluation of a dialogue training," *Cogent Social Sci.*, vol. 6, no. 1, Jan. 2020, Art. no. 1821981.
[139] 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.
[140] M. F. Abrar, M. S. Arefin, and Md. S. Hossain, "A framework for analyzing real-time tweets to detect terrorist activities," in *Proc. Int. Conf. Electr., Comput. Commun. Eng. (ECCE)*, Feb. 2019, pp. 1–6.
[141] *Newsguard. Rating Process and Criteria*. Accessed: Sep. 5, 2023. [Online]. Available: https://www.newsguardtech.com/solutions/newsguard/
[142] *Perspective API Web Service*. Accessed: Sep. 5, 2023. [Online]. Available: https://perspectiveapi.com/



OMRAN BERJAWI received the bachelor's degree in computer and communication engineering from the Islamic University of Lebanon, in 2019, and the master's degree in computer and communication engineering, in 2022. He is currently pursuing the Ph.D. degree in cyber security with the IMT School for Advanced Studies in collaboration with the University of Salerno. His research interests include network security, security architecture, intrusion detection systems, open-source intelligence, and information disorder. He is also working on cybersecurity solutions to counteract radicalization in digital media using computational intelligence methods. His research interests include detecting and countering this phenomenon and containing its impact.

GIUSEPPE FENZA (Member, IEEE) received the Graduate degree and the Ph.D. degree in computer sciences from the University of Salerno, Italy, in 2004 and 2009, respectively. He is currently an Associate Professor in computer science with the University of Salerno. He has over 60 publications in fuzzy decision making, knowledge extraction and management, situation and context awareness, semantic information retrieval, service oriented architecture, and ontology learning. More recently, he has worked in automating open source intelligence and big data analytics for counterfeiting extremism and supporting information disorder awareness. His research interests include computational intelligence methods to support semantic-enabled solutions and decision-making.

VINCENZO LOIA (Senior Member, IEEE) received the Graduate degree in computer science from the University of Salerno, Italy, in 1985, and the Ph.D. degree in computer science from Université Pierre & Marie Curie Paris VI, France, in 1989. He is currently a Computer Science Full Professor with the University of Salerno, where he was a Researcher, from 1989 to 2000, and an Associate Professor, from 2000 to 2004. He is the Co-Editor-in-Chief of *Soft Computing* and the Editor-in-Chief of *Ambient Intelligence and Humanized Computing*. He serves as an editor for 14 other international journals.

Open Access funding provided by 'Università degli Studi di Salerno' within the CRUI CARE Agreement