American Journal of Psychiatric Rehabilitation

Expert Opinion Article

# Prediction of Pre-Radicalism Leading to Hate Speech in Social Media Accounts Using Machine Learning and Intelligence Data Gathering Frameworks

**Ala Berzinji[1]*, Sharo Karwan[2]**

[1]Department of Computer Science, College of Science, University of Sulaimani, Sulaimani, Iraq
[1]Department of Systems Sciences, Stockholm University, Stockholm, Sweden
[1]Cyber Security Department, IQ Group Holding, Sulaymaniyah 46001, Iraq
[2]Cyber Security Department, IQ Group Holding, Sulaymaniyah 46001, Iraq

***Corresponding Author:** Ala Berzinji
*Email: alabe@dsv.su.se

**Abstract**
Social media presents concerns about extremist content spreading which produces anxiety about radicalization that occurs through online means. Digital platforms provide two advantages to evil organizations by helping them recruit new members and conducting their activities while distributing propaganda. BERT (Bidirectional Encoder Representations from Transformers) within the designed system detects indicators of pre-radicalization to prevent detecting the onset of extremist behavior and hateful speech. The system draws from Twitter data collected by intelligence data gathering frameworks and conducts ongoing training updates to enhance its operational efficiency as well as adaptability capabilities. Text normalization begins a preprocessing sequence which is followed by tokenization operations before companies perform feature extraction. Evaluations of the model happen through precision and recall measurements as well as accuracy assessment followed by F1-score evaluation. The identification of radicalization warning signs by separate deep learning algorithms proves more effective than traditional systems at early stages of development. AI systems for radicalisation detection need continuous training which results in improved detection capabilities and allows their application in both counterterrorism operations and political extremism support programs and online hate speech monitoring.

**Keywords:** Prediction, Pre-Radicalism, Hate Speech, Machine Learning, BERT, Model

## 1. Introduction

Extremist groups use social media channels to distribute extreme ideologies for recruiting new members while spreading their messages along with coordinating activities between members. The organizations ISIS and Al-Qaeda along with far-right nationalist groups extensively exploit platforms like Facebook and Twitter and YouTube and Telegram to distribute their ideology with violent messaging [1]. Extremist propaganda benefits from these platforms since they enable inexpensive distribution of radical content to a broad spectrum of people in a single operation [2]. The excessive growth of extremist content available online drives greater incidents of hate speech alongside false information and digital-based radicalization of individual users [3].

Terrorism countermeasures from the past center their efforts on law enforcement actions that begin after radicals complete their extremist activities [4]. Early-stage radicalization detection requires a proactive strategy since it allows security operators to prevent violent extremism from escalating. NLP combined with machine learning has achieved successful identification of radical communication patterns through automated processing of digital content [5].

Local research dedicated to detecting explicit hate speech and fully radicalized content continues to face difficulties identifying pre-radicalization signals since these signals generally display subtle behaviors before moving to explicit extremism [6].

The proposed study implements a model built using BERT technology to recognize pre-radioicalization markers which appear in social media content. The study avoids the same approach followed by previous researches which studied only complete radicalized speech and instead focuses on detecting early signs of radicalization for timely intervention [7]. This model uses Twitter data obtained from OSINT (Open-Source Intelligence) frameworks through continuous learning features that provide flexibility for adapting to changing extremist communication patterns [8].

This approach has broader applicability, as it can also be utilized to detect political extremism, hate speech, and ideological radicalization across different contexts [9].

The structure of this paper is as follows: In Section 2, there is a presentation of related work. Section 3 presents the research methodology. The experimental results and analysis in Section 4 are for evaluating effectiveness of the proposed system.

Finally, Section 5 offers the last explanations and suggests some data for additional study.

## 2. Related Work

In this section some of the studies will be presented which explore automated radicalization detection through machine learning and NLP techniques.

