Received November 16, 2021, accepted November 28, 2021, date of publication December 3, 2021,
date of current version December 14, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3132651

Exploratory Data Analysis and Classification of a New Arabic Online Extremism Dataset

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

[1]Management Information Systems Department, College of Business Administration, King Saud University, Riyadh 11451, Saudi Arabia
[2]Information Systems Department, College of Computer and Information Sciences, King Saud University, Riyadh 11451, 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 The dissemination of extremist ideas and causes online has intensified over the last decade. Extremist organizations use social media to gain publicity and new recruits, often with little interference from network providers. New techniques are being developed to identify extremist content, ensuring it can be promptly removed and its authors blocked from network access. However, most techniques are only compatible with the English language, despite the fact that extremist propaganda is frequently shared in other languages, including Arabic. Since the most effective methods for automated linguistic analysis use deep learning and require large, high-quality datasets, creating specialised data samples containing examples of extremist communication is an essential step toward a practical solution. In this paper, we present a dataset compiled for this purpose and discuss the classification methods that can be used for extremism detection. The manually annotated Arabic Twitter dataset consists of 89,816 tweets published between 2011 and 2021. Using guidelines, three expert annotators labelled the tweets as extremist or non-extremist. Exploratory data analysis was performed to understand the dataset's features. Classification algorithms were used with the dataset, including logistic regression, support vector machine, multinominal naïve Bayes, random forest, and BERT. Among the traditional machine learning models, support vector machine with term frequency-inverse document frequency features achieved the highest accuracy (0.9729). However, BERT outperformed the traditional models with an accuracy of 0.9749. This dataset is expected to enhance the accuracy of Arabic online extremism classification in future research, and so we have made it publicly available.

INDEX TERMS Extremism, radicalization, benchmark dataset, exploratory data analysis, machine learning, bidirectional encoder representations from transformers.

I. INTRODUCTION

Extremism on social media is a growing problem [1]. Extremists use these channels to promote their ideologies and gain recruits, exerting their influence and extending their operations beyond physical space. The specific features of online networks enable extremists to use them to contact other groups or individuals anonymously [2]. Therefore, social media platforms such as Twitter often serve as an ideal place for extremist individuals, groups, or organisations to gather substantial audiences, recruit cost-effectively, and engage in extremist discourse with limited restrictions [3].

In 2021, the number of social media users reached approximately 3 billion [4]. With this large and easily accessible audience, online social networks have become a useful platform for extremist propaganda. For example, recent events surrounding the 2020 US Presidential Election and the Black Lives Matter movement reflect how the distribution of violent or inflammatory content on social networks can instigate violence in the streets [5]. This reflects the power of social media networks to influence public opinion. Therefore, guarding against the use of this power by online extremists has become

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

a key topic of interest for many governments, organisations, and social media platforms.

The Islamic State of Iraq and Syria (ISIS) emerged in 2014 [6]. At the same time, it became clear that extremists can effectively use social media networks to pursue their agendas. In response to this, companies such as Twitter and Facebook responded with various initiatives to prevent extremism on their platforms. The methods used for this task were both manual and automatic, with the latter often involving the detection of violent or extreme keywords. Recently, many researchers have sought to devise systems to automatically detect and predict extremist online content. Although progress has been made, most researchers have studied English content, and only a few have explored Arabic content [7]. Significantly, this lack of studies can be attributed to the limited availability of public Arabic datasets.

Since extremists are extensive social media users, it is worthwhile to identify and detect extremist content automatically to assist in restricting its spread. Researchers from different areas, including computer science, social science, and psychology, have collaborated to develop initiatives to counter online extremism. Most of these initiatives have aimed to detect and classify extremist content on social media [8]–[10], and substantial contributions have been made in this area. Such efforts hold considerable value for governments, counter-terrorism agencies, and social network operators because they may contribute to controlling crime, limiting the spread of extremist ideologies, and preventing terrorist recruitment.

Due to the value of automated systems for online extremism detection, the need for research in the domain of computer science has increased. Various domains have used natural language processing (NLP) and machine learning techniques to solve problems such as movie and product reviews, authorship, and sentiment analysis. These techniques have also been applied to detect and predict extremism or radicalisation in social media content (e.g., Twitter posts), usually in the context of ISIS groups.

The number of active Twitter users reached 397 million users in July 2021 [11], which makes it an important data source for researchers. Twitter is a real-time public platform used by members of various social strata, ranging from ordinary people to celebrities to international organisations. Since November 2017, Twitter users have been able to post short, 280-character text messages, or tweets, which other users can interact with. The short form of the text messages makes extremism detection challenging because it is difficult to build contextual meaning from short sentences.

The Arabic language is widely used on Twitter, with estimates from 2014 indicating that there are approximately 5 million active Arabic-speaking Twitter users [12]. At present, the Arab world includes 22 countries and millions of individuals who speak and write Arabic. However, despite the availability of online extremism detection techniques and datasets in the English language, only a limited body of literature exists on classifying Arabic content.

A basic overview of existing scientific resources revealed that relatively few datasets that have been built specifically to explore the spread of extremism are publicly available at this time. Two datasets were released in 2016 and contain 17,000 and 122,000 messages related to ISIS activities, respectively [13], [14]. Another was published by Gupta *et al.* [15] in 2017 with 48,000 messages related to several terrorist groups, including Al Qaeda and the Taliban. All of these datasets are in the English language and, as such, cannot be used as training tools for other languages, which limits their practical utility. The only existing dataset of this kind in the Arabic language was created by Fraiwan, with 24,000 messages annotated by knowledgeable experts [16].

Regarding the classification methods that can be used to differentiate between extremist content and normal messages, machine learning methodologies have proven to be very effective. Some of the most commonly used methods include random forest (RF), support vector machines (SVM), and long short-term memory (LSTM) networks. These are trained using features related to message text, profile of the sender, and the time sent. These methods have proven to be relatively successful, recognising extremist messages with up to 85% accuracy, but they still leave too large a margin of error for practical use.

The primary contributions of this study are as follows:

- A novel open-access Arabic language benchmark dataset for online extremism detection consisting of 89,000 labelled tweets has been created. Expert annotation and data validation were performed using different techniques to ensure the quality of the proposed dataset.
- Exploratory data analysis using in-depth statistical analyses was performed to understand and visualise the proposed online extremism dataset.
- Different classification models for online extremism detection are presented. To boost the accuracy of the classification models, N-gram features along with different feature sets were evaluated.

The remainder of this paper is organised as follows. Section II provides an overview of related work, including existing datasets. Section III presents the proposed methodology for the detection of extremist content. Section IV explains the implementation setup for evaluating the proposed method, and Section V offers concluding remarks and discusses pathways for future research.

II. RELATED WORK

This section provides an analysis of the existing literature on extremism detection with respect to the datasets and classifier techniques used.

A. DATASETS

Researchers have sought to detect online extremism by applying artificial intelligence techniques to social media content datasets (e.g. tweets from Twitter). However, research in this area faces a challenge in terms of the availability and quality of datasets containing extremist content. Aldera *et al.* [7]

reported that very few datasets are publicly available because of data regulations. We conducted a review of publicly available datasets and identified only four. An overview of these datasets is provided in this section.

The Kaggle data science community [13] published two English language datasets in 2016. The first dataset, 'How ISIS Uses Twitter', scraped over 17,000 tweets from more than 100 pro-ISIS Twitter users worldwide after the November 2015 Paris Attacks. The second dataset, 'Tweets Targeting ISIS', served as a counterpart to the first, containing 122,000 tweets from 95,725 distinct users; these were general tweets about ISIS and related words [14]. These two datasets have been used in many studies [1], [17], [18], [19], [20], [21], and their availability has enabled the development of extremism detection techniques.

