DOI: 10.24996/ijs.2024.65.11.35

ISSN: 0067-2904

A Review for Arabic Extremism Detection Using Machine Learning

Redhaa Fadhil Sabri *, Nada A. Z. Abdullah
Department of Computer Science, College of Science, University of Baghdad,Baghdad, Iraq.

Received: 16/2/2023 Accepted: 26/9/2023 Published: 30/11/2024

Abstract

The detection of extremism is a field of research aimed at discovering extremism and hate speech on social networking sites through the use of natural language processing tools. Although this field is wide, there is little research in Arabic that has dealt with this topic compared to research in English, although a very large percentage of social media users are Arabs and interact in Arabic. In this research, the most important research works related to the field of extremism detection in the Arabic language using machine learning algorithms will be reviewed, and related works will be compared in terms of methodology and results.

Keywords: Extremism Detection, Dialect Arabic, Machine Learning, hate speech, NLP.

مراجعة لكشف التطرف في النصوص العربية باستعمال التعلم الآلي

رضاء فاضل صبري * , ندا عبد الزهرة عبدالله
قسم الحاسوب، كلية العلوم، جامعة بغداد، بغداد، العراق

الخلاصة

كشف التطرف هو مجال بحثي الغرض منه اكتشاف التطرف وخطاب الكراهية على مواقع التواصل الاجتماعية وذلك من خلال استعمال ادوات معالجة اللغات الطبيعية. على الرغم من ان هذا المجال واسع الا انه هناك القليل من البحوث باللغة العربية التي قد تناولت هذا الموضوع مقارنه مع البحوث باللغة الانكليزية مع ان نسبة كبيرة جدا من مستخدمي مواقع التواصل الاجتماعية عرب ويتفاعلون باللغة العربية . في هذا البحث , سنستعرض اهم الاعمال البحثية التي تخص مجال كشف التطرف في اللغة العربية باستعمال خوارزميات التعلم الالي كما ستقارن الاعمال ذات الصلة من حيث المنهجية والنتائج.

1. Introduction

Nowadays, compared to the past, the Internet has become accessible to everyone, and the number of users of social networking sites has increased. There are many applications spread on the Internet, such as Facebook, Twitter, Instagram, and other applications that contain a huge number of users who can express their opinions and feelings towards several fields and things. And spreading their ideas through these platforms, and despite the fact that these applications are useful, they contain a bad side, which is that some extremist groups and people who try to use these platforms to spread their sectarian ideas and hate speech in society, and an example of these extremist groups is the terrorist organization ISIS [1].

* Email: ridhaa.fadel2101m@sc.uobaghdad.edu.iq

Hate speech is the use of insulting, violent, and sectarian phrases and words directed towards a specific person or minority group. With the increase in Arab social media users, this increase has been accompanied by a very large rise in hate speech on the Internet [2]. Despite the facts of the development taking place in the technologies present within the applications, there are no available technologies that identify posts on social media that contain extremism. The first example of this is the ISIS messages that it published and misused on these platforms to promote extremism and violence [3]. The most worrying thing about this very important issue is how difficult it is to manually monitor and check all daily publications. Because there are hundreds of thousands of publications in a very few minutes that users publish on social media platforms, it has become important that we develop techniques that automatically detect content hate and extremism without human intervention [4].

The remainder of the paper will have the following structure: The second section will be a simple background, the third section will explore a review of the research on Arabic extremism detection using machine learning, the fourth section will illustrate the Arabic language challenges, the fifth section will explain the extremism detection model, which includes subsections, and the sixth section will present a brief comparison of the annotated research. The conclusion and future work will be in the last section.

2. Background

2.1 Definition of Extremism

Extremism via the Internet is defined as the attempt of some people or groups to spread their extremist ideas through social media platforms in order to influence other people, where they promote their extremist ideas, behaviors, or opinions that relate to some social, political, and religious matters. Accordingly, the definition of extremism in this study is an interaction between extremists and members of the community. Where the first party seeks to effect the second party through its messages and ideas that it shares on these platforms [5].

2.2 Arabic Language Overview

There are many difficulties facing the Arabic language, as Arabic contains 28 letters written from right to left. One of the most important challenges that Arabic faced was that the shapes of the letters differed according to their location within the same word compared to other languages, such as the English language, whose letters are similar in all their locations within the same word. Also, one of the most important difficulties facing the Arabic language is that most Arabic words are derived from certain roots. Moreover, users of communication sites prefer to write in their native dialect rather than in Standard Arabic [6]. The fifth most widely used language in the world is Arabic [7]. Words in the Arabic language are often derived from simple verbs called roots, which usually consist of three letters. We can drop one letter from the roots, or sometimes more than one letter in some derivations. But sometimes tracing the root derived from a particular word can be very problematic [8].

