Article Review #1: LLM-Enhanced OSINT and Cyberterrorism Detection: A Social Science Perspective

Nicolas Ryan Stephens

School of Cybersecurity, Old Dominion University

CYSE 201S: Cybersecurity and the Social Sciences

Diwakar Yalpi

February 2nd, 2026

Introduction/BLUF

Berzinji and Abdalmajid (2024) examine the integration of Large Language Models (LLMs) with Open-Source Intelligence (OSINT) strategies to enhance the detection of extremist rhetoric/content on Twitter. The authors used a Retrieval-Augmented Generation (RAG) model integrated with the LLaMA3 open-source LLM to retrieve relevant external information from a sample of Twitter posts, enabling the model to analyze tweets in a dynamically, contextually informed environment and detect subtle, distinct signs of radicalism that span cultures and geopolitical boundaries.

The authors concluded that integrating a Large Language Model with an OSINT-based Retrieval-Augmented Generation methodology did, in fact, significantly enhance the detection of subtle extremist patterns and linguistic markers in content posted on Twitter.

Relation/Connection to Social Science Principles

The study by Berzinji and Abdalmajid (2024) exemplifies all seven core principles of the social sciences: relativism, objectivity, determinism, ethical neutrality, skepticism, parsimony, and empiricism.

Relativism is one of the most evident principles of this paper. The authors clearly demonstrated that radical and extremist rhetoric cannot be correctly comprehended without situational, political, and cultural context. Extremist language is often riddled with hidden meanings, historical references, and geopolitical ideas, which may appear innocent without further investigation. However, through the integration of Open-source Intelligence with Retrieval-Augmented Generation, the LLaMA3 LLM is able to analyze contextual data from various sources, as Berzinji and Abdalmajid (2024) state, “The rationale behind incorporating

LLMs is to enable the automatic access, inference, validation, and condensation of information from various sources validation, such as threat reports, academic papers, and technical scripts." Additionally, the authors mention that additional sources, such as news articles and government reports, can broaden the context and information available to the LLM. The specificity in ensuring the LLM has cultural and political context reinforces the relativist principle that meaning must be understood within cultural and political bounds rather than a universal standard.

Objectivity is also deeply embedded in the study. The authors, rather than relying on subjective judgment, evaluated performance through measurable statistics. The study also compared the LLaMA3 and Gemma2 LLM models in a controlled environment, examining their performance with and without OSINT-based Retrieval-Augmented Generation. Additionally, Berzinji and Abdalmajid (2024) carefully considered their data collection, which was used as input for LLaMA3 and Gemma2. Using various pre-filtering and random sampling techniques, the authors developed separate datasets for non-extremist and extremist tweets. This systematic, methodical approach to data collection further reduces potential human bias that could affect model performance and, in turn, enhances the study's internal validity by ensuring that the observed results are properly attributed to the experimental conditions rather than to other factors.

Similar to relativism and objectivity, determinism is another crucial aspect embedded in the theoretical assumptions underlying this study. Determinism in the social sciences suggests that human behavior is the result of recognizable patterns rather than spontaneous choice. This idea is reinforced by the author's assumption that online extremist rhetoric can be identified

through linguistic patterns. Thus, machine learning can be employed to identify and recognize these patterns and subsequently measure and predict them.

The authors maintained an ethically neutral stance throughout the study, which provides additional validity for their techniques and scientific principles, even though extremism and terrorism are morally and politically charged topics. The research presented in the study is methodologically sound and scientifically rigorous, presenting empirical data and scientific conclusions grounded in evidence rather than a critique of extremist ideology and morality.

The authors also maintain a healthy dose of skepticism, as evidenced by the study's design. Rather than hypothesizing that OSINT-based RAG models could improve extremist detection on social media, the authors used empirical data collection and scientific methods to compare models, techniques, and performance. By evaluating model performance across different variables, the authors avoided overstating their scientific findings and identified strengths and weaknesses, including variations in model accuracy, precision, recall, and F1 scores. The study also acknowledged the limitations of the dataset, which consisted of approximately 1,000 tweets each for the two groups, extremist and non-extremist. By critically examining their own results and recognizing research limitations, the authors exemplify the social science principle of skepticism through rigorously tested claims and conclusions backed by evidence.

Parsimony is another principle of the social sciences that favors simplicity in explaining a phenomenon over complexity. In this case, the researchers simplified the rather complex task of classifying extremist tweets by categorizing them as either extremist or non-extremist and focusing on measurable linguistic patterns rather than attempting to analyze every behavioral and psychological nuance related to commitment to or indulgence in terrorism, extremism, or

cyberterrorism. The authors' streamlined approach embodies the spirit of parsimony, enabling them to draw concise conclusions that will further advance research in the field of cyberterrorist and extremist detection.

Empiricism has been mentioned multiple times thus far as it's the central principle displayed throughout this study. Empiricism within the realm of the social sciences is simply the idea that knowledge must be gathered from observable and measurable evidence. Berzinji and Abdalmajid (2024) ground their conclusion that LLMs integrated with OSINT-based RAG models increase the effectiveness of detecting subtle extremist rhetoric in empirical data, rather than hypothetical speculation. By using quantitative performance metrics, dataset filtering, and analysis, the authors effectively ensure their findings are supported by measurable outcomes, which embodies the principle of empiricism as a foundation for social science research.

