IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 10, 2021, accepted March 1, 2021, date of publication March 8, 2021, date of current version March 23, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3064178

Online Extremism Detection in Textual Content: A Systematic Literature Review

SAJA ALDERA[1], (Member, IEEE), AHMAD EMAM[2], MUHAMMAD AL-QURISHI[3,4], (Member, IEEE), MAJED ALRUBAIAN[3], (Member, IEEE), AND ABDULRAHMAN ALOTHAIM[2]

[1]Department of Management Information Systems, College of Business Administration, King Saud University, Riyadh 11451, Saudi Arabia
[2]Department of Information Systems, College of Computer and Information Sciences, King Saud University, Riyadh 11451, Saudi Arabia
[3]College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia
[4]Department of Research and Innovation, Elm Company, Riyadh 12382, Saudi Arabia

Corresponding author: Saja Aldera (saaldera@ksu.edu.sa)

This work was supported by the Deanship of Scientific Research, King Saud University, through the initiative of DSR Graduate Students Research Support (GSR).

ABSTRACT Social media networks such as Twitter, Facebook, YouTube, blogs, and discussion forums are becoming powerful tools that extremist groups use to disseminate radical ideologies and propaganda, and to recruit people to their cause. Identifying extremist social media content and profiles is a top priority for counter-terrorist agencies, technology companies, and governments. The main objective of this paper is to provide a better understanding of the definition of extremism, and a detailed review of the current research regarding online extremism in text. To identify gaps in the literature, a systematic literature review (SLR) of 45 studies published between 2015 and 2020 was undertaken, which revealed challenges, technical pitfalls in previous studies, and opportunities for extending and improving prior results in meaningful ways. The systematic review indicates the need for better understanding of the landscape and directions of the online extremism. This study offers a critical analysis of the new area of research.

INDEX TERMS Extremism detection, radicalism detection, terrorism detection, artificial intelligence, machine learning, natural language processing, systematic literature review.

I. INTRODUCTION

The recent emergence of social media networks has fostered a broad exchange of ideas and opinions. This stems from the unique characteristics of the medium, which include anonymity, minimal barriers to publishing, and the negligible cost of publishing or accessing content [1]. The three major social media networks of today – Facebook, Twitter, and YouTube – have a combined total of 3.2 billion users, a staggering number that amounts to around 42% of the world's population [2]. Since an average user spends 142 minutes daily on social media networks, engaging in activities such as logging in, viewing content, and interacting with friends [2], these networks clearly play a fundamental role in daily life. This situation also highlights the significant challenge social media networks face in terms of the need to monitor large amounts of traffic and user-generated content while evolving to meet user demand and remain competitive.

Social media networks are often exploited by so-called "extremist" groups in order to operate anonymously, promote radicalized views, and recruit new members worldwide, and young people are especially at risk [3]. For example, the Islamic State of Iraq and the Levant (ISIS), a terrorist organization, has used social media networks extensively for recruitment purposes, migrating almost all of its propaganda campaigns from physical venues to online networks, thereby raising more funds and attracting more manpower in the digital environment [3]. Due to the sheer volume of social media traffic, ISIS-generated online content, as well as online content from other extremist groups, is difficult to detect and remove, especially using the manual detection methods that most social media networks rely on [4]. Therefore, developing an effective and fully-automated extremism detection technique is a top priority for governments, counter-terrorist organizations, cybersecurity companies, and other

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

stakeholders, all of whom have an interest in limiting extremist activity and content online [5].

Unfortunately, terrorist groups exploit the characteristics of social media networks. Currently, online channels are intensively used by terrorist groups to promote their extremist views and attract new recruits around the world, and young people are frequently targeted [3]–[5]. With the availability of social media networks, terrorist organizations, particularly ISIS, have migrated their operations from physical locations such as mosques and houses to these platforms to disseminate propaganda, raise funds, and recruit manpower [3]. Since the amount of data on social media has increased dramatically, it is difficult to detect radical content manually. Therefore, an effective automatic method to achieve extremism detection is urgently required. Moreover, automatic detection of extremist profiles on social media has become a central interest and top priority for governments and counter-terrorism organizations in their fight against terrorist social network accounts. Hence, the creation of information technology resources to identify cyber terrorists will help counter online extremism [6].

The primary research contributions of this paper are as follows:

1. Offers a systematic literature review (SLR) and taxonomy of social media extremism detection and prediction. The literature is categorized into multiple taxonomies, each of which has a set of related features. Hence, the paper presents multiple perspectives on online extremism detection methods.
2. Identifies challenges, open issues, and future research opportunities, and compares and contrasts existing approaches used to analyze and detect online extremism.
3. Highlights the datasets used for social media extremism detection and prediction.
4. Provides a guideline for researchers to select the most suitable techniques and methods for social media extremism detection and prediction.

The rest of this paper is organized as follows: the next section provides a background of extremism definitions and text analysis techniques; the third section explains the literature review process used to select and analyze research articles; the fourth section presents the SLR's results, followed by fifth section classify the selected articles into different facets and briefly outline the approaches. Finally, the sixth section presents a discussion of the review and provides concluding remarks.

II. BACKGROUND

A. DEFINITIONS OF EXTREMISM

Within the literature, there are different definitions of the term ''extremism'', as well as a number of overlapping views, and there is no universally-accepted definition in academia or government [7]. This is unsurprising in view of the fact that, among researchers and governments, no consensus has been established in reaching even a definition of the ubiquitous term ''terrorism'', which is a significantly less nuanced concept when compared to the related concept of extremism [8]. However, governments and other agencies have offered definitions of extremism that share the following features: firstly, an emphasis on the political, social, and religious dimensions of extremism [9]; secondly, a positioning of the ''extremist'' as an individual or entity that stands in active opposition to a context-dependent set of values and norms (e.g., in the UK, democracy) [10]; and thirdly, a recognition that, to some extent, extremism – as well as the related concept of ''radicalization'' – can lead to violence and, in particularly severe cases, acts of so-called ''terrorism'' [11].

One of the key difficulties associated with defining extremism relates to the different contexts in which extremists are perceived. For example, the UK Home Office's [10] *Counter Extremism Strategy* defines extremism as ''the vocal or active opposition to our fundamental values, including democracy, the rule of law, individual liberty, and respect and tolerance for different faiths and beliefs''. By contrast, Sharma [12] defined extremism as ''the extent of support for the use of violence against outgroup members on the basis of their ...affiliation ...to achieve ...[religious, political, or social] objectives''. Notably, the latter definition differs from the former in that it sees extremism as something inherently linked to violence (and, more specifically, to terrorism) [7]. By contrast, the former definition recognizes extremist acts that do not necessarily lead to violence. Both definitions also highlight that, depending on the nature of the outgroup and ingroup, as well as the normative values of the ingroup, what one conceptualizes at extremist behavior is likely to change.

With the above issues in mind, it is clear that the diversity on display in the available definitions of extremism is reflective of the complex, nuanced, context-dependent, and multifaceted nature of the term. Another challenge in defining these terms is that, for some authors, the term ''extremism'' is used interchangeably with ''terrorism'' and ''radicalization'', while for other authors, these terms are viewed as related but distinct [13]. For example, by Sharma's [12] definition, all extremists are terrorists – or, at the very least, individuals who support terrorism – due to the link between violence and terrorism [14]. However, by the UK Home Office's [10] definition, extremism is one of many possible antecedents to terrorism. Regarding the link between extremism and radicalization, some scholars use these terms interchangeably [13], whereas others suggest that radicalization refers to the process by which an individual acquires behaviors, emotions, and beliefs that can be considered extremist [7].

Since the construction of profiles for online extremism is a fundamental aim of this research, it is also worth drawing attention to the religious, social, and political dimensions of extremism. These dimensions are immediately noteworthy because, in some cases, individuals radicalize (i.e., acquire extremist beliefs) along political lines, while others radicalize along religious lines, or even those relating to social issues [15]. Although each of these extremism dimensions shares features (i.e., because the religious, social, or political beliefs, emotions, and behaviors an individual acquires

will be ''extreme'') [7], the nature of each dimension is distinct [15].

