A Smart Transformer-Based Model to Classify Extremist Behavior on Social Media

Tamer Salah*, Hazem M. El-Bakry, Amira Rezk
Information Systems Department-Faculty of Computers and Information, Mansoura University, Mansoura, Egypt

***Abstract*—In recent years, social media has been used as a fertile environment for spreading radical ideologies through extremist organizations and groups to recruit new individuals capable of implementing their schemes. This has led to an enormous amount of extremist data that can be used by research agencies to analyze and understand the behavioral and linguistic characteristics associated with extremist language, thereby contributing to the development of policies and laws that help to protect societies from this threat. This study proposes a smart transformer-based model for collecting and analyzing data on extremism from social media to be classified based on whether or not it relates to extremism. The proposed framework begins by collecting data from Twitter by using keywords related to extremism and then filtering and reprocessing the data to implement the word embedding process that relies on a bi-directional LSTM network to provide the best representation vector for each word in the input sentence that can provide the true meaning of each word based on its position in the input sentence. Here comes the importance of extracting linguistic and semantic features of the text by building a matrix of features that can be extracted from the text, divided into three categories related to surface features, polar features, and specific domain features. Finally, after effectively preparing data and extracting features, the vector representing each sentence is inserted into the proposed smart transformer-based model to carry out the classification task, as the development steps focus on increasing the model's ability to raise the standard weight of words related to extremism in the inserted sentence, thereby contributing to increasing the accuracy of the model to classify the text. The results demonstrated the ability of the proposed model to achieve a high accuracy rate on the F1-score benchmark of 91.6% as well as 96.7% on the accuracy standard. The proposed model was able to outperform the accuracy of 5 models of other algorithms that achieved high results in the classification task, exceeding the closest competition of SVM-Linear by 0.7% at the F1-score benchmark and 0.5% at the overall rating accuracy level.**

***Keywords*—Social media; transformer-based model; word embedding; Bi-LSTM network; additive attention**

I. Introduction

In recent years, technology has evolved significantly, helping the data sector grow exponentially. With so much data, especially on social media, there is a need to develop models, algorithms, and methods to analyse these data and discover hidden relationships between them, which helps exploit these data to help decision makers in many sectors of business and medicine, as well as political decisions for the benefit of the beneficiaries.

The spread of extremist ideology in recent years in many countries of the world has led many researchers to try to understand how extremists think [12] to try to deal with them more efficiently and provide safety for societies. The rebels' use of technology has led to greater sympathy and the dissemination of their thinking, resulting in increased availability of extremist data on social media sites, a dynamic that has also been linked to the fund-raising and disruption strategies used by extremist organisations [13].

The sheer volume of available data has increased the need to develop models and algorithms for data analysis, especially text data, a need that a systematic review of the online extremism detection literature has traced across datasets, classification techniques, and validation methods [11]. The development of these algorithms aims to help the machine to understand texts like humans through natural language processing techniques (NLP). Deep learning models have evolved to achieve excellent results in terms of model performance of models, especially for text classification tasks.

A new model based on a pre-trained transformer-based learning model is designed to analyse extremism data collected from the Twitter platform, where the dataset was collected using several keywords associated with extreme behaviour. The dataset has been reprocessed to help extract semantic and linguistic features, helping to provide the best possible representation vector for the input text by using a bidirectional LSTM network to pay more attention to the meaning of the context of each word in the text. Ultimately, after discovering and expressing the features of the text well, each word's representation vector was introduced into the deep learning network based on smart transformers to provide the best classification of the text in terms of extremism into three main classes: extremist, non-extremist, and neutral.

