OBSIDIAN: An OSINT-Driven NLP Framework for Detecting Cyber-Physical Threats in Arabic Social Media

Abdullah Saeed Almalki[1], Salmane Chafik[3], Saad Ezzini[1,2*]
[1]Information and Computer Science Department, KFUPM
[2]IRC for Intelligent Manufacturing and Robotics, KFUPM
Dhahran, Saudi Arabia
[3] Mohammed VI Polytechnic University (UM6P), Morocco
[*]saad.ezzini@kfupm.edu.sa

Abstract

Social networks have evolved into rich sources of Open-Source Intelligence (OSINT), enabling analysts to monitor unrestrained content expressing user activities, sentiments, and emerging behaviors. The immense use of these platforms has made it essential for cybersecurity and threat intelligence professionals to analyze and classify such content to proactively detect cyber-physical threats. While significant research has been conducted on the Arabic language regarding Hate Speech (HS) and Cyberbullying (CB), limited work has addressed Cyber Threat Intelligence (CTI) and OSINT-driven security classification in Arabic, despite their critical importance for early-warning systems and crisis response. In this paper, we introduce OBSIDIAN-AR, a novel real-world, large-scale dataset designed for detecting cyber-physical threats, comprising over 15,000 social media posts primarily from the Gulf region. The dataset is manually curated and annotated into five OSINT-relevant categories: Violence, Threat, Distress, Complaint, and Neutral. Using OBSIDIAN-AR, we fine-tune an Arabic BERT-based model named OBSIDIAN. This context-aware framework acts as a digital early-warning system, leveraging pretrained language representations and domain-specific knowledge derived from our high-quality dataset. Experimental results demonstrate that OBSIDIAN achieves strong performance, reaching up to 98% accuracy on unseen data, proving its viability for modern cybersecurity operations.

1 Introduction

Social media platforms have evolved into global hubs, attracting billions of users who utilize these digital spaces for diverse purposes, ranging from personal expression and political discourse to the real-time reporting of critical events. Increasingly, these platforms serve as primary channels for documenting eyewitness accounts of violence, terrorist attacks, and immediate security threats, such as bomb scares or human traffic. Furthermore, users frequently leverage social networks to broadcast personal distress or report complaints related to cyber-physical threats and critical infrastructure. As a result, the rapid spread of such information requires an equally rapid response from security analysts and threat intelligence operations. This creates a pressing need for automated OSINT-driven analytical tools capable of categorizing and interpreting these signals in real time, thereby providing decision-makers with actionable situational awareness.

Previous work has extensively explored hate speech detection in Arabic social media (Ousidhoum et al., 2019; Mulki et al., 2019), with a primary focus on identifying abusive and toxic content targeting religious groups (Albadi et al., 2018), minorities, and gender-based discrimination (Mulki and Ghanem, 2021). In parallel, several studies have addressed offensive language detection (Husain and Uzuner, 2021), aiming to distinguish between general profanity and more harmful expressions. Other research efforts have focused on cyberbullying (ALBayari et al., 2021; Aljalaoud et al., 2025) and cyber harassment (Kanan et al., 2020), highlighting the growing concern over harmful interactions in online environments. These works typically rely on annotated datasets and supervised learning approaches to identify patterns of aggressive or harmful behavior in user-generated content. Building on these resources, recent studies have proposed models capable of automatically detecting and classifying posts into categories such as hate speech (HS), offensive language (OL), cyberbullying (CB), and cyber harassment (CH) (Mubarak et al., 2022; Haidar et al., 2017).

However, even with the availability of multiple datasets and models addressing these categories, existing approaches remain limited to relatively narrow and often overlapping definitions of harm-

Proceedings of the Second International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security (NLPAICS'2026)
11–12 June, 2026.

ful content. As a result, broader security-oriented classification tasks, such as detecting threats, violent intent, or distress signals, remain largely underexplored in the context of Arabic social media. This limitation is critical, as such classifications are essential for enabling timely intervention and effective response in real-world emergency scenarios.