3. Related Work

This section summarizes the most important studies that have been conducted on extremism detection in Arabic texts, focusing on those using machine learning approaches.

The first study on Arabic extremism detection was by Nuha Albadi et al. in 2018 [4]. They created the first Arabic dataset that was used to detect hate speech and also created an Arabic dictionary containing religious hate terms. It was the first Arabic dictionary that they made available to the public in order to encourage future research on this topic. Their method was based on using different classification models. They found that the best results were 0.84 with the recurrent neural network (RNN) based on the gated recurrent unit (GRU) based on certain

measures, while the Arabic hate speech lexicon based on pointwise mutual information (Arahate-PMI) was the best in terms of F1, recall, and accuracy. The Arabic hate speech lexicon based on bi-normal separation (Arahate-BNS) had the best performance in terms of accuracy compared to AUCROC. While the n-gram-based models, logistic regression, and support vector machine (SVM) performed similarly, they outperformed the lexicon-based models, especially in terms of accuracy.

Nuha Albadi et al. [9] presented in 2019 a study that was a continuation paper from a conference published in ASONAM 2018 (Albadi et al. 2018 [4]). They employed four methods for detecting religious hate speech, which include a lexicon-based approach, an Ngram-based approach, GRU+ word embedding, and four manual embedding features of GRU+. They concluded that GRU-based RNNs with word embedding pre-trained models outperform other lexicon-based and n-gram classifiers. Their training of the GRU model focused on some features such as user, content, and temporal. As well as including pre-trained words for tweets and user descriptions, this resulted in a speech recall performance of 0.84.
Mohammed A. Al Ghamdi et al. introduced a system in 2020 [10]. They used datasets that were tweets to train a classifier to detect suspicious activity using supervised machine learning algorithms. In the testing stage, the data from the unlabeled tweets was processed by the system to determine whether they were suspicious or not. They use six supervised machine learning algorithms to test the system: decision tree (DT), k-nearest neighbors (KNN), linear discrimination algorithm (LDA), SVM, artificial neural networks (ANN), and long short-term memory networks. ANN has the slowest execution speed, while SVM outperforms all other classifiers in predicting correct results, with an average accuracy of 86.72%.

Amal Rekik et al. [11] introduced a recursive method to detect extremism on social media in 2020. Their method is mainly based on the repeated extraction of a group of profiles that are extremist based on suspicious interactions and dangerous posts. Then, they used the mining of item groups, N-grams, and the degree of violence, after which they analyzed the combined textual data in order to extract the dangerous vocabulary. The results showed that the method achieves good discrimination performance, which indicates its effectiveness in identifying extremism on Twitter.

Ibrahim Al Jarah et al. in 2020 [2], They extracted several different types, such as features and emotions, and arranged them in 15 different datasets. So, four classification models were used to test the dataset: SVM, Naive Bayes (NB), DT, and Random Forest (RF). The Term Inversion Frequency Document (TF-IDF) feature set gave the best results, as it included RF and profile-related features. In addition, they performed a feature significance analysis using the RF classifier in order to determine the predictive power of the features in relation to the hated category.

Ahmed I. A. Abd-Elaal et al. [1] introduced a new architecture with an intelligent system that independently detects pro-ISIS accounts on Twitter in 2020. The system involves two sub-systems: the crawling system and the query system. The two kernel subsystems are smart detectors, which have characteristics such as linguistic and behavioral characteristics. The Smart Detector kernel was built for the crawl and query subsystems using supervised machine learning methods. The results were as follows: the linear SVM algorithm with TF-IDF embedding got the best accuracy of 89% for the ISIS content detector. It also showed that the ISIS computation detector is 94% accurate based on the f1 score and the Skip-gram linear-modulated SVM algorithm.

Rawan Abdullah Alraddadi et al., 2021 [12], They proposed a model built on a supervised machine learning (ML) method, and they used support vector machine classifiers, multinomial naive bayes (MNB), and feature extraction (TF-IDF). They applied it to two sets of data. Different experiments were conducted. These experiments are word-level and trigram-level, and the results were compared. They found that the supervised machine learning method with word level works best for two sets of data, with 97% accuracy on the balanced set of data when the SVM algorithm with TF-IDF was used. To detect and classify hate speech such as Islamic text content on the Internet, they developed and built a prototype of an interactive web application.

Norah Al-Harbi et al. [13] proposed an efficient text classifier in 2021 by using machine learning to automatically identify tweets. They chose AdB_SAMME, AdB_SAMME.R, linear SVM, NB, and LR as classifiers and then executed these classifiers on three types of features: S1 (unigram), S2 (bigram), and S3 (tri-gram). This is based on 346 pre-processed tweets. The SVM linear classifier had the best results, with a classification accuracy of 99.7% on S3 among all the other classifiers tested. When they relied on accuracy and time, the NB classifier was performed on S1 with an accuracy of 99.4%, which is comparable to linear SVM.

