RedditOSINT: Automated OSINT Collection from Reddit and Analysis Using Machine Learning

Kowsar Mahmud Pappu[1], Sheikh Hasanul Banna[1], Md Shohrab Hossain[1], and Md Abdur Rahman[2]

[1] Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, Bangladesh,
kowsarmahmud156@gmail.com, sheikhevan0@gmail.com, shohrab@cse.buet.ac.bd
[2] Cyber Security and Forensic Computing Department, University of Prince Mugrin, Medina, KSA
m.arahman@upm.edu.sa

Abstract. Open Source Threat Intelligence (OSINT) that uses public information to detect cyber threats is crucial in cybersecurity. While Twitter has led real-time security discussions, Reddit is becoming a valuable OSINT source due to its topic-focused subreddits, where experts share insights on vulnerabilities and incidents. Despite this, Reddit remains underused in OSINT research. Existing research often prioritizes public data and mainstream social media platforms. The use of Natural Language Processing (NLP) in OSINT is also very limited, offering room for innovation. This paper presents an automated OSINT collection and analysis system that uses machine learning techniques to extract security-related information from Reddit. The core of our approach is a predictive model which is trained to classify Reddit posts based on their relevance to cybersecurity threats. It filters out general discussions and highlights posts containing actionable threat intelligence. Using historical cybersecurity data, the model is continuously refined to improve its accuracy in identifying high-value information. Notably, our system employs lightweight machine learning models due to hardware limitations, yet still achieves high-confidence threat detection. The findings suggest that even resource-efficient models can significantly enhance OSINT capabilities. The findings demonstrate that our approach enhances early threat identification, reduces analyst workload, and enables real-time situational awareness—offering a practical, resource-efficient solution for improving OSINT capabilities in cybersecurity.

1 Introduction

Open-source intelligence refers to the process of gathering, analyzing, and disseminating information from publicly available sources to gain intelligence. By utilizing diverse open sources such as websites, social media platforms, public records, government reports, news articles, and other freely accessible resources,

OSINT enables organizations and individuals to obtain valuable insights in a legal and ethical manner [10]. Social media holds valuable OSINT data through metadata and user discussions on cybersecurity threats. Cybercriminal traces and threat-related communities make it a rich source. Monitoring these discussions and posts could give us great insight about the threats and give us a chance to defend the victims in a more advantageous manner [3]. These posts and discussions are written in natural human language. Therefore, applying modern natural language processing approaches to find verdict about them could show us some light in this regard.
Existing OSINT research often uses government records and social media metadata. Some studies apply machine learning to extract insights from platforms like Twitter and Facebook. While many sites focus on casual content, **Reddit** stands out for its technical discussions on cybersecurity incidents, emerging threats, and technological advancements in dedicated subreddits. Its structured communities allow professionals, researchers, and enthusiasts to exchange valuable insights on security vulnerabilities, exploits, and mitigation strategies [1]. With NLP tools like BERT, LLaMA, and GPT, meaningful insights and sentiment analysis can be extracted. This study used Reddit data and NLP to identify cyber threats, demonstrating strong potential for automated intelligence gathering. The key contributions of this research are as follows:

- **Dataset Creation**: We introduce a novel dataset curated from security-related discussions on social media, specifically designed to identify potential cybersecurity threats. To the best of our knowledge, no such publicly available dataset existed prior to this work.
- **Utilizing Reddit and Large Language Models (LLMs) for Vulnerability Analysis**: This research extends the application of LLMs and NLP techniques to the detection of security vulnerabilities within Reddit discussions, an area that has received limited attention in previous studies.
- **Vulnerability Detection Using RNN, CNN, and LSTM models**: We leveraged advanced deep learning models, including Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and Long Short-Term Memory (LSTM) networks, to process and classify security-related posts.

This study highlights the effectiveness of combining NLP-based classification methodologies with deep learning to extract valuable cybersecurity insights from social media, contributing to the advancement of automated cyber threat intelligence.

Rest of the paper is organized as follows. Section 2 introduces related terminologies and existing works. Section 3 describes our proposed methodology. Section 4 summarizes our findings and reasonings. Section 5 explains the contribution of this research and possible future improvements.

2 Background and Literature Review

In this section, various related terminologies are explained in brief and a few existing works are presented.

