Neural Networks and Sentiment Features for Extremist Content Detection in Arabic Social Media

Hanen Himdi
Department of Computer Science and Artificial Intelligence, University of Jeddah, Saudi Arabia
hthimdi@uj.edu.sa

Fatimah Alhayan
Department of Information Systems Princess Nourah bint Abdulrahman University, Saudi Arabia
fnalhayan@pnu.edu.sa

Khaled Shaalan
Faculty of Engineering and Information Technology, The British University United Arab Emirates
khaled.shaalan@buid.ac.ae

Abstract: *The proliferation of extremist content on social media poses critical threats to societal stability, necessitating advanced detection mechanisms. Despite substantial research on extremist content detection in various languages, Arabic remains significantly underexplored. Recognizing the pivotal role of social media, this study introduces a novel approach to detecting extremist posts in Arabic by leveraging neural networks. The proposed models utilize Arabic Bidirectional Encoder Representations from Transformers (AraBERT), Multi-Layer Perceptron (MLP), and Sentiment Features (SFs). Among the tested models, the optimal configuration-fine-tuning AraBERT with integrated MLP and SF-achieved an impressive 98% accuracy in detecting extremist Arabic tweets. Additionally, the model demonstrated robust performance when evaluated on real-world extremist posts from VKontakte, achieving 81% accuracy. These findings underscore the effectiveness of combining AraBERT, MLP, and SF in improving extremist content detection and highlight the potential of neural network-based solutions in combating harmful online content.*

Keywords: *BERT, sentiment analysis, deep learning, transformer-based models.*

Received July 9, 2024, 2024; accepted March 16, 2025
https://doi.org/10.34028/iajit/22/3/8

1. Introduction

The rapid proliferation of extremist content has become a pressing concern, evident across a wide range of online platforms [5]. Social media, in particular, has emerged as a critical breeding ground for the dissemination of radical ideologies, with the potential to radicalize vulnerable individuals and exacerbate societal tensions. Beyond fostering ideological divides, such content can incite violence, posing severe threats to both individual safety and societal stability [2].

As defined by Lipset [27], extremism encompasses ideologies (whether religious or political) that are considered unacceptable by society. To put it in simple terms, extremism is defined as the advocacy of 'extreme' beliefs that significantly deviate from mainstream societal norms [22]. In a broader context, extremism can broadly be defined as a term that represents the adoption and promotion of radical ideologies that advocate for the pursuit of goals through violent or intolerant means. Whether rooted in political, religious, or social beliefs, extremist ideologies prioritize uncompromising adherence to a particular worldview, often rejecting pluralism, diversity, and democratic principles.

For example, in light of the January 6, 2021, Capitol incident in the US, a number of news articles emphasize the pressing necessity to confront and destroy extremist organizations. The uprising underscored the ability of these groups to adjust and develop, constantly presenting serious dangers, even in the aftermath of the incident. According to the Atlantic Council, online communities had difficulties in handling the increase in extremist material, since several users advocated for more violence. Following the conflict, there was an upswing of messages on social media sites expressing anger and urging for more action. Analyzing these sentiments provides insight into the motivations and probable actions of the rioters. This highlights the vital role of sentiment analysis in identifying extreme content. Although there have been numerous studies conducted in other languages to address this issue [12, 20, 31, 34], there has been a dearth of research done in Arabic. So, it is imperative to develop effective techniques for detecting extremist content in Arabic. This underscores the crucial function of sentiment analysis in detecting extreme content.

Certainly, the imperative of detecting extremist posts and content in Arabic social media is particularly acute given the region's complex geopolitical dynamics and the prevalence of Arabic as a widely used language online. Extremist groups leverage social media and online forums to amplify their messages and influence public discourse, often exacerbating existing societal and political fault lines. This is particularly pronounced

in Arabic-speaking regions, where extremist movements have exploited linguistic and cultural nuances to spread their ideologies, posing significant challenges to peace and reconciliation [41].

To detect such harmful text on online platforms such as hate speech, it was found that Sentiment Features (SFs) contributed greatly to identifying such texts [42, 44]. Though extremism is broader in terms of context, it encompasses entire ideologies that may incorporate a spectrum of hate speech [23]. Based on that, we investigate the ability of these features to facilitate the detection of a specific type of hate speech, i.e. extremist posts. However, analyzing textual content poses significant challenges due to the absence of metadata that could facilitate examination. The nuanced distinctions between extremist and non-extremist posts, particularly when addressing similar topics and expressive narrative tones, often prove difficult to identify [17]. For that, we opt for innovative methods of Artificial Intelligence (AI), which play a crucial role in addressing such threats, specifically, approaches equipped with transformer-based models and Natural Language Processing (NLP) techniques.

SFs have been found to greatly aid in detecting harmful text, such as hate speech, on online platforms [42, 44]. In the context of Arabic, where extremism can manifest itself through a spectrum of ideologies embedded in linguistic subtleties, these characteristics are especially relevant. Extremist content often includes hate speech as a subset, making sentiment analysis a crucial tool for detecting these narratives [23]. However, the analysis of Arabic textual content presents unique challenges due to the language's complexity, including rich morphology, diacritics, and contextual dependencies. Distinguishing extremist posts from non-extremist ones, particularly when they share similar topics or narrative tones is a nuanced task [17].

To address these challenges, this study employs advanced AI techniques, specifically transformer-based models and NLP approaches. These methods are particularly well-suited to handling the complexities of the Arabic language, offering a robust framework for detecting extremist content.

1.1. Problem Statement and Objective

The extensive spread of extremist content on Arabic social media platforms presents a significant challenge to maintaining a safe and respectful online environment. Traditional techniques for identifying such content often struggle to capture the nuanced, context-specific features associated with extremist language, especially in Arabic, where semantic and contextual intricacies are profound.

Advanced NLP techniques, particularly transformer-based models such as BERT, offer a promising solution. BERT's ability to deeply understand context and semantics is especially valuable for identifying and categorizing extremist tweets [25]. The present study aims to develop effective and robust detection models based on BERT to accurately identify extremist Arabic content. This includes fine-tuning the pre-trained BERT model and integrating SFs with neural networks, such as Multi-Layer Perceptron (MLP), to improve classification performance. The study also examines the model's effectiveness in real-world extremist contexts.

1.2. Contributions

The primary contributions of this work are:

- Developing Arabic Bidirectional Encoder Representations from Transformers based (AraBERT-based) models specifically tailored for detecting extremist Arabic textual content on social media.
- Highlighting the importance of SFs in identifying extremist content by integrating them with neural network models, such as MLP.
- Proposing a novel model that combines AraBERT, SFs, and MLP to enhance the detection of extremist content in Arabic.

The rest of the paper is organized as follows: Sections 2 related work on extremist content detection. Section 3 outlines the proposed methodology, including dataset and model development. Section 5 presents the experimental setup, while section 6 shares the results and discusses them. The last two, sections 7 and 8, provide limitations, future directions, and conclusions of the proposed work.

2. Related Work

A framework for analyzing extremist-related texts and content was introduced in a study by Ahmad *et al*. [2]. In this study, the researchers adopted a simple approach that classifies tweets into extremist and non-extremist categories. Using Deep Learning (DL)-based sentiment analysis techniques on user-generated posts from Twitter, the research findings demonstrate promising results.