Research Question /Hypothesis/ Independent Variable/Dependent Variable

The main research question is whether integrating Large Language Models with an Open-Source Intelligence-based Retrieval-Augmented Generation framework improves the detection of extremist content, specifically on Twitter. Additionally, the authors examine the effects of contextual augmentation enabled by the RAG framework on detection and classification performance against standalone Large Language Models, which face limitations in pre-existing LLM training data.

The authors' hypothesis is that Large Language Models integrated with OSINT-based RAG frameworks demonstrate improved ability to detect and classify extremist tweets.

The study includes two primary independent variable conditions. The first condition was the type of LLM used, either LLaMA3 or Gemma2. The second condition for the independent

variable was the utilization or absence of OSINT-based RAG integration with the LLM. These independent-variable conditions enabled the researchers to assess the impact of OSINT-based RAG and LLM structures on extremist classification performance.

The dependent variables of the study include all measurable statistics regarding classification performance, specifically accuracy, precision, recall, and F1 score. Additionally, the authors implemented confusion matrices, such as true positives, true negatives, false positives, and false negatives, to examine how each model analyzed non-extremist and extremist tweets. The measures were used to evaluate the effects of the independent variables and were thus the primary measures for determining whether OSINT-based RAG integration improved extremist detection on social media.

Types of Research Methods used

Berzinji and Abdalmajid (2024) employed a quantitative experimental research design using an artificially constructed balanced dataset consisting of approximately 1,000 extremist and 1,000 non-extremist tweets. The non-extremist tweets were gathered from the Internet Archive (IA) and consisted of 9.4 billion tweets from 2013 to 2023; for this reason, the authors used random sampling and pre-filtering techniques to obtain a sample set that consisted of accounts that weren't suspended by Twitter, weren't flagged by Twitter's algorithm, had no emoji symbols, and were written in English. Regarding the extremist dataset, the authors used a sample from the Kaggle dataset, which included posts from "ISIS fanboys", and the dataset collection was only on posts from July 4th and July 11th, 2016 (Berzinji and Abdalmajid, 2024).

After dataset construction and filtering were complete, the authors conducted a lab-controlled comparative experiment utilizing two of the most advanced Large Language Models

available at the time, LLaAM3 and Gemma2. The LLMs were tested with and without OSINT-based Retrieval-Augmented Generation and evaluated using quantitative criteria.

Types of Data Analysis used

The authors used Twitter posts labeled as extremist or non-extremist as the primary dataset. This dataset consisted of approximately 2,000 posts, which were equally divided between the non-extremist and extremist groups. The authors also ensured the datasets were balanced to make performance comparisons fair and ensure that each category is equally represented.

Regarding performance, the researchers used quantitative metrics such as accuracy, precision, recall, and F1 score. Additionally, the authors implemented confusion matrices, which altogether allowed them to accurately compare model performance and determine whether OSINT-based RAG integration improved extremist classification.

Connections to other Course Concepts

The study by Berzinji and Abdalmajid (2024) effectively employs multiple research methods, specifically archival research and experimental design.

The study very closely follows the principles of archival research, which involve using archives such as written records, social media posts, and or websites to conduct scientific research. The authors of the study relied heavily on records from the Internet Archive and Kaggle to gather and pre-filter the data for the experiment.

Additionally, the study employs the concept of experimental design, meaning that researchers introduce a manipulation, an independent variable, or a treatment to examine its effects on the dependent variable. In this study, the independent variable was the presence or absence of OSINT-based RAG frameworks in the tested LLMs, and the dependent variable was the LLM's effectiveness, measured by its performance score. This controlled experimental comparison is exactly what researchers strive to achieve when implementing experimental design into their studies.

Connections to the Concerns or Contributions of Marginalized Groups

This study has critical implications for marginalized groups, as they are frequently the subject of virtual extremist rhetoric. Extremist rhetoric promotes hate, violence, and discrimination towards political, ethnic, religious, and minority groups, so by improving detection methods for extremist rhetoric on social media, companies can work to prevent harmful language from reaching these groups, thus reducing the impact of extremism on marginalized groups across the globe.

Overall societal contributions of the study/Conclusion

This research paper has several important and meaningful contributions to society. One of its primary contributions is advancing efforts to detect and classify subtle forms of extremist rhetoric on social media. Through LLM integration with open-source intelligence-based Retrieval-Augmented Generation, the study successfully demonstrated a measurable improvement in detecting extremist linguistic markers.

In addition to improving the detection of extremist rhetoric, this study pushes the boundaries of the use of artificial intelligence in cybersecurity research and applications. Through the authors' innovative approach of integrating LLMs with OSINT-based RAG frameworks, the study provides scalable methods and empirical data that will further enhance contextual analysis for years to come.

Overall, this study represents an invaluable step forward in the application of AI-driven solutions to socially complex cyber-related challenges.

Reference

Berzinji, A., & Abdalmajid, M. F. (2024). *Utilisation of large language models (LLMs) in OSINT-based cyberterrorism detection on social media*. Cybercrimejournal.com. https://www.researchgate.net/publication/389652606_Utilisationof_Large_Language_Models_LLMs_in_OSINT-Based_Cyberterrorism_Detection_on_Social_Media