Saja Aldera et al. [14] collected a data set in 2021 to discuss which classification methods can be used to detect radicalization. The data set consisted of 89,816 tweets in Arabic published from 2011 to 2021. A group of experts ranked the tweets based on specific guidelines as to whether they were extremist or not. An exploratory analysis of the data was performed to understand the data set's features. Then they used classification algorithms such as logistic regression, support vector machines, naive Bayes polynomials, RF, and bidirectional encoder representations from transformers (BERT). The SVM TF-IDF feature realized the highest accuracy among conventional machine learning models (0.9729). While BERT outperformed the conventional models with 0.9749 accuracy.

Emad M. Al-Shawakfa et al. in 2021 [15], They used SVM, Nave Bayes (NB), and RF classifiers. They used a group of users' opinions on applications on the Google Play Store, and there were 1500 reviews in Arabic. Then, they rated the reviews using a two-stage rating process. First, they used a binary classifier to distinguish between positive and negative reviews. Second, based on some rules that classify extreme positive reviews from positive reviews and severely negative reviews from negative reviews, they used a binary classification mechanism. A total of four experiments were performed with ten different sub-experiments using different X validation schemes and the TF-IDF for select features. They got results showing that the best performance during Phase 1 was SVM with 30% of the test data, and NB performed better with 20% of the test data. In the Phase II classification results, SVM performed better at identifying positive reviews when dealing with the positive data set, with an accuracy of 68.7%, while NB had the best performance at identifying negative reviews when dealing with the negative data set, with 72.8% accuracy.

Khalid T. Mursi et al. [16] in 2022 used machine learning (ML) methods and sentiment analysis, as well as Word2Vec. They classify and analyze 100,000 tweets using the proposed model. This work encourages future researchers to analyze Arab hate speech using a hand-categorized Arab dataset. The trained model realized 92% accuracy and thus can be used as a primary tool by governments, ISPs, and social applications to detect radicalization in posts before they spread to a larger audience.

Mohammad Fraiwan in 2022 [17], The study was based on classifying tweets as terrorist-related, general religious, or unrelated using artificial intelligence (AI) and ML classification algorithms. The obtained results achieved the accuracy of K-nearest neighbors (KNN), Bernoulli Naive Bayes (BNN), and SVM [one-against-all (OAA) and all-against-all (AAA) algorithms]. At SVM-OAA, it has a highly rated F1 score of 83%.

4. Brief Comparison between Methods

In this section, a summary and brief comparison of the studies reviewed in this paper show their advantages and weaknesses, as illustrated in Table 1.

Table 1: Comparison between studies, showing their advantages and weaknesses

Studies	Advantages	Weaknesses
Nuha Albadi et al. (2018) [4]	Uses automated tools to detect hate speech, saving time and effort. A unique dataset and lexicon were created to improve model accuracy and try different ways to classify it.	Look at religious hate speech only; this might miss other types. might not catch new ways people express hate over time. Manual labeling for hate speech can be subjective, causing errors.
Mohammed A. AlGhamdi et al (2020) [10]	Helps analyze Arabic tweets, deal with the challenge of understanding complicated Arabic sentences, and create a tool to spot suspicious messages, which is helpful for safety.	The tool's performance could change with different data. It doesn't say how well the tool works with new tweets.
Amal Rekik et al. (2020) [11]	Focuses on stopping radical groups from using social media. Comes up with a new way to find these groups.	Looking at violent words only might miss some dangerous content. Depending on one expert might not be accurate enough.
Ibrahim Aljarah et al. (2020) [2]	Use technology to find and stop hate speech, and look at different kinds of data to understand the issue better. Explains which features are most important for spotting hate speech.	Saying some features are most important might not always be true, and just because a method works on one dataset doesn't mean it will work everywhere.
Ahmed I. A. Abd-Elaal et al. (2020) [1]	Deals with the problem of radical groups to make social media safer by spotting these groups. Use a smart system to do this automatically.	Assuming that certain words and actions are only used by radical groups could be wrong because these groups might change their ways to avoid getting caught. Finding accounts automatically could lead to mistakes and invade privacy.
Rawan Abdullah Alraddadi et al. (2021) [12]	Uses technology to automatically find and classify these harmful websites. Compares different methods to find the best one.	Might work differently for other languages and look only at written content, might miss harmful stuff. Stopping websites automatically might take away free speech.
Norah AL-Harbi et al. (2021) [13]	Use computers to figure out if tweets are bad, and try ways to figure out the tweets. Check different types of word patterns to see which works best.	Only made for Arabic, not other languages, and looks only at words that might miss other bad stuff. Also, might not catch new ways terrorists talk online.
Saja Aldera et al. (2021) [14]	Focuses on stopping extremist ideas online. They create a big dataset of Arabic tweets and use different methods to find extremist content.	Experts might disagree on what's extremist, and it misses other types of extremist content. One method is complex and needs lots of resources.