Motivated by these challenges, we introduce a novel large-scale dataset composed of more than 15,000 Arabic social media posts, annotated into five categories relevant to OSINT and Cyber Threat Intelligence (CTI): **Violence**, **Threat**, **Distress**, **Complaint**, and **Neutral**. The dataset is manually curated following well-defined annotation guidelines to ensure consistency and quality. Based on this dataset, we fine-tune an Arabic BERT-based model that leverages both pretrained language representations and domain-specific knowledge to accurately classify social media posts. Our contributions are summarized as follows:

- **OBSIDIAN-AR**: A novel real-world, large-scale dataset comprising more than 15,000 Arabic social media posts, manually annotated into five categories: **Violence**, **Threat**, **Distress**, **Complaint**, and **Neutral**.
- **OBSIDIAN**: A security-oriented classification model trained on OBSIDIAN-AR, designed to effectively identify and categorize posts relevant to proactive detection of cyber-physical threats.
- A comprehensive evaluation comparing traditional machine learning approaches with transformer-based fine-tuning methods for Arabic OSINT-driven threat detection.

The remainder of this paper is structured as follows. Section 2 reviews prior work on the classification of harmful social media content. Section 3 describes the construction of OBSIDIAN-AR while Section 4 details the training procedure of the OBSIDIAN model. Section 5 outlines the experimental setup and presents the evaluation results. Section 6 discusses the main findings, and Section 7 concludes the paper and highlights directions for future work.

2 Related Work

Natural Language Processing for Arabic social media has grown significantly, driven by the need to detect harmful or sensitive content such as hate speech, abusive language, and violent or threatening posts.

Previous works focused on hate speech on textual posts that insult or degrade groups or individuals based on protected characteristics like race, religion, gender, or disability. (Mubarak et al., 2022) introduce a large scale dataset for hate speech composed of 1,339 samples in different categories such as 'Social Class', 'Ideology', and 'Race'. In parallel, (Mulki et al., 2019) introduced another dataset named L-HSAB focused solely on Levantine dialects leveraging only three categories 'Hate', 'Abusive' and 'Normal' for neutral posts. In the same context, (Mulki and Ghanem, 2021) focused on hate speech against females to be the first benchmark dataset for Arabic misogyny, where (Albadi et al., 2018) focused on hate speech against religious groups focusing on the six most common religions in the Arab world.

Other research has explored offensive language as a broader category that includes profanity, insults, and vulgar expressions. Such language may be disrespectful or hurtful, but it does not necessarily target a specific group. (Alakrot et al., 2018) focused on comments rather than post content, primarily from YouTube channels and videos, and developed a dataset specifically designed for detecting offensive language across several Arabic dialects. In parallel, (Chowdhury et al., 2020) constructed another dataset of dialectal Arabic comments collected from multiple social media platforms, including Twitter, Facebook, and YouTube. Both studies aim to classify text into two categories: 'Offensive' and 'Non-offensive'.

In the same context, several research works have focused on cyberbullying, which is typically characterized by repeated, targeted, and intentional aggression directed at an individual, often with the aim of causing psychological harm. (Shannag et al., 2022) introduced a new Arabic dataset for cyberbullying detection, named ArCybC, which consists of 4,505 samples manually annotated as either cyberbullying (CB) or non-cyberbullying. Similarly, (Haidar et al., 2017) collected 35,273 cyberbullying posts from various sources and applied two machine learning models, namely Naive Bayes and Support Vector Machines (SVM), achieving an F1-score of 92.7% using the SVM model. Finally, cyber harassment has also received attention in Arabic research, referring to texts that express intimidation or persistent unwanted contact, and