Based on the abovementioned considerations, the operational definition of online extremism adopted for the proposed research is the following: online extremism refers to an interaction that occurs among extremist groups or individuals and their society, using social media platforms, in which the groups or individuals promote extremist thoughts, attitudes, or behaviors with the aim of segmenting society using political, social, and/or religious topics, and changing the nature of others' thoughts, attitudes, or behaviors. Hence, this study's definition views extremism as a two-way interaction that exists between extremists and members of society, where the former seek to influence the latter in promoting their extreme political, social, and/or religious ideologies.

B. SOCIAL MEDIA NETWORKS AND TEXT ANALYSIS

Social media networks create new ways for people from diverse cultures and backgrounds to communicate and share information. The most popular social media platforms are Twitter, Facebook, and YouTube. On these platforms, the written text that users produce may be unstructured or semi-structured, and it may also contain spelling and grammatical errors that generate semantic ambiguity. For this reason, extracting accurate information from unstructured or semi-structured text data of this kind is challenging.

Text mining can solve the abovementioned problem. With the growing amount of digital information emerging from social media platforms, blogs, and other online sources, text mining has become increasingly important. Text mining is a knowledge discovery process that leverages computational algorithms to identify patterns and relationships in large amounts of text [16]. Text mining exploits practices from multidisciplinary fields such as natural language processing (NLP), information retrieval, information classification, and text analysis. Notable differences between text mining and text analysis are stated as follows:

- Text Analysis
  - Access to unstructured text data
    - Transforms unstructured data into structured data
    - Indexing and search operations
    - NLP
      - Entity extraction
    - Classify relationships among entities
      - Sentiment Analysis
- Text Mining
  - Semantic determinations
  - Key term identification
  - Document categorization

The two most commonly used text analysis techniques in social networking are the following: firstly, the use of classification (supervised) with machine learning algorithms or ontology-based text classification; and secondly, the use of clustering (unsupervised). However, information extraction has become more challenging due to the availability of multilingual textual data in social media platforms. Privacy and trust in online communication are also major issues [16]. To overcome these challenges, researchers must apply different text analysis techniques that can filter out relevant information from large text corpora.

C. NATURAL LANGUAGE PROCESSING APPROACHES

NLP is one of the techniques employed in extremism detection. NLP is a collection of techniques that enables computers to understand the natural, unprocessed language that humans use for communication [17]. This capability relies on the identification of grammatical structure and meaning in an input text [18]. The main tasks of NLP are semantic analysis and syntax analysis, where the former includes lexical analysis, sentiment analysis, language recognition, entity recognition, and others; and the latter includes tasks such as stemming, lemmatization, and speech tagging [18]. Many computational techniques are used for semantic and syntax analysis, but NLP is not straightforward. This is because language operates at several levels (e.g., the phoneme, the syllable, the sentence, and so on) [17]. To function effectively, NLP algorithms must parse and unpack the relationships and rules that define the structure and meaning of language. For this reason, NLP is always combined with artificial intelligence (AI) and machine learning (ML) techniques such as the nearest neighbor search (NNS), support vector machines (SVMs), and decision trees (DTs) [17].

D. MACHINE LEARNING APPROACHES

ML is a powerful technique used in extremism detection. ML is defined as the ability of a computer to teach itself how to make decisions using available data [19]. In this context, the data used for ML are known as training data. The types of decisions a computer may make in ML include classification and prediction, where a computer may classify a new piece of data by training models using learning algorithms.

If the learning algorithm (or training model) depends on labeled data (i.e., human-classified data), then it is referred to as a supervised algorithm. In extremism and hate speech detection, a corpus of data could be labeled manually as either containing hate/extreme speech or not [20]. When the training data are unlabeled, then the learning algorithms are referred to as unsupervised algorithms [21]. These algorithms learn how to classify autonomously, and this process is based on similarities and differences within the data. When both supervised and unsupervised learning are combined, the algorithm is known as a semi-supervised learning algorithm [19]. Several famous ML algorithms have been devised, including naïve Bayes classifiers, the k-nearest neighbor (KNN) algorithm, SVMs, and DTs [20]. ML algorithms are widely incorporated into online extremism detection systems due to the vast amount of data transmitted over social media networks, where it is infeasible for manpower alone to process the available data.

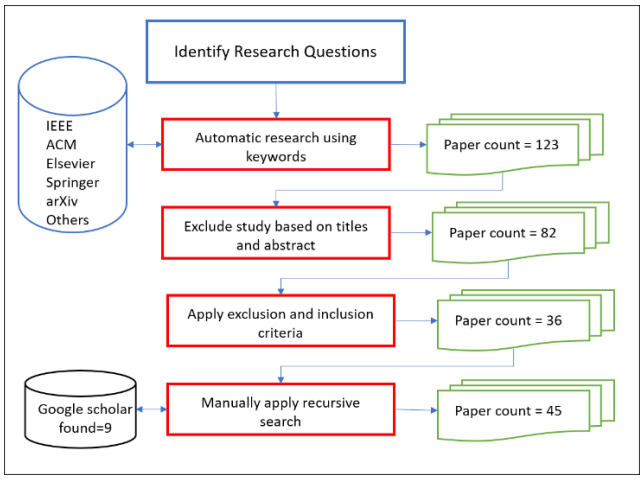

FIGURE 1. Study selection process.

In recent years, deep learning (DL) techniques have been used in extremism detection, and many proposed systems have achieved remarkable accuracy [22], [23]. Deep learning is a form of ML that can learn in an unsupervised way from unstructured datasets. Multiple hidden layers are used to progressively extract higher level features from the input. The two main deep learning architectures used in text classification are convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Noteworthily, deep learning algorithms require significantly greater amounts of training data when compared to traditional ML algorithms [17].

III. METHODS

We sought to synthesize evidence from all relevant papers in the area of extremism. For this reason, a systematic literature review (SLR) method was adopted based on the guidelines of [24]. This involved comprehensively searching and mapping the relevant literature to address problems by identifying, critically evaluating, and integrating the most relevant works (e.g., high-quality individual studies addressing one or several research questions). The procedure used to review, evaluate, and synthesize available evidence is described in this section and the next. Figure 1 illustrates the review search strategy, which involved both automatic and manual search.

A. SEARCH STRATEGY

1) STAGE 1: DEFINITION OF RESEARCH QUESTIONS

The main goal of systematically mapping studies is to provide an overview of a research area and to identify the quantity and type of available studies and results. To achieve this goal, the following research questions (RQs) were established:

RQ1: What research has been conducted to date regarding online extremism?

RQ2: Is there a benchmark dataset? What is the most common social media network used to detect online extremism?

RQ3: What techniques have researchers adopted for online extremism detection on social media networks?

RQ4: What key challenges and limitations currently face online extremism detection on social media networks?

TABLE 1. Inclusion and exclusion criteria.

Included articles were:	• Published in the period between 2015 to 2020 • Available as full text • Written in English • Relevant to the research questions • In the domain of computer science
Excluded articles were:	• Outside the search timeframe • Not available in full text • Not written in English • Not relevant to the research questions • Not relevant to the field of computer science • Duplicated studies

2) STAGE 2: CONDUCT SEARCH

Scientific databases and services, including ACM, arXive, IEEE, Springer, Elsevier, and Google Scholar, were included in the literature search. The intention was to achieve the greatest possible coverage in the areas of computer science and software engineering. The search was undertaken based on the research questions and keywords specified in the first step, and all retrieved articles were collected into a knowledge base repository classified according to the database name. These keywords included ''extremism or radicalism or terrorism detection'', ''extremism or radicalism or terrorism'', ''extremism on social media''. The review involved extracting relevant data from journal articles, conference papers, and workshops written in English and published in electronic databases between 2015 and 2020. After the first and second stages of the SLR procedure, duplicate articles were removed, which yielded a total of 123 research articles.

3) STAGE 3: TITLE AND ABSTRACT SCREENING

Title and abstracts of papers were examined to eliminate irrelevant research articles. Out of 123 research articles, 41 were eliminated, leaving 82 in total at the end of this stage.