The contribution of the proposed model is to provide the best possible representation vector for each word in the input text, which reflects the best meaning of each word based on its order in the sentence. This process begins with the effective filtering and reprocessing of the text, then extracts the extremists' behaviours by analysing the linguistic and semantic characteristics of the input text based on many algorithms that extract the surface and polar features as well as specific features of the domain. The second contribution is architectural: rather than computing full pairwise self-attention across all token pairs as in standard transformer architectures, the proposed model compresses the query, key, and value representations into single global vectors via additive attention weighting, reducing the attention computation from the standard quadratic complexity to a linear one while retaining the ability to prioritize words most relevant to extremism. This global-vector additive-attention design is combined with a three-category engineered feature

*Corresponding author.

representation — surface, polarity, and domain-specific features, the latter derived from the HertLex extremism lexicon — that is extracted independently of the transformer's learned attention weights. This combination of an efficiency-oriented additive-attention mechanism with an explicit, domain-lexicon-grounded feature representation distinguishes the proposed architecture from prior transformer-based approaches to extremism and hate-speech detection ([6], [7], [10]), which rely on the transformer's learned representations alone without an explicit domain-lexicon feature layer.

This study is organized as follows. Section II presents an overview of previous work on dealing with the detection and classification of extremist text. Section III describes the materials and methods used to analyse extremist text classification tasks. Section IV presents our results, and Section V discusses them. Section VI outlines the limitations of the present study. Finally, Section VII offers our conclusions.

II. Related Work

A dynamic matrix has been proposed in [2] to classify extremism and terrorism (DMET) that is capable of analysing and classifying extremist data on social media based on identifying user behavior and cognitive signals and determining the level of extremism based on the dissemination of their ideologies and ideas. DMET is designed to help platform decision-makers who deal with extremist content in terms of preventing or disseminating it according to its degree and risks. The machine learning (ML) model was also proposed in [3] to analyze text content on social media platforms to discover the psychological and behavioral characteristics of the text. This model relies on adding improvements to the word embedding process to ensure a good representative vector to represent the input text that contributes to building a highly efficient ML model; the experimental results show that the proposed model has achieved high performance after being applied on the Twitter platform to detect extremist activities.

In [1], a calendar-correlation analysis (CCA) hybrid algorithm based on analysis of online communities was proposed to detect far-right communities by defining distinctive changes in the activities of these communities based on key dates associated with these communities. This hybrid algorithm was applied to the Russian social network VKontakte, and the results demonstrated the algorithm's ability to accurately identify far-right societies up to 84.3% compared to other models that performed less well in identifying these communities. The Ontology Framework to Facilitate Early Detection of 'Radicalization' (OFEDR) was developed in [4] for the early detection of extremist persons by detecting the link between the political extremism of persons and its compatibility with known criminal files. The framework was applied to five deep learning models and tested on the Protégé program designed by the University of Stanford for this purpose. The results demonstrated the ability of these models to achieve outstanding results with accuracy ranging from 86.3% to 91.2%. The results also demonstrated this framework's high ability to detect early signs of extremism using probability estimates for extremely rare events.

A broader survey of extremism analysis using natural language processing [5] has documented the range of methodological approaches applied to this problem, from qualitative ontology-building efforts drawing on populations of data collected across multiple social networking platforms such as Twitter, Facebook, and YouTube, to fully automated deep learning classification pipelines, and highlights persistent challenges around dataset scarcity, label disagreement, and the rapidly evolving vocabulary of extremist discourse, along with the corresponding trends toward data augmentation strategies designed to alleviate these limitations.

More recent work has continued to push transformer-based extremism and hate-speech detection forward. RADAR#, an ensemble approach combining a hybrid CNN-Bi-LSTM framework with the AraBERT transformer through a weighted strategy, achieved 98% accuracy and a 97% F1-score on a dataset of 89,816 Arabic tweets, and additionally reported ablation studies and attention visualizations to support interpretability [6]. A related hybrid deep learning study compared Bi-LSTM and DistilBERT-based transformer architectures for extremism classification, reporting binary classification accuracy of 97.33% and multiclass transformer accuracy of 91.07% [7]. Other recent studies have advanced the representation side of the problem, proposing multi-dimensional information representations for offensive language detection [8] and deep active learning strategies for identifying hate and offensive content across multilingual social media posts [9]. Closest in spirit to the attention-based contribution of the present study, an ontology-guided attention mechanism integrating RoBERTa embeddings with graph convolutional networks was shown to outperform baseline RoBERTa implementations on multiclass hate speech detection, with particularly strong gains on demographically challenging categories [10]. Collectively, these studies confirm that hybridizing transformer embeddings with either structured domain knowledge or additional attention mechanisms, as proposed in this study, remains an active and productive direction for extremist and hateful content classification.