2.1 Terminologies

OSINT: Open-Source INTelligence (OSINT) refers to intelligence derived from publicly accessible information that is systematically collected, processed, analyzed, and disseminated to relevant stakeholders to fulfill specific intelligence needs [4]. Publicly available information sources include mass media, social media, forums and blogs, public and government data and research publications [11].

Reddit and Subreddit: Reddit is a social media platform that functions as a hub for a wide range of forums. The forums are called subreddits [8]. A Subreddit usually contains a specific interest and users who engage in discussions of that interest.

AI-Driven OSINT: AI-Driven OSINT refers to the use of Artificial Intelligence (AI) methods to analyze publicly available data to gather intelligence. It employs various Large Language Models (LLMs) and Natural Language Processing (NLP) methods. OSINT is used for social sentiment analysis, cybercrime investigation, and cybersecurity defense. The evolution of OSINT has seen a shift from manual collection to AI-driven analysis [7].

2.2 Literature Review

In this section, we have provided an overview of recent literature relevant to our work and gaps in those studies.

OSINT for cyber crime and security analysis OSINT has always been a very important part of cyber security academics and interests. In [6], the researchers analyzed different use cases of OSINT from social media in the military. Different military organizations used Instagram, Facebook and Twitter posts and media metadata to find out different security threats.

Threat prediction using OSINT One method of utilizing OSINT for cybersecurity was suggested using open-source intelligence (OSINT) from both the surface web and deep web hacker forums to identify cyber threat-related texts [2]. The approach combined threat indicators from these sources to provide cybersecurity experts and law enforcement with actionable intelligence and achieved 82% accuracy in detecting cyber threats. Additionally, the authors analyzed over 10 billion breached records from 8,000+ reported cases (2005-2019, US) using data from the Privacy Rights Clearinghouse (PRC).

AI-Driven OSINT on Social Media Data Using social media is comparatively new approach in the OSINT arena. One research work was done over Twitter posts to find imminent cyber security threats which inspired our works, the tool named TwitterOSINT [5]. This was a tool designed to extract and visualize publicly available data from Twitter in near-real time.

OSINT with ML For Intrusion Detection System Another approach introduced IDS-ML, an open-source code repository for developing Intrusion Detection Systems (IDSs) using machine learning (ML) techniques [12]. The repository provides implementations of multiple IDS frameworks optimized for detecting various cyberattacks, including signature-based, anomaly-based, and hybrid IDSs. This approach implemented tree-based, ensemble-based (LCCDE), and hybrid IDS (MTH-IDS) models.

GPT-based OSINT Framework One recent work proposes a GPT-powered framework for automated feature extraction from cyber threat-related texts, replacing traditional manual methods [9]. It enables non-technical cyber strategists to analyze cyber incidents using natural language queries, with AI-driven insights presented in plain language.

Gap Analysis Existing works have elucidated some important limitations:

- **Limited Use of NLP**: Use of NLP frameworks in OSINT is still limited at present. NLP methods can understand the semantic meaning of language. NLP methods have the potential to accurately filter out irrelevant data, which is still a big challenge in OSINT.
- **Scarcity of Reddit-based OSINT**: Reddit has not been widely used for OSINT. Especially with LLM methods, it is often ignored. However, reddit has a few advantages over other social media, such as:
  - Reddit hosts specialized subreddits (e.g., r/cybersecurity, r/netsec) where professionals and enthusiasts actively share updates, vulnerabilities, and discussions. This does not generally apply to other social media.
  - Compared to platforms like Twitter/X and Facebook, where hashtags and trending topics can dilute valuable cybersecurity content, Reddit's structured subreddits keep discussions focused. So, there is a low noise-to-signal ratio.

3 Proposed Methodology

We decided to analyze Reddit forum posts to detect imminent cyber threats using deep learning approaches. The lineup of our work was to fine-tune some renowned NLP models for classification tasks with our dataset made from relevant Reddit posts and then observe their metrics. To achieve this, we had several milestones to pass through, which are shown in Fig. 1.