4) STAGE 4: SCREENING OF PAPERS FOR INCLUSION AND EXCLUSION CRITERIA

Inclusion and exclusion criteria were used to exclude studies that are not relevant to the research questions. We included only the most recent articles (within the last six years), and we focused on all papers, books, and other literature with abstracts proposing a contribution to the research area. We excluded any articles that were outside the domain of computer science. Table 1 shows the criteria for this review. At this stage, 46 research articles were excluded, leaving 36 in total.

5) STAGE 5: RECURSIVE SEARCH

We used Google Scholar to locate the articles that cited the core papers, thereby reducing the time required to find

TABLE 2. Data extraction form.

Extracted data	Description
ID	Unique identifier for each paper
Study title	Paper name
Publication date	Year of publication (2015-2020)
Publication platform	Publisher Platform
Type of paper	Journal, conference, or workshop
Research type	Experimental paper, analytic paper, survey paper, etc.
Problem	Problem the paper addresses
Context	Study area (extremism, radicalism, or terrorism)
Methodology/algorithm	Research methodology used in paper
Social network	Social media platform used as data source
Dataset	Size and details about the dataset used
Dataset language	Dataset language (English, Arabic, or other)
Performance metrics	Final results of the study

related papers. At this stage, 9 additional research articles were found. The final result of the systematic review included 45 core studies.

B. DATA EXTRACTION STRATEGY

Data extraction is a critical element in any systematic literature review (SLR). Therefore, to ensure effective data extraction, we read the 45 included studies and summarized them in a template. This aided in understanding the subject matter and identifying gaps for future research. Excel spreadsheets were used to store extracted data according to study title, year of publication, social network source, performance metrics, and others (see Table 2).

IV. RESULTS

This section presents the SLR's search results, focusing on statistical data regarding the percentage of included papers per article type, year of publication, publication platform database, and research context. Following this, the SLR's results for the four main research questions are presented.

A. SEARCH RESULTS

1) PUBLICATION PLATFORM DATABASE

The included studies were retrieved from different online databases, as shown in Figure 2. The most common platform was IEEE with 11 studies. 15 studies were published on "Other" platforms, including ResearchGate and the Open University.

2) ARTICLE TYPE

Included studies were classified based on each paper's methodology. As shown in Figure 3, 76% were experimental papers aiming to explain a particular outcome or phenomenon based on specific actions or interventions. 7% were survey papers, which gathered and analyzed questionnaire data to draw conclusions. 13% were analytic papers, which collected data from various researchers to analyze contrasting viewpoints. Finally, 10% were solution proposal papers, in which a solution for a specific problem was introduced.

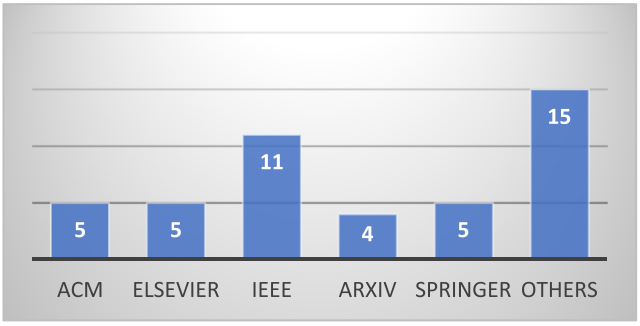

FIGURE 2. Publication platform.

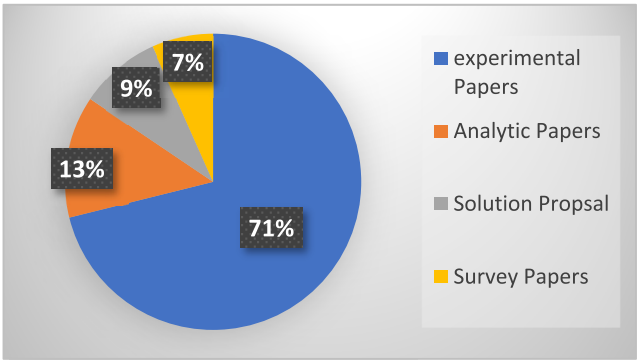

FIGURE 3. Article type.

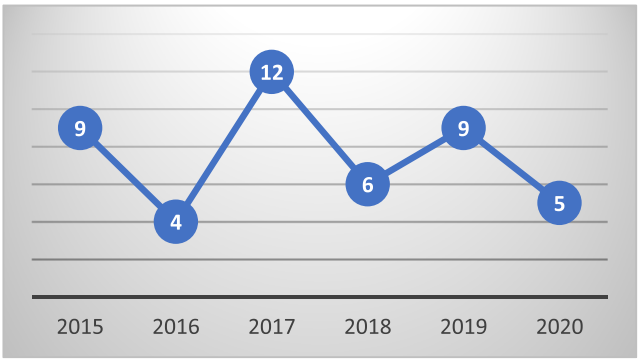

FIGURE 4. Number of publications per year.

3) YEAR OF PUBLICATION

Figure 4 shows the spread of study publication dates between 2015 and 2020. Among the 45 publications identified, 13 were published in either 2015 or 2016. 12 studies were published in 2017, and this figure decreased to 6 in 2018. 9 studies were published in 2019, and in 2020, 5 studies were published. This result indicates that the availability of studies addressing online extremism detection has gradually increased since 2015.

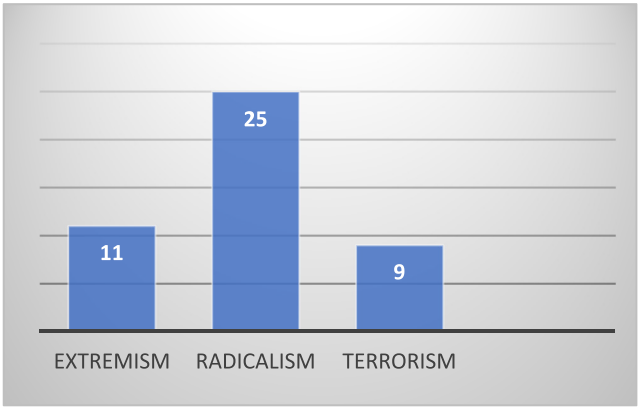

FIGURE 5. Publication context.

4) CONTEXT

Figure 5 shows the distribution of the studies in terms of the research context and focus. The radicalism context had the greatest number of studies (n = 25), while 11 and 9 studies were published in the extremism and terrorism contexts, respectively.

B. RESEARCH QUESTION RESULTS

1) RESEARCH QUESTION 1

The issue of online extremism detection has been approached from different perspectives in the literature. Multidisciplinary teams consisting of experts in the social sciences, computer science, and psychology have sought to devise solutions [25]. Most studies have focused either on the analytic examination of online behavior or on practical proposals that describe solutions for the early detection or prediction of extremist content in online environments [25].

a: ANALYSIS OF ONLINE EXTREMISM

Researchers who have analyzed online extremism have pursued different objectives. For example, Klausen [26], Carter *et al.* [27], and Chatfield *et al.* [28] adopted an analytic approach, and they examined the behavior of extremist social media users via proxy parameters (e.g., posting frequency and user mentions). Rowe and Saif [29] studied the social media actions and interactions of Europe-based Twitter users before, during, and after they exhibited pro-ISIS behavior. The researchers found that, before becoming extremists, users typically exhibited uncharacteristic behavior (e.g., communicating with new users and adopting new terms), a fact that can be exploited for detection and prediction purposes.

The studies undertaken by Klausen *et al.* [26], Chatfield *et al.* [28], Vergani and Bliuc [30], and Rowe and Saif [29] points the way towards potential solutions for online radicalization detection, principally because they all studied the types of language that ISIS members and supporters typically use to communicate and recruit new members. Specifically, the studies mainly focused on how terrorists communicate with their followers, the terms that terrorists typically use in their communication, mentions of ISIS supporters and fighters, emotional language, and the high relevance of social homophily on the diffusion of pro-ISIS terminology. It is important to recognize that these studies focused on understanding the online radicalization process rather than automatically detecting it.