Nouh et al. [10] analyzed linguistic and psychological cues in ISIS propaganda and demonstrated that sentiment analysis and text classification could effectively detect radical discourse. Their findings showed that extremist messages often contain ideological reinforcement, recruitment appeals, and calls to action. Gaikwad et al. [11] reviewed machine learning techniques for extremism detection, identifying limitations in traditional classifiers such as Support Vector Machine (SVM), Decision Trees, and Naïve Bayes. Their study emphasized that deep learning models particularly BERT and transformer-based architecture offer superior accuracy due to their ability to understand context in language processing.

Mussiraliyeva and her colleagues [12] tested six different machine learning models for radicalization detection which showed that BERT-based models surpassed Machine Learning traditional methods in performance. RoBERTa managed to reach state-of-the-art performance levels after being trained on extremist content according to Araque et al. [13] for the identification of jihadist and far-right propaganda on Twitter. Social media data monitoring for detecting potential security risks is facilitated through the use of OSINT (Open-Source Intelligence) tools by counterterrorism agencies. Around 75% of Twitter users seek help from European Union Internet Referral Unit (EU IRU) and Twitter's AI systems to manage extremist content through automatic detection tools as both units experience limitations associated with false detection alerts and changes in extremist communication methods. According to Almusaylim et al. machine learning models need ongoing retraining processes to stay up-to-date with emerging extremist literature as well as the techniques used for radicalization.

## 3. Methodology

The research adopts systematic methods for data processing which integrates data acquisition with data refinement and feature development until the training phase ends in an adaptable learning mechanism to track changing patterns of radicalization. The successive stages represent vital steps which advance system accuracy while making it more capable of discovering early warning signs of radicalization in social media platforms. The study methodology flowchart can be seen in Figure 1.

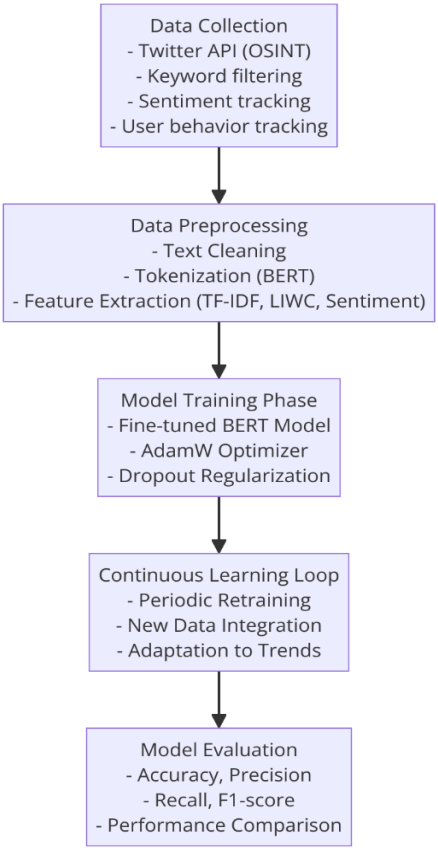


***Figure 1. methodology flowchart***

### 3.1 Data Collection

The main dataset derived from Twitter was collected through OSINT (Open-Source Intelligence) frameworks. Specific criteria within the data collection guide users to obtain content that indicates initial radicalization behavior signs. To detect extremist speech with slurs and radical terms that characterize extremist speech patterns a keyword-based filter operates. Sentiment tracking mechanisms assist in spotting intensely polarized and intense negative or hostile expressions because these point toward the possibility of radicalization. The data-gathering process includes built-in user behavior tracking which enables system monitoring of interactions with extremist accounts together with ideological content retweets and engagement in radical discussions.

The dataset quality assurance system implements multiple validation steps including domain expert reviews for pieces of collected data. These processing techniques remove worthless detection findings while keeping training data with appropriate indicators of radicalization behavior. New data samples accumulate in the dataset perpetually while the model maintains its ability to adapt to emerging behavioral and linguistic patterns of online extremism.