Gupta *et al.* [15] released a data repository on GitHub [22] to assist in the identification of radical social media posts using machine learning. The authors used the Twitter Search REST API to extract public tweets that were posted between mid-February 2017 and mid-March 2017. The targeted tweets contained hashtags associated with radical groups, including #ISIS, #Taliban, #AlQaeda, #Wahhabism, and #Daesh (as seed hashtags); in turn, the frequencies of all the hashtags were calculated from the extracted tweets for use as a new search query. This process yielded approximately 48,000 unique English tweets. The content of the tweets was cleaned and pre-processed using tokenising and lemmatisation. Finally, approximately 25,000 tweets from the initial 48,000 were manually labelled as radical or not radical.

Fraiwan [23] published the first dataset of annotated ISIS radical tweets in Arabic, which consisted of 24,000 tweets from 174 accounts related to ISIS. The author developed crawler software to collect tweets from suspected ISIS accounts. The annotation process evaluated whether a given tweet was radical, religious but not radical, or unrelated to the subject matter (e.g. sports). Two experts in religion from the armed forces performed the annotation and found that the dataset contained 45% radical, 43% religious but not radical, and 11% unrelated tweets.

In their survey paper, Gaikwad *et al.* [24] identified three main challenges in online extremism datasets. The first is data imbalance, where the extremist class is smaller than the non-extremist class. The second challenge is that of data validation/verification: data availability becomes a challenge due to the suspension of extremist accounts, which makes replication of results difficult. Additionally, manual data validation suffers from bias as few experts label the data. The third challenge is that the data are collected from specific events or groups, which introduces a bias into the classification algorithms.

As our analysis indicates, there is a limited number of datasets available for online extremism detection study. Moreover, certain datasets are incomplete and suffer from bias due to unclear or low-quality annotation processes; most of the datasets are in English, and only one is in Arabic. Thus, obtaining and annotating more data, specifically in the Arabic language, is essential for continued research on online extremism.

B. CLASSIFICATION

Recently, interest in text classification models for extremism detection, particularly in the context of social media networks, has been growing. Over the last few years, experts from different disciplines (e.g. computer science, social science, and psychology) have collaborated to develop solutions by applying artificial intelligence technology to the issue of online extremism [25]. This section presents the results of a literature review on automatic detection and classification of extremism on social media networks.

Multidisciplinary research in online extremism detection has focused on the analysis of online extremism to understand the processes underlying it [18], [26] and the examination of how propaganda spreads online [27]. Researchers have also sought to devise systems to automatically detect extremist users and radical content online. In recent years, using machine learning techniques with textual features has become a popular practice.

Traditional machine learning techniques and, more recently, deep learning techniques have been used by researchers to detect extremism on social media networks. Table 1 summarises prior studies that have applied machine learning techniques to online extremism detection.

As shown in Table 1, the most commonly implemented algorithms were support vector machine (SVM), random forest (RF), and long short-term memory (LSTM). In several studies, including [28], [10], and [29], the accuracy of SVM exceeded 90%, and in [25], [15], and [30], the accuracy of the RF algorithm was higher than that of the SVM. Recently, deep learning techniques, particularly convolutional neural networks (CNN), originally developed in [31], and recurrent neural networks (RNN), proposed in the late 80s [32], [33], have yielded notable results. Researchers have also used LSTM networks to devise systems for extremist content detection on social media, as well as techniques such as SVM, RF, and maximum entropy [25]. The results in Table 1 indicate that LSTM outperformed most traditional machine learning techniques in terms of precision (85.9%).

Most prior studies addressed extremism detection using three main categories of features: textual (NLP features), time, and profile. Textual features were primarily used in the classification task, which may involve the use of techniques such as term frequency-inverse document frequency (TF-IDF), N-gram, part-of-speech, and bag-of-words. Some studies combined time, profile, and network features to classify extremist content. In addition, some authors used more advanced features, such as psychological and behavioural features. However, to the best of our knowledge, no comparative study of the literature has identified which features and classification models perform with greater accuracy than others.

TABLE 1. Popular Machine Learning Techniques Used in Online Extremism Detection.

Ref	Year	Algorithm	Feature Selection	Data Source	Performance Metric
[30]	2020	SVC, RF, MNB, LR	TF-IDF, Word2Vec	Vkontakte	SVC (Accuracy = 61%) RF (Accuracy = 83%) MNB (Accuracy = 81%) LR (Accuracy = 70%)
[34]	2020	Linear SVC	Sentiment and lexicon features	Facebook	F1-score = 0.81
[16]	2020	SVM-OAA	N/A	Twitter	F1-score = 83.2%, Accuracy = 82.6%
[35]	2020	Linear SVM	Semantic, lexicon, and emotions	Magazine, Kaggle	F1-score = 94.02%
[36]	2019	SVM	N-gram, TF-IDF	Twitter	Accuracy = 0.84
[37]	2019	CNN + LSTM	Embedding layer	Twitter, Dark Web forums	Accuracy = 92.66
[38]	2019	RF	Textual, psychological, and behavioural features	Twitter	F1-score=1.0
[25]	2019	LSTM SVM RF MaxEnt	Word2Vec	News, Articles, Blogs	LSTM (F1-score = 0.65) SVM (F1-score = 0.45) RF (F1-score = 0.65) MaxEnt (F1-score = 0.68)
[39]	2019	Char-LSTM SVM, LabelSpreading (RBF)	Twitter handle-related, profile-related, and content-related features	Twitter	Char-LSTM (F1-score= 0.76), SVM (F1-score= 0.65), LabelSpreading - RBF (F1-score= 0.76)
[1]	2017	LR	Topic sensitivity, post-effectiveness indicators, and emotion indicators	Twitter	Accuracy = 80%, F1-score = 0.95
[2]	2017	RF	Topic modelling, tone analysis, and semantic features	Tumblr	Precision = 0.81, Recall = 0.84
[40]	2017	GR-Learnt	Unigram, Word2Vec, sentiment-based features, emotions, and political terms	Forms	Accuracy = 0.27
[15]	2017	SVM RF	Stylometric and time-based features	Twitter	SVM (Accuracy = 98.03) RF (Accuracy = 98.43)
[8]	2016	RF	Time-based, profile, and network features	Twitter	ROC-AUC = 0.93
[28]	2015	SVM	TF-IDF, Glasgow, entropy	DWFP, OSAC	Accuracy = 93.6%
[27]	2015	AdaBoost	Data-independent and data-dependent features	Twitter	Arabic Tweets (Accuracy = 0.86), English Tweets (Accuracy = 0.99)
[10]	2015	SVM, AdaBoost	Stylometric, time-based, and sentiment-based	Twitter	SVM (Accuracy = 0.97), AdaBoost (Accuracy = 0.100)
[29]	2015	SVM	Religious, war-related, offensive words, negative emotions, and Internet slang	Twitter	F1-score = 0.83, Accuracy = 0.97

AdaBoost: adaptive boosting; CNN: convolutional neural network; LR: logistic regression, RF: random forest; LSTM: long short-term memory; MaxEnt: maximum entropy; MNB: multinomial naïve Bayes; RBF: radial basis function; SVC: support vector clustering; SVM: support vector machine; TF-IDF: term frequency-inverse document frequency.

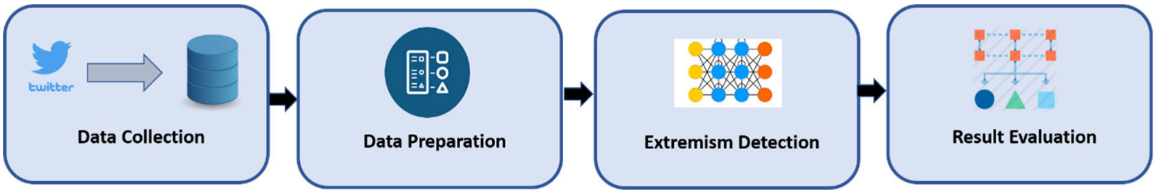

FIGURE 1. Proposed architecture.

III. METHODS

This section outlines the architecture for the proposed extremism detection module, which has a four-part architecture, as shown in Figure 1.

In the first part, data are obtained from Twitter using the Twitter API. Following this, standard NLP pre-processing techniques (e.g., tokenisation and lemmatisation) are applied to generate a dataset. Thereafter, the tweets are manually labelled as either extremist or non-extremist.