Dataset	Domain	Categories	Languages/Dialects	Sources	Size
(Ousidhoum et al., 2019)	Hate Speech	Origin, Gender, SexOrient, Religion, Disability, Other	English, French, Arabic	Twitter	13,000
(Albadi et al., 2018)	Hate Speech	Negative/Positive against (Islam, Christianity, Judaism, Christianity, Shia, Atheism)	Arabic	Twitter	6,000
(Mulki and Ghanem, 2021)	Hate Speech	Misogynistic Language (Non-Misogynistic, Discredit, Derailing, Dominance, Stereotyping & Objectification, Threat of Violence, Sexual Harassment, Damning)	Levant Dialects	Twitter	6,550
(Alshaalan and Al-Khalifa, 2020)	Hate Speech	Racist, Regional, Tribal, Religious-Inter, Ideological	Saudi Dialect	Twitter	9,316
(Mulki et al., 2019)	Hate Speech & Abusive Language	Normal, Abusive, Hate	Levant Dialects	Twitter	5,846
(Mubarak et al., 2022)	Hate Speech & Offensive Language	Offensive, Not-Offensive, Hate Speech (Gender Race Ideology, Social Class, Religion, Disability)	Arabic	Twitter	12,698
(Chowdhury et al., 2020)	Offensive Language	Offensive or Not-Offensive	Egyptian & Levantine	Twitter, Facebook and YouTube	4,000
(Alakrot et al., 2018)	Offensive Language	Negative or Positive	Arabic Dialects	Youtube	10,715
(Shannag et al., 2022)	Cyber Bullying	CB or Non-CB	Arabic	Twitter	4,505
(Haidar et al., 2017)	Cyber Bullying & Cyber Harassment	Yes or No	English and Arabic	Twitter	35,273
Ours	**Security-Focused**	**Violence, Threat, Complaint, Distress and Neutral**	**Arabic (Middle East)**	**Twitter**	**15,000**

Table 1: Comparison of Existing Arabic Datasets for Hate Speech, Offensive Language, Cyberbullying, and Cyber Harassment Against OBSIDIAN-AR.

is mainly associated with adult contexts. (Kanan et al., 2020) employed multiple machine learning models, including KNN, SVM, Naïve Bayes, Random Forest (RF), and J48, achieving an accuracy of 94.9% using the Random Forest model. The study also applied various NLP preprocessing techniques, such as stemming and stop-word removal.

Even with these advances in Arabic NLP tailored for detecting hate speech, offensive language, and cyberbullying, existing work remains largely confined to narrowly defined and often overlapping categories of harmful content. These approaches primarily focus on linguistic toxicity rather than broader security implications, leaving critical aspects such as the detection of violent intent, explicit threats, or expressions of distress insufficiently explored. This gap limits the applicability of current systems in real-world scenarios where timely identification of potentially dangerous situations is essential for OSINT and proactive cyber-physical threat intelligence. To address this limitation, we introduce **OBSIDIAN-AR**, a large-scale dataset of more than 15,000 Arabic social media posts annotated across five security-relevant categories: **Violence**, **Threat**, **Distress**, **Complaint**, and **Neutral**. Furthermore, we propose **OBSIDIAN**, a fine-tuned Arabic BERT-based model to enable accurate and practical classification of cybersecurity-critical content. Together, these contributions aim to bridge the gap between traditional harmful content detection and real-world OSINT and security operations in Arabic social media analysis.

Table 1 provides a comprehensive comparison between **OBSIDIAN-AR** and existing Arabic datasets for harmful content detection, highlighting differences in scope, annotation granularity, and dataset scale.

3 Dataset

This section provides a detailed overview of the dataset construction pipeline, including data sources, pre-processing, and annotation procedures.

3.1 Dataset Collection

To construct a diverse and representative Arabic corpus, we adopted a multi-source data collection strategy that combines API-based collection, large-scale scraping, and curated public resources. Tweets were primarily collected using the `Twitter API (v2)`, which enabled controlled retrieval through keyword-based queries and thematic filters. To increase coverage and capture high-volume, trending content, we additionally employed `Apify` for large-scale scraping. Finally, we incorporated publicly available Arabic datasets from `Hugging Face` to enrich linguistic diversity and facilitate benchmarking against existing resources.