III. Materials and Methods

A. Methodology

This study presents the proposed methodology for building a new framework based on transformer-based models to extract and analyse several text features to identify extreme views on social media platforms. This framework is based on the use of a bi-directional network of LSTM to provide a well-represented vector for input data capable of reflecting the best meaning of each of the listed sentence's words. The proposed architecture follows a two-stage pipeline: the Bi-LSTM layer first produces a context-aware embedding for each input word, and this embedding matrix then serves as the input to the transformer block, in which standard scaled dot-product self-attention is replaced by the additive-attention formulation described in detail below. On the other hand, the process of extracting features from text depends on three main axes: the extraction of surface features, the features of polarity, as well as the specific features of the domain.

As for the process of extracting surface features of the input text, it depends on identifying the existence of specific punctuation marks, especially the exclamation mark, or not in the text, as well as URLs. Polarity feature extraction is also

aimed at identifying and classifying texts into three main categories: positive, negative, and neutral. On the other hand, extracting the specific features of the domain depends on comparing the text of the input to the HertLex dictionary and determining whether or not it is identical to the text.

The proposed framework begins with the collection of data on extremist behaviour from Twitter using many words related to extremism. The data collected is then subject to a reprocessing step to filter the data by going through many steps such as removing punctuation, removing stop words, lowering the text, tokenizing words, stemming text, and lemmatization. The overall structure of the proposed framework is clearly explained in Fig. 1.

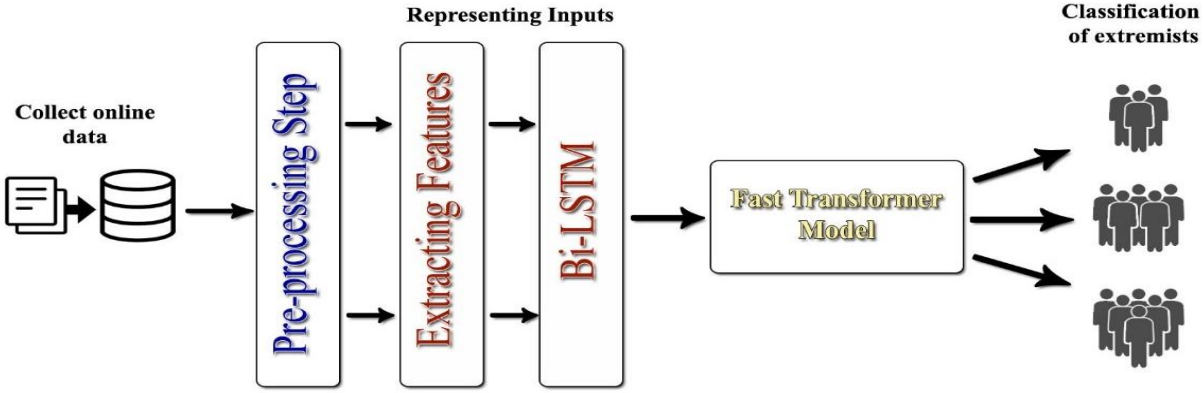

Fig. 1. General architecture for classifying extremist content.

This follows the process of representing the input text by relying on a bi-directional LSTM network that processes in both directions at the same time to focus more on the meaning of each of the entry words to provide a suitable vector that clarifies the relationship between each word and the surrounding words in the sentence. After the completion of the adequate text representation process comes the role of extracting features from the text discussed above. This process is done by relying on several algorithms capable of implementing them skilfully. The Linguistic Inquiry and Word Count (LIWC) algorithm was used to extract various additional features related to linguistic and semantic features of the input data. These features included 93 linguistic features that varied from simple word counts to complex features associated with emotional analysis of text.