In this study, we performed exploratory data analysis (EDA) to understand the dataset. Afterwards, we evaluated the dataset using various traditional machine learning models with different NLP features. We also evaluated the dataset using the bidirectional encoder representations from transformers (BERT) deep learning model. Finally, we evaluated the model's performance using metrics such as accuracy, F1-score, and area under the receiver operating characteristic curve (AUC).

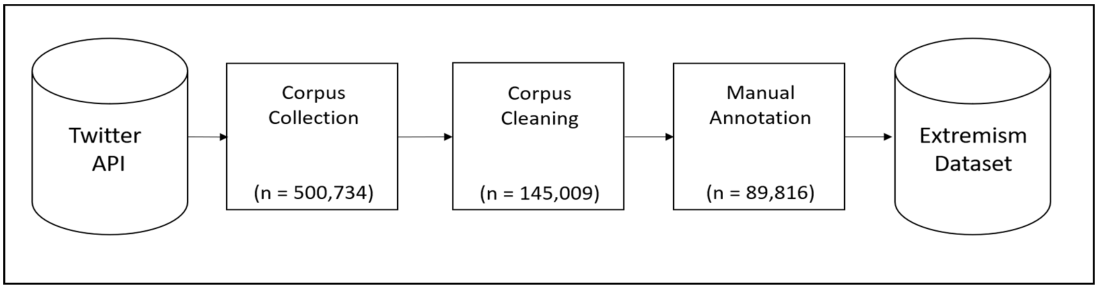

FIGURE 2. Corpus collection methodology.

The following subsections outline the key stages of our module, including data collection and preparation. In turn, feature extraction, EDA, and predictive model development are discussed.

A. DATA COLLECTION AND PREPARATION

Figure 2 demonstrates the methodology adopted to collect and construct the corpus. The methodology comprises three primary phases: corpus collection, corpus cleaning, and data annotation, which are described in detail in the following subsections.

AdaBoost: adaptive boosting; CNN: convolutional neural network; LR: logistic regression, RF: random forest; LSTM: long short-term memory; MaxEnt: maximum entropy; MNB: multinomial naïve Bayes; RBF: radial basis function; SVC: support vector clustering; SVM: support vector machine; TF-IDF: term frequency-inverse document frequency.

1) CORPUS COLLECTION

By the third quarter of 2020, the average number of active daily Twitter users reached 187 million, and based on an average posting frequency of 500,000 tweets every day, approximately 200 billion tweets are published on the platform per year, each with a 280-character length [41]. Moreover, Twitter allows researchers to access its public data for research purposes via the Twitter API, with some limitations.

As mentioned in Section II, there is a limited number of datasets, particularly Arabic language datasets, available on the Internet for online extremism detection. Therefore, we collected new data from Twitter. The Twitter API allows developers to collect real-time tweets with different parameters based on given query terms. Our final query was as follows:

$$\text{Data } [] = \text{Search } \{\text{Search\_Term, longitude, latitude, lang}\}$$

The API returned the text of tweets, along with user information (e.g. username, user location, number of friends, number of followers, number of likes, and user description). Moreover, a list of search terms was prepared for data collection based on trending Twitter topics in the Arab world.

2) CORPUS CLEANING

After corpus collection and before its annotation, a cleaning step is required to prepare the tweets for processing. We removed duplicates and empty tweets from the corpus and excluded non-Arabic tweets. At this stage, the number of tweets was reduced to approximately 145,000.

3) MANUAL ANNOTATION

The annotation process is critical because it directly influences model accuracy. Furthermore, annotating large datasets can be costly and time-consuming. Given the costly nature of the annotation process, most researchers use automatic annotation techniques that are based on dictionary resources, including *WordNet* and *SentiWordNet*. However, a limitation of these techniques is their potentially low accuracy. Owing to this and other limitations of automated annotation, we opted for manual annotation; specifically, we annotated our dataset by checking each tweet and considering the occurrence of each word and the meaning and context of the tweet. Although it is lengthy, manual annotation is accurate and reliable compared to automated annotation.

In this research, three different raters manually labelled the collected tweets. Precautions were taken to minimise bias; these included establishing clear guidelines and validating the results using various techniques (as discussed in Section IV).

4) FEATURE EXTRACTION

Before classification, our proposed system processes tweets in the form of vectors, enabling the classification models to perform statistical operations. Initially, an NLP pre-processing technique is followed (e.g. stop word removal, lowercasing, tokenisation, and lemmatisation to obtain unigrams). To create feature vectors, different feature extraction techniques were used in this study:

- **N-grams:** These are the basic features in detection problems. A sequence of n words can be referred to as a unigram (one word), bigram (two words), trigram (three words), and so on, depending on the value of n.
- **TF-IDF:** This is a statistical measure obtained by first counting the occurrences of a word in a document and then calculating the inverse of the number of documents in which the word appears. It is possible to combine

N-gram features with TF-IDF features to increase model accuracy.

- **Word2Vec:** This involves the use of neural networks to learn word vectors. Words are represented by distributed vector representations to preserve the relationships between words.

B. EXPLORATORY DATA ANALYSIS

EDA is a technique used to explore datasets so as to extract useful and actionable information, identify relationships among the explanatory variables, detect mistakes, and preliminarily select appropriate models. It uses descriptive statistics and graphical tools to develop an understanding of the data [42]. EDA is used primarily to maximise insight into a dataset, detect outliers and anomalies, and test underlying assumptions [43]. In this study's EDA, we used graphical methods to summarise the data visually and diagrammatically. We applied a univariate graphical method that examines one variable at a time (e.g., using histograms, boxplots, or pie charts) for categorical data, whereas a multivariate graphical method was applied to consider two or more variables at a time and explore relationships. In the latter case, correlation analysis was used, which is a technique that can calculate the overall correlation for two or more numerical variables. In the section IV, we discuss how categorical and numerical features were explored and visualised using several Python libraries, including the Plotly Python graphing library, NumPy, and pandas. Using EDA, we acquired detailed insights into the dataset and meaningful information about the dataset's characteristics.

C. CLASSIFICATION MODELS

There are many available classification models, and their effectiveness depends on the problem domain. Choosing the right model is critical for building a robust detection system. For our problem, the following algorithms were evaluated.

- **Logistic regression (LR):** A linear model that estimates the probability that a variable belongs to a class. LR is usually used in binary classification problems. Due to the strength of LR in binary classification, it is used in many fields, including spam, cyberbullying, and fraud detection.
- **SVM:** A supervised learning model that underlines two different classes in a high dimensional space. The advantages of SVM include high speed, scalability, and the capacity to detect intrusions in real time and dynamically update training patterns. Variations of SVM, known as kernels, which include linear and polynomial support vector clustering and the kernel trick, are available for classification of high dimensional data.
- **Multinomial naïve Bayes (MNB):** A probabilistic learning algorithm typically used with NLP problems and textual data classification. It is based on Bayes' theorem, which calculates the probability of an event occurring based on prior knowledge of conditions related to it. The algorithm is simple, easy to implement, and can easily handle large datasets.
- **RF:** An ensemble learning classifier developed from several decision trees. Each decision tree is built using a random subset of features to create a 'forest' of trees, which reduces overfitting; the classifier achieves high prediction accuracy.
- **BERT:** A transformer-based machine learning model developed by Google in 2018 for NLP tasks. BERT is the first NLP approach that applied self-attention, as facilitated by the bidirectional transformers that lie at the core of its design. BERT was trained on over 3 billion tokens. Two different types of BERT were introduced by [44], namely, BERT base, which is 12 layers deep, and BERT large, which is 24 layers deep. One of the key features of BERT is that it randomly masks the words in a sentence and then predicts them, which is dissimilar to previous language models (i.e., those that predict the next word in a sequence such as, Word2Vec). BERT has achieved state-of-the-art performance on most NLP tasks. Several BERT multilingual models have been released to support different languages, as found on the Hugging Face website [45].