Data collection spans the period from **2024** to **2025**, using a curated list of keywords related to security, complaints, and help requests (see Table 2). The resulting corpus exceeds 200,000 Arabic tweets and includes a wide range of dialects

Category	Keywords
Security	عملية، مطاردة، اشتباك، خلية، جماعة، خطر، دم، تفجير، رهاب، ممنوع، محرم، مسروقات تهديد، قنبلة، خطر، انفجار، هجوم، تفجير، مسلح، إطلاق، سلاح، رصاصة، بلاغ أمني، حزام ناسف
Complaints	ظلم، شكوى، عنصرية، فساد، تمييز، مقهور، حسبي الله، وين الرقابة، ليش كده أين المسؤول، فساد، سرقونا، طفح الكيل، مافي ضمير، ذلونا، تعبنا، الواسطة
Help Requests	استغاثة، مفقود، نداء، ساعدوني، الحقوني، الله يرحمكم، تكفون، وين الشرطة، مافي أمان ساعد، بلغوا الشرطة، بسرعة، محتاج مساعدة، افتحوا الكاميرات، انقذوني

Table 2: Categories and Associated Keywords

and subdialects, such as **Modern Standard Arabic, Saudi (Najdi, Hijazi ...)**, reflecting substantial variation in writing style, orthography, and sociolinguistic context. This diversity is critical for modeling real-world, user-generated Arabic text, which is often informal and highly heterogeneous.

3.2 Data Pre-processing

Given the noisy nature of social media text, we applied a standard normalization and cleaning pipeline. This process included removing URLs, emojis, duplicated entries, spam content, and non-linguistic symbols.

A key contribution of this work is the subsequent human-centered curation stage. Instead of relying solely on automated labeling or distant supervision, we employed manual annotation and expert validation to ensure high semantic quality. From the cleaned corpus, we selected approximately 15,000 tweets for annotation.

Annotation Protocol. Each tweet was annotated according to a predefined schema consisting of five classes: *Threat*, *Violence*, *Distress*, *Complaint*, and *Neutral*. Annotators were provided with detailed guidelines, including: (i) clear definitions of each category, (ii) multiple positive and negative examples per class, (iii) instructions to focus on the overall context of the tweet rather than isolated keywords, taking into account tone, intent, and implicit meaning, and (iv) a requirement to assign a single label only when the semantic intent was unambiguous.

The annotation team was composed of six Saudi native Arabic speakers from the western, eastern, central, northern, and southern regions of the Kingdom, providing broad linguistic coverage of major Saudi dialects, including Hejazi, Najdi, Gulf, and southern dialectal varieties.

Ambiguous, borderline, or multi-label cases were systematically excluded to prioritize label precision over coverage. This conservative filtering strategy ensures high inter-class separability and reduces annotation noise.

3.3 Dataset Characteristics

The final dataset consists of approximately 15,000 manually verified tweets with high semantic clarity. By design, only instances with strong alignment to a single category were retained. While this reduces dataset size, it substantially improves label reliability and downstream model performance.

The dataset captures nuanced linguistic phenomena, including implicit threats, dialectal variation, and context-dependent expressions that are often overlooked in automatically labeled corpora. Example instances for each category are illustrated in Figure 1. To ensure optimal class distribution, the 15,000 instances are perfectly balanced, with 3,000 tweets per category. Furthermore, the textual posts maintain a substantive length, averaging approximately 141 words each.

3.4 Qualitative Dataset Analysis

To better understand the linguistic characteristics of the collected corpus, we generated word clouds for each category based on token frequency. These visualizations highlight dominant lexical patterns and commonly occurring expressions associated with different semantic classes.

Figure 2 presents a word cloud generated from the *threat* category. Frequently occurring terms such as راح (I will), أقتلك (I will kill you), and أحرقكم (I will burn you) reveal the dominance of threat expressions within this category. The visualization provides qualitative insight into the lexical characteristics of threatening Arabic social media

Figure 1: Samples from different categories from OBSIDIAN-AR

Figure 2: Threat Word Cloud

Figure 3: Violence Word Cloud

content and demonstrates the semantic consistency of the annotated samples. While Figure 3 presents the word cloud from the *violence* category illustrating different terms such as ضرب مبرح (A severe beating) and قتل (murder) reveal the dominance of violence expressions.