Badawy and Ferrara [31] collected around 1.9 million tweets from 25,000 ISIS-associated Twitter accounts. They successfully classified more than 50% of the tweets under the following categories: violence, theological, sectarian, and name. Moreover, two topics dominated the dataset: namely, theological issues and violence. Looking at the share of these topics in the whole dataset, as well as among those classified, suggests the importance of these two topics for ISIS. In particular, theological and violence-related issues accounted for over 30% of all tweets and more than 50% of the classified tweets. Moreover, the authors noticed a sharp increase both in theological content and in Twitter discussion among ISIS members after the organization's self-proclamation of its caliphate status. Additionally, the authors discussed multiple examples indicating a close connection between real-time events and the ISIS Twittersphere. Perhaps most importantly, the finding that radical propaganda tends to revolve around four independent types of messaging (theological, violence, sectarianism, and influential actors and events) may serve as a starting point for online extremism detection applications.

Lara-Cabrera *et al.* [32] derived a set of computational features (mostly sets of keywords) from a collection of indicators proposed in social science theories of radicalization (e.g., feelings of frustration, introversion, and perceptions of discrimination). Having identified these computational features, the researchers noted that they could automatically extract them from the data. The authors concluded that, while the proposed metrics showed promising results, the fact that they were based mainly on keywords was a limitation. As such, more refined metrics can be proposed to map social science indicators.

b: PREDICTION AND DETECTION OF ONLINE EXTREMISM

Many recent research papers have focused on the automatic detection of radical content online. Most of these works are based on certain textual features, and they frequently exploit existing machine learning (ML) techniques. Both Ashcroft *et al.* [33] and Kaati *et al.* [34] constructed classifiers to detect radical messages using different types of data-dependent features (e.g., most common hashtags, word bigrams, and frequent words) and data-independent features (e.g., stylometric features and time features). Both papers demonstrated that combining data-dependent and data-independent features is advantageous for enhancing classifier performance.

Agarwal *et al.* [35] and Magdy *et al.* [36] proposed a ML approaches involving the use of one-class classifiers to predict radical tweets based on the KNN and SVM algorithms. The classifiers were trained while considering different features (e.g., religious terms, negative emotions, emoticons, hashtags, and bag of words). Rowe and Saif [29] constructed

a lexicon for pro-ISIS language, which contained the words most commonly used by pro-ISIS accounts (e.g., ''caliphate'' and ''Islamic State''). The authors also established a lexicon for anti-ISIS language (e.g., ''Isil'' and ''Daesh''). Using these lexicons, the researchers identified two radicalization signals for detecting a user's adoption of a pro-ISIS position. It is noteworthy that no radicalization knowledge base ontologies have been proposed in the literature for text mining social networks.

Most studies in this field have focused on answering multiple fundamental questions, the most important of which is how to detect extremist content online. Arpinar *et al.* [37] sought to address this issue, which was also covered by Kaati *et al.* [34]. Another priority is to predict which users are most susceptible to Islamist propaganda and, thus, are in danger of radicalization. Anwar and Abulaish [38] and Ferrara *et al.* [1] addressed this problem by studying interactions between extremist users and non-extremist users. Perhaps even more importantly, the seminal works by Ashcroft *et al.* [33] and Scanlon and Gerber [39] sought to identify platforms that allow the exchange of extremist content. Several authors have attempted to create comprehensive models for measuring the level of extremist influence to which a person is exposed, and Fernandez *et al.* [3] provided one of the best such solutions. However, no existing solution can ensure the reliable and accurate recognition of extremist content across all platforms, which has motivated a new wave of research to incorporate multiple domain-specific techniques that have been proven useful in isolation. The work of Saif *et al.* [40] proposed a semantic graph-based approach to identify pro-ISIS and anti-ISIS social media accounts. The authors developed multiple classifiers and showed that the classifier trained for semantic features outperformed those trained from lexical, sentiment, topic, and network features by 7.8% on an average F1-measure.

c: LITERATURE REVIEWS OF ONLINE EXTREMISM

From a computer science perspective, the scientific study of extremism is in its infancy, and there are limited studies in the field. We found only one survey article during the literature review process. In the study [25], the authors reviewed papers from 2003 to 2011. Except for a paper published in 2013, there have been no survey papers dedicated to this area in the last nine years. In this paper, we report the first recent research effort to survey the literature related to radical and extremism analysis, detection, and prediction on social media networks to identify gaps that need to be filled. We analyze the various radicalization detection mechanisms that researchers have previously implemented, and we propose a categorization method that can be followed to manipulate these methods. Our main goal is to identify the gaps that should be examined, to highlight the limitations of existing approaches, and to identify the means that can be used to design effective, accurate, and reliable radicalization detection techniques. Moreover, we expect this paper to become the gateway for future contributions in this area of research. In a book chapter [41], the researchers provided an overview of current AI technologies addressing the problem of online extremism, and they also identified limitations in existing approaches, highlighting opportunities for future research.

2) RESEARCH QUESTION 2

a: SOCIAL MEDIA WEBSITES AS DATA SOURCES

To address the problem of detecting extremism on social media networks, the SLR identified the number of publications using different data sources over the last six years (see Figure 3). Based on Figure 6, it is clear that most studies have focused on the best-known social media networks, video sharing platforms, microblogging websites, and forums. Over the last six years, social network datasets were most commonly obtained using Twitter. Blogs were the second most popular source for each of the years, and only one publication was identified using WhatsApp.

b: BENCHMARK DATASETS

The most popular benchmark dataset used in the SLR's included studies was Kaggle (1), which was employed by [42]–[44], [5], and [32]. Kaggle (2) was the second most popular, featuring in [3], [42], [43], and [44]. Research groups in the remaining articles created custom datasets from various social media networks, only a few of which were made publicly available. Table 3 provides an overview of the available datasets. Most of the available datasets are English and compiled from Twitter, apart from one Arabic publicly available dataset [46]. Clearly, there is a significant need for a novel and more recent dataset.

c: DATASET LANGUAGES

Exclusively English datasets were used in 86% of the papers, while 7% used only Arabic datasets. A further 2% used a combination of English and Arabic datasets, while 5% used a combination of English and other languages (e.g., Dutch, Italian, and Urdu).

3) RESEARCH QUESTION 3

The SLR revealed that interest has recently been growing in the field of social media text analysis. Earlier in this section, we discussed the analysis of online extremism and offered a brief overview of literature on online extremism detection. This section focuses on papers that have attempted to develop techniques for automatic online extremism detection. Figure 7 shows the methods and techniques that were used in the included studies to facilitate the detection of extremism on social media networks, as well as the related concepts used in the studies within different subfields of AI.

a: MACHINE LEARNING APPROACHES

Machine learning (ML) and deep learning (DL) were the most commonly used techniques in studies addressing the detection of extremism on social media networks. 21 out of 32 experimental papers used the ML approach. The most

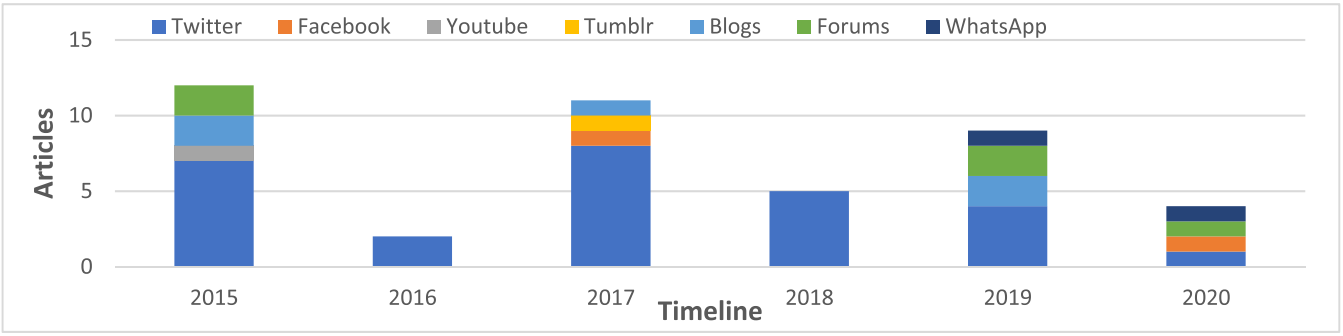

FIGURE 6. Social network datasets used in the literature.

TABLE 3. Publicly available datasets.