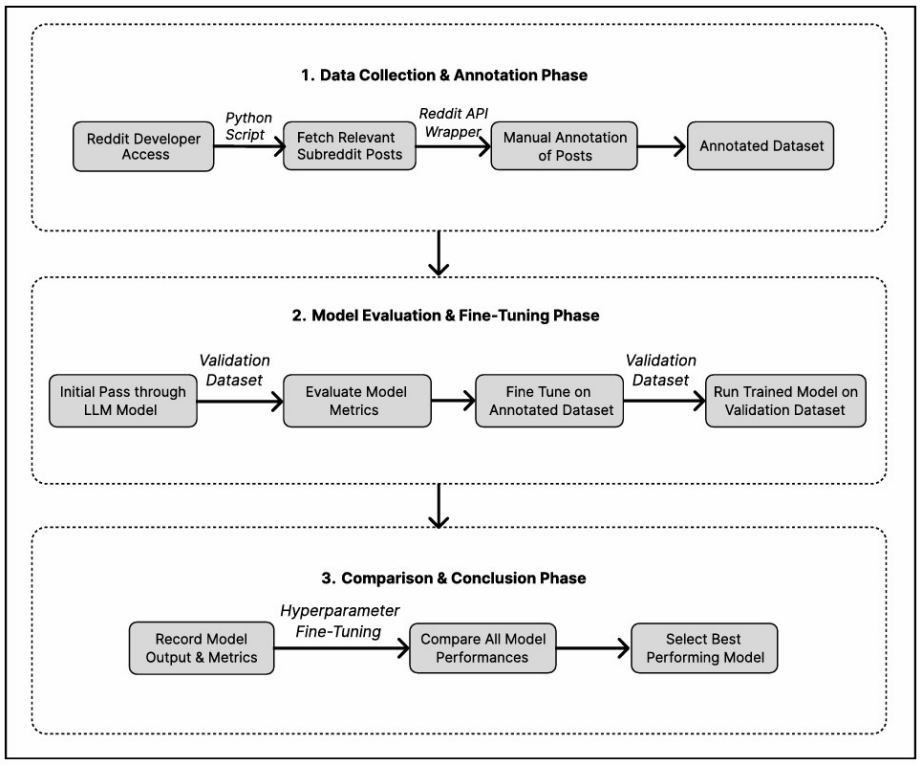

Fig. 1: Overview of the Methodology

3.1 Overview of the Workflow

1. **Data Collection**: As there is no publicly available, standardized dataset specifically focused on cybersecurity-related discussions on Reddit, we created our own dataset using Reddit's official API. As shown in Fig. 2, we began by registering a developer application through Reddit's API portal, which granted us access credentials necessary for programmatic data retrieval.

   Using the Python Reddit API Wrapper (PRAW), an official and API-compliant Python interface for Reddit, we wrote customized scripts to collect posts from a curated list of subreddits known for active discussions on cybersecurity topics. These included `r/cybersecurity`, `r/Malware`, `r/CyberCrime`, `ransomwarehelp` etc. The subreddits were selected based on their relevance, activity level, and history of hosting informative content related to real-world threats, vulnerabilities, and industry trends.

   To ensure the dataset reflected current threat intelligence discourse, we focused on posts published within the past year. For each post, we retrieved key fields such as the subreddit name, post title, full text, creation timestamp,

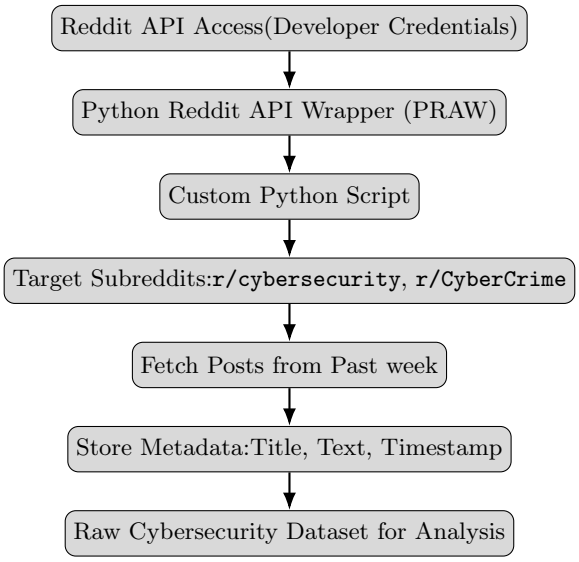

Fig. 2: Reddit Data Collection Workflow

number of comments, and upvote count. All data was collected in accordance with Reddit's terms of service and API usage policies. The collected posts were stored in structured formats suitable for downstream analysis and modeling.