4 Approach

This section describes the methodologies employed for the classification of Arabic Twitter posts into five categories: *violence*, *threat*, *complaint*, *distress*, and *neutral*. We explore both traditional machine learning models and transformer-based architectures.

4.1 Task Definition

Given a social media post $x \in \mathcal{X}$, the objective is to learn a function $f : \mathcal{X} \to \mathcal{Y}$ that maps the input text to one of the five predefined classes:

$$\mathcal{Y} = \{\text{violence}, \text{threat}, \text{complaint}, \text{distress}, \text{neutral}\}$$

This is formulated as a multiclass classification problem.

4.2 Machine Learning Methods

We consider one of the famous classical machine learning algorithms for text classification due its efficiency and strong baseline performance.

Text Preprocessing and Representation. Tweets are preprocessed using Arabic NLP techniques, including normalization by handling diacritics and letter variants, removal of URLs, emojies, mentions, and non-Arabic characters.

We represent text using TF-IDF features. For a term t in document d, the TF-IDF score is defined as:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \log\left(\frac{N}{\text{DF}(t)}\right)$$

where N is the total number of documents and $\text{DF}(t)$ is the document frequency of term t.

Logistic Regression (LR): For multiclass classification, Logistic Regression uses the softmax

function:

$$P(y = c \mid x) = \frac{e^{w_c^T x}}{\sum_{j=1}^{C} e^{w_j^T x}}$$

where $C = 5$ is the number of classes. We apply L2 regularization and optimize using the `saga` solver.

4.3 Transformer-based Methods

We leverage a pre-trained Arabic language model, `aubmindlab/bert-base-arabertv02`, based on the BERT architecture (Vaswani et al., 2017). This model is specifically designed for Arabic text and incorporates language-specific preprocessing.

Given an input tweet $X = (x_1, x_2, \ldots, x_n)$, the model produces contextualized representations:

$$H = \text{BERT}(X)$$

We use the representation of the special `[CLS]` token (h_{CLS}) for classification:

$$\hat{y} = \text{softmax}(W h_{\text{CLS}} + b)$$

Fine-tuning Details. The model is fine-tuned end-to-end using the cross-entropy loss:

$$\mathcal{L} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c)$$

where $C = 5$. We train for $E = 3$ epochs using the Adam optimizer with a learning rate of $\eta = 2 \times 10^{-5}$ and batch size $B = 16$ and a weight decay of 0.01. Input sequences are padded or truncated to a maximum length $L = 128$ while padding short sentences.

We evaluate model performance using standard metrics for multiclass classification. Macro-averaging is used to ensure equal importance across all classes, particularly in the presence of class imbalance.

5 Evaluation

In this section, we evaluate the approaches introduced in Section 4 on the proposed OBSIDIAN-AR dataset. The evaluation aims to assess the effectiveness of both classical machine learning and transformer-based methods for Arabic social media classification, with particular emphasis on detecting sensitive categories such as threats, violence, distress, and complaints.

5.1 Research Questions

To guide the experimental analysis, we investigate the following research questions:

1. **RQ1: Do transformer-based models outperform traditional machine learning methods for Arabic tweet classification, and how well do both paradigms generalize to out-of-distribution test sets with distributions that differ from the training data?** This question evaluates whether fine-tuned transformer architectures achieve better classification performance than classical machine learning approaches, such as Logistic Regression, while also evaluating their robustness under distribution shift.

2. **RQ2: Does OBSIDIAN achieve consistent performance across all categories?** This question examines whether the proposed dataset and classification framework perform uniformly across the five classes, or whether certain categories remain more challenging due to linguistic ambiguity, or implicit meaning.