In a study by Ul Rehman *et al*. [43], an attempt was made to identify extremist and radical discourse on social media. The researchers introduced a novel dataset tailored for radicalization detection. Also, the novelty of the research is that an innovative classifier method was proposed that integrates religious and radical features to train and classify the data by analyzing the use of violent and offensive language. The findings show that integrating religious texts into the training model enhances classifier performance metrics in terms of accuracy, precision, recall, and F1-score. Further, the findings of the study also highlight the significant impact of utilizing new datasets on determining the classifier efficacy-basis the variations observable in extremist narratives. Additionally, the findings of the

study also reveal that the presence of violence and offensive language serves as a distinguishing factor between radical and random users.

In yet another study, Gaikwad *et al*. [22] conducted a thorough survey that employed the PRISMA methodology by gathering data from 64 studies on extremism research. Using the Snowballing technique, extant studies were sourced from a wide variety of reputable databases. One of the primary findings is the scarcity of publicly available, well-balanced, and unbiased datasets that are considered essential for the accurate detection and classification of social media extremism texts and content. Regarding the validation techniques capable of assessing the accuracy and quality of custom datasets without human intervention, the findings show that there is a notable gap, apart from highlighting a significant bias towards research focused on the ideology of the Islamic State of Iraq and Syria (ISIS)-a violent group of Sunni jihadis. Despite all these limitations, there is a sufficient indication in the findings highlighting that automated extremism detection techniques or models that are based on Deep Learning (DL) show the tendency to outperform other methods.

Concurrently, a study by Mussiraliyeva *et al*. [32] introduces a specialized corpus aimed at detecting religious extremism in social networks by focusing on the Kazakh language. In so doing, the researchers evaluate the effectiveness of six machine learning algorithms for binary classification. The findings show that it achieves a remarkable accuracy of 98% in identifying extremist messages within Kazakh texts. It is important to note that by incorporating various features and balancing techniques, the study demonstrates robust performance across diverse datasets commonly encountered in daily life. In hindsight, the overall findings underscore the potential of the proposed model or approach in real-world applications, especially for detecting extremist content.

Further, a study by Aldera *et al*. [6] introduces a comprehensive dataset designed for extremism detection in Arabic Twitter texts and content. Overall, the dataset contains 89,816 Arabic tweets annotated as extremist or non-extremist. Various classification methods were explored for conducting the analysis. For example, the classification methods involve traditional models like logistic regression and support vector machines, including advanced techniques like Bidirectional Encoder Representations from Transformers (BERT). Based on these premises, the findings of the study highlight the efficacy of support vector machines using term frequency-inverse document frequency features that achieved an impressive accuracy of 0.9729 among traditional machine learning models. However, the highlight of the findings shows the transformative power of BERT, which surpassed traditional models with a remarkable accuracy of 0.9749.

Likewise, a recent study by Ahmed *et al*. [3] introduces a novel text detection mechanism for identifying extremist orientations in Arabic text. Leveraging Rough Set theory, the findings show that the approach enhances model accuracy. The same approach also was found to be reliable in identifying text orientation. Besides, when the proposed model was compared with existing algorithms, the experimental results demonstrated superior performance with accuracies ranging from 71.95 % to 90.85 %.

3. Methodology

Designing effective models involves a sequence of important steps that embarks from domain specific datasets, followed by training after text preprocessing and feature extraction to afterward be tested and evaluated. In this study, preprocessing refers to applying different methods to get clean text, feature extraction employs the sentimental label of each tweet using Arabic NLP tools, and the training stage is to train models on the different processed features, Figure 1.

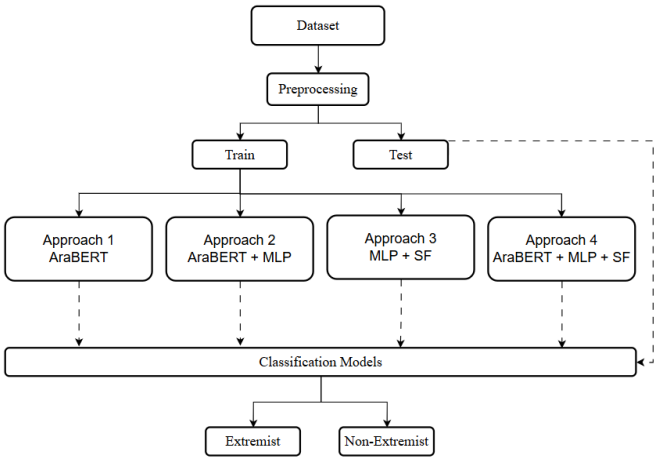

Figure 1. Proposed framework.

3.1. Datasets

This study included three datasets from studies that tackled the issue of extremism. The first and second datasets named "Arabic Extremist Dataset," and "Annotated ISIS Radical Tweets Dataset", respectively, were employed to train and test the developed models. On the other hand, the third dataset, named the" Cross-Platform Test Dataset"in this study, was used to test the performance of the optimum model from real-world extremist tweets aimed to generalize the model's efficacy.

3.1.1. Arabic Extremist Dataset

The dataset was an organized work of Aldera *et al*. [6], which originated by using the X Streaming API and Search API to accumulate real-time data on tweets, applying certain filters such as Arabic keywords and locations. The data collected was posted between May 2011 and March 2021. With an emphasis on political

and religious terminology, Arabic search terms were implemented in the query. The data was manually annotated to determine its class as extremists or non-extremists. All in all, 52,929 distinct users were the authors of the collected 89,816 tweets. The number of tweets classified as extremist was 50,279 (56%) from 22,858 unique users, while 39,537 (44%) from 30,911 unique users were classified as non-extremist. The dataset included highly extreme phrases related to violence, aggression, and extremist ideas, such as العدو enemy, داعش ISIS, and اللعنة cursing, as shown in Figure 2. On the other hand, though some of the terms were mutually found in extreme and non-extreme texts, such as الله Allah and الشعب the public, it could be that the texts discussed topics relating to these terms, as found in the common words in non-extreme texts in Figure 3.

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
الشعب	The public
معك	With you
احفظ	Save
ضد	Against
امريكا	America
عهده	Under
العراق	Iraq

Figure 2. Most frequent words in extremist Tweets [6].

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
العرب	Arabs
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
لعنة	Curse

Figure 3. Most frequent words in non-extremist Tweets [6].

Given the tweets in the dataset were provided by tweet ID, Twitter's content redistribution policy restricts the sharing of tweet information besides the tweet IDs and user IDs. In light of the challenges associated with retrieving tweets using tweet IDs, we coded a Python script to extract the first 1660 tweets from the IDs pertaining to the extremist class and another 1660 tweets corresponding to the non-extremist class from the dataset.

3.1.2. Annotated ISIS Radical Tweets Dataset

For a more comprehensive and diverse analysis of extremist and non-extremist online content, a second publicly available dataset was utilized, containing tweets that advocated for and promoted ISIS, categorizing such content as extremist. The dataset, initially presented by Fraiwan [21], included 24,078 tweets extracted from 174 accounts linked to ISIS. It includes over 10,000 tweets labeled as radical and terror-related, another 10,000 tweets classified as religious but non-terror-related, and approximately 5,000 randomly selected tweets unrelated to any religious topics.