D. PERFORMANCE EVALUATION

Different performance metrics were used to evaluate the classifier performance. Accuracy is the simplest and most widely used metric to evaluate a classifier.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \tag{1}$$

where TP is the number of correctly classified extremism tweets, TN is the number of correctly classified non-extremism tweets, FN is the number of incorrectly classified extremism tweets, and FP is the number of incorrectly classified non-extremism tweets.

Precision is defined as the ratio of true positives to all tweets identified as positive.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

Recall is defined as the ratio of correctly classified positives to total positives. In our study, recall is a measure of the proportion of the detected extremism tweets.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

The trade-off between recall (false negatives) and precision (false positives) is captured by the F1-measure. The F1-score is defined as the weighted average of recall and precision; it balances precision and recall in a single value and therefore is commonly used as a classification evaluation metric.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{4}$$

The receiver operating characteristic curve is obtained by plotting the true positive rate relative to the false positive rate,

TABLE 2. Annotation Guidelines.

After reading the tweet, please classify it as 'extreme' or 'not-extreme' using the following table focusing on the context of its text:	
Extremist (T)	**Classify a tweet as extremist if it has one of the following traits:** 1. Concept of burglary 2. Intellectual unilateralism 3. The 'grievance' fallacy 4. The Islamophobia fallacy 5. Dismantling and building chaos **The most important 'ideas' and 'sayings' of extremist discourse are:** 1. Atonement for others who are different 2. Advocating and inciting violence 3. Reliance on references to extremist groups 4. The logic of fatal divisions 5. The idea of a single model
Non-extremist (F)	Classify a tweet as non-extremist if it does not contain the features mentioned above even if it contains religious, political, ethnic, or social themes or if its topic is unrelated to extremism (e.g. sports or fashion).

and the AUC measure, which is bounded by zero and one, typically exceeds 0.5.

IV. EXPERIMENTS

The framework was implemented in three parts: first, collection and annotation of data; second, EDA; and third, development of classification models and evaluation on the newly proposed dataset.

A. DATA COLLECTION

After registering as a developer on the Twitter developer platform, we obtained user status data with more than 40 attributes. We used the Twitter Streaming API and Search API to collect data relating to tweets in real time, filtering with specific settings (e.g., based on keywords, location, and Arabic language). The crawled data were published between May 2011 and March 2021, and a final set of 2 million tweets with their associated metadata was extracted. The dataset size was reduced to 500,000 after excluding retweets (i.e., considering original tweets only). Arabic search terms were used in the query, focusing on religious and political terms. A list of specific query search terms, including جمعة الغضب, مقاطعة المنتجات, رعاة الإرهاب, هيئةعدو الله, تنظيم داعش, كبار العلماء, داعشداعشالعراق, كفار, was used to identify and collect tweets from public Twitter profiles.

1) DATA CLEANING

Before labelling the dataset, the data were cleaned by removing empty tweets and tweets containing fewer than seven words (excluding hashtags and user mentions). Then, duplicate tweets were removed, using the dedupe Pythonlibrary [46]. The library applies machine learning for rapid de-duplication and entity resolution on structured data, using human-annotated training data. After data cleaning, the dataset consisted of 145,000 tweets.

2) DATA ANNOTATION

As discussed in Section III, manual annotation is more reliable and accurate than automatic annotation. Therefore, the data annotation technique followed in this study was that of Wosom [47], which is specialised to handle data annotation and supports annotation of text, audio, image, and video among others.

It is difficult to identify extremism based on an individual's judgment about whether a given text (e.g. a tweet) is extremist or not. For this reason, a predefined process is required to identify extremism and to avoid personal judgment and bias of the annotators. In this research, three expert annotators reviewed the tweets individually and labelled them as extremist or non-extremist. The guidelines listed in Table 2 were used for the annotation process.

The final label for each tweet was decided by a majority vote. An odd number of annotators were employed to prevent a tie. At the end of this stage, 89,816 tweets were labelled using the majority vote method. Moreover, during the annotation process, a sample of 1000 tweets was used to validate the annotation process. A fourth annotator was invited to check the annotators' work, resulting in more than 80% agreement between the average of the three raters and the validating fourth rater. This procedure was undertaken ten times during the annotation process. Our dataset has been made publicly available on IEEE DataPort [48].

3) ANNOTATION EVALUATION

The inter-annotator agreement was calculated using Gwet's AC1 measure, which is one of the main statistical measures of agreement and can be used when the outcome is ordinal or nominal in nature [49]. In this study, nominal weights were used to take into account the nominal nature of the data [49].

The initial study sample included 89,816 observations. The overall value of Gwet's AC1 was 0.6, which indicates substantial agreement between raters.

4) POTENTIAL BIAS

Most datasets have a risk level in terms of demographic bias [50], and the risk increases when using manual search terms. Researchers must be aware of potential biases in their datasets and address them. Gender and ethnicity are common sources of bias that are often identified in datasets. Therefore, to explore the potential biases in our dataset, we initially checked whether gender bias is a feature of the dataset. Since Twitter does not record user gender information, we used the Gender API [51] to detect user gender from the first name. We were able to process the gender of 52,929 unique users and inferred that 66.5% of the users were identified as male, 25.6% as female, and 7.9% as unknown (cannot be identified). This suggests a male bias in terms of the users included in the dataset. However, this approach is limited because names are not a reliable and truthful way to determine gender identities.

Regarding ethnicity bias, we generalised the search keywords during the dataset collection phase in order to mitigate potential ethnicity bias. However, it should be noticed that the popularity of certain keywords may still give rise to an unintended bias. For example, most of the tweets were posted from Saudi Arabia, Egypt, and Yemen due to the use of specific keywords such as 'جمعة الـغضب', 'هيئة كبار العلماء'.

5) FEATURE EXTRACTION

An NLP pre-processing was performed, which involved lemmatisation, stop word removal, and tokenisation. To create word vectors, various feature extraction techniques were used, including unigrams, bigrams, and trigrams with TF-IDF and Word2Vec.

B. EXPLORATORY DATA ANALYSIS

After a preliminary examination of the dataset, we identified a series of steps to perform at the outset:

- Capitalise all column names.
- Split user dictionary column into spread columns using Python pandas and Contractions libraries.
- Cast columns to their appropriate data types.
- Map target classes to extremist and non-extremist.
- Create a new feature for the tweet text length and word count.
- Split the date column into multi-features (e.g. hour, day, month) by processing the associated timestamp for each tweet record.
- Drop irrelevant columns: Conversationid, Outlinks, Tcooutlinks, etc.
- Check for duplicates and remove them.
- Check for missing values.

1) METADATA ANALYSIS

We performed further analysis of the tweet metadata to identify important features. In the extremism dataset, there were 89,816 tweets in total published by 52,929 unique users. Figure 3 shows the percentage of extremist tweets and non-extremist tweets identified in the dataset. A total of 50,279 tweets (56%) from 22,858 unique users were labelled as extremist, whereas 39,537 tweets (44%) from 30,911 unique users were labelled as non-extremist. We applied Shannon's entropy measure to check the dataset's balance, deriving a result of 0.98, which indicates that the dataset is well balanced.

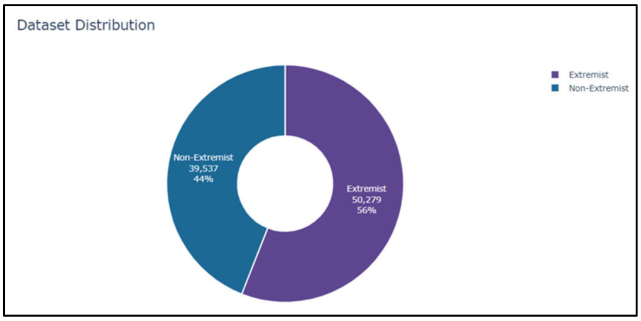

FIGURE 3. Dataset class distribution.