Emad M. Al-Shawakfa et al. (2021) [15]	Use two stage of classification process, and applies different classifiers for accuracy. Incorporates rules to differentiate extreme opinions.	Assume that negative opinions are extreme and might not catch nuanced, extreme opinions. X-validation schemas may not cover all data, and TFIDF might miss context.
Mohammad Fraiwan (2022) [17]	Focuses on Arabic-speaking ISIS members and identifies common markers and keywords of ISIS rhetoric.	Limited to analyzing Twitter data from ISIS members, this might not capture all the nuances of terror-related content. Doesn't account for potential evasion techniques used by extremists.

5. Extremism Detection Model

Most studies adopt a model that detects extremism in texts, and it consists of several steps, illustrated in Figure 1 below:

Figure 1: Extremism detection: main steps

5.1 Dataset for Extremism Detection

It was found through what was briefly discussed in the third section that research that detects extremism in the Arabic language does not depend on standard data. Most of them collected data from social networking sites, either through manual collection or through the use of tools that help collect data, such as the tool provided by Twitter to developers for data collection, the Application Programming Interface (API).

5.2 Pre-processing

As the previous works analyzing Arabic data are similar [5], [18], and [19], the data cleaning step consists of many important steps such as removing usernames, punctuation, numbers, links, English letters, duplicated tweets, and (\n, R, -,). The main goal of natural language processing is to teach machines how to understand human language and then analyze and classify data. There are several steps available for Arabic (NLP):

1) *Normalization*

The letters in Arabic have different shapes, such as the letter "Alef," which has many shapes. There are some letters that are used as extra letters in the Arabic language, which are other forms of primary characters depending on their location in the word; for instance, (إ,آ,أ,ا) Alef maksora (ى) is one of Alefs' forms but is regularly muddled when writing with the letter yaa (ي) Taa Marbotta (ة), which is regularly muddled with ha (ه) [20].

2) *Tokenization*

It is a technique that splits the text into words; each word stands alone and distinguishes the following word by using a space; each division is referred to as a token [21].

3) *Stemming*

In Arabic, the word has many parts: suffixes, prefixes, antefixes, and postfixes. The main goal of stemming is to remove these parts from the word and return it to its root [19]. For example (ليسمعونهم): antefix (ل), prefix (ي), root (سمع), suffix (ن), postfix (هم).

4) *Remove Arabic Stop Words*

Stopwords are words that must be removed before we can process the text to make the analysis process faster. And stopping the word removal step doesn't change the meaning of the sentence, and it should be removed because it may mislead the results, so we should ignore them in order to improve the research process [21].

5.3 Feature Extraction

After the preprocessing phase is completed, the feature extraction process begins, which represents the text in a format that machines can read by replacing it with numbers, i.e., a numerical representation of a word. Word embedding is a feature-gaining-to-know method wherein the contextual hierarchy of phrases is used to map phrases to vectors, in which the word's meanings and semantic relationships are captured. Word embedding is the process of digitally representing words and documents by converting them into digital vectors, and these vectors represent the word in low-dimensional space. This process allows similar words to have the same numerical representation [22]. Feature vectors for similar phrases might be the same. Characteristic vectors were created either by using building fashions, such as libraries, with special techniques or extraordinary vector dimensions, or by using pre-trained models. Feature extraction is divided into three categories: the BoW, TF, and TF-IDF models in [3], [16], and [14], which are written in the Python programming language [23]. In [17], CountVectorizer was used to convert each tweet into a token count matrix. The CountVectorizer matrix was then normalized to TFIDF representation using Term-Frequency Times Inverse Document-Frequency (TFIDF). Some studies use N-grams, TF-IDF, and Word2Vec to extract features [15], [13]. Violence degree and N-grams are used in [12]. Some of these studies use word embedding models such as BoW bag-of-words models [11]. Chi-square, PMI, and BNS are the methods used to select features in some studies [10], [5]. We have explained in a simple way the most commonly used feature extraction models in the studies that we have presented [24]:

1) Term Frequency (TF)

The principle of this model is that it calculates the frequency of the feature appearing in the document, and it is the number of times it is recurrent in the document that represents the importance of this feature.

2) Inverse Document Frequency (IDF)

IDF is used to evaluate the importance of a word in a document that is often less visible, in contrast to TF.

3) Term Frequency – Inverse Document Frequency TF-IDF

It is based on the Bag of Words (BoW) model, and this model is a mixture between TF and IDF, as it provides greater importance to the most frequent words as well as around the less frequent words, giving them more weight, which are important in the document.