Due to the dataset's size, extracting SFs from all tweets was computationally intensive. For analytical purposes, we randomly selected a subset: 1,700 tweets from the radical and terror-related category, labeled as 'extremist,' and 1,700 tweets from the religious but non-terror-related category, labeled as 'non-extremist.'

3.1.3. Cross-Platform Test Dataset

Due to the limited availability of Arabic extremist datasets, we utilized a publicly available extremist dataset in the Kazakh language [31]. This dataset contains both extremist messages and neutral texts. The extremist messages explicitly reference engaging in extremist activities, providing financial support to extremist groups, and expressing interest in weapons. These texts were specifically selected as examples of extremist content. The neutral texts were chosen from commonly used words that are generally understood by the Kazakh audience and do not contain any religious references.

The dataset includes 1,200 extremist messages and non-extremist texts, totaling over 140,000 words. To facilitate the translation process, we extracted a small subset from the dataset: 50 extremist texts and 50 neutral texts, each limited to a maximum of 100 words. In order to ensure accurate translation into Arabic, we used Google Translate. As a way to warranty the accuracy of the translation, we recruited the expertise of two Khaza-Arabic translators via Upwork, a freelancer services platform. Each reviewer received 50 messages. Within a duration of two weeks, both reviewers submitted the 100 translated and reviewed texts, 50 categorized as 'extremist' and 50 as 'non-extremist.' In the remainder of the research, the assembled dataset is referred to as 'cross-platform test dataset.' Table 1 shows a translated non-extremist text from Kazakh to Arabic.

Table 1. Sample translated non-extremist text in Kazakh and translated Arabic.

Class	Sample
Kazakh	Ислам күнтізбесінде қасиетті төрт ай бар. Олардың үшеуі қатарынан келеді: Зул-Қа'да, Зул-Хиджа және Мухаррам. Төртінші айға келсек, бұл жалғыз келетін Ережеп айы.
Arabic	ذو وهي متتابعة تأتي منها ثلاثة .مقدسة أشهر أربعة هناك الإسلامي، التقويم في .منفردًا يأتي الذي رجب شهر فهو الرابع الشهر أما .ومحرم الحجة، ذو القعدة،

3.2. Preprocessing

Three common preprocessing techniques for Arabic text including tokenization, removal of unnecessary words, and normalization, were applied to the reviews using the Tasaheel tool [24]. The previous tool provides several integrated Arabic NLP tools that offer tasks such as

tokenization, part-of-speech tagging, normalization, and stemming. A brief description of the preprocessing approaches conducted on the dataset using this tool is detailed below:

- **Tokenization**: it involves breaking text into smaller units called tokens.
- **Normalization**: in Arabic text, normalization aims to standardize the data for uniformity and consistency. This typically involves several steps, such as removing diacritics, which are marks representing vowels and phonetic information in Arabic words. Simplifying text processing by eliminating diacritics is a common practice. Additionally, normalization ensures that Arabic letters are represented in a standardized form, as these letters can have multiple forms depending on their position within a word. For example, normalization helps maintain a consistent representation of the Arabic letter alef ا, which can appear in different forms depending on its position in a word.
- **Remove Numerical Data, Non-Alphabetic Characters, and Stop Words**: The removal of numerical data has been shown to enhance the quality of datasets, as highlighted by Sudheesh *et al*. [35]. Therefore, all numerical data were eliminated from the text, as they do not contribute meaningful information to the decision-making process. Furthermore, to improve dataset quality and model performance, non-alphabetic characters such as punctuation marks, special characters [?, @, #, /, &, %], and URLs were removed. This preprocessing step aims to streamline the dataset by eliminating irrelevant elements. Additionally, Arabic stop words were excluded during preprocessing, aligning with findings by Sudheesh *et al*. [35]. Eliminating stop words not only enhances model accuracy and training efficiency by retaining only relevant information but also allows for more in-depth analysis, particularly beneficial for a limited dataset [26].

4. Models and Features

This section provides a background of the techniques used in this study, which are BERT-based models, MLPs and SFs.

4.1. Bidirectional Encoder Representations from Transformers (BERT)

BERT was developed and first introduced in 2018 by Google [19]. It is a language model pre-trained on large text datasets and was built adopting the transformer architecture and is available in many versions such as BERT Base and BERT Large. The BERT Base model is composed of 12 layers, each with 12 attention heads, and a hidden size of 768. In all, it has 110 million parameters. Alternatively, BERT Large consists of 24 layers, 16 attention heads, and a hidden size of 1024, resulting in a total of 340 million parameters. Both versions have the capability to manage a sequence length of up to 512 tokens.

In the architecture perspective, the first phase of BERT consists of a pair of primary objectives: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). These actions serve the model's comprehension of word context within a sentence and the correlation between subsequent sentences. BERT understands context bidirectionally. This approach aids in improving language understanding. As a language model, BERT can be fine-tuned for specific tasks to achieve state-of-the-art results, especially in various NLP tasks. This study incorporates the BERT transformer-based models, as they achieve the best results in several NLP-based studies [4, 9, 16]. Consequently, several pre-trained versions of BERT, tailored for specific languages were developed, such as AraBERT for Arabic [11], CamemBERT for French [29], BETO for Spanish [14], and German BERT for German [15].

4.2. Bidirectional Encoder Representations from Transformers (BERT)

An MLP is a feed-forward artificial neural network composed of an input layer, one or more hidden layers, and an output layer. MLPs are based on the principle of modeling complicated connections in data using non-linear transformations over many layers. Every layer consists of neurons that use activation functions to process inputs from previous layers and transfer outputs to the following layer [18]. MLPs, often referred to as universal approximators, can be used to solve a variety of problems, such as non-linearly separable classification, regression, and prediction tasks [13]. MLPs' efficacy is derived from their capacity to identify patterns and relationships in data by adjusting weights during the training process, which is typically achieved using a back propagation algorithm [40].

4.3. Sentimnt Features (SFs)

Sentiment analysis, or opinion mining, is a field of NLP that focuses on identifying the polarity of textual data [37]. The objective is to discern the sentiments conveyed in a text, often classifying them as positive, negative, or neutral. The area has expanded considerably due to the growth of social media and big data, offering significant insights for organizations, marketers, and academics [30]. Sentiment analysis may be approached via two main ways: lexicon-based methods, which rely on preset dictionaries of words with known sentiment, and machine learning techniques, which include training algorithms on annotated datasets to identify patterns of sentiment [28]. Development in DL and emotion detection has improved sentiment analysis, allowing for advanced and specific interpretations of textual data. The produced

data indicating the sentiment of the text is often referred to as SFs. This phenomenon has been used in several Arabic studies [1, 39].