Also, the Levenshtein algorithm was applied to compare input words to the HertLex dictionary by using a distance metric that measures the difference between two words. This algorithm is based on measuring the distance between two words by calculating the minimum required adjustments (delete - add - replace) on one word to reach the other, and this is done at the level of the character instead of the level of the word, where the distance between similar words decreases and vice versa.

In the end, the data that has been filtered, reprocessed, and extracted surface features from enters the last phase, which aims to extract the data polarity and also deal with punctuation, as this is done through the design of a pre-trained deep learning network based on smart transformers to classify texts into three classes (positive, negative, and neutral) and detect extreme users. The pre-trained Smart Transformer Network is based on paying more attention to the important words in the input text, which is the cornerstone of determining whether the text is extreme or not through the use of the additive transformation model that has proved highly efficient in dealing with long texts.

This model is based on converting the input word vector into three main matrices representing query, key, and value. Through the interaction between these matrices, the query weights attention(αi) is calculated to adjust the value of attention for each entry word to calculate the global query vector. It is done by calculating the appropriate attention weights based on the query matrix and the learnable parameter. The global key vector is also calculated by adjusting the key matrix weights (βi) as well as relying on the global query vector. The global value vector is also calculated by adjusting the appropriate weights and relying on the interaction between the global key vector and the adjusted weights to calculate global key vector. finally, the query, key and value global vectors are assembled using a single transformation layer suitable for representing the interactions and then passed to a dense layer based on the activation function of sigmoid to exit the appropriate classification in terms of extremes (positive, negative, and neutral) of the input text.

More formally, given the input word embedding matrix X produced by the Bi-LSTM representation layer, the query, key, and value matrices are obtained through learned projection matrices such that $Q = XW_Q$, $K = XW_K$, and $V = XW_V$. The additive attention weight for the query of the i-th word is computed as $\alpha_i = softmax\left(v^{\mathrm{T}} tanh(W_a Q_i + b_a)\right)$, where v and W_a are learnable parameters, giving the global query vector as the attention-weighted sum $q = \Sigma_i \alpha_i Q_i$. The global key vector is obtained analogously, with the key weight $\beta_i = softmax(v^{\mathrm{T}} tanh(W_b K_i + b_b)$ conditioned on q), giving $k = \Sigma_i \beta_i K_i$, while the global value $vector v_g = \Sigma_i \gamma_i V_i$ is computed using an analogous learnable weighting γ_i conditioned on k. The three global vectors q, k, and v_g are concatenated and passed through a single transformation layer, $W_o[q; k; v_g] + b_o$, before the final sigmoid-activated dense layer produces the classification output.

B. Data

The data were aggregated from Twitter by using the Twitter API tool during the period from September 2022 to February 2023 by relying on several right-wing extremism-related

keywords such as ("White Pride", "Ethnonationalism", "Anti-Globalism", "National Sovereignty", "Red Pill", and "Western Civilization"), where the dataset consists of 17,392 tweets. The dataset consists of TweetID, Username, timestamp for Tweets, and Tweets. Based on a team of three language professionals, the data was divided into three main categories: positive to extremism, negative to extremism, and neutral. Table I shows the composition of the dataset and the number of tweets in each category.

TABLE I. DETAILED DESCRIPTION OF THE DATASET

Class Name	No. of Tweets	Percentage %
Positive (Extremist)	9,451	54.3%
Negative (Non-Extremist)	5,287	30.4%
Neutral	2,654	15.3%

C. Data Preprocessing

Based on the nature of the data collected from Twitter, as it is irregular and unstructured data, it needs several steps to be processed so that it can be analysed in a good way to get out with high accuracy results in the classification task. Data processing begins with the deletion of detailed data that we do not need, such as removing non-alphabetical characters, hashtags, links, and whitespaces (tabs, newlines), converting text to lowercase, tokenizing text, using an emoji library in Python to replace emojis with short text, and finally normalizing numbers.