### 3.2 Preprocessing and Feature Engineering

Preprocessing operations convert natural text into processed data which structured deep learning models need for their operation. During this phase text cleaning processes eliminate both insignificant semantic characters and URLs and special symbols. Standardization processes for the dataset require case normalization as well as stop word removal procedures to achieve uniformity. The WordPiece tokenizer of BERT performs text cleaning operations to create meaningful wordlike parts from textual data for enhanced word context interpretation. The BERT tokenization method preserves word semantics to help the model identify deep semantic meanings which appear in radical discourse. After tokenization the model receives stronger computational signals while enriched data extraction processes make it possible to handle unprocessed text. The model includes TF-IDF scoring with sentiment analysis results and LIWC psycholinguistic markers among its three variable classifications. The model's discrimination abilities increase for detecting radicalization patterns at the on-set when engineers analyze vocal intensity effects and psychological and emotional traits.

### 3.3 Model Training and Continuous Learning

The suggested system functions with a BERT classifier that uses binary classification to distinguish radical and non-radical content. The basic method in training systems is supervised learning since labeled Twitter data helps enhance model parameter effectiveness. The training process depends on AdamW optimizer using 2e-5 learning rate for enhancing stability and convergence performance. Multiple regularization techniques using drop rates between 0.3 to 0.5 help prevent overfitting thus making the model accurately classify new unknown data. The study makes continuous learning its main innovation. The model gets updated through retraining with new Twitter data collections instead of maintaining a set dataset. The method allows better control of changing radical narratives and extremist terminology through time. Real-time model updates prevent new threats to allow accurate detection of threats. Data sets categorized since the last retraining period use incremental learning techniques to fuse current information with prior training databases while maintaining equal proportions of old and fresh samples. The continuous learning approach gains advantages from this method to prevent abrupt memory losses labeled as catastrophic forgetting that data-related issues can trigger.

## 4. Experimental Results and Analysis

The system effectiveness assessment depends on accuracy measurements and precision rates combined with recall metrics which get evaluated through the F1-score performance metric. The initial baseline measurement takes place in training and evaluates the improvement of detection abilities during subsequent training cycles. Table 1 Presents the module results.

**Table 1. Module Results**

| Iteration | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Initial Model | 85.4% | 83.1% | 87.2% | 85.1% |
| After Retraining | 91.2% | 89.3% | 93.5% | 91.4% |

### 4.1 Performance Improvements Through Continuous Training

The performance evaluation demonstrated better metrics throughout all measurements between the initial model and its retrained variant. The model demonstrates improved accuracy statistics because it better detects radical along with non-radical material from 85.4% to 91.2%. The model achieved

better precision in preventing incorrect labels by detecting genuine extremist content as its precision went from 83.1% to 89.3%. The modified model now better identifies genuine radical material because recall metrics progressed from 87.2% to 93.5%. This decrease in missed urgent threats becomes more likely with this enhancement. The F1-score showed the greatest transformation from its starting value at 85.1% until reaching 91.4% within the retrained model. The system exhibited better authentic radical detection capability while simultaneously reducing faulty assessments based upon this outcome. Through its retraining system the model can continuously learn so it stays reliable by adapting to fresh radical content.

### 4.2 Error Analysis and Challenges

The system manages its operations better while continuously confronting new difficulties. The system occasionally fails to identify non-extreme political language that feels radical even though it remains incorrect thus highlighting the distinction between extremist content and early signs of radicalization. Extremist speech produces classification problems during times that are not characterized by violence. Extremist language evolution makes it challenging for identifying extremist contents. The anti-social parties modify their language to prevent detection making keyword-based filtering strategies require consistent updates. To address this issue properly the necessary actions include user behavior analysis and outside intelligence cross-checks along with contextual response capabilities that transcend language-based assessment.