Data source	Year	Language	Dataset size	Link
Kaggle (1)	2015	English	Over 17,000 tweets from around 100 pro-ISIS users worldwide	https://www.kaggle.com/fifthtribe/how-isis-uses-twitter
Kaggle (2)	2016	English	122,000 tweets from 95,725 distinct users	https://www.kaggle.com/activegalaxy/isis-related-tweets
[45]	2017	English	25,482 tweets	https://github.com/prabhakar267/extremism-TT-data set
[46]	2020	Arabic	24,078 tweets from 174 accounts connected to ISIS	https://data.mendeley.com/datasets/8kftmw7rct/draft?a=0fd4d8fc-42e8-478b-bd17-ba61996aad61

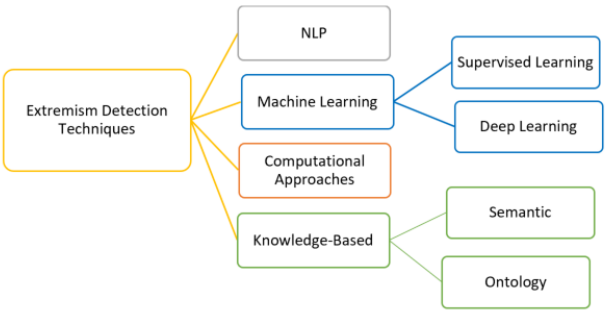

FIGURE 7. AI techniques for extremism detection.

frequently used algorithms were SVM, random forest (RF), and LSTM. The SVM algorithm yielded results with an accuracy level greater than 90% in many studies [20], [33], [35], [47], [48]. The RF algorithm also yielded high accuracy in many studies, as shown in Table 4. Deep learning algorithms such as RNN and CNN were first used in 2016, and since then, they have shown promising results in many studies. In [49], the authors used an approach based on long short-term memory (LTSM) networks to identify extremist content on social media networks. The approach outperformed most ML techniques (e.g., SVM and RF) in terms of precision (85.9%). Table 4 summarizes related studies in extremism detection using machine learning and deep learning.

In [49], the proposed architecture consists of two parts: the first part is a data preparation system comprising data collection, annotation, and cleaning the raw data; and the second part is a training and visualization system that is concerned with generating word embeddings, and then classifying texts as radical or non-radical. For the first part of the proposed architecture, the data were obtained from multiple sources, including Twitter, news articles, and blogs, and 61,601 records were collected. The researchers provided these collected records to domain-specific experts for annotation into three categories based on specific criteria: namely, radical, non-radical, and irrelevant. Afterwards, they measured the agreement between two experts on the annotated data using Cohen's kappa coefficient, yielding 0.79 inter-expert agreement, which is a sign of substantial agreement. The researchers then pre-processed the text to train their model, which involved selecting English records only from articles and blogs and removing stop words. The second part of the architecture –training and visualization – uses word embeddings to help carry words with similar meanings that occur in similar context. Word2vec was used and vectors were neutrally trained separately for radical and non-radical texts to generate word embeddings using a Levenberg training mechanism. Finally, a feedforward neural network supported by LSTM was used for weight training as it assumes texts are interrelated. The proposed approach outperformed most existing machine learning techniques, including SVM, RF, and MaxEnt, in terms of precision (85.9%).

In [23], the researchers presented a scheme to determine if a given Twitter handle could belong to an extremist using three groups of information related to the Twitter handle, profile, and textual content of users. The framework first uses highly indicative patterns related to extremism to filter out unlikely extremists. Ultimately, a likely extremist is identified

TABLE 4. Studies on machine learning and deep learning.

Ref	Year	Algorithm	Feature selection	SN	Dataset size	Performance metric
[20]	2015	SVM	TF-IDF, Glasgow, Entropy	DWFP, OSAC	Thousands of Arabic web pages	Acc = 93.6%
[47]	2015	SVM	Linguistic Features	Twitter	Two publicly available datasets	P = 0.78, R = 0.88, F = 0.83, Acc = 0.97
[34]	2015	AdaBoost	Data-independent Features, Data-dependent Features	Twitter	6,729 profiles	Arabic Tweets (Acc = 0.86), English Tweets (Acc = 0.99)
[33]	2015	SVM, AdaBoost	Stylometric Features, Time-based Features, Sentiment-Based Features	Twitter	2,000 pro-ISIS tweets, 2,000 anti-ISIS tweets	SVM (Acc = 0.97), AdaBoost (Acc = 0.100)
[13]	2015	Best-first Search	Profile Features, Contextual Metadata, N-gram	YouTube	612 videos	Acc = 0.69
[35]	2015	SVM	Religious, War-related, Bad Words, Negative Emotions, Internet Slang	Twitter	45,344,958 tweets	P = 0.78, R = 0.88, F = 0.83, Acc = 0.97
[50]	2015	Rule-based Classifier	Topic Modeling	Twitter	N/A	N/A
[1]	2016	RF	Time-based Features, Profile Features, Network Features	Twitter	25,000 suspended Twitter accounts with 3 million tweets	AUC = 0.93
[44]	2017	Logistic Regression	Topic Sensitivity, Post Effectiveness Indicators, Emotion Indicators	Twitter	Kaggle 17,000 tweets	Acc = 80%, F1 = 0.95
[51]	2017	RF	Topic Modeling, Tone Analysis, Semantic Features	Tumblr	3,228 posts extracted	P = 0.81, R = 0.84
[52]	2017	Naïve Bayes	Sentiment-based Features, Lexicons	Twitter	1,480 tweets	N/A
[53]	2017	GR-Learnt	Unigram, Word2vec, Sentiment-based Features, Emotions, Political Terms	Forms	3,938 profiles	Acc = 0.27
[48]	2017	SVM, RF	Stylometric Features, Time-based Features	Twitter	48,644 tweets	SVM (Acc=98.03), RF (Acc = 98.43)
[54]	2019	SVM	N-gram, TF-IDF	Twitter	A total of 7,500 tweets	Acc = 0.84
[21]	2019	RF	Textual Features, Psychological Features, Behavioral Features	Twitter	Kaggle 17,000 tweets	P=1.0, R=1.0, F1=1.0
[49]	2019	LSTM	Word2vec	News, articles, blogs	61,601 total records	P = 0.85, R = 0.53, F = 0.65
[23]	2019	Char-LSTM SVM, LabelSpreading (RBF)	Twitter Handle-related Features, Profile-related Features, Content-related Features	Twitter	300,000 tweets	Char-LSTM (F1= 0.76), LabelSpreading - RBF (F1= 0.76), SVM (F1= 0.65)
[55]	2020	Linear SVC	Sentiment Lexicon Features	Facebook	20,000 rows (476,050 words) and 4300 lexicons word count	P=0.81, R=0.82, F1=0.81
[56]	2020	SVM-OAA	N/A	Twitter	173 accounts having 24,078 tweets	F1=83.2%, Acc=82.6%
[57]	2020	SVM	Semantic Similarity, Emotion signals	Magazine, Kaggle	N/A	different for each dataset
[58]	2020	Semi-supervised ML	Term Frequency, Pattern	WhatsApp	N/A	N/A

Precision (P), Recall (R), F-measure (F1), Accuracy (Acc), Area Under the Curve (AUC).

using only features related to their usernames. The authors first demonstrated that extremists on Twitter are inclined towards adopting handles with similar patterns. To that end, they used the well-known Levenshtein ratio as a measure of distance between two Twitter handles and performed two-sample t-tests to demonstrate that, compared to normal users, extremists tended to choose similar handles. The first step involved feature engineering, where the researchers used three major groups: Twitter handle-related features, profile-related features, and content related features. The second step for each pair of extremist users involved computing the similarity between their corresponding handles. To determine whether it was possible to infer the labels (i.e., extremist versus non-extremist) of unseen handles based on their proximity to the labeled instances, the researchers first obtained the feature spaces associated with the labeled and unlabeled instances. This was achieved by converting each handle into a vector of five features: length of handle, maximum number of occurrences of a character in handle, number of unique characters in handle, number of digits at beginning of handle, and handle complexity. Ultimately, the two feature spaces were used as inputs to the semi-supervised and supervised

learning algorithms, including SVM, KNN, character-based LSTM (Char-LSTM), and RF. SVM achieved the highest precision of 0.96 in identifying violent online extremists, which shows the significance of the proposed feature set. The semi-supervised LabelSpreading (radial basis function [RBF]) approach performed as effectively as Char-LSTM, both of which achieved the highest F1-score. Along with the fact that Char-LSTM showed promising results in the literature without using hand-crafted features, the results further demonstrate the effectiveness of the proposed feature engineering scheme. For Char-LSTM, the precision was 0.77, and a high recall of 0.76 was maintained on the positive class. This suggests that the memory module in LSTM assists in minimizing the number of false negatives. Moreover, the most significant feature was the number of unique characters in the username, while the least important feature was the maximum number of occurrences of a character in the username. This observation further demonstrates that the frequency and importance of features in a labeled dataset are not necessarily consistent with each other and, in fact, are inversely related in [23].