The data processing process is based on several tools from Python, where the Scikit-learn Library was used to implement the reprocessing steps; also, TensorFlow and Keras 2.4 tools were used to build the proposed deep learning model. The proposed deep learning model was implemented in an operational environment based on an eight-core CPU due to the small volume of data. The model's hyperparameter metrics are defined as follows: the aggregated data is divided into 80% training data and 20% test data, and the proposed model is also trained for 5 epochs using a 3e-4 learning rate, with Adam identified as the optimizer of the proposed model. On the other hand, the data is divided into 64-size batches, with a maximum sequence size of 128 and a 0.5 threshold for classification tasks.

On the other hand, the hyperparameters of the Bi-LSTM network are as follows: the hidden state of the network is set to 200, divided into 100 hidden states forward as well as 100 backward, also set 0.4 as a dropout rate in the LSTM layers; and stochastic gradient descent (SGD) is also used to handle the loss function through the back-propagation step. Finally, a sigmoid activation function in the pretrained steps for the multilabel classification tasks.

IV. RESULTS

The proposed framework was applied, and the classification function was evaluated on the basis of accuracy of results using precision, recall, and the F1-score of the positive data class. The results were also compared with several pre-trained models. Table II includes a comparison of the results of the proposed model assessment with a set of 5 algorithms, all implemented on the same collected data, as these algorithms have achieved promising results in the data classification task. The evaluation results of our model were compared with the results of several Machine learning algorithms, such as Support Vector Machine (SVM-linear) [14], Support Vector Machine with Radial Basis Function (SVM-RBF), Random Forest, Logistic Regression, and Decision Tree (DT).

TABLE II. COMPARISON OF THE PROPOSED MODEL WITH FIVE LEADING ALGORITHMS

Model	Precision	Recall	F1-score
SVM-Linear	0.934	0.886	0.909
SVM-RBF	0.951	0.835	0.889
Random Forest	0.976	0.801	0.880
Logistic regression	0.934	0.855	0.892
Decision Tree(DT)	0.896	0.912	0.904
Our Model	0.922	0.916	0.916

There are many hyperparameters set up for machine learning algorithms; for the SVM algorithm relied on a linear kernel, adjusted the penalty parameter to 1, and set the stopping criteria to 0.001. As for the logistic regression algorithm, the stopping criteria were set to 0.01 with reliance on L2 penalty. On the other hand, the random forest algorithm used the entropy standard with 200 estimators. In addition, the decision tree algorithm used default settings. Table III explains the percentage accuracy rate in the implementation of the classification task based on the confusion matrix.

TABLE III. TEXT CLASSIFICATION ACCURACY RESULTS OF COMPARING THE PROPOSED MODEL WITH 5 MACHINE LEARNING ALGORITHMS

Model	Accuracy %
SVM-Linear	96.2%
SVM-RBF	95.5%
Random Forest	95.3%
Logistic regression	95.6%
Decision Tree (DT)	95.8%
Our Model	96.7%

V. DISCUSSION

The experimental results in Table II showed that there was strong competition between our model and machine learning algorithm models in the implementation of the classification task, as the lowest rate of F1-score was achieved by the Random Forest algorithm at 88% while the highest rate achieved by our model is 91.6%, up to 3.6%, although Random Forest's algorithms are capable of achieving the highest correct positive predictions rate for extreme tweets(Precision) by 97.6%, while the Recall ratio was the lowest on all models at 80.1%, reflecting the overall inaccuracy of the model.

This reflects the low accuracy rate of the Random Forest algorithm, shown in Table III at a rate of 95.3%. The other models while our model was able to achieve the highest accuracy rate of 96.7%. This is because the superiority of our model over other machine learning algorithms in the standard F1- score has led to another outperformance in the overall rating accuracy level, leading to an increase in the accuracy of the model up to 1.4% from the lowest model, the Random Forest

algorithm, while the increase in accuracy is up to 0.5% from the nearest competing model, SVM-Linear.