5.4 Machine Learning Algorithms

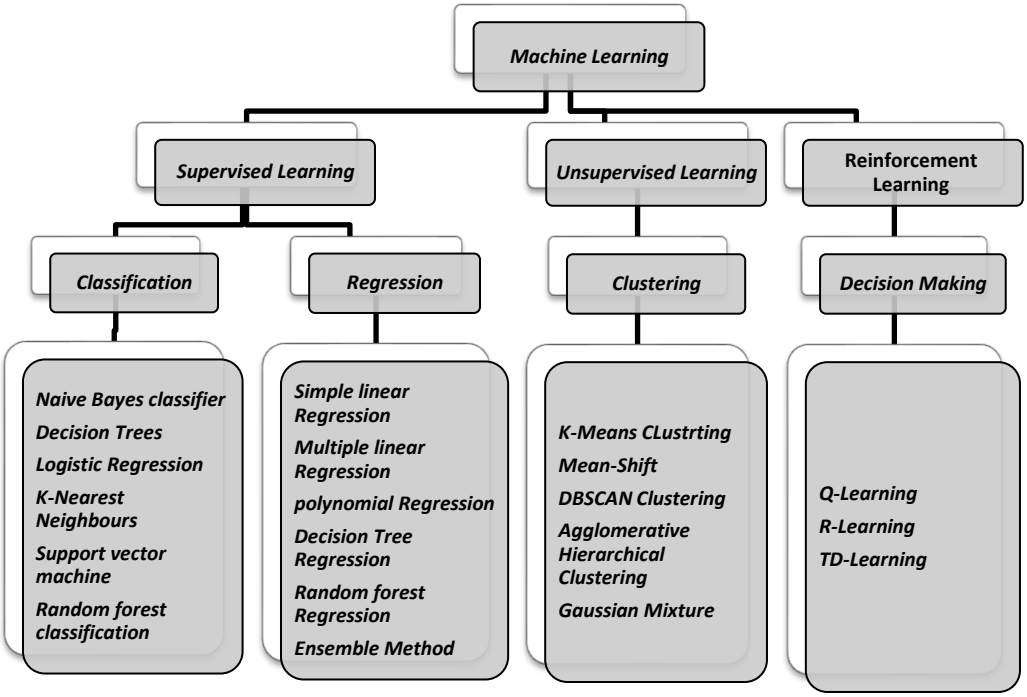

Figure 2: Machine Learning Algorithms [23]

ML is a powerful technology used by researchers in multiple fields, the most important of which is the detection of extremism. ML is the ability of a computer to learn itself how to make decisions based on data [23]. In addition to machine learning, it revolutionizes several areas of life, not only specialized in a specific field, as it has become an important part in the medical fields and is used in many medical applications that help diagnose diseases, so that it has proven that it can be more accurate in diagnosing than doctors [25]. The data that is used by ML researchers is referred to as training data. In ML, the decisions that the computer makes are either classification or prediction, whereby the computer can classify new data by training models using learning algorithms. A learning algorithm or training model that relies on human-labeled data is denoted as supervised. In the field of extremism and hate speech detection, we can manually classify a set of data as containing hate speech/extremist or not [26]. When the training data is not labeled, the learning algorithms are known as unsupervised algorithms [27]. Algorithms learn to classify on their own, depending on similarities and differences in the dataset. A semi-supervised learning algorithm is one that combines supervised learning with unsupervised learning [23]. Several well-known machine learning (ML) algorithms have been developed, including naive Bayes classifiers, the k-nearest neighbors (KNN) algorithm, SVMs, and DTs [26]. Because of the huge amount of data that is published daily and in very few minutes via social networks, it is impossible for these platforms alone to process the available data. The algorithms of ML are widely present in online radicalization detection systems. The techniques of deep learning (DL) have been used in the detection of radicals in recent years, with several proposed systems achieving remarkable accuracy [28], [29]. DL is a type of machine learning that can learn unsupervised from unstructured data sets. In deep learning, there are multiple hidden layers that are used to select higher-level features from the input. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are two of the most important algorithms used in text processing, especially RNNs. When compared to traditional machine learning algorithms, DL algorithms require significantly more data for training [30]. Most of the research discussed in the third section of this research used machine learning algorithms, and some others combined machine learning and deep learning, but machine learning algorithms were more commonly

used, as shown in Table 1, which presents a comparison to the research related to Arabic extremism detection based on machine learning algorithms. We will explain some algorithms in a simple way [31]:

1) *Naïve Bayes*

This algorithm depends on the hypotheses that affect the variables and is considered one of the most comfortable ways, but it is less useful compared to other methods because it is not a simple explicit model.