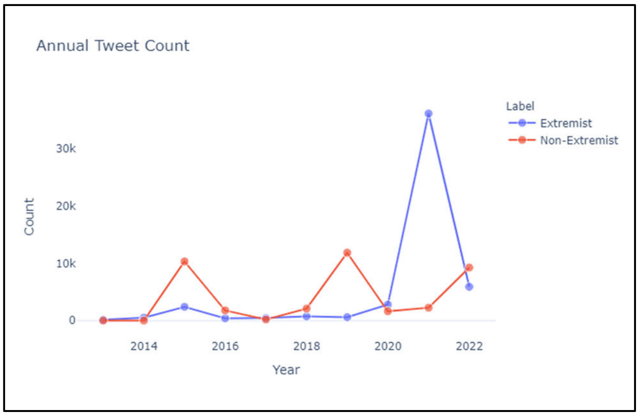

FIGURE 4. Annual tweet count data.

Figure 4 illustrates the number of tweets published annually. We observed increased content publication in certain months, for example in October 2018. In this month, the publication of extremist content increased by 100%, which coincided with the death of Saudi journalist Jamal Khashoggi; this evidently sparked significant social media activity. As another example, the April 2021 publication of the American CIA report on the Khashoggi case also appears to have caused an increase in Twitter social media content engagement.

Figures 5 and 6 show the day and time of extremist and non-extremist tweet publication, respectively. The time for extremist tweets was typically later in the day, which is likely because most people are busy or working during the daytime. By contrast, the days of publication are almost uniformly distributed over the week with a slight increase in non-extremist tweets during the weekend. Based on these observations, it was concluded that these features are not associated with the variable of extremism, and thus they were excluded from the classification model.

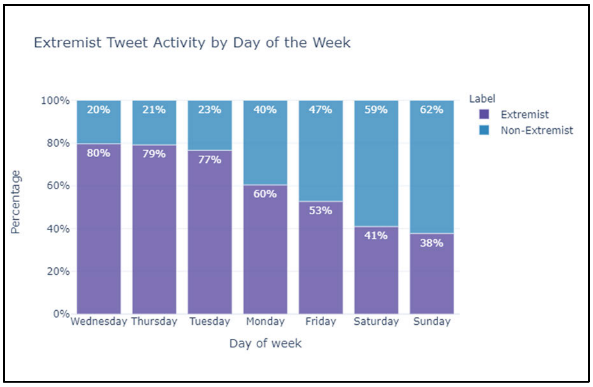

FIGURE 5. Extremist activity by day of the week.

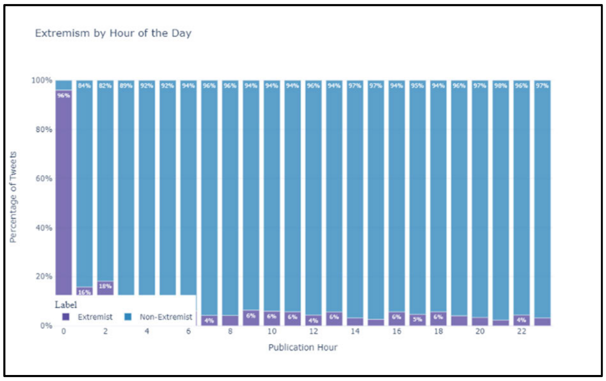

FIGURE 6. Extremist activity by hour of the day.

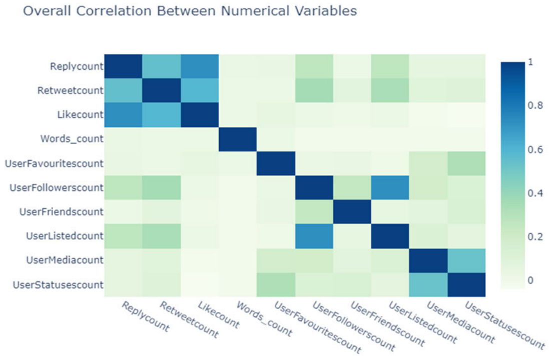

FIGURE 7. Correlations between numerical variables.

We also examined the correlations between numerical variables to check for any relationships among them. Figure 7 shows a strong correlation between reply count, retweet count, and like count. Moreover, a relationship was identified between a user's follower count and listed count. Finally, a correlation existed between user media count and user status count. Therefore, we can use these correlations to build new features, such as favourite/follower distribution and retweet/like distribution.

Figures 8 and 9 show that the favourite count was higher in extremist tweets compared with non-extremist tweets

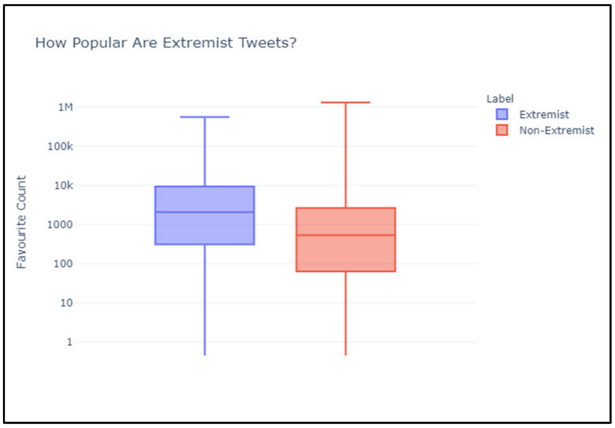

FIGURE 8. Favourite count distribution.

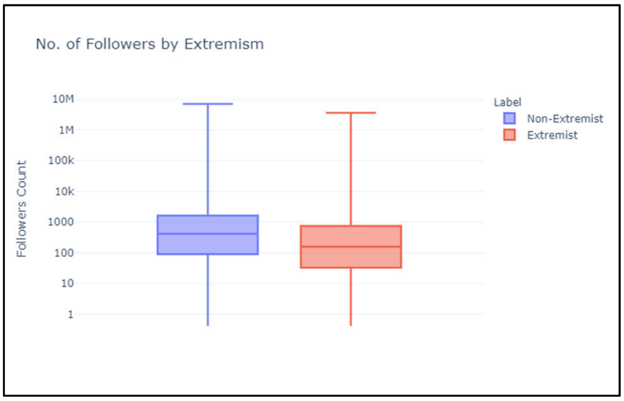

FIGURE 9. Follower count distribution.

(non-extremist median=2084, extremist median=540). However, the follower count of extremist tweets slightly exceeded that of non-extremist tweets (non-extremist median=419, extremist median=157). This may be attributable to the fact that users who have refrained from following an account may agree with the idea and mark the tweet as a favourite.

Figures 10 and 11 show that the lengths and word count of the tweets were similar across extremist and non-extremist tweets, respectively. However, as mentioned earlier, tweets with fewer than seven words were excluded from the dataset.

2) NLP ANALYSIS

To analyse the tweets in our Arabic language online extremism dataset in depth, we investigated the top 10 unigrams and bigrams based on TF-IDF for extremist and non-extremist tweets after removing Arabic stop words. In TF-IDF, words are assigned numerical weights that represent their relative importance in a particular document within a set of documents (i.e., a corpus). Figures 12 and 13 show the top-ranked unigrams based on TF-IDF for extremist and non-extremist tweets. Tables 3 and 4 provide the English translations for the words shown in Figures 12 and 13, respectively. As the

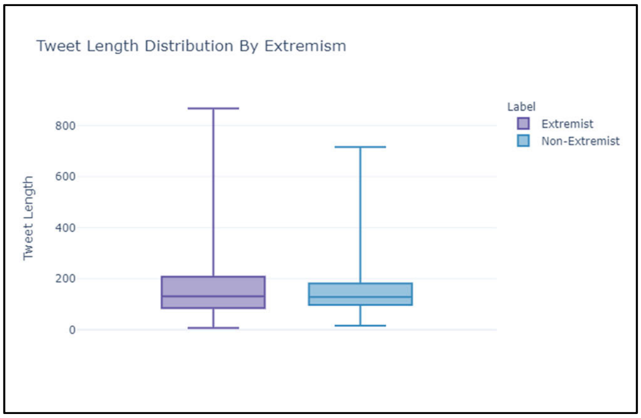

FIGURE 10. Tweet length distribution.