In this research, the Arabic NLP tool from the CaMeL Lab, named CaMeL tools, was used [33]. This is a complete set of NLP tools that have been developed specifically for the Arabic language. The CAMeL Lab, located at New York University Abu Dhabi, is dedicated to enhancing comprehension and analysis of Arabic texts using cutting-edge NLP methods. It has been widely adapted in Arabic classification tasks, proposing beneficial performance [7, 8, 10]. The system incorporates a sentiment analysis module that assesses the emotional tone of the text, categorizing it as either positive, negative, or neutral. A Python script was composed to leverage the CAMeL Tools library in performing the sentiment labeling for each tweet. Through utilizing the script, specifically, the pre-trained SentimentAnalyzer model recognized the sentiment of the text into three distinct types: positive, negative, and neutral. Further, a sentiment score of 1.0 is assigned to the defined sentiment category, while the scores for the other two categories are set to 0.0. The results showcase how each text entry is examined for sentiment, with scores being assigned accordingly. For instance, if a tweet is identified as having a positive sentiment, it will display scores such as: Positive: 1.0, Negative: 0.0, and Neutral: 0.0. This process highlights the efficiency of CaMeL tools in analyzing the sentiment within Arabic text, transforming the data into valuable insights by categorizing it according to sentiment. Figure 4 displays a sample of tweets with their sentiment results produced by CaMeL. Uniquely, it has the ability to recognize and distinguish between various Arabic dialects, which is specifically significant for processing the tweets in the dataset that contain various dialects. To ensure an extensive evaluation, SF were attentively extracted for each tweet of the three datasets.

tweet	label	Positive	Negative	Neutral
المسلم فقط دمه مباح افيقوا باقوم البعض ايضا لام سليمان الحلبي ع	EXTR	0	1	0
سيبقى الوطن شامخاً بك وبطموحك الذي يعانق السماء سلمان سعد	No_EXTR	1	0	0
وصلت التكفير ليش عاد شنو بدر منك بس لانك ما تحب الوهابية ص	No_EXTR	0	1	0
الناس تسلم شرقا وغربا ونحن تولى أمرنا خونة يحاربون الدين السيسي	EXTR	0	1	0
انا عراقي ومصلحة العراق فوق أي اعتبار انا لا احب أمريكا ولا حكومة ا	EXTR	0	1	0
ضد شخص ماكرون مثل ما فعلوا مع السيسي و قالوا انه عدو الله لان	EXTR	0	1	0
ضغطته اكثر ورسلت له صورة محمد بن سلمان عجبي مسجد من ال	No_EXTR	0	1	0

Figure 4. Sentiment features extraction results.

4.4. Model Development

In this section, an overview of the four developed models' structural details to detect Arabic extremist posts are explained:

Approach 1: AraBERT

To employ the extremist detection task, we opted for the pre-trained AraBERT base model [11]. It was trained on 8.6 billion words and contains a size range of 136 million parameters. Moreover, it consists of 12 transformer layers integrated into the encoder stack, along with 12 attention heads and a hidden size of 768. The following steps describe the AraBERT compilation:

- *Tokenization and Padding*: are two important techniques in NLP. Tokenization refers to the process of breaking down a text into individual words or tokens. Padding, on the other hand, involves adding extra characters or tokens. The BERT tokenizer from the transformer's library was used to tokenize and pad textual data. Specifically, the 'aubmindlab/bert-base-arabertv02'from the Hugging Face platform was applied for this task. The sequences were standardized to have the same length, with a maximum length of 100 tokens.
- *Encoding Categorical Variables Using Label Encoding*: the categorical labels were converted into numerical values using the 'LabelEncoder' from 'sklearn.preprocessing' to enable numerical processing for the classification model.
- *Architectural Design*: the 'TFBertModel,' based on the BERT model, was opted as the main technique for sequence processing. The inputs consisted of tokenized sequences (input_ids) and attention masks (attention_mask) to emphasize the relevant tokens for classification tasks.
- *Multi-Head Attention Layer*: the model's ability to recognize connections among tokens was enhanced by the inclusion of a multi-head attention layer. The number of attention heads was set at six, with a key dimension of 64. The model would simultaneously concentrate on numerous components of the sequence due to its self-attention mechanism.
- *Layer Normalization and Pooling*: a normalization layer was incorporated after the attention mechanism to enhance the stability of the training process and facilitate convergence. Subsequently, a global average pooling layer was used to combine the sequence embedding into a constant-sized vector, thereby capturing all components of the input sequence.
- *Classification Layer*: the final classification layer included a dropout layer with a dropout rate of 0.3 to mitigate overfitting. The output layer consisted of a dense layer that applied a Softmax activation function, which generated probabilities for each class within the label set.

Approach 2: AraBERT Incorporates MLP

This model is similar to the model in approach 1, however, following the normalizing and pooling layers, an MLP was added to increase the model's performance. The flattened attention vector was sent through a sequence of dense layers of the MLP. The first dense layer consisted of 32 units and used a Rectified Linear Unit (ReLU) activation function. This was then followed by a dropout layer with a dropout rate of 0.3, which was implemented to address the issue of

overfitting. Subsequently, an additional dense layer with 8 units and ReLU activation was applied, followed by another dropout layer.

Approach 3: MLP Incorporated SF

In this approach, a simple MLP classifier trained by SF was compiled. Specifically, SF were transformed into a numeric array for employment, where the MLP was developed to process the numerical features. The input layer of the MLP was configured to incorporate the same dimensions as the numerical feature array. In terms of compilation, the first dense layer consisted of 64 units and used a ReLU activation function. This was then followed by a dropout layer with a dropout rate of 0.1. The output layer consisted of a dense layer and integrated a Softmax activation function, yielding a probability distribution across the various classes.

Approach 4: AraBERT Incorporated with SF and MLP

Similar to the earlier AraBERT model discussed in approach 1, we incorporate the SF and an MLP layer with AraBERT in this model, illustrated in Figure 5. Notably, the pooled attention vector was concatenated with the numerical features following the normalization and pooling layer to provide a composite feature representation. The combined representation was then fed into multiple dense layers inside the MLP. The first dense layer consisted of 32 units and used a ReLU activation function. This was then followed by a dropout layer with a dropout rate of 0.1, which was implemented to mitigate the risk of overfitting. Additionally, a second layer with a dense of 8 units and ReLU activation was implemented, which was then followed by a dropout layer. The ultimate classification layer included a dense layer fitted with a Softmax activation function, which produced the probability distribution across the various classes.

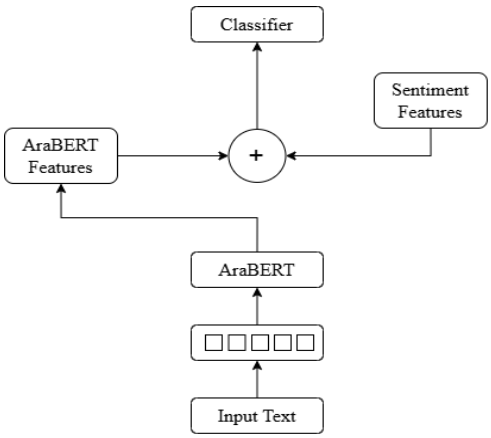

Figure 5. A visualization for approach 4 model.

5. Experimental Setup

The models underwent training and testing by employing both the Arabic Extremist and Annotated ISIS Radical Tweets datasets. The first dataset included 3320 tweets divided into 1660 tweets labeled as 'extremist' and 1660 labeled as 'non-extremist.' A ratio of 70%, 2352 instances, was set for the training of the dataset. Further, a ratio of 15%, 484 instances, for validation, and 15%, 484 instances, for testing was set to ensure fair evaluations. On the other hand, the latter dataset contained 1700 tweets labelled as 'extremist' and 1700 labelled as 'non-extremist.' Similar to the first dataset, a ratio of 70%, 2380 tweets, was set for training 15 % for validation, and 15% for testing, 510 tweets for each, respectively. For all datasets, early stopping to monitor the validation loss and minimize overfitting was configured. The experiments were conducted on a computer running 64-bit Windows 13.04.3 LTS with an Intel (R) Xeon (R) 2.00 GHz processor, 12 GB RAM, and an NVIDIA T80 GPU with 12 GB memory. All codes for this study were developed utilizing Keras 2.3.1 in the Python programming language.