2) *Support Vector Machine*

The SVM algorithm was created to solve problems that occur with the classification process and has been developed over time. It is considered one of the most used algorithms, especially in problems that require a binary classification.

3) *Logistic Regression*

This algorithm is used to predict something that is very useful, especially since the dependent variable takes values in a finite group. This algorithm was established in the 1960s, but it was used in the 1980s due to its computational facilities.

4) *K-Nearest Neighbors*

This algorithm classifies query points whose class is unknown and assumes that every example of the learning group is a random variable. This algorithm is used when the information is small or nonexistent.

5) *Random Forest*

This algorithm is one of the most used in the field of analysis and prediction of data and is considered a group of groups, meaning that it consists of a large number of individual decision groups and is a tree, with every tree containing random samples from the data set.

6. Brief Comparison of Arabic Extremism Detection Researches

In this section, we present a summary and brief comparison between the studies reviewed in this paper in terms of data sets, processing methods, feature extraction methods, classification models, and the highest accuracy obtained by the model. As illustrated in Table 2, note that the values of the fields with the highest accuracy are written in bold.

Table 2: Comparison of Arabic Extremism Detection Research

Researchers and year	Datasets	Dialect/MSA	Pre-processing	Text representation	Machine Learning model	accuracy
Nuha Albadi et al. (2018) [4]	6000 Arabic tweets in 2017 and 600 tweets in 2018	Dialect and MSA	Clean data Remove stopwords Tokenize Stemming Normalizing	chi-square, PMI, and BNS web_CBOW Wikipedia_CBOW	Approaches based on lexicons, N-grams, and deep learning	GRU-based RNN performs best, with 0.79 accuracy and 0.84 AUROC.
Nuha Albadi et al. (2019) [9]	6000 Arabic tweets	Dialect and MSA	Clean data Remove stopwords Tokenize Stemming Normalizing	Chi-square, PMI, and BNS web_CBOW Wikipedia_CBOW	AraHate-PMI AraHate-Chi AraHateBNS logistic regression SVM GRU + word embeddings GRU + word	Training a GRU and pre-trained word embeddings performs in terms of recall (0.84)

					embeddings + handcrafted features	
Mohammed A. AlGhamdi et al. (2020) [10]	1555 tweets	MSA	Clean data, Stemming, and Lemmatization	Bag-of-words (BoW) model and word embedding	DT , k-NN , LDA, SVM , ANN, and Long short-term memory networks (LSMN)	SVM was the best performance with 86.72% mean accuracy.
Amal Rekik et al. (2020) [11]	(3325 profiles) For dangerous users	MSA	remove letters are not Arabic, diacritics removal, remove punctuation marks, remove numbers and stop words	each data was represented in n-grams by the non-radical community	N-grams and itemsets mining violence degree	0,88 for the proposed methodology
Ibrahim Aljarah et al. (2020) [2]	3696 tweets	Dialect and MSA	Cleaning the data, Normalization, Tokenization	BoW, TF, and TF-IDF models,	(SVM, NB, DT, and RF)	best result that achieved by RF at TF-IDF with accuracy equals to 0.882
Ahmed I. A. Abd-Elaal et al. (2020) [1]	21,000 tweets and three datasets in Kaggle "How ISIS Uses Twitter", "Religious Texts Used By ISIS", "Tweets Targeting ISIS"	Dialect and MSA	Remove URL links and mentions , Discarding non-alpha letters removal, Normalization, Stop words removal, Tashkeel removal, Prefix/suffix removal	TF-IDF and Skip-gram “Mazajak”	BNN, DT C, K-NN, SVM, LR and RF Classifiers	best accuracy 94% by linear SVM with Skip-gram word embedding
Rawan Abdullah Alraddadi et al. (2021) [12]	9000 data collected from articles, journals and personal blogs	MSA	Stop-words removal, Normalization, Stemming and Lemmatization	TF-IDF	(SVM) and (MNB)	high accuracy with 97% SVM algorithm
Norah AL-Harbi et al.	135,069 Tweets	Dialect and MSA	Remove stop words,	TF-IDF	AdB_SAM M E,	SVM performed