Both of the abovementioned studies, [49] and [23], used ML and DL techniques, but different results were obtained. In [49], the authors found that DL outperformed other ML techniques, whereas [23] reported that ML techniques would outperform DL techniques if the correct features were used properly. Therefore, it is clear that feature engineering plays a critical role in the performance of an ML algorithm. Hence, the features used in the included articles were also analyzed, focusing on extremism detection. Finding the correct features in a classification problem is often one of the most challenging tasks. Therefore, a description of the three categories of features that were commonly used in the literature is given as follows:

- *Textual Features (NLP features):* These were the most commonly used type of feature in the included studies. Textual features include part-of-speech (POS) tagging, named entity, N-grams, term frequency–inverse document frequency (TF-IDF), keywords, sentiment analysis, and topic modeling. In total, 75% of the papers combined different types of contextual features. However, TF-IDF, N-grams, and topic modeling were the most used.
- *Time Features:* 11% of the papers used time features (e.g., average number of tweets posted per day).
- *Profile Features:* 14% of the papers used features related to social media profiles (e.g., number of tweets, followers, and friends associated with each profile; the frequency of adoption of hashtags, mentions, and URLs; and profile descriptors).

Table 4 indicates that the key factor in the ML model's performance was the utilized features group. In particular, using the multi features group increased model performance. For example, the authors in [21] used a combination of textual features, psychological features, and behavioral features, and they reached an F-measure of 1.0 using the random forest model. We recommend that every researcher should use a combination of different features depending on the utilized dataset, after which they should evaluate the features that are the most effective and give a better performance.

b: KNOWLEDGE-BASED APPROACHES

Knowledge-based approaches are concerned with knowledge-based entity and relation extraction and analysis. To improve the accuracy of existing algorithms for online extremism detection, Saif *et al.* [40] and Fernandez and Alani [59] identified the semantic context in which terms are used, or the context in which certain entities are mentioned. Additionally, Masmoudi *et al.* [60] proposed an ontology-based approach through a reasoning mechanism for mining extremism indicators from online messages. This relied on the use of semantics and domain knowledge. The researchers also evaluated the effectiveness of the proposed approach by selecting a subdataset from Kaggle, which contained 2,317 messages, both extremist and neutral, from 20 extremist Twitter profiles. After comparing the results to a baseline method, it was found that the use of semantic and domain knowledge improved the detection of extremism indicators compared to the baseline method (F1 $= 0.44$ and F1 $= 0.22$, respectively). However, the accuracy of the results is still considered low because the authors focused on ontologies only, neglecting to combine ontologies with other methods.

c: NATURAL LANGUAGE PROCESSING APPROACHES

Researchers have examined natural language processing (NLP), as well as the use of lexicons, to interpret text. Drawing on a lexicon-based approach to classify tweets as violent, theological, sectarian, or other, Badawy and Ferrara [31] and Fernandez *et al.* [3] explored the use of social media networks by ISIS to spread propaganda and recruit militants. The researchers found that content driven by violent, theological, or sectarian themes plays a crucial role in ISIS messaging.

d: COMPUTATIONAL APPROACHES

Regarding computational approaches, it is notable that, from this reviewer's perspective, Kursuncu *et al.* [42] proposed the most effective technique for online extremism detection. This can be attributed to the technique's high accuracy, as well as its coverage of three fundamental areas of extremism: religion, ideology, and hate. The authors also sought to address one of the most pressing issues in the online world, namely, the growing presence of websites and social media accounts that systematically spread aggressive Islamist propaganda. Early recognition of such online users is paramount, and this is why, in their study, the researchers presented a novel method based on word representations and N-gram analysis. Relying on the most commonly mentioned words, the authors defined three dimensions of extremist thought (religion, ideology, and hate), and they tried to determine how essential each component is. To this end, they conducted N-gram analysis with hierarchical clustering, followed by topical analysis. Using this method, it was possible to detect

whether a profile was an extremist or not in 97% of cases where all three dimensions were tracked. However, the ratio of false positives was high. It was also found that ideology is a more reliable predictor of extremist attitudes than the other two dimensions, and this factor alone was sufficient for an impressive 90% accuracy. Considering the importance of this issue and the relative effectiveness of the proposed method, this work must be considered a substantial success. As a matter of fact, the paper could end up influencing an entirely new wave of research in this field.

4) RESEARCH QUESTION 4

This SLR reviewed a large number of studies in order to illuminate the current state of extremism detection on social media networks. The results indicate that the concept of online extremism detection has evolved rapidly in recent years, and it has prompted increased attention and interest among organizations and researchers. However, the SLR also demonstrated that, within the current body of knowledge on extremism detection on social media networks, there is much work that remains to be done. Specifically, three critical challenges and gaps in the literature were identified: firstly, the lack of a common definition of extremism; secondly, the lack of a commonly adopted dataset; and thirdly, methodology-specific challenges, including (i) the inadequacies of the proposed AI and deep learning systems for online extremism detection; (ii) the lack of studies focusing on profile similarity detection; and (iii) the lack of studies that have exploited deep learning models with multi-level features (e.g., the combined use of text/NLP features, timing features, and profile-related features).

Regarding the first challenge, Section 2.1 outlined the multi-faceted and nuanced nature of the concept of ''extremism'', demonstrating that, despite the availability of more than one hundred definitions, there is little consensus about what extremism actually is. This problem is also reflected in the studies examined in this SLR, where – although each one focuses on the phenomenon of online extremism – subtle differences in the definitions of extremism adopted in each one undermine direct comparisons of the results; and, even more importantly, undermine confidence in the degree to which the proposed systems actually detect online extremism in a valid and reliable way. For this reason, it will be worthwhile to address this gap in the literature by establishing a systematic and evidence-based operational definition of the term extremism, around which consensus may form in future research. It is anticipated that future researchers in this field will be able to adopt the same definition in their explorations of AI applications for online extremism detection.

Based on the results of this SLR, it is reasonable to conclude that all of the datasets used to study online extremism in prior studies have suffered from definite challenges that limit the significance, generalizability, and relevance of their findings. As a case in point, for datasets made publicly available by the Kaggle data science community, labelling by human annotators casts doubt on the quality of the data. Additionally, for private datasets, limited access to these datasets means that the nature of the results published using them cannot be guaranteed or reproduced [61]. As the SLR indicated, and as discussed more generally in [61] and [62], problems arising from datasets include the potential biases that may affect available datasets (e.g., bias arising from geographical location of extremist and normal users), as well as the potential for incomplete data (e.g., in this case, incomplete user profiles). Therefore, it is worthwhile to create a benchmark/golden and balanced dataset on extremism that can serve as ground truth for the proposed model, as well as a testing and training ground for further research.

As for the three methodology-specific challenges and gaps that were identified in this SLR, the first of these relates to the inadequacies of the proposed AI and deep learning systems for online extremism detection. Although the SLR revealed that applications such as that of Kursuncu *et al.* [42] yielded substantial accuracy rates of up to 97% in detecting extremism on social media networks (which, to the best of the researcher's knowledge, is the most effective system proposed to date), the rate of false positives was high. Therefore, there is significant room for improvement. Specifically, a novel deep learning model exploiting multi-level features (in this case, the combined use of text/NLP features, timing features, and profile-related features) is recommended to counter the gap identified in the literature, namely, that only lexicon-based features and social media network features have previously been considered.