The models employing AraBERT undertook 3 epochs, and the one employing MLP with SF undertook 50 epochs. The reason for the number of epochs variance between the latter models is due to the average dataset number and the fact that AraBERT is highly pre-trained, hence higher epochs might cause overfitting. All models utilized a varying number of batches for training for each dataset. We maintain the batch size at 32, being observant of the GPU memory limits, with consideration of sizes of training data, training accuracy, and loss using validation data. The hyperparameters tuning involved selecting the Adam optimizer and 1e-5 learning rate, to enable AraBERT to tackle the forgetting problem, which is justified by that the use of higher learning rates, such as 4e-4, can result in the failure of consolidation on the training set [36]. Utilizing a reduced learning rate might allow the model to acquire a more ideal set of weights. Table 2 summarizes the model's hyperparameters.

Table 2. Hyperparameters of compiled models.

Model	Hyperparameters
AraBERT	Loss=sparse categorical_crossentropy, Learning rate=1e-5. Optimizer=Adam, Epochs=3, Batch_size=32
MLP	Loss=sparse categorical_crossentropy, Learning rate=1e-5. Optimizer=Adam, Epochs=50, Batch_size=32

To assess the models' performance in all experiments, we employed standard metrics such as accuracy, precision, recall, and F1-scores, presented in formulas displayed in equations 1 to 4. These metrics were calculated by comparing expected and measured results, enabling the analysis of prediction accuracy within the training sample. The classification into four groups-True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN)-facilitated the derivation of these measures. Moreover, to thoroughly assess the models' performance in terms of TP rate versus FP rate across different threshold settings, we display the Receiver Operating Characteristic (ROC)

curves for each model when tested on the unseen dataset.

6. Results and Discussion

This section delineates the achieved results for the Arabic extremist and ISIS Radical tweets datasets, and then for the Kazakh extremist dataset as a cross-platform test dataset. This is followed by a discussion based on the results of the proposed models. Additionally, it will encompass a comparison between our proposed model and other models aimed at solving the same issue.

6.1. Results on Arabic Extremist

This section compares several approaches for classifying tweets into Extremist (E) and Non-Extremist (NE) categories. According to Tables 3 and 4, the AraBERT model performed adequately when defining extremist content, with a precision of 94%, recall of 92%, F1-score of 94%, and total accuracy of 94%. For non-extremist labeling, it maintained a 94% accuracy, 93% recall, and 93% F1-score. Moreover, the model developed by AraBERT integrated MLP resulted in considerably improved performance. This model attained a precision of 96%, recall of 98%, F1-score of 96%, accuracy of 96% for extremist content, and a constant accuracy of 94%, recall of 93%, and F1-score of 93% for non-extremist labels. A valid explanation for this increase would be that MLPs might have the ability to convey additional textual features in furtherance to the embeddings supplied by AraBERT. This interaction implies the model to capitalize on a richer range of features, possibly recognizing nuances in the text that AraBERT's embeddings individually may overlook. With this broader perspective, the model's performance to effectively classify tweets may be beneficial.

Table 3. Extremist and non-extremist detection models performance for Arabic extremist dataset.

Models	Class	Precision	Recall	F1-score	Accuracy
AraBERT	Extremist	94	92	94	94
	Non-Extremist	94	93	93	
AraBERT+MLP	Extremist	96	98	96	96
	Non-Extremist	94	93	93	
MLP+SF	Extremist	87	81	84	84
	Non-Extremist	82	88	85	
AraBERT+SF +MLP	Extremist	97	96	96	98
	Non-Extremist	96	96	98	

Interestingly, the MLP model converged with SF, performed more prevalent than AraBERT-based models. It scored accuracy of 84% and 82% for the extremist and non-extremist categories, respectively. This could be due to the fact although sentiment assessments may be challenging owing to the details and intricacies of human language, these simple sentiments may identify the general sentiment albeit fail to recognize nuanced expressions such as idiomatic and sarcastic terms that might be found in extremist material. So, excluding the AraBERT from this model reduces the model's rich contextual knowledge, lowering its performance.

Inevitably, the model compiled by concatenating AraBERT with SF, and further integrated MLP outperformed all other models. This model obtained 97% precision, 96% recall, 96% F1-scores, and 98% accuracy for extremist material. As for the non-extremist class, it maintained 96% accuracy, 96% recall, and 98% F1-score. In devoutly, this model demonstrated superior performance compared to all other models, a justification to this would be that the combined approach of AraBERT, SF, and MLP optimizes the individual potential of each component. AraBERT presents the foundation for interpreting context, SF incorporates specific domain features, and the MLP exploits this rich collection of features to provide effective predictions. This synergy generates a model that is more robust to capture nuanced variances in the language found in extremist and non-extremist tweets. Our models performed better than the models employed by CamemBERT in the study by Dragos and Constable [20], which reached an accuracy of 75%. Moreover, it outscored the best-performing model in [31] of Gradient Boost with Word2Vec, whose F1-score achieved 86%. As for models to detect extremist content in Arabic, our optimal model achieved higher accuracy than the BERT model compiled by Aldera *et al*. [5], which scored 97.4. In its entirety, these findings indicate that incorporating SF with AraBERT and MLP, may successfully strengthen the models' performance to detect extreme textual content.

6.2. Results on Annotated ISIS Radical Tweets Datasets

This section evaluates multiple models in classifying tweets from the Annotated ISIS Radical Tweets Dataset into Extremist (E) and Non-Extremist (NE) categories. As illustrated in Table 4, the AraBERT model provided moderate performance in detecting extremist content, achieving a precision of 76%, recall of 69%, F1-score of 75%, and an overall accuracy of 75%. For non-extremist classification, it maintained an accuracy of 74%, recall of 80%, and F1-score of 77%. While AraBERT demonstrated reasonable effectiveness, its performance indicates potential limitations in capturing subtle linguistic variations within this dataset.

Table 4. Extremist and non-extremist detection models performance for annotated ISIS radical Tweets dataset.

Models	Class	Precision	Recall	F1-score	Accuracy
AraBERT	Extremist	76	69	75	75
	Non-Extremist	74	80	77	
AraBERT+MLP	Extremist	87	81	85	86
	Non-Extremist	78	83	79	
MLP+SF	Extremist	84	81	80	81
	Non-Extremist	81	80	81	
AraBERT+SF +MLP	Extremist	87	82	84	88
	Non-Extremist	91	88	90	

Significantly, integrating AraBERT with MLP resulted in noticeable improvements across all metrics. This combined model attained an accuracy of 86% in detecting extremist content, with precision of 87%, recall of 81%, and an F1-score of 85%. For non-extremist tweets, it reached an accuracy of 78%, recall of 83%, and an F1-score of 79%. The performance enhancement suggests that MLP introduces an additional layer of abstraction, allowing the model to identify intricate textual features that AraBERT alone might not effectively capture. This highlights the advantage of leveraging neural network architectures to refine contextual embeddings and optimize classification.