(2021) [13]			Punctuation removal, blank spaces, and Diacritic marks, Tokenization		AdB_SAMM E.R, Linear SVM, NB, and LR	exceptionally with **99.7%** accuracy the NB classifier perform on S1 with 99.4% accuracy
Saja Aldera et al. (2021) [14]	89,816 tweets published between 2011 and 2021	Dialect and MSA	Lemmatization, Stop-words removal, Tokenization	TF-IDF and Word2Vec	LR, MNB, SVM, RF, and BERT	SVM using TF-IDF achieved accuracy (0.9729) , while BERT model outperformed SVM, achieve 0.9749.
Emad M. Al-Shawakfa et al. (2021) [15]	Dataset of 1500 unique reviews 750 positive and 750 negative reviews	Dialect and MSA	Clean data Tokenization POS-Tagging Stopword Removal	(TF-IDF), (BOW)	SVM, NB and RF	SVM accuracy of 68.7%.
Khalid T. Mursi et al. (2022) [16]	AJGT dataset & manually collected 103,000 samples	MSA	Cleaning the dataset, Unifying the dialect of the tweets to Modern Standard Arabic (MSA), Removing the stop words, Word segmentation	TF-IDF, Word2Vec	(SVM), Multi-layer perceptron (MLP)	SVM with an accuracy of 0.92 and MLP with an accuracy of 0.91
Mohammad Fraiwan (2022) [17]	24,078 tweets	Dialect and MSA	Filtering the duplicate tweets, Tokenizing, Removing diacritic marks, Normalization, Lemmatizing	Word embedding	KNN, BNB and SVM linear Kernel OAO and OAA classifiers	achieved F1 score of 83\% in SVM-OAA

7. Conclusion and Future Work

- Most of the studies discussed that dealt with extremism specifically dealt with Standard Arabic, while most users on social media platforms express their opinions using their own dialects.

- There are few studies related to the detection of extremism in Arabic dialects, so we must focus on this topic, in addition to the fact that most researchers either collect little data or the data on extremism written in Arabic dialects available on the Internet is very small.
- The Arabic language is rich and frequently used and needs to be handled with good care.
- Most of the pre-processing steps were not standardized, such as normalization and derivation.
- It is difficult to obtain data on extremism written only in Standard Arabic because most users of social media platforms use their own dialects.
- The difference in dialects and his way of writing from one person to another led to the weakness of the results obtained.
- Most of the studies used special machine learning algorithms, and most of the good results were obtained through SVM.

Finally, in order to improve the detection of extremism in the Arabic language, we suggest in the future that the study be conducted on accurate and large data sets, in addition to paying attention to dialects such as the Iraqi dialect, as it is considered the most difficult and most diverse of the Arabic dialects. Focusing on supervised machine learning algorithms because they perform better with data related to extremism, as well as using modern word embedding techniques in order to obtain more accurate results by representing words in numerical vectors with semantic meaning.

References

[1] A. I. A. Abd-Elaal, A. Z. Badr and H. M. K. Mahdi, "Detecting Violent Radical Accounts on Twitter," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 8, pp. 516-522, 2020.

[2] I. Aljarah, M. Habib, N. Hijazi, H. Faris, R. Qaddoura, B. Hammo, M. Abushariah and M. Alfawareh, "Intelligent detection of hate speech in Arabic social network: A machine learning approach," *Journal of Information Science*, vol. 47, no. 4, pp. 483–501, 2020.

[3] Y. Wei, L. Singh and S. Martin, "Identification of extremism on Twitter," *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 1251-1255, 2016.

[4] N. Albadi, M. Kurdi and S. Mishra, "Are they Our Brothers? Analysis and Detection of Religious Hate Speech in the Arabic Twittersphere," *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2018.

[5] 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, p. 42384–42396, 2021.

[6] T. Kanan, O. Sadaqa, A. Aldajeh, H. Alshwabka, W. AL-dolime, S. AlZu'bi, B. Hawashin and M. A. Alia, "A Review of Natural Language Processing and Machine Learning Tools Used to Analyze Arabic Social Media," *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, 2019.

[7] N. A. Abdullah and F. A. Abdulghani, "A Survey on Arabic Text Classification Using Deep and Machine Learning Algorithms," *Iraqi Journal of science*, vol. 63, no. 1, pp. 409–419, Jan. 2022.

[8] N. A. Abdullah and N. T. Jaboory, "Arabic Keywords Extraction using Conventional Neural Network," *Iraqi Journal of Science*, vol. 63, no. 1, pp. 283–293, Jan. 2022.

[9] N. Albadi, M. Kurdi and S. Mishra, "Investigating the effect of combining GRU neural networks with handcrafted features for religious hatred detection on Arabic Twitter space," *Social Network Analysis and Mining*, vol. 9, p. 41, 2019.

[10] M. A. AlGhamdi and M. A. Khan, "Intelligent Analysis of Arabic Tweets for Detection of Suspicious Messages," *Arabian Journal for Science and Engineering*, vol. 45, no. 8, pp. 6021–6032, 2020.

[11] A. Rekik, S. Jamoussi and A. Ben Hamadou, "A recursive methodology for radical communities' detection on social networks," *Procedia Computer Science,* vol. 176, p. 2010–2019, 2020.

[12] R. A. Alraddadi and M. I. El-Khalil Ghembaza, "Anti-Islamic Arabic Text Categorization using Text Mining and Sentiment Analysis Techniques," *International Journal of Advanced Computer Science and Applications,* vol. 12, no. 8, pp. 776-785, 2021.