2. **Data Annotation**: Due to the absence of any publicly available dataset that met the specific requirements of our research problem, it was necessary to construct a custom dataset from scratch. This process involved the manual annotation of each data sample to ensure the accuracy and relevance of the labels. Each entry was carefully reviewed and categorized based on predefined criteria aligned with the objectives of the study.
   The annotation task was particularly time-consuming, as it required both domain knowledge and consistent judgment to maintain labeling quality across the dataset. To minimize bias and improve the reliability of the annotations, we followed a systematic labeling strategy that included:
   - Establishing clear labeling guidelines before the annotation began.
   - Periodic reviews of the annotations to identify and correct inconsistencies.
   - Cross-validation of samples by multiple annotators where ambiguity was present.

   This manual annotation process shown in Fig. 3 significantly contributed to the quality and trustworthiness of the dataset, forming a strong foundation for the subsequent stages of model development and evaluation.
3. **Classification Models**: At first we chose some big models renowned for text-generation tasks as they are more capable to reason with human languages. These included Llama-3.2-1B, Mistral, GPT-3, GPT-3.5 & GPT-4.

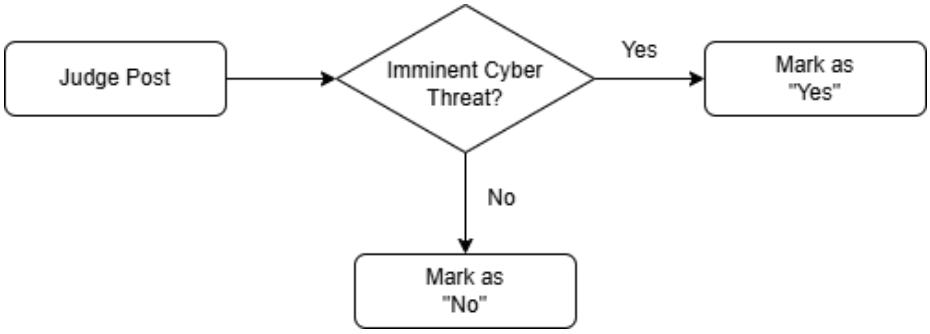

Fig. 3: Data Annotation Workflow

But these models turned out to be pretty much expensive in terms of hardware requirements. Therefore, we had to look for some lightweight models that would do very well with our tasks. Fortunately we found some NLP models those are very much capable of classification tasks with natural language. These included AWD-LSTM, ULMFiT, DistilBERT, TextCNN and TinyBERT.

4. **Fine-Tuning**: We fine tuned the NLP models using our custom made dataset. The dataset contained 9320 samples which were processed posts from Reddit. 80% of the dataset was used for training, 10% was used for validation & 10% was used for testing.
5. **Record & Compare**: We fine-tuned different NLP models in the same manner and recorded their performances to compare them.

3.2 Strength of our approach

The primary strength of our approach lies in its novel integration of Reddit-based OSINT with lightweight yet effective natural language processing models, enabling accurate and efficient detection of cybersecurity-related threats. Unlike prior studies that predominantly rely on metadata or mainstream platforms such as Twitter, our work focuses on Reddit - an underutilized but content-rich platform - where subject-matter experts and practitioners actively engage in structured, technical discussions within dedicated subreddits.

Our methodology has several key strengths:

1. **Domain-Specific Data Collection:** By targeting cybersecurity-focused subreddits, we ensure the collection of highly relevant and context-specific data, thereby reducing noise and improving the quality of inputs for downstream analysis.
2. **Custom Dataset Creation and Annotation:** In the absence of any publicly available Reddit-based OSINT dataset, we constructed a novel, manually annotated dataset tailored to threat detection, significantly enhancing model training and evaluation fidelity.

3. **Resource-Efficient Model Selection:** Recognizing hardware constraints, we adopted and fine-tuned lightweight NLP models such as DistilBERT and TinyBERT. These models demonstrated strong performance metrics while maintaining low computational requirements, making our approach suitable for deployment in resource-limited environments.
4. **Fine-Tuning for Contextual Relevance:** The models were fine-tuned on a dataset specifically curated for cybersecurity discourse, enabling them to capture domain-specific terminology, context, and linguistic patterns more effectively than general-purpose models.

3.3 Limitations & Challenges

Throughout this study, we encountered several challenges that influenced our methodologies. These include:

- **Computational Resource Constraints**: The deployment of certain Large Language Models (LLMs) necessitates significant computational resources, which were beyond our available infrastructure. For instance, LLaMA 3.2 models require approximately 32GB of GPU memory, which was not accessible for this research.
- **Dataset Limitations**: No publicly available dataset existed for security-related discussions on Reddit. Consequently, our synthesized dataset was relatively small, potentially impacting model generalization.
- **Subjectivity in Labeling**: The annotation labels in our dataset were manually assigned, introducing the possibility of human bias, which could affect the consistency and reliability of the model's training data.