V. CONCLUSION

This paper presented an SLR to provide an overview of state-of-the-art online extremism detection techniques over the past six years. First, we analyzed the concept of extremism in different contexts. 45 relevant research articles were identified that concentrated on online extremism analysis and detection, and these covered the following research contexts: extremism, radicalism, and terrorism. Most of the reviewed research articles (71%) were experimental papers (71%), followed by analytical papers (13%), solution proposal papers (9%), and survey papers (7%).

Among the studies that focused on automatic online extremism detection, most used ML techniques and achieved high accuracy results. The most frequently used algorithms were SVM, RF, and LSTM. The SLR revealed that feature selection plays a critical role in ML technique performance. Textual features were the most commonly used features, appearing in 75% of the included studies, and the most common among these were TF-IDF, N-grams, and topic modeling. 14 % of the studies used profile-related features, while 11% used time-related features. Regarding the dataset, most studies used private datasets. Twitter was the most commonly used social media platform for dataset construction, and almost all of the used datasets used the English language.

Significantly, this SLR highlights a series of limitations that point towards future research possibilities. Firstly, the issues related to datasets and, in particular, the lack of a

reliable benchmark dataset with solid verification is a key limitation; most of the datasets used were inaccessible private datasets. Secondly, several AI approaches were used to detect online extremism, but there is no comparison between these approaches to address the strengths and limitations of each one. Finally, different techniques and solutions should be developed in order to adapt different features and contexts.

In this SLR, we focused only on formally published research literature and disregarded the large volume of ''grey'' literature generated by experts in different non-academic sources, including social networks, videos, and Internet blogs. As a future work, a multivocal literature review (MLR), which is a type of SLR, will be conducted to retrieve more relevant data by searching the grey literature and social networks. This will be important to identify research topics that emerge from the software industry.

REFERENCES

[1] E. Ferrara, W.-Q. Wang, O. Varol, A. Flammini, and A. Galstyan, ''Predicting online extremism content adopters and interaction reciprocity,'' in *Proc. Int. Conf. Social Inform.*, 2016, pp. 22–39.

[2] D. Chaffey, ''Global social media research summary August 2020,'' in *Proc. Smart Insights*, Aug. 2020. [Online]. Available: https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/

[3] M. Fernandez, M. Asif, and H. Alani, ''Understanding the roots of radicalisation on Twitter,'' in *Proc. 10th ACM Conf. Web Sci. WebSci*, May 2018, pp. 1–10.

[4] R. Borum and T. Neer, ''Terrorism and violent extremism,'' in *Handbook of Behavioral Criminology*. Cham, Switzerland: Springer, 2018, pp. 729–745, doi: 10.1007/978-3-319-61625-4_41.

[5] R. Lara-Cabrera, A. G. Pardo, K. Benouaret, N. Faci, D. Benslimane, and D. Camacho, ''Measuring the radicalisation risk in social networks,'' *IEEE Access*, vol. 5, pp. 10892–10900, 2017.

[6] E. Bodine-Baron, T. Helmus, M. Magnuson, and Z. Winkelman, *Examining ISIS Support and Opposition Networks on Twitter*. Santa Monica, CA, USA: RAND Corporation, 2016, pp. 29–30.

[7] 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.

[8] 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.

[9] J. M. Berger, *Extremism*. Cambridge, MA, USA: MIT Press, 2018.

[10] *Counter-Extremism Strategy*, UK Home Office, Gov.U.K., London, U.K., 2015.

[11] H. El-Said and R. Barrett, ''Radicalisation and extremism that lead to terrorism,'' in *Globalisation, Democratisation and Radicalisation in the Arab World*, J. Harrigan and H. El-Said, Eds. London, U.K.: Palgrave Macmillan, 2011, doi: 10.1057/9780230307001_11.

[12] K. Sharma, ''What causes extremist attitudes among Sunni and Shia Youth? Evidence from northern India,'' in *Evidence from Northern India*. Washington, DC, USA: Program on Extremism, Nov. 2016.

[13] S. Agarwal and A. Sureka, ''A focused crawler for mining hate and extremism promoting videos on YouTube.'' in *Proc. 25th ACM Conf. Hypertext Social Media*, Sep. 2014, pp. 294–296.

[14] D. A. Alexander and S. Klein, ''The psychological aspects of terrorism: From denial to hyperbole,'' *J. Roy. Soc. Med.*, vol. 98, no. 12, pp. 557–562, Dec. 2005.

[15] A. Rink and K. Sharma, ''The determinants of religious radicalization: Evidence from Kenya,'' *J. Conflict Resolution*, vol. 62, no. 6, pp. 1229–1261, Jul. 2018.

[16] R. Irfan, C. K. King, D. Grages, S. Ewen, S. U. Khan, S. A. Madani, J. Kolodziej, L. Wang, D. Chen, A. Rayes, N. Tziritas, C.-Z. Xu, A. Y. Zomaya, A. S. Alzahrani, and H. Li, ''A survey on text mining in social networks,'' *Knowl. Eng. Rev.*, vol. 30, no. 2, pp. 157–170, 2015.

[17] Y. Tsuruoka, ''Deep learning and natural language processing,'' *Brain Nerve*, vol. 71, no. 1, pp. 45–55, 2019.

[18] V. Mallamma and M. Hanumanthappa, ''Semantical and syntactical analysis of NLP,'' *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 3, pp. 3236–3238, 2014.

[19] O. Pentakalos, ''Introduction to machine learning,'' in *Proc. C. Impact*, 2019.

[20] T. Sabbah, A. Selamat, M. H. Selamat, R. Ibrahim, and H. Fujita, ''Hybridized term-weighting method for dark Web classification,'' *Neurocomputing*, vol. 173, pp. 1908–1926, Jan. 2016.

[21] 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.

[22] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, ''Deep learning for hate speech detection in tweets,'' in *Proc. 26th Int. Conf. World Wide Web Companion-WWW Companion*, no. 2, 2017, pp. 759–760.

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

[24] K. Petersen, S. Vakkalanka, and L. Kuzniarz, ''Guidelines for conducting systematic mapping studies in software engineering: An update,'' *Inf. Softw. Technol.*, vol. 64, pp. 1–18, Aug. 2015.

[25] D. Correa and A. Sureka, ''Solutions to detect and analyze online radicalization: A survey,'' Jan. 2013, *arXiv:1301.4916*. [Online]. Available: https://arxiv.org/abs/1301.4916

[26] J. Klausen, ''Tweeting the *Jihad*: Social media networks of Western foreign fighters in Syria and Iraq,'' *Stud. Conflict Terrorism*, vol. 38, no. 1, pp. 1–22, 2015.

[27] J. A. Carter, S. Maher, and P. R. Neumann, *#Greenbirds: Measuring Importance and Influence in Syrian Foreign Fighter Networks*. London, U.K.: The International Centre for the Study of Radicalisation and Political Violence, King's College, 2014.

[28] A. T. Chatfield, C. G. Reddick, and U. Brajawidagda, ''Tweeting propaganda, radicalization and recruitment: Islamic state supporters multi-sided Twitter networks,'' in *Proc. 16th Annu. Int. Conf. Digit. Government Res.*, May 2015, pp. 239–249.

[29] M. Rowe and H. Saif, ''Mining pro-ISIS radicalisation signals from social media users,'' in *Proc. 10th Int. Conf. Web Soc. Media (ICWSM)*, 2016, pp. 329–338.

[30] M. Vergani and A. Bliuc, ''The evolution of the ISIS'language: A quantitative analysis of the language of the first year of Dabiq magazine,'' *Sicurezza, Terror. E Societa*, vol. 2015, pp. 7–20, Oct. 2017.

[31] A. Badawy and E. Ferrara, ''The rise of Jihadist propaganda on social networks,'' *J. Comput. Social Sci.*, vol. 1, no. 2, pp. 453–470, Sep. 2018.

[32] R. Lara-Cabrera, A. Gonzalez-Pardo, and D. Camacho, ''Statistical analysis of risk assessment factors and metrics to evaluate radicalisation in Twitter,'' *Future Gener. Comput. Syst.*, vol. 93, pp. 971–978, Apr. 2019.

[33] M. Ashcroft, A. Fisher, L. Kaati, E. Omer, and N. Prucha, ''Detecting jihadist messages on Twitter,'' in *Proc. Eur. Intell. Secur. Informat. Conf.*, Sep. 2015, pp. 161–164.