5.2 Implementation Details

To investigate the proposed research questions, we conducted a series of experiments using Google Colab with 12 GB of GPU memory. The classical machine learning experiments were implemented using the `scikit-learn` library, while transformer-based experiments were developed using the Hugging Face ecosystem[1], a widely adopted open-source framework for natural language processing research and development.

For the transformer-based pipeline, we relied on several core Hugging Face libraries: (1) *Transformers*, which provides tools for loading, fine-tuning, and evaluating pretrained language models; (2) *Datasets*, which facilitates efficient dataset management and preprocessing; and (3) *Evaluate*, which offers standardized implementations of common evaluation metrics for machine learning and NLP tasks.

5.3 Experiments

To answer the proposed research questions, we conducted the following experiments:

`EX1`: In this experiment, we compare the performance of a traditional machine

[1] https://huggingface.co/

learning baseline, namely Logistic Regression (LR), against a transformer-based model, `aubmindlab/bert-base-arabertv02`, on the OBSIDIAN-AR dataset.

The dataset was divided into training, validation, and test subsets using a stratified split to preserve the original class distribution across all subsets. Specifically, 70% of the data (10,500 samples) was allocated for training, while 15% (2,250 samples) and 15% (2,250 samples) were reserved for validation and testing, respectively.

To ensure experimental reliability, we maintained identical category distributions during splitting. The Logistic Regression model was trained using TF-IDF representations, whereas AraBERT was fine-tuned end-to-end using the hyperparameters described in Section 4.

To assess model generalization under distribution shift, we further evaluated both models on a newly collected and manually annotated test set, denoted as *Test 2* and composed of 500 samples. Test 2 was also sourced from Twitter but constructed using a different set of keywords, resulting in a distribution that differs from the original training data. This setting enables the evaluation of the robustness and transferability of both traditional machine learning and transformer-based approaches when exposed to unseen keywords and varying linguistic patterns in Arabic tweets. Performance was evaluated using Accuracy, Precision, Recall, and Macro F1-score.

`EX2:` In this experiment, we analyze the performance of the best-performing model (OBSIDIAN) from the previous experiment on each category individually to determine whether certain classes are inherently more challenging than others.

More specifically, we report class-wise Precision, Recall, and F1-score for the five categories: *violence*, *threat*, *complaint*, *distress*, and *neutral*. In addition, we analyze the confusion matrix to identify common misclassification patterns and semantic overlaps between categories. This experiment provides deeper insight into the robustness of OBSIDIAN-AR in handling implicit threats, emotionally charged expressions, and dialectal variations commonly observed in Arabic social media text.

5.4 Answers to RQs

5.4.1 RQ1

Table 3 presents the performance comparison between the proposed transformer-based approach (OBSIDIAN using AraBERT) and the classical Logistic Regression baseline on the development (Dev) and test sets. The results show that both models achieve strong performance overall, with OBSIDIAN slightly outperforming Logistic Regression on the test set. Specifically, Logistic Regression achieves a macro F1-score of 98.221% on the Dev set and 97.693% on the test set, while OBSIDIAN obtains 97.825% and 98.129%, respectively.

On the other hand, Table 4 reports results in a more challenging out-of-distribution setting (Test 2), where the data are collected using different keywords and present a distribution shift from the training data. In this scenario, OBSIDIAN clearly outperforms Logistic Regression across all evaluation metrics. The transformer-based model achieves an F1-score of 90.5% compared to 56% for Logistic Regression, corresponding to a substantial improvement of approximately 34.5 percentage points. This performance gap is also reflected in accuracy, precision, and recall, demonstrating the superior robustness of transformer-based representations under distribution shift.

These findings suggest that while traditional models such as Logistic Regression can perform competitively, they struggle to generalize to unseen lexical and contextual variations. In contrast, OBSIDIAN benefits from contextualized representations learned by AraBERT, enabling better generalization to out-of-distribution samples and more robust handling of linguistic variability in Arabic tweets.

5.4.2 RQ2

To investigate whether the model performs consistently across all categories, Table 5 presents the class-wise Precision, Recall, and F1-score achieved by OBSIDIAN on the test set.