Experimental results in Table II show that the proposed model has produced promising results in the text classification task, where it can outperform all machine learning models that have managed to achieve high results in this task before. In more detail, our model achieved 91.6% on the F1-score benchmark, outperforming the strongest classical machine-learning baseline evaluated in this study, the SVM-Linear algorithm, by 0.7%; although the SVM-Linear algorithm outperformed the correct positive prediction rate for extreme tweets by 1.2%, it failed in sensitivity ratio (Recall) by 3%. This close competition between deep learning and classical machine learning approaches echoes findings from comparative classification studies conducted on extremist content in other languages [16], suggesting that the margin of improvement achievable through architectural changes alone may be inherently limited without complementary gains in feature representation or data quality.

VI. LIMITATIONS

Although the proposed model has achieved promising results, several limitations of the present study should be acknowledged. On the dataset side, the aggregated data was collected using keywords associated with a single ideological strand of extremism, right-wing extremism, over six months, and from a single platform, Twitter, so the ability of the proposed model to generalise to other forms of extremism, other platforms, or other time periods has not yet been tested. On the other hand, the three classes are notably imbalanced (54.3%, 30.4%, and 15.3% of the dataset, respectively), and the effect of this imbalance on the model's ability to correctly classify the minority neutral class has not been separately analysed. As for the labelling process, although a team of three language professionals was relied on to annotate the dataset, a formal inter-annotator agreement score was not computed, so the reliability of the ground-truth labels cannot yet be quantified. In terms of evaluation, the proposed model was assessed using a single 80%/20% train-test split rather than k-fold cross-validation, and the reported improvements over the closest competing algorithm, SVM-Linear, are relatively small (0.7% at the F1-score benchmark and 0.5% at the accuracy level), so statistical significance testing would be needed to confirm that these gains are not attributable to chance. A paired test such as McNemar's test, applied to the per-instance predictions of the two models on the shared test set, would be a natural next step to establish this formally. Finally, the comparison in this study was restricted to classical machine learning algorithms and did not extend to other pre-trained transformer architectures such as BERT or RoBERTa, although prior work applying transformer-based approaches to related extremism and hate-speech classification tasks has reported F1-scores in the 91–97% range [6,7], broadly consistent with the 91.6% F1-score achieved here; nor did it include an ablation study capable of isolating the individual contribution of the feature-extraction step from that of the attention-weight modification.

VII. CONCLUSION

A new framework has been developed in this study based on the extraction of the text's connotative and linguistic features by paying attention to several steps that directly help to improve the representation of the data entered, including attention to the process of preparing and reprocessing the data collected to extract important features from the text. Relying on the LIWC extraction algorithm to extract many language features, as well as those associated with word counts and repeated words in the input text, contributes to the introduction of the text representation vector capable of reflecting as many text features as possible.

The process of extracting these linguistic features contributed to the presentation of a well-equipped text for the process of embedding words, in which a Bi-LSTM network was used to increase attention to the importance of the order of words in the input sentences, thereby contributing to further clarification of the meaning of each word in the input sentence. On the other hand, after paying attention to the good representation of the data, the data enters a pre-trained deep learning model based on smart transformers. The power of the smart transformer is to use matrices of global queries, keys, and values by calculating the required adjustments to the attention weights that are assembled in one transformation layer to interact with each other and then using a suitable dense layer with a sigmoid activation function to classify the input text into three categories: extremist, non-extremist, and neutral.

After the proposed model was trained on 80% of the aggregated data and tested on 20% of the same data, the results of the performance evaluation of the proposed model appeared in Table II and Table III, where Table II showed the results of the evaluation of the accuracy of the proposed model compared to 5 machine learning models that had previously achieved promising results in the data classification task. In addition, Table III shows the accuracy of each model in general in the data classification task.

The proposed model clearly outperformed all algorithms in comparison based on the F1-Score benchmark by an estimated 3.6% increase over the lowest competitors and Random Forest and 0.7% higher than the nearest competing algorithm, SVM-Linear. At the level of accuracy of the models as a whole, our model was able to achieve the highest accuracy with an estimated 1.4% increase from the lowest accuracy achieved by the Random Forest algorithm and an estimated 0.5% increase from the nearest rival's SVM-Linear algorithm.