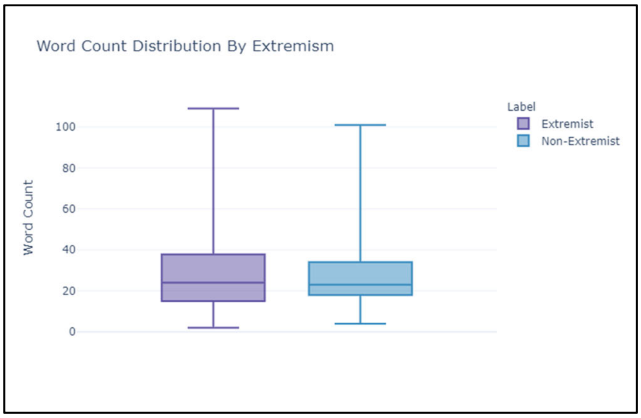

FIGURE 11. Word count distribution.

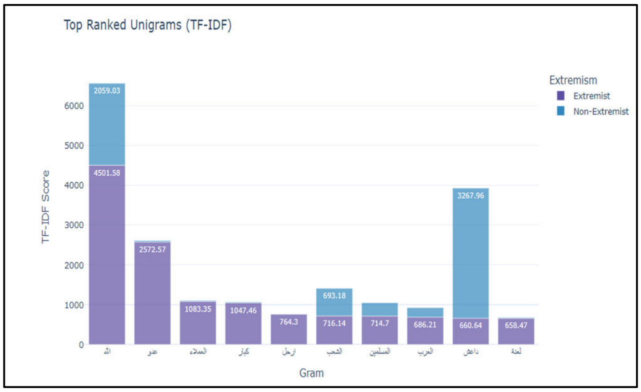

FIGURE 12. Top 10 extremist unigrams based on TF-IDF.

TABLE 3. English translations of top 10 extremist unigrams based on TF-IDF.

Arabic Word	Translation/Description
الله	Allah/God
عدو	Enemy
العملاء	Those who cooperate with the US or Israeli alliance
كبار	Senior
ارحل	Get out
الشعب	The public
المسلمين	Muslims
الـعـرب	Arabs
د ا عش	Daesh (Arabic acronym for ISIS). The group itself does not use that name; Daesh is used by many Muslims, who believe it distinguishes the group from their faith
لـعـنـة	Curse

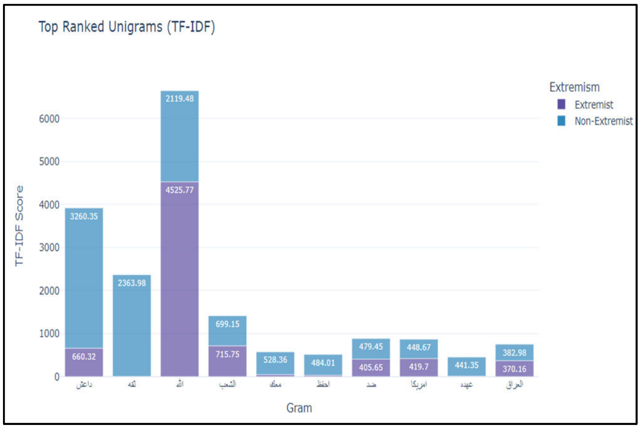

FIGURE 13. Top 10 non-extremist unigrams based on TF-IDF.

TABLE 4. English translations of top 10 non-extremist unigrams based on TF-IDF.

Arabic Word	Translation/Description
داعش	Daesh (Arabic acronym for ISIS). The group itself does not use that name; Daesh is used by many Muslims, who believe it distinguishes the group from their faith
ثقة	Trust
الله	Allah/God
الـشـعـب	The public
مـعـك	With you
احـفـظ	Save
ضـد	Against
امـريـكـا	America
عـهـده	Under
العراق	Iraq

figures indicate, the word ‘Allah’ was used in both types of tweets but more frequently in extremist tweets. Furthermore, the most frequent terms in extremist tweets tended to be more violent and dominant compared to those of non-extremist tweets.

Figures 14 and 15 show the bigrams based on TF-IDF values sorted for extremist and non-extremist tweets, respectively. Tables 5 and 6 provide English translations for the words shown in the respective figures.

C. CLASSIFICATION MODELS

Different classification algorithms were used to classify tweets as extremist or non-extremist. The dataset consisted of TF-IDF features computed from the unigram, bigram, and trigram terms, where all the tweets had pre-assigned class labels (i.e. extremist or non-extremist). Five classifiers were considered: LR, MNB, SVM, RF, and BERT. Each classification model was trained in a supervised environment using binary labelled tweets.

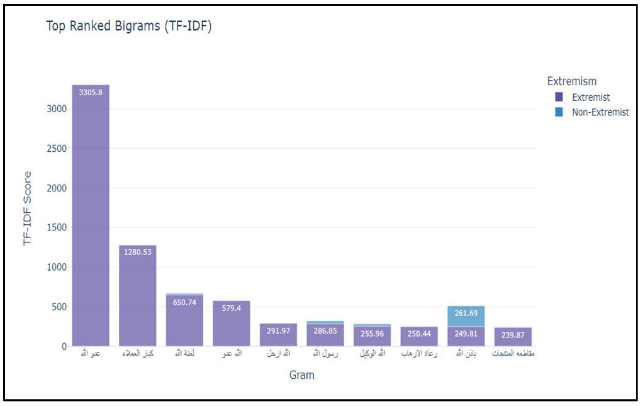

FIGURE 14. Top 10 non-extremist bigrams based on TF-IDF.

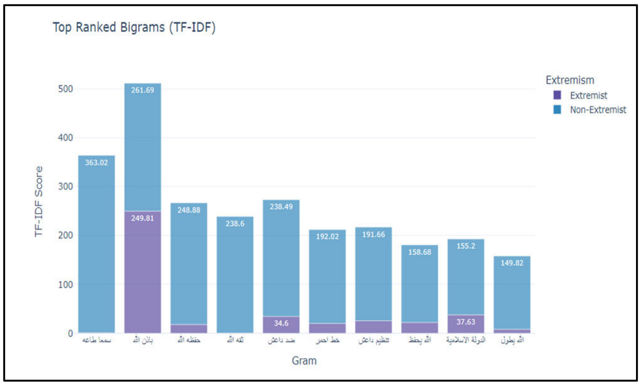

FIGURE 15. Top 10 non-extremist bigrams based on TF-IDF.

TABLE 5. English translations of top 10 non-extremist Bigrams based on TF-IDF.

Arabic Word	Translation/Description
عدو الله	Enemy of God
كبار العملاء	Senior for those who cooperate with the US or Israeli alliance
لعنة الله	Cursed by God
عدو الله	God's enemy
الله ارحل	God leave
رسول الله	Messenger of God
الله وكيل	God's agent
رعاة الإرهاب	Sponsors of terrorism
باذن الله	God willing
مقاطعة المنتجات	Boycott products

Supervised classification was performed by splitting the dataset into training, validation, and test sets because the dataset is sufficiently large to have an independent test set to verify the performance of the models; sufficient data also remained available for training and validation. We tuned the hyperparameters on the validation set, and after achieving optimal results, evaluated the real performance on the test set. We used four machine learning models, namely LR, SVM, MNB, and RF, and all experiments were performed using a combination of feature sets extracted from our Twitter dataset. This consisted of a TF-IDF feature set with N-gram and Word2Vec embeddings to determine the most accurate and effective model. Moreover, we used the BERT model, which is a popular language model, that can be fine-tuned for a specific NLP classification task; we included different Arabic BERT models in our research. Finally, we trained and tested 17 models with our Twitter extremism dataset.

TABLE 6. English translations of top 10 non-extremist bigrams based on TF-IDF.

Arabic Word	Translation/Description
سمعا وطاعه	Hearing and obeying
باذن الله	God willing
حفظه الله	God save him
تقة الله	God's trust
ضد داعش	Against ISIS
خط احمر	Red line
تنظيم داعش	ISIS group
الله يحفظه	God bless him
الدولة الاسلامية	Islamic country
الله يطول	God give you long life

D. RESULTS

This section presents the performance evaluation results of our models, in terms of correct classification of extremist and non-extremist tweets. Table 7 lists the obtained values for accuracy, F1-score, and AUC. As our dataset was balanced, accuracy, F1-score, and AUC measures are considered optimal for evaluating our models.