[13] N. AL-Harbi and A. Bin Kamsin, "An Effective Text Classifier using Machine Learning for Identifying Tweets' Polarity Concerning Terrorist Connotation," *International Journal of Information Technology and Computer Science,* vol. 13, no. 5, pp. 19–29, 2021.

[14] S. Aldera, A. Emam, M. Al-Qurishi, M. Alrubaian and A. Alothaim, "Exploratory Data Analysis and Classification of a New Arabic Online Extremism Dataset," *IEEE Access,* vol. 9, pp. 161613–161626, 2021.

[15] E. M. Al-Shawakfa and H. . H. Husni, "A Two-Stage Machine Learning Classification Approach to Identify Extremism in Arabic Opinions," *International Journal of Advanced Trends in Computer Science and Engineering,* vol. 10, no. 2, pp. 736–745, 2021.

[16] K. T. Mursi, M. D. Alahmadi, F. S. Alsubaei and A. S. Alghamdi, "Detecting Islamic Radicalism Arabic Tweets Using Natural Language Processing," *IEEE Access,* vol. 10, pp. 72526–72534, 2022.

[17] M. Fraiwan, "Identification of markers and artificial intelligence-based classification of radical Twitter data," *Applied Computing and Informatics*, 2022.

[18] A. Alharbi, M. Kalkatawi and M. Taileb, "Arabic Sentiment Analysis Using Deep Learning and Ensemble Methods," *Arabian Journal for Science and Engineering,* vol. 46, no. 9, pp. 8913–8923, 2021.

[19] H. Froud, R. Benslimane, A. Lachkar and S. O. El Alaoui, "Stemming and similarity measures for Arabic Documents Clustering," *2010 5th International Symposium on I/V Communications and Mobile Network,* pp. 1-4, 2010.

[20] M. N. Al-Kabi, A. Gigieh, I. Alsmadi and H. A. Wahsheh, "Opinion Mining and Analysis for Arabic Language," *International Journal of Advanced Computer Science and Applications,* vol. 5, no. 5, pp. 181-195, 2014.

[21] A. Kumar, M. Abdel-Basset, I. M. El-henawy and A. Fakhry, "Arabic text clustering using improved clustering algorithms with dimensionality reduction," *Cluster Computing,* vol. 22, no. S2, pp. 4535–4549, 2018.

[22] A. H. Wadud, M. F. Mridha and M. M. Rahman, "Word Embedding Methods for Word Representation in Deep Learning for Natural Language Processing," *Iraqi Journal of Science,* vol. 63, no. 3, pp. 1349-1361, 2022.

[23] A. Aldahiri, B. Alrashed, and W. Hussain, "Trends in Using IoT with Machine Learning in Health Prediction System," *Forecasting*, vol. 3, no. 1, pp. 181–206, Mar. 2021, doi: 10.3390/forecast3010012. [Online]. Available: http://dx.doi.org/10.3390/forecast3010012

[24] S. K. Adnan and N. A. Abdullah, "Arabic Query Expansion Using Wordnet And Cuckoo Algorithm," *ARPN Journal of Engineering and Applied Sciences*, vol. 14, pp. 1898-1903, 2019.

[25] A. Al-Imam and F. Lami, "Machine Learning for Potent Dermatology Research and Practice," *Journal of Dermatology and Dermatologic Surgery*, vol. 24, no. 1, pp. 1-4, 2020.

[26] T. Sabbah, A. Selamat, M. H. Selamat and R. Ibrahim, "Hybridized term-weighting method for Dark Web classification," *Neurocomputing*, vol. 173, pp. 1908–1926, 2016.

[27] M. Nouh, M. Goldsmith and R. J. Nurse, "Understanding the Radical Mind: Identifying Signals to Detect Extremist Content on Twitter," *2019 IEEE International Conference on Intelligence and Security Informatics (ISI),* pp. 98-103, 2019.

[28] P. Badjatiya, S. Gupta, M. Gupta and V. Varma, "Deep Learning for Hate Speech Detection in Tweets," *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion, p*p. 759–760, 2017.

[29] H. Alvari, S. Sarkar and P. Shakarian, "Detection of Violent Extremists in Social Media," *2019*

2nd International Conference on Data Intelligence and Security (ICDIS), pp. 43-47, 2019.

[30] Y. Tsuruoka, "Deep learning and natural language processing," *Brain Nerve*, vol. 71, no. 1, pp. 45-55, 2019.

[31] M. M. Mijwil and B. S. Shukur, "A Scoping Review of Machine Learning Techniques and Their Utilisation in Predicting Heart Diseases," *Ibn Al-Haitham Journal for Pure and Applied Sciences*, vol. 35, no. 3, pp. 175–189, Jul. 2022.