Interestingly, the standalone MLP combined with SF model delivered comparable performance, achieving an accuracy of 81% for both extremist and non-extremist tweets. With a precision of 84% and recall of 81% for extremist content, and an F1-score of 80%, this model slightly lower than the AraBERT+MLP combination. A plausible explanation for this is that while Sentiment-based Features (SFs) contribute valuable indicators of radical content, their effectiveness is constrained by the complexity of extremist rhetoric, which often employs implicit or coded language. Without AraBERT's contextual embeddings, the model may struggle to fully comprehend nuanced expressions.

Ultimately, the best-performing model was AraBERT combined with SF and MLP, attaining an overall accuracy of 88%. It exhibited superior precision 87% and recall 82% for extremist content, leading to an F1-score of 84%. The robustness of this model underscores the advantages of a hybrid approach, where AraBERT extracts deep contextual information, SF contributes domain-specific insights, and MLP synthesizes these inputs into more refined predictions. The synergy among these components strengthens the model's capability to differentiate between extremist and non-extremist language effectively.

Overall, our findings reaffirm that integrating SF with AraBERT and MLP enhances model performance in detecting extremist textual content, particularly in challenging datasets such as the Annotated ISIS Radical Tweets dataset.

6.3. Results on Cross-Platform Test Dataset

To further evaluate the robustness and generalizability of the model, we conducted additional testing using a cross-platform test dataset. To ensure data compatibility between both datasets, this dataset endured the same preprocessing and SF extraction steps as the Arabic extremist dataset. Testing the performance of these models on this dataset allows us to verify the model's performance beyond the training and validation. It validates its applicability to real-world extremist tweets.

The whole dataset was inputted into each model as a cross-platform test dataset and each model was evaluated. The models' detection performance achieved accuracies ranging from the lowest produced by MLP trained on SF reaching 59% and the highest achieving 81% of the AraBERT assimilated with SF and MLP. In support, ROC is highest in the model AraBERT with embedded SF and MLP and the model AraBERT incorporated with MLP, reaching 91%. Further, the AraBERT model generates ROC of 85%, in contrast to the MLP trained with SF which produced the lowest ROC of 55%, as exhibited in Figure 6.

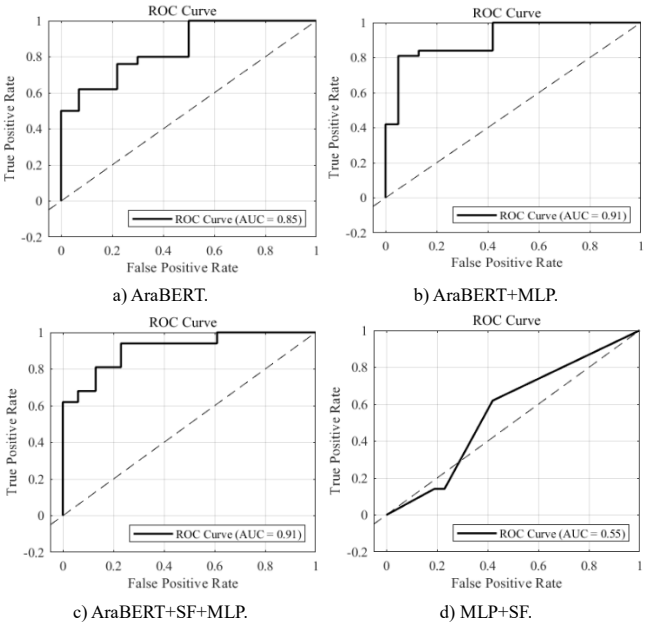

Figure 6. ROC of models on unseen test dataset.

6.4. Important Features

We tested the previous findings by calculating the Information Gain (IG) to determine the important features. IG is a metric used in decision trees to estimate the efficacy of a feature when dividing the dataset into distinct groups. The calculation determines the decrease in entropy, which represents the level of uncertainty of the target variable (class labels) when a feature has been identified. Essentially, IG serves to determine the rate at which a certain feature affects the accuracy of predictions in a model [38]. Features with higher IG are considered to be more effective on the models' performance.

According to Figure 7, the top 20 important features, ranked according to the highest IG score were various words that pertained to insights into the tone found in the extremist text. First, the frequent use of explicit words denoting negativity such as boycott مقاطعه, infidel كافر, and enemy عدو. These terms underscore the antagonistic essence of extreme discourse, commonly used in such negative posts expressing anger. On the other hand, encouragement positive words such as trust ثقه and covenant العهد focus on loyalty and trustworthiness among extremist societies, which is often utilized to foster solidarity and support ideological narratives.

Also, an expected finding was terms that referred to political figures such as Salman سلمان, AlSisy (Egyptian president) السيسي, Mohammed محمد with instances of indirect references to some figures to evade extremist text detection.

In addition to these highly used terms, powerful features in the form of sentiment classes, specifically the features of positive and negative sentiment, are found to be vital in identifying extremist text. As demonstrated earlier, words denoting violence and anger found to express negativity would fully fit with negative sentiment. Moreover, words that express glorification and loyalty, would be aligned with the positive sentiment. Through comprehensive sentiment definition for extremist content, the analysis thoroughly recognizes the complex nature of extremist content by collecting both explicit and implicit terms, together with sentiment-driven context.

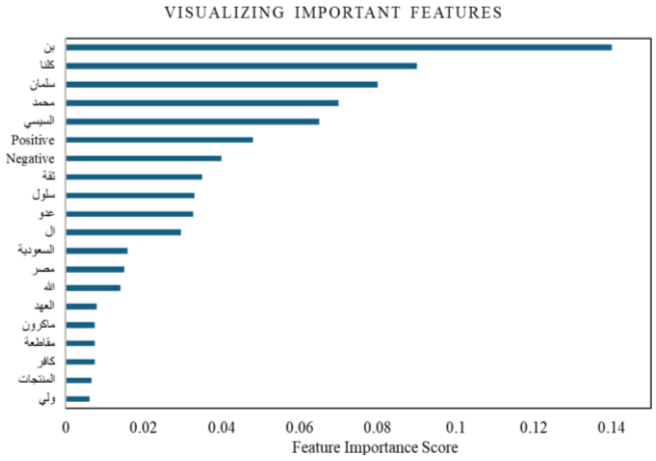

Figure 7. Top 20 important features.

7. Limitations and Future Works

Although the research has shown encouraging findings, it is important to acknowledge that there are some limitations. Firstly, the dataset, while substantial, may not include all forms of extremist language, especially those involving implicit and encoded extremism. Additionally, the dependability on SFs may indicate that extremist content always expresses explicit sentiments expresses views. However, with the rapid advances in technology, images, and speech may be incorporated to convey their messages, which makes the SF inadequate to tackle this issue. Further research is needed to conquer these limitations and improve the model's ability to be employed in wider extremist detection platforms. For future research, we also plan to explore various datasets and enhance our approach using AI techniques to detect extremist content on social media.