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

[35] S. Agarwal and A. Sureka, ''Using KNN and SVM based one-class classifier for detecting online radicalization on Twitter,'' in *Distributed Computing and Internet Technology*, vol. 8956. Cham, Switzerland: Springer, 2015.

[36] W. Magdy, K. Darwish, and I. Weber, ''FailedRevolutions: Using Twitter to study the antecedents of ISIS support,'' First Monday, 2016.

[37] I. B. Arpinar, U. Kursuncu, and D. Achilov, ''Social media analytics to identify and counter islamist extremism: Systematic detection, evaluation, and challenging of extremist narratives online,'' in *Proc. Int. Conf. Collaboration Technol. Syst. (CTS)*, Oct. 2016, pp. 611–612.

[38] T. Anwar and M. Abulaish, ''Ranking radically influential Web forum users,'' *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 6, pp. 1289–1298, Jun. 2015.

[39] J. R. Scanlon and M. S. Gerber, ''Automatic detection of cyber-recruitment by violent extremists,'' *Secur. Informat.*, vol. 3, no. 1, pp. 1–10, Dec. 2014.

[40] H. Saif, T. Dickinson, L. Kastler, M. Fernandez, and H. Alani, ''A semantic graph-based approach for radicalisation detection on social media,'' in *The Semantic Web* (Lecture Notes in Computer Science), vol. 10249, E. Blomqvist, D. Maynard, A. Gangemi, R. Hoekstra, P. Hitzler, and O. Hartig, Eds. Cham, Switzerland: Springer, 2017.

[41] M. Fernandez and H. Alani, ''Artificial intelligence and online extremism: Challenges and opportunities,'' in *Predictive Policing and Artificial Intelligence*, J. McDaniel and K. Pease, Eds. 2020, pp. 1–31. [Online]. Available: http://oro.open.ac.uk/69799/

[42] 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.

[43] M. Fernandez and H. Alani, "Contextual semantics for radicalisation detection on Twitter," in *Proc. Workshop Semantic Web Social Good (SW4SG), 17th Int. Semantic Web Conf. (ISWC)*, Monterey, CA, USA, Oct. 2018. [Online]. Available: http://ceur-ws.org/Vol-2182/
[44] S. Das Bhattacharjee, B. V. Balantrapu, W. Tolone, and A. Talukder, "Identifying extremism in social media with multi-view context-aware subset optimization," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2017, pp. 3638–3647.
[45] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on Twitter," in *Proc. NAACL*, 2016, pp. 88–93.
[46] M. Fraiwan, "Annotated ISIS radical tweets," *Mendeley Data*, vol. 1, 2021, doi: 10.17632/8kftmw7rct.1.
[47] A. Sureka and S. Agarwal, "Learning to classify hate and extremism promoting tweets," in *Proc. IEEE Joint Intell. Secur. Informat. Conf. (JISIC)*, Sep. 2014, p. 320.
[48] P. Gupta, P. Varshney, and M. P. S. Bhatia, "Identifying radical social media posts using machine learning," Tech. Rep., 2017, doi: 10.13140/RG.2.2.15311.53926.
[49] A. Kaur, J. K. Saini, and D. Bansal, "Detecting radical text over online media using deep learning," 2019, *arXiv:1907.12368*. [Online]. Available: https://arxiv.org/abs/1907.12368
[50] P. Wadhwa and M. P. S. Bhatia, "Discovering hidden networks in online social networks," *Int. J. Intell. Syst. Appl.*, vol. 6, no. 5, pp. 44–54, Apr. 2014.
[51] S. Agarwal and A. Sureka, "Characterizing linguistic attributes for automatic classification of intent based racist/radicalized posts on Tumblr micro-blogging website," 2017, *arXiv:1701.04931*. [Online]. Available: https://arxiv.org/abs/1701.04931
[52] S. A. Azizan and I. A. Aziz, "Terrorism detection based on sentiment analysis using machine learning.pdf," *J. Eng. Appl. Sci.*, vol. 12, no. 3, pp. 691–698, 2017.
[53] D. Preoţiuc-Pietro, Y. Liu, D. Hopkins, and L. Ungar, "Beyond binary labels: Political ideology prediction of Twitter users," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2017, pp. 729–740.
[54] 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.
[55] 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.
[56] M. Fraiwan, "Identification of markers and artificial intelligence-based classification of radical Twitter data," *Appl. Comput. Inform.*, vol. 16, no. 1, Apr. 2020.
[57] 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.
[58] K. Deb, S. Paul, and K. Das, "A framework for predicting and identifying radicalization and civil unrest oriented threats from WhatsApp group," in *Advances in Intelligent Systems and Computing*. Cham, Switzerland: Springer, 2020.
[59] H. Saif, Y. He, M. Fernandez, and H. Alani, "Contextual semantics for sentiment analysis of Twitter," *Inf. Process. Manage.*, vol. 52, no. 1, pp. 5–19, Jan. 2016.
[60] A. Masmoudi, M. Barhamgi, N. Faci, Z. Saoud, K. Belhajjame, D. Benslimane, and D. Camacho, "An ontology-based approach for mining radicalization indicators from online messages," in *Proc. Int. Conf. Adv. Inf. Netw. Appl. (AINA)*, May 2018, pp. 609–616.
[61] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *J. Big Data*, vol. 2, no. 1, pp. 1–21, Dec. 2015.
[62] B. K. Olorisade, P. Brereton, and P. Andras, "Reproducibility in machine learning-based studies: An example of text mining," in *Proc. ICML*, Sydney, NSW, Australia, 2017.

SAJA ALDERA (Member, IEEE) is currently pursuing the Ph.D. degree with the College of Computer and Information Sciences, King Saud University. She is currently working as a Lecturer with the Department of Management Information System, College of Business Administration, King Saud University. Her research interests include social media analysis, NLP, and deep learning.

AHMAD EMAM received the B.Sc. degree in computer science from Ain Shams University, Cairo, Egypt, the M.Sc. degree from Menoufia University, Menoufia, Egypt, and the Ph.D. degree from the Department of Computer Science and Computer Engineering, Speed Engineering School, University of Louisville, KY, USA, in summer 2001. He is currently a Professor of Information Systems with the College of Computer and Information Systems, King Saud University, where he teaches database systems, data mining, and big data analytics.

MUHAMMAD AL-QURISHI (Member, IEEE) received the Ph.D. degree from the College of Computer and Information Sciences (CCIS), King Saud University (KSU), Riyadh, Saudi Arabia, in 2017. He was a Postdoctoral Researcher with the Chair of Pervasive and Mobile Computing (CPMC), CCIS, KSU. He is one of the founding members of CPMC. He is currently a Data Scientist working with the Research and Innovation Department, ELM Company. He has published several articles in refereed journals (IEEE, ACM, Springer, and Wiley). His research interests include data science, big data analysis and mining, pervasive computing, and machine learning. He received an Innovation Award for a Mobile Cloud Serious Game from KSU 2013 and the Best Ph.D. Thesis Award from CCIS, KSU, in 2018. He got the IBM Data Science Professional Certificate and the Deep Learning Certification from deep learning.ai.

MAJED ALRUBAIAN (Member, IEEE) received the Ph.D. degree from the Department of Information Systems, College of Computer and Information Sciences (CCIS), King Saud University (KSU), Riyadh, Saudi Arabia, in 2015. He has authored several papers in the refereed IEEE/ACM/Springer journals and conferences. His research interests include social media analysis, data analytics and mining, social computing, information credibility, and cyber security. He is a Student Member of ACM.

ABDULRAHMAN ALOTHAIM received the M.Sc. degree in information systems from the University of Maryland and the Ph.D. degree in information technology from the University of Nebraska. He is currently working as an Assistant Professor of Information Systems with the College of Computer and Information Sciences, King Saud University. He is also a College Representative for the distinguished and talented students' program, attending to honors students' issues, and cultivating their talents and skills. He works as a Consultant in digital banking with Alinma Bank. His research interests include data science (deep learning and NLP) and artificial intelligence, crowdsourcing, electronic governments, digital transformation, and digital banking.