The results indicate that the proposed model achieves strong and stable performance across all five categories, with F1-scores ranging from 96.752% to 99.558%. Among all categories, the *Threat* class achieves the highest performance, reaching an F1-score of 99.558%, with perfect recall (100%) and very high precision (99.120%). This suggests that threatening language in the

Model	Dev				Test			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
OBSIDIAN (BERT-based)	97.822	97.836	97.822	97.825	98.134	98.130	98.133	98.129
Logistic Regression	98.222	98.232	98.222	98.221	97.689	97.730	97.689	97.693

Table 3: Different approaches performance on OBSIDIAN-AR test and development sets

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	62.0	58.0	55.0	56.0
OBSIDIAN	92.5	91.0	90.0	90.5

Table 4: OBSIDIAN and Logitic Regression Performance on Test 2.

Category	Precision	Recall	F1-score
Threat	99.120	100.00	99.558
Violence	98.013	98.666	98.338
Distress	97.516	96.000	96.752
Complaint	97.986	97.333	97.658
Neutral	98.013	98.666	98.338

Table 5: Precision, Recall and F1-score of OBSIDIAN performance on each category from the Test set.

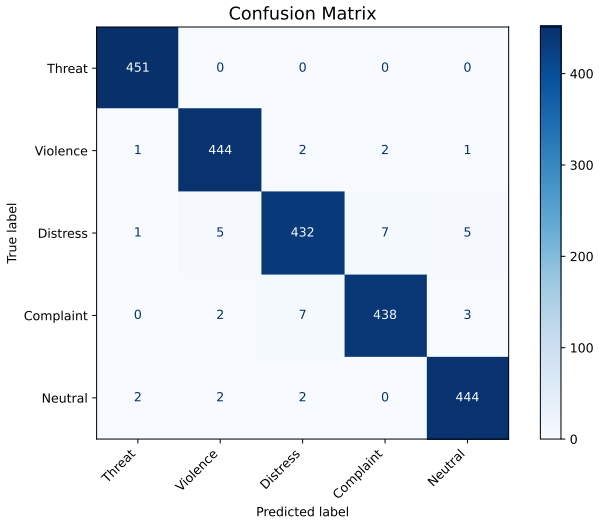

Figure 4: Confusion Matrix on the test set

dataset contains highly distinctive lexical and semantic patterns that are effectively captured by the transformer-based model.

The *Complaint*, *Violence*, and *Neutral* categories also demonstrate consistently strong performance, achieving F1-scores of 97.658%, 98.338%, and 98.338%, respectively. In particular, the high recall values across these classes indicate that the model is effective at correctly identifying instances of each category with few missed predictions.

Slightly lower performance is observed for the *Distress* category, which achieves an F1-score of 96.752%. This drop can be attributed to the more subtle and context-dependent nature of distress expressions, which often overlap semantically with both neutral and violence-related content. This is further supported by the confusion matrix in Figure 4, where a small number of distress samples are misclassified as violence or complaint.

Overall, the results demonstrate that OBSIDIAN generalizes effectively across all semantic categories while maintaining high and balanced performance. The relatively small variance between classes further confirms the robustness of the model and the quality of the dataset annotation process.

6 Discussion

Our evaluation results demonstrate that transformer-based approaches consistently outperform traditional machine learning methods for Arabic cybersecurity-oriented content classification. OBSIDIAN consistently outperforms Logistic Regression across all evaluation metrics, which is likely due to the ability of AraBERT to capture contextualized semantic representations rather than relying solely on sparse lexical features such as TF-IDF. This advantage becomes particularly important in Arabic social media text, where meaning is often expressed implicitly through dialectal language, informal writing styles, sarcasm, and context-dependent expressions.

The category-level analysis further shows that performance varies across classes depending on their linguistic characteristics. The *Threat* category achieves the highest scores, likely because threatening content frequently contains explicit and highly

discriminative vocabulary, whereas categories such as *Distress* and *Violence* remain more challenging due to semantic overlap and implicit emotional expressions.