8. Conclusions

This study efficiently developed an innovative model that detects extremist texts on social media by leveraging AraBERT and employing SF and MLP. The model demonstrated superior results, with a noteworthy accuracy score of 98% for detecting extremist texts. The model's generalizability and robustness were further validated by testing it on a cross-platform test dataset that featured extremist messages in the Kazakh language that were translated into Arabic. The model achieved good performance, with an accuracy of 81%, therefore corroborating its applicability for real-world extremist texts. The incorporation of SF had an important impact in enhancing the model's capability to recognize sentiment tones in extremist messages, thereby strengthening its detection. Additionally, the study highlights the effectiveness of various compiled models such as AraBERT and improving its performance by integrating MLP, reaching an accuracy of 96%. In conclusion, this research provides vital findings and effective models that contribute to safety online and deter the dissemination of extremist textual content. We aim for this study to provide a valuable foundation for future research, especially in developing accurate models for detecting online extremism in the Arabic language.

References

[1] Abdul-Mageed M., Diab M., and Kubler S., "SAMAR: Subjectivity and Sentiment Analysis for Arabic Social Media," *Computer Speech and Language*, vol. 28, no. 1, pp. 20-37, 2014. https://doi.org/10.1016/j.csl.2013.03.001

[2] Ahmad S., Asghar M., Alotaibi F., and Awan I., "Detection and Classification of Social Media-based Extremist Affiliations Using Sentiment Analysis Techniques," *Human-Centric Computing and Information Sciences*, vol. 9, pp. 1-23, 2019. https://doi.org/10.1186/s13673-019-0185-6

[3] Ahmed A., Hasan M., Jaber M., Al-Ghuribi S., Abd D., Khan W., Sadiq A., and Hussain A., "Extremism Arabic Text Detection Using Rough Set Theory: Designing a Novel Approach," *IEEE Access*, vol. 11, pp. 68428-68438, 2023. DOI:10.1109/ACCESS.2023.3278272

[4] Alatawi H., Alhothali A., and Moria K., "Detection of Hate Speech using BERT and Hate Speech Word Embedding with Deep Model," *Applied Artificial Intelligence*, vol. 37, no. 1, pp. 384-405, 2023. https://doi.org/10.1080/08839514.2023.2166719

[5] Aldera S., Emam A., Al-Qurishi M., Alrubaian M., and Alothaim A., "Exploratory Data Analysis and Classification of a New Arabic Online Extremism Dataset," *IEEE Access*, vol. 9, pp. 161613-161626, 2021. DOI:10.1109/ACCESS.2021.3132651

[6] Aldera S., Emam A., Al-Qurishi M., Alrubaian M., and Alothaim A., Annotated Arabic Extremism Tweets, IEEE Dataport,

https://dx.doi.org/10.21227/g9c0-1t21, Last Visited, 2024.

[7] Aldumaykhi A., Otai S., and Alsudais A., "Comparing Open Arabic Named Entity Recognition Tools," *in Proceedings of the 24th International Conference on Information Reuse and Integration for Data Science*, Bellevue, pp. 46-51, 2023. https://ieeexplore.ieee.org/document/10229342

[8] Alfaidi A., Alwadei H., Alshutayri A., and Alahdal S., "Exploring the Performance of Farasa and CAMeL Taggers for Arabic Dialect Tweets," *The International Arab Journal of Information Technology*, vol. 20, no. 3, pp. 349-356, 2023. https://doi.org/10.34028/iajit/20/3/7

[9] Al-Khalifa H., Magdy W., Darwish K., Elsayed T., and Mubarak H., "Overview of OSACT4 Arabic Offensive Language Detection Shared Task," *in Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, Marseille, pp. 48-52, 2020. https://aclanthology.org/2020.osact-1.0/

[10] Alluhaibi R., Alfraidi T., Abdeen M., and Yatimi A., "A Comparative Study of Arabic Part of Speech Taggers Using Literary Text Samples from Saudi Novels," *Information*, vol. 12, no. 12, pp. 1-13, 2021. https://doi.org/10.3390/info12120523

[11] Antoun W., Baly F., and Hajj H., "AraBERT: Transformer-based Model for Arabic Language Understanding," *in Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, Marseille, pp. 9-15, 2020. https://aclanthology.org/2020.osact-1.2.pdf

[12] Berhoum A., Meftah M., Laouid A., and Hammoudeh M., "An Intelligent Approach Based on Cleaning up of Inutile Contents for Extremism Detection and Classification in Social Networks," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 5, pp. 1-20, 2023. https://doi.org/10.1145/3575802

[13] Bisong E., *Building Machine Learning and Deep Learning Models on Google Cloud Platform, A Comprehensive Guide for Beginners*, Apress, 2019. https://link.springer.com/chapter/10.1007/978-1-4842-4470-8_31

[14] Canete J., Chaperon G., Fuentes R., Ho J., Kang H., and Perez J., "Spanish Pre-Trained BERT Model and Evaluation Data," *arXiv Preprint*, vol. arXiv:2308.02976v1, pp. 1-9, 2020. https://doi.org/10.48550/arXiv.2308.02976

[15] Chan T., Schweter S., and Moller T., "German's Next Language Model," *arXiv Preprint*, vol. arXiv:2010.10906, pp. 1-9, 2020. https://arxiv.org/pdf/2010.10906

[16] Chouikhi H., Chniter H., and Jarray F., "Arabic Sentiment Analysis Using BERT Model," *in Proceedings of the 13th International Conference on Advances in Computational Collective Intelligence*, Kallithea, pp. 621-632, 2020. https://doi.org/10.1007/978-3-030-88113-9_50

[17] Cohen K., Johansson F., Kaati L., and Mork J., "Detecting Linguistic Markers for Radical Violence in Social Media," *Terrorism and Political Violence*, vol. 26, no. 1, pp. 246-256, 2014. https://doi.org/10.1080/09546553.2014.849948

[18] Da Silva I., Spatti D., Flauzino R., Liboni L., Dos Reis Alves S., *Artificial Neural Networks: A Practical Course*, Springer, 2017. https://link.springer.com/chapter/10.1007/978-3-319-43162-8_5

[19] Devlin J., Chang M., Lee K., and Toutanova K., "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," *in Proceedings of the* NAACL-HLT, Minneapolis, pp. 4171-4186, 2019. https://aclanthology.org/N19-1423.pdf

[20] Dragos V. and Constable Y., "Comparison of Classification Techniques for Extremism Detection in French Social Media," *in Proceedings of the 26th International Conference on Information Fusion*, Charleston, pp. 1-7, 2023. https://hal.science/hal-04313505

[21] Fraiwan M., "Identification of Markers and Artificial Intelligence-based Classification of Radical Twitter Data," *Applied Computing and Informatics*, pp. 1-13, 2022. https://doi.org/10.1108/ACI-12-2021-0326

[22] Gaikwad M., Ahirrao S., Phansalkar S., and Kotecha K., "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

[23] Gelber K., "Terrorist-Extremist Speech and Hate Speech: Understanding the Similarities and Differences," *Ethical Theory and Moral Practice*, vol. 22, no. 3, pp. 607-622, 2019. https://doi.org/10.1007/s10677-019-10013-x

[24] Himdi H. and Assiri F., "Tasaheel: An Arabic Automative Textual Analysis Tool-All in One," *IEEE Access*, vol. 11, pp. 139979-139992, 2023. DOI:10.1109/ACCESS.2023.3340520

[25] Jamil M., Pais S., Cordeiro J., and Dias G., "Detection of Extreme Sentiments on Social Networks with BERT," *Social Network Analysis and Mining*, vol. 12, no. 1, pp. 1-16, 2022. https://doi.org/10.1007/s13278-022-00882-z