For the detection of extremist tweets using traditional machine learning algorithms, Table 7 shows that SVM with TF-IDF features achieved the highest F1-score, accuracy, and AUC score with values of 0.9730, 0.9729, and 0.9909, respectively. Notably, the BERT model outperformed the traditional machine learning models and achieved values of 0.9749, 0.9749, and 0.9948 for the F1-score, accuracy, and AUC score, respectively.

We also used the multi-dialect Arabic BERT model downloaded from the Hugging Face website [52], which yielded better results than AraBERT [53], because the former has been trained on 10 million Arabic tweets covering all 21 Arab countries compared by, AraBERT which pre-trained on 200 million modern standard Arabic sentences gathered from different sources. To fine-tune BERT, we used ktrain, a lightweight wrapper for the TensorFlow deep learning library, to help build, train, and deploy neural networks and other machine learning models. The final hyperparameters for BERT included 24 hidden layers, a batch size of 16, 3 epochs, a learning rate of 2e-5, the Adam BERT optimiser, and the pre-trained BERT model 'bashar-talafha/multi-dialect-bert-base-arabic'. Additionally, to reduce training time, we used a GPU from Google Colab for model training.

Table 3 confirms that the inclusion of bigrams and trigrams does not improve the performance of the models. Furthermore, the performance of certain models degraded when trigram features were included. The only model that benefitted from the inclusion of bigrams and trigrams was the MNB. Additionally, the inclusion of Word2Vec resulted in improvements for RF only.

TABLE 7. Performance Evaluation.

No.	Algorithm	Features	Metrics		
			F1-Score	Accuracy	AUC
1	LR	TF-IDF	0.9724	0.9723	0.9919
		TF-IDF + Bigrams	0.9715	0.9714	0.9907
		TF-IDF + Tri-grams	0.9705	0.9705	0.9896
		Word2Vec	0.9650	0.9645	0.9920
2	MNB	TF-IDF	0.9032	0.9046	0.9867
		TF-IDF + Bigrams	0.9037	0.9052	0.9890
		TF-IDF + Trigrams	0.9011	0.9026	0.9890
		Word2Vec	0.8647	0.8645	0.9614
3	SVM	TF-IDF	**0.9730**	**0.9729**	**0.9909**
		TF-IDF + Bigrams	0.9726	0.9725	0.9900
		TF-IDF + Tri-grams	0.9721	0.9720	0.9891
		Word2Vec	0.9664	0.9655	0.9888
4	RF	TF-IDF	0.9653	0.9671	0.9902
		TF-IDF + Bigrams	0.9620	0.9596	0.9899
		TF-IDF + Tri-grams	0.9619	0.9567	0.9876
		Word2Vec	0.9664	0.9655	0.9888
5	BERT	BERT Embedding	**0.9749**	**0.9749**	**0.9948**

V. CONCLUSION AND FUTURE WORK

The easy accessibility and widespread nature of social media networks provide extremist individuals, groups, and organisations with an easy means to attract large audiences, disseminate propaganda, and recruit members. In this study, our objective was to compile a dataset of Arabic language tweets obtained from Twitter and automatically detect extremist content using machine learning algorithms. Many online extremism detection systems have been proposed in the literature, often achieving accuracies of around 90%. However, there is an insufficient number of publicly available extremism datasets, particularly in the Arabic language; most of the available Arabic language datasets are limited to ISIS and do not extend to political and other types of extremism.

In this research, we present an Arabic Twitter dataset for online extremism detection consisting of 89,816 tweets and associated metadata. The dataset was manually annotated by three experts and achieved a Gwet's AC1 score of 0.6, indicating substantial inter-annotator agreement. A two-step analysis was performed: first, EDA to understand the dataset and provide insights into the features; and second, the classification modelling process, wherein 17 different classification models were used. Among the traditional machine learning models, SVM achieved the best accuracy (0.9729) using TF-IDF features extracted from the tweet content. Notably, the BERT deep learning model outperformed SVM, achieving an accuracy of 0.9749.

This research applies some of the achievements in NLP to a pivotal social issue and seeks to create tools that other researchers and stakeholders can use to prevent the proliferation of extremist ideas. The compilation of our Arabic language dataset, consisting of annotated Tweets that may or may not include terrorist rhetoric, is hugely important and may have a wide-ranging impact on this field. Data collection and pre-processing were conducted systematically, while the annotation process was performed manually and with a full understanding of the context. This approach resulted in a very reliable dataset that is expected to provide a solid foundation for future works, specifically those aiming to detect extremist content quickly and accurately on social media.

In this research, we also compared several simple machine learning classification algorithms, along with a more complex deep learning approach based on the BERT model, obtaining results that illustrate the advantages of the latter. Classification methods were evaluated experimentally using metrics such as accuracy and F1-score to assess their effectiveness.

This study is notable for being one of the first to address the issue of extremist communication in the Arabic language, which is widely used on the Internet and constitutes one of the major global languages. However, this study has a limited scope, and it fails to account for regional variations of Arabic or consider that various groups could be using different vocabularies or code words. This is why it is necessary to continue building similar datasets and make them publicly available to other researchers, thereby strengthening the digital defences against hateful ideologies and dangerous individuals and organisations.

It is worth highlighting that monitoring extremist tweets and users can help establish early warning systems and create opportunities for predictive and preventative actions against extremism. In future works, different features and different combinations of features will be studied to improve the performance of our model. We intend for this study to serve as a useful basis for future research, particularly for building accurate models for online extremism detection in the Arabic language.

DATA AVAILABILITY STATEMENT

The dataset are made available on doi: https://dx.doi.org/10.21227/g9c0-1t21. (Accessed on November 16, 2021).

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

REFERENCES

[1] S. D. 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)*, Boston, MA, USA, Dec. 2017, pp. 3638–3647, doi: 10.1109/BigData.2017.8258358.

[2] S. Agarwal and A. Sureka, "Characterizing linguistic attributes for automatic classification of intent based racist/radicalized posts on Tumblr micro-blogging website," Jan. 2017, *arXiv:1701.04931*.

[3] 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. IEEE 32nd Int. Conf. Adv. Inf. Netw. Appl. (AINA)*, Kraców, Poland, May 2018, pp. 609–616, doi: 10.1109/AINA.2018.00094.

[4] B. J. Carter, B. L. Bullock, and B. D. Chaffey. *Global Social Media statistics Research Summary 2021 | Smart Insights*. Accessed: Sep. 5, 2021. [Online]. Available: https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/

[5] M. Mundt, K. Ross, and C. M. Burnett, ''Scaling social movements through social media: The case of black lives matter,'' *Social Media*, vol. 4, no. 4, Oct. 2018, Art. no. 205630511880791.

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

[8] E. Ferrara, W.-Q. Wang, O. Varol, A. Flammini, and A. Galstyan, ''Predicting online extremism, content adopters, and interaction reciprocity,'' in *Social Informatics* (Lecture Notes in Computer Science). Cham, Switzerland: Springer, 2016, pp. 22–39, doi: 10.1007/978-3-319-47874-6_3.

[9] M. Hashemi and M. Hall, ''Visualization, feature selection, machine learning: Identifying the responsible group for extreme acts of violence,'' *IEEE Access*, vol. 6, pp. 70164–70171, 2018, doi: 10.1109/access.2018.2879056.

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

[11] Statista. (2021). *Most Used Social Media 2021*. Accessed: Nov. 10, 2021. [Online]. Available: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

[12] Dubai School of Government Governance and Innovation Program. (2011). *Twitter in the Arab Region*. [Online]. Available: https://www.arabsocialmediareport.com/Twitter/LineChart.aspx

[13] Fifth Tribe. (Jan. 2014). *How ISIS Uses Twitter: Analyze How ISIS Fanboys Have Been Using Twitter Since 2015 Paris Attacks*. [Online]. Available: https://www.kaggle.com/fifthtribe/how-isis-uses-twitter

[14] ActiveGalaXy. (Jul. 26). *Tweets Targeting ISIS: General Tweets About ISIS & Related Words*. [Dataset] Kaggle v24. Accessed: Nov. 2, 2021. [Online]. Available: https://www.kaggle.com/activegalaxy/isis-related-tweets/activity

[15] P. Gupta, P. Varshney, and M. Bhatia, ''Identifying radical social media posts using machine learning,'' Tech. Rep., 2017, doi: 10.13140/RG.2.2.15311.53926.