In addition, the strong and balanced performance across categories highlights the effectiveness of the manual annotation strategy adopted in OBSIDIAN-AR. By filtering ambiguous and multi-label instances, the dataset achieves high semantic consistency and reduced annotation noise, which likely contributes to the robustness of the resulting models. Overall, the findings confirm the importance of contextual language modeling and high-quality human annotation for reliable Arabic cyber-physical threat detection in real-world OSINT environments.

7 Conclusion

In this paper, we introduced OBSIDIAN-AR, a manually curated Arabic dataset for OSINT-driven cyber-physical threat detection on social media platforms. The dataset was constructed using a multi-source collection strategy and carefully annotated into five semantic categories: *threat*, *violence*, *distress*, *complaint*, and *neutral*. We evaluated both classical machine learning and transformer-based approaches on the proposed benchmark and demonstrated that AraBERT-based models significantly outperform traditional Logistic Regression baselines, achieving strong and consistent performance across all categories. The experimental findings further highlight the importance of contextualized language modeling for capturing implicit meanings, dialectal variations, and emotionally nuanced expressions commonly found in Arabic social media text. Overall, the results confirm the effectiveness of the proposed dataset and establish OBSIDIAN-AR as a valuable benchmark resource for future research on Arabic OSINT, Cyber Threat Intelligence (CTI), and proactive cybersecurity operations.

References

Azalden Alakrot, Liam Murray, and Nikola S. Nikolov. 2018. Dataset construction for the detection of anti-social behaviour in online communication in arabic. *Procedia Computer Science*, 142:174–181. Arabic Computational Linguistics.

Nuha Albadi, Maram Kurdi, and Shivakant Mishra. 2018. Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 69–76.

Reem ALBayari, Sharif Abdullah, and Said A. Salloum. 2021. Cyberbullying classification methods for arabic: A systematic review. In *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2021)*, pages 375–385, Cham. Springer International Publishing.

Huda Aljalaoud, Kia Dashtipour, and Ahmed Y. Al-Dubai. 2025. Arabic cyberbullying detection: A comprehensive review of datasets and methodologies. *IEEE Access*, 13:69021–69038.

Raghad Alshaalan and Hend Al-Khalifa. 2020. Hate speech detection in saudi twittersphere: A deep learning approach. In *Proceedings of the fifth Arabic natural language processing workshop*, pages 12–23.

Shammur Absar Chowdhury, Hamdy Mubarak, Ahmed Abdelali, Soon-gyo Jung, Bernard J. Jansen, and Joni Salminen. 2020. A multi-platform Arabic news comment dataset for offensive language detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6203–6212, Marseille, France. European Language Resources Association.

Batoul Haidar, Maroun Chamoun, and Ahmed Serhrouchni. 2017. Multilingual cyberbullying detection system: Detecting cyberbullying in arabic content. In *2017 1st cyber security in networking conference (CSNet)*, pages 1–8. IEEE.

Fatemah Husain and Ozlem Uzuner. 2021. A survey of offensive language detection for the arabic language. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 20(1):1–44.

Tarek Kanan, Amal Aldaaja, and Bilal Hawashin. 2020. Cyber-bullying and cyber-harassment detection using supervised machine learning techniques in arabic social media contents. *Journal of Internet Technology*, 21(5):1409–1421.

Hamdy Mubarak, Hend Al-Khalifa, and AbdulMohsen Al-Thubaity. 2022. Overview of osact5 shared task on arabic offensive language and hate speech detection. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 162–166.

Hala Mulki and Bilal Ghanem. 2021. Let-mi: An Arabic Levantine Twitter dataset for misogynistic language. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 154–163, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. 2019. L-HSAB: A Levantine Twitter dataset for hate speech and abusive language. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 111–118, Florence, Italy. Association for Computational Linguistics.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Fatima Shannag, Bassam H Hammo, and Hossam Faris. 2022. The design, construction and evaluation of annotated arabic cyberbullying corpus. *Education and Information Technologies*, 27(8):10977–11023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.