As future work, the proposed framework could be extended beyond text-only inputs to incorporate multimodal feature extraction techniques, such as those developed for meme and image-based sentiment classification [15], in order to capture the full range of multimedia content through which extremist material is shared on social media.

REFERENCES

[1] Karpova, A., Savelev, A., Vilnin, A. & Kuznetsov, S. Method for detecting far-right extremist communities on social media. Soc. Sci. 11, 200 (2022). https://doi.org/10.3390/socsci11050200

[2] Marten, R., Kevin, M. B., Susilo, W., Rita, J.-M. & Winnifred, L. Dynamic matrix of extremisms and terrorism (DMET): a continuum approach towards identifying different degrees of extremisms. Glob. Internet Forum to Counter Terrorism, Article ID 7531980, 13 pages (2021).

[3] Chadha, C., Gupta, V., Gupta, D. & Khanna, A. Classifying text using machine learning models and determining conversation drift. arXiv preprint arXiv:2211.08365 (2022).

[4] Wendelberg, L. An ontological framework to facilitate early detection of 'radicalization' (OFEDR) – A three world perspective. J. Imaging 7, 60 (2021).

[5] Torregrosa, J., Bello-Orgaz, G., Martinez-Camara, E., Ser, J. D. & Camacho, D. A survey on extremism analysis using natural language processing: definitions, literature review, trends, and challenges. J. Ambient Intell. Humaniz. Comput. https://doi.org/10.1007/s12652-022-04025-5 (2022).

[6] Al-Shawakfa, E. M., Alsobeh, A. M. R., Omari, S. & Shatnawi, A. RADAR#: an ensemble approach for radicalization detection in Arabic social media using hybrid deep learning and transformer models. Information 16(7), 522 (2025). https://doi.org/10.3390/info16070522

[7] Suleiman, H. A. & Ali, Z. H. The extremism detection using hybrid deep learning. Mustansiriyah J. Pure Appl. Sci. 3(4), 173–189 (2025). https://doi.org/10.47831/MJPAS.V3I4.292

[8] Yang, P., Li, B., Zhao, H., Huang, H. & Wu, D. Multi-dimensional information representation for offensive language detection. IEEE Trans. Emerg. Top. Comput. Intell. 10(1), 130–141 (2026). https://doi.org/10.1109/TETCI.2025.3559094

[9] Kumari, K., Singh, J. & Kumar, A. Deep active learning for identifying hate and offensive content in multilingual social media posts. Knowl. Inf. Syst. 68(1) (2026). https://doi.org/10.1007/s10115-026-02682-9

[10] Abusaqer, M. & Saquer, J. Multiclass hate speech detection with RoBERTa-OTA: integrating transformer attention and graph convolutional networks. arXiv preprint arXiv:2603.04414 (2026).

[11] Gaikwad, M., Ahirrao, S., Phansalkar, S. & Kotecha, K. Online extremism detection: a systematic literature review with emphasis on datasets, classification techniques, validation methods, and tools. IEEE Access 9, 48364–48404 (2021).

[12] Shaffer, R. Psychological perspectives on radicalization. J. Policing Intell. Counter Terror. 17, 229–230 (2022). https://doi.org/10.1080/18335330.2021.1880021

[13] Keatinge, T. & Florence, K. Social media and (counter) terrorist finance: a fund-raising and disruption tool. Stud. Confl. Terror. 42, 178–205 (2019).

[14] Mammone, A., Turchi, M. & Cristianini, N. Support vector machines. Wiley Interdiscip. Rev. Comput. Stat. 1, 283–289 (2009).

[15] Ouaari, S., Tashu, T. M. & Horváth, T. Multimodal feature extraction for memes sentiment classification. In Proc. 2022 IEEE 2nd Conf. Inf. Technol. Data Sci. (CITDS) 285–290 (IEEE, 2022).

[16] Dragos, V. & Constable, Y. Comparison of classification techniques for extremism detection in French social media. Fusion (20223)