[26] Kadhim A., "An Evaluation of Preprocessing Techniques for Text Classification," *International Journal of Computer Science and Information Security*, vol. 16, no. 6, pp. 22-32, 2018.

https://www.academia.edu/36998792/An_Evaluation_of_Preprocessing_Techniques_for_Text_Classification

[27] Lipset M., "Social Stratification and 'Right-Wing Extremism," *The British Journal of Sociology*, vol. 10, no. 4, pp. 346-382, 1959. https://doi.org/10.2307/587800

[28] Liu B., *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*, Cambridge University Press, 2015. https://books.google.jo/books?id=PdX7DwAAQBAJ&printsec=frontcover&redir_esc=y#v=onepage&q&f=false

[29] Martin L., Muller B., Suarez P., Dupont Y., Romary L., De la Clergerie E., Seddah D., and Sagot B., "CamemBERT: A Tasty French Language Model," *in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Seattle, pp. 7203-7219, 2020. https://aclanthology.org/2020.acl-main.645.pdf

[30] Mohd M., Javeed S., Nowsheena, Wani M., and Khanday H., "Sentiment Analysis Using Lexico-Semantic Features," *Journal of Information Science*, vol. 50, no. 6, pp. 1449-1470, 2020. https://doi.org/10.1177/01655515221124016

[31] Mussiraliyeva S., Bolatbek M., Omarov B., and Bagitova K., "Detection of Extremist Ideation on Social Media Using Machine Learning Techniques," *in Proceedings of the 12th International Conference* on *Computational Collective Intelligence*, Da Nang, pp. 743-752, 2020. https://link.springer.com/chapter/10.1007/978-3-030-63007-2_58

[32] Mussiraliyeva S., Omarov B., Yoo P., and Bolatbek M., "Applying Machine Learning Techniques for Religious Extremism Detection on Online User Contents," *Computers, Materials and Continua*, vol. 70, no. 1, pp. 915-934, 2022. https://doi.org/10.32604/cmc.2022.019189

[33] Obeid O., Zalmout N., Khalifa S., Taji D., Oudah M., Alhafni B., Inoue G., Eryani F., Erdmann A., and Habash N., "CAMeL Tools: An Open Source Python Toolkit for Arabic Natural Language Processing," *in Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, pp. 7022-7032, 2020. https://aclanthology.org/2020.lrec-1.868.pdf

[34] Rajendran A., Sahithi V., Gupta C., Yadav M., Ahirrao S., Kotecha K., Gaikwad M., Abraham A., Ahmed N., and Alhammad S., "Detecting Extremism on Twitter During U.S. Capitol Riot Using Deep Learning Techniques," *IEEE Access*, vol. 10, pp. 133052-133077, 2022. DOI:10.1109/ACCESS.2022.3227962

[35] Sudheesh R., Mujahid M., Rustam F., Mallampati B., Chunduri V., De la Torre Diez and I., Ashraf I., "Bidirectional Encoder Representations from Transformers and Deep Learning Model for Analyzing Smartphone-Related Tweets," *PeerJ Computer Science*, vol. 9, pp. e1432, 2023. https://doi.org/10.7717/peerj-cs.1432

[36] Sun C., Qiu X., Xu Y., and Huang X., "How to Fine-Tune BERT for Text Classification?," *in Proceedings of the 18th China National Conference* on *Chinese Computational Linguistics*, Kunming, pp. 194-206, 2019. https://doi.org/10.1007/978-3-030-32381-3_16

[37] Taboada M., "Sentiment Analysis: An Overview from Linguistics," *Annual Review of Linguistics*, vol. 2, pp. 325-347, 2016. https://doi.org/10.1146/annurev-linguistics-011415-040518

[38] Tangirala S., "Evaluating the Impact of GINI Index and Information Gain on Classification Using Decision Tree Classifier Algorithm," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 2, pp. 612-619, 2020. DOI:10.14569/IJACSA.2020.0110277

[39] Tartir S. and Abdul-Nabi I., "Semantic Sentiment Analysis in Arabic Social Media," *Journal of King Saud University-Computer and Information Sciences*, vol. 29, no. 2, pp. 229-233, 2017. https://doi.org/10.1016/j.jksuci.2016.11.011

[40] Taud H. and Mas J., *Geomatic Approaches for Modeling Land Change Scenarios*, Springer, 2018. https://doi.org/10.1007/978-3-319-60801-3_27

[41] Torregrosa J., Bello-Orgaz G., Martinez-Camara E., Del Ser J., and Camacho D., "A Survey on Extremism Analysis Using Natural Language Processing: Definitions, Literature Review, Trends and Challenges," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 8, pp. 9869-9905, 2023. https://doi.org/10.1007/s12652-021-03658-z

[42] Torregrosa J., Thorburn J., Lara-Cabrera R., Camacho D., and Trujillo H., "Linguistic Analysis of Pro-ISIS Users on Twitter," *Behavioral Sciences of Terrorism and Political Aggression*, vol. 12, no. 3, pp. 171-185, 2020. https://doi.org/10.1080/19434472.2019.1651751

[43] Ul Rehman Z., Abbas S., Khan M., Mustafa G., Fayyaz H., Hanif M., and Saeed M., "Understanding the Language of ISIS: An Empirical Approach to Detect Radical Content on Twitter Using Machine Learning," *Computers, Materials and Continua*, vol. 66, no. 2, pp. 1075-1090, 2021. https://doi.org/10.32604/cmc.2020.012770

[44] Watanabe H., Bouazizi M., and Ohtsuki T., "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection," *IEEE Access*, vol. 6, pp. 13825-13835, 2018.

DOI: 10.1109/ACCESS.2018.2806394

Hanen Himdi is an Assistant Professor of Computer Science and Artificial Intelligence in the College of Computer Science and Engineering, University of Jeddah, KSA. She is a computer scientist with a Ph.D. degree in Computer Science from the University of Strathclyde, Scotland, UK. Her research interests are machine learning, natural language processing, and textual analysis. Her current research interests lie in the area of deep learning and the creation of AI models that make use of cutting-edge learning techniques.

Fatimah Alhayan is an Assistant Professor of Computer Science in the College of Computer and Information Science at Princess Noura University, Saudi Arabia. Holding a Ph.D. in Computer Science from the University of Strathclyde, Scotland, UK. Her research interests include Information Credibility, Data Mining, Computational Social Science, Machine Learning, and Natural Language Processing (NLP) in both English and Arabic languages.
Photo:

Khaled Shaalan is a Prof. Khaled Shaalan currently occupies the Co-Chair of the Faculty of Engineering and IT position at The British University in Dubai, UAE. He is currently holding the rank of a Full Professor of Computer Science and AI. He has gained significant academic experience and insights into understanding complex ICT issues in many industrial and governmental domains through a career and affiliation spanning for more than 30 years. Areas of interest are Artificial Intelligence (AI), Natural Language Understanding, Knowledge Management, Health Informatics, Education Technology, E-businesses, cybersecurity, and Smart Government Services. He is ranked among the worldwide 2% top scientists till now according to a study led by Dr Ioannidis and his research team at Stanford University. He is also ranked as one of the Top Computer Scientists in the UAE according to the Research.comindex.