### 4.3 Future Directions for Enhancement

The upcoming system versions will include multi-modal analysis through text together with image and behavioral analysis to improve contextual understanding. Expanding data analysis across Telegram and Facebook in addition to Twitter will generate a wider perspective about radicalization patterns that occur online. The team will research ways to make models more understandable because this will aid in justifying flagged content by delivering clear, interpretive reasoning.

## 5. Conclusion

An approach based on machine learning served to identify pre-radicalization signs in social media text content. The BERT mechanics with OSINT data gathering frameworks provide a system able to detect developing radicalization signs before they transform into extremist activities. The model demonstrates better performance after each successive retraining session because the technique enables it to detect emerging extremist discourses and novel linguistic methods. Model reliability benefits from regular updates since its accuracy metrics increased from 85.4% to 91.2%. The research emphasizes that detection models for radicalization need to execute ongoing updates. Extremist language together with online behaviors undergo regular changes which make fixed models ineffective for detection. The system stays pertinent to current threats through its performance of continuous learning. A major challenge exists for the system to differentiate between the language of pre-radicalization and politically oriented speech. Additional improvements need to handle both ethical issues and bias factors together with refining how choices are made by the model. Researchers need to study interconnected detection across various platforms together with multi-channel analysis since it will produce a better understanding of radicalization patterns by examining text data alongside images and user activity. The detection capabilities will benefit from expanding the available dataset into multiple languages as well as cultural contexts. The system will enhance its effectiveness to detect early radicalization patterns through these improvements thus serving counterterrorism operations and online content moderation functions.

## 6. References

1. M. Conway, “Routing the Extreme Right: Challenges for Online Counter-Narratives,” Stud. Confl. Terror., vol. 42, no. 1, pp. 1–23, 2019.
2. A. Bruns, “Filtering Radicalization: Social Media Content Moderation and Free Speech,” New Media Soc., vol. 23, no. 3, pp. 534–551, 2021.
3. J. Berger and J. Morgan, The ISIS Twitter Census: Defining and Describing the Population of ISIS Supporters on Twitter, The Brookings Project on U.S. Relations with the Islamic World, 2015.
4. P. N. Howard and B. Kollanyi, “Bots, Social Media, and the Spread of Misinformation,” J. Inf. Technol. Polit., vol. 15, no. 2, pp. 91–108, 2018.
5. C. O’Callaghan, D. Greene, M. Conway, J. Carthy, and P. Cunningham, “Down the (White) Rabbit Hole: The Extreme Right and Online Radicalization,” Internet Policy Rev., vol. 6, no. 1, 2017.
6. L. Scrivens, S. Gill, and R. Conway, “The Role of the Internet in Radicalization: A Systematic Review of Research,” Terror. Polit. Violence, vol. 34, no. 3, pp. 545–564, 2022.
7. S. Weimann, Terrorism in Cyberspace: The Next Generation, New York, NY, USA: Columbia Univ. Press, 2015.
8. F. Benigni, K. Joseph, and K. Carley, “Online Extremism and the Amplification of Violence,” Comput. Math. Organ. Theory, vol. 27, no. 1, pp. 32–49, 2021.

9. B. Sayyid and S. Zac, Thinking Through Islamophobia: Global Perspectives, New York, NY, USA: Columbia Univ. Press, 2011.
10. M. Nouh et al., "Understanding the radical mind," IEEE ISI, 2019.
11. M. Gaikwad et al., "Online extremism detection," IEEE Access, 2021.
12. S. Mussiraliyeva et al., "Detecting violent extremism," IJACSA, 2023.
13. O. Araque et al., "Hate speech detection using RoBERTa," Comput. Commun., 2023.
14. European Union Internet Referral Unit, EUROPOL Online Radicalization Report, 2021.
15. Twitter, "Automated moderation systems," Twitter Transparency Report, 2022.
16. M. Almusaylim et al., "Automated extremism detection," Applied Sciences, 2021.