[16] M. Fraiwan, ''Identification of markers and artificial intelligence-based classification of radical Twitter data,'' *Appl. Comput. Inform.*, vol. 16, no. 1, Apr. 2020. [Online]. Available: https://app.dimensions.ai/details/publication/pub.1126497488.

[17] 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, p. 151, Nov. 2019, doi: 10.1145/3359253.

[18] R. Lara-Cabrera, A. G. Pardo, K. Benouaret, N. Faci, D. Benslimane, and D. Camacho, ''Measuring the radicalisation risk in social networks,'' *IEEE Access*, vol. 5, pp. 10892–10900, 2017, doi: 10.1109/access.2017.2706018.

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

[20] R. Lara-Cabrera, A. Gonzalez-Pardo, and D. Camacho, ''Statistical analysis of risk assessment factors and metrics to evaluate radicalisation in Twitter,'' *Future Gener. Comput. Syst.*, vol. 93, pp. 971–978, Apr. 2019, doi: 10.1016/j.future.2017.10.046.

[21] M. Fernandez and H. Alani, ''Contextual semantics for radicalisation detection on Twitter,'' in *Proc. Workshop Semantic Web Soc. Good*, vol. 2182, Oct. 2019, pp. 1–14. [Online]. Available: http://ceur-ws.org/Vol-2182/paper_4.pdf

[22] P. Gupta, P. Varshney, and M. Bhatia. (Jun. 2017). *Dataset Repository for Identifying Radical Social Media Posts Using Machine Learning*. [Online]. Available: https://git.io/vHTUP

[23] M. Fraiwan. (Aug. 2021). *Annotated ISIS Radical Tweets*. [Dataset] Mendeley v1. [Online]. Available: https://data.mendeley.com/v1/datasets/8kftmw7rct/draft?a=0fd4d8fc-42e8-478b-bd17-ba61996aad61

[24] M. Gaikwad, S. Ahirrao, S. Phansalkar, and K. Kotecha, ''Online extremism detection: A systematic literature review with emphasis on datasets, classification techniques, validation methods, and tools,'' *IEEE Access*, vol. 9, pp. 48364–48404, 2021, doi: 10.1109/access.2021.3068313.

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

[26] R. Borum and T. Neer, ''Terrorism and violent extremism,'' in *Handbook of Behavioral Criminology*. Cham, Switzerland: Springer, 2017, pp. 729–745, doi: 10.1007/978-3-319-61625-4_41.

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

[28] 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, doi: 10.1016/j.neucom.2015.09.063.

[29] 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*. Cham, Switzerland: Springer, 2015, pp. 431–442, doi: 10.1007/978-3-319-14977-6_47.

[30] S. Mussiraliyeva, M. Bolatbek, B. Omarov, and K. Bagitova, ''Detection of extremist ideation on social media using machine learning techniques,'' in *Computational Collective Intelligence*. Cham, Switzerland: Springer, 2020, pp. 743–752, doi: 10.1007/978-3-030-63007-2_58.

[31] Y. L. Cun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, ''Handwritten digit recognition with a back-propagation network,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 1990, pp. 396–404.

[32] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, ''Learning internal representations by error propagation,'' *Parallel Distrib. Process.*, vol. 1, pp. 318–362, 1986.

[33] J. L. Elman, ''Finding structure in time,'' *Cogn. Sci.*, vol. 14, no. 2, pp. 179–211, Mar. 1990.

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

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

[36] W. Sharif, S. Mumtaz, Z. Shafiq, O. Riaz, T. Ali, M. Husnain, and G. S. Choi, ''An empirical approach for extreme behavior identification through tweets using machine learning,'' *Appl. Sci.*, vol. 9, no. 18, p. 3723, Sep. 2019, doi: 10.3390/app9183723.

[37] 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, p. 24, Jul. 2019, doi: 10.1186/s13673-019-0185-6.

[38] 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)*, Shenzhen, China, Jul. 2019, pp. 98–103.

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

[40] 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*, Vancouver, BC, Canada, 2017, pp. 729–740.

[41] J. Zote. (2021). *12 Essential Twitter Stats to Guide Your Strategy in 2021*. Accessed: Nov. 10, 2021. [Online]. Available: https://sproutsocial.com/insights/twitter-statistics/

[42] K. Sahoo, A. K. Samal, J. Pramanik, and S. K. Pani, ''Exploratory data analysis using Python,'' *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 12, pp. 4727–4735, Oct. 2019, doi: 10.35940/ijitee.l3591.1081219.

[43] A. Kulkarni and A. Shivananda, *Natural Language Processing Recipes: Unlocking Text Data With Machine Learning and Deep Learning Using Python*. New York, NY, USA: Apress, 2019, doi: 10.1007/978-1-4842-4267-4.

[44] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ''BERT: Pre-training of deep bidirectional transformers for language understanding,'' in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (NAACL HLT)*, vol. 1, 2019, pp. 4171–4186.

[45] *Hugging Face*. Accessed: Aug. 19, 2021. [Online]. Available: https://huggingface.co/

[46] F. Gregg and D. Eder. (2019). *Dedupe Python Library*. [Online]. Available: https://github.com/dedupeio/dedupe

[47] *Wosom*. Accessed: Aug. 19, 2021. [Online]. Available: https://wosom.ai/

[48] S. Aldera, A. Emam, M. Al-Qurishi, M. Alrubaian, and A. Alothaim, "Annotated Arabic extremism tweets," [Dataset] IEEE Dataport, Aug. 2021, doi: 10.21227/g9c0-1t21.

[49] K. L. Gwet, "Computing inter-rater reliability and its variance in the presence of high agreement," *Brit. J. Math. Statist. Psychol.*, vol. 61, no. 1, pp. 29–48, May 2008, doi: 10.1348/000711006x126600.

[50] D. Hovy and S. L. Spruit, "The social impact of natural language processing," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 591–598.

[51] Gender API. *Gender API—Determines the Gender of a First Name*. Accessed: Nov. 11, 2021. [Online]. Available: https://gender-api.com/

[52] B. Talafha, M. Ali, M. E. Za'ter, H. Seelawi, I. Tuffaha, M. Samir, W. Farhan, and H. T. Al-Natsheh, "Multi-dialect Arabic BERT for country-level dialect identification," Jul. 2020, *arXiv:2007.05612*.

[53] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based model for Arabic language understanding," in *Proc. 12th Int. Conf. Lang. Resour. Eval. (LREC)*, Marseille, France, 2020.

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

AHMED EMAM received the B.Sc. degree from Ain Shams University, Cairo, Egypt, the M.Sc. degree from Menoufia University, Menoufia, Egypt, and the Ph.D. degree from the Computer Science and Computer Engineering Department, Speed Engineering School, University of Louisville, Louisville, KY, USA, in Summer 2001. He is currently a Professor of information systems at 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 at the Research and Innovation Department, ELM Company. He has published several papers in refereed journals, such as 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, in 2013, and the Best Ph.D. Thesis Award from CCIS, KSU, in 2018. He also received the IBM Data Science Professional Certificate and Deep Learning Certification from deeplearning.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, College Park, MD, USA, and the Ph.D. degree in information technology from the University of Nebraska. He is currently working as an Assistant Professor of information systems at the College of Computer and Information Sciences, King Saud University. He also works as a Consultant in digital banking at Alinma Bank. He is a college representative for the distinguished and talented students' program, attending to honors students' issues and cultivating their talents and skills. His research interests include data science (deep learning and NLP) and artificial intelligence, crowdsourcing, electronic governments, digital transformation, and digital banking.