4 Results

4.1 Performance Metrics

We evaluated our models using four standard classification metrics: accuracy, precision, recall, and F1-score, which together provide a balanced understanding of each model's performance in identifying cybersecurity-related posts from Reddit.

- **Accuracy** measures the overall correctness of the model by calculating the proportion of correctly classified instances out of all predictions. While useful, it can be misleading in imbalanced datasets, which is why we also consider other metrics.
- **Precision** refers to the proportion of true positive predictions among all positive predictions made by the model. In the context of cybersecurity, high precision ensures that flagged posts are indeed relevant, reducing the burden of false alarms on analysts.
- **Recall**, also known as sensitivity, measures the proportion of actual positive cases (i.e., relevant threat-related posts) that the model successfully identifies. A high recall is critical in OSINT applications, where missing potential threats could lead to security oversight.

– **F1-score** is the harmonic mean of precision and recall, offering a single metric that balances both. It is especially valuable when there is a trade-off between precision and recall, providing a more comprehensive assessment of model performance.

These metrics were selected to ensure that our evaluation captures not only how often the models are correct but also how reliably they can identify and isolate actionable threat intelligence from general discussions. This is particularly important for OSINT systems intended to support real-world cybersecurity operations.

4.2 Hyperparameter Fine-Tuning

To optimize the performance of the models we worked with on the OSINT classification task, we fine-tuned several key hyperparameters. These include learning rate, batch size, number of training epochs, warmup steps, weight decay, and mixed-precision training (FP16). The objective was to balance precision, recall, and F1-score while preventing overfitting or underfitting. The following adjustments were made:

– **Learning Rate:** We experimented with values in the range of 1×10^{-5} to 5×10^{-5}, selecting a lower learning rate to allow more stable convergence, which resulted in improved precision but slightly reduced recall.
– **Batch Size:** Batch sizes of 8, 16, and 32 were tested. A batch size of 16 offered a good trade-off between training stability and resource efficiency.
– **Number of Epochs:** We varied epochs from 2 to 5. While more epochs improved overall accuracy and precision, excessive training caused a slight drop in recall due to overfitting on the majority class.
– **Warmup Steps:** Varying warmup steps, we ex helped prevent unstable updates in early training, contributing to smoother convergence.
– **Weight Decay:** Several decay factors were applied to test which configuration regularizes the model better and avoid overfitting, with a modest improvement in generalization performance.
– **FP16 Training:** Mixed-precision training was enabled to accelerate training on compatible GPUs without significantly affecting model metrics.

4.3 Result Summary

After fine-tuning, the best-performing model, **DistilBERT**, achieved an accuracy of 98%, a precision of 97%, a recall of 98%, and an F1-score of 98%. These results indicate that DistilBERT is not only accurate but also highly effective at identifying relevant cybersecurity content, as reflected by its high recall. This is particularly important in threat detection tasks, where failing to detect relevant threats (i.e., false negatives) can lead to significant security risks.

TinyBERT, despite being a smaller and more lightweight model, also delivered competitive results. It achieved an accuracy of 88%, a precision of 86%, a

recall of 90%, and an F1-score of 88%. While its overall performance was slightly lower than DistilBERT across all metrics, TinyBERT offers a favorable trade-off between performance and computational efficiency, making it well-suited for deployment in resource-constrained environments.

TextCNN also performed exceptionally well. It achieved an accuracy of 95%, a precision of 95%, a recall of 94%, and an F1-score of 95%. These results demonstrate that TextCNN is highly capable of correctly classifying relevant cybersecurity content, maintaining a strong balance between precision and recall. Its slightly lower recall compared to DistilBERT suggests a marginally higher rate of missed relevant threats, but overall, TextCNN remains a robust and reliable model for threat detection tasks. Moreover, its architecture allows for faster training and inference compared to larger transformer-based models, making it a practical choice for scenarios where computational efficiency is important.

As shown in Table 2, these models significantly outperformed their respective pre-fine-tuning versions reported in Table 1, demonstrating that task-specific fine-tuning can substantially enhance model effectiveness in OSINT-related classification. The confusion matrices in Fig. 4, Fig. 5 and Fig. 6 further illustrate these improvements, showing a better distribution of true positives and a notable reduction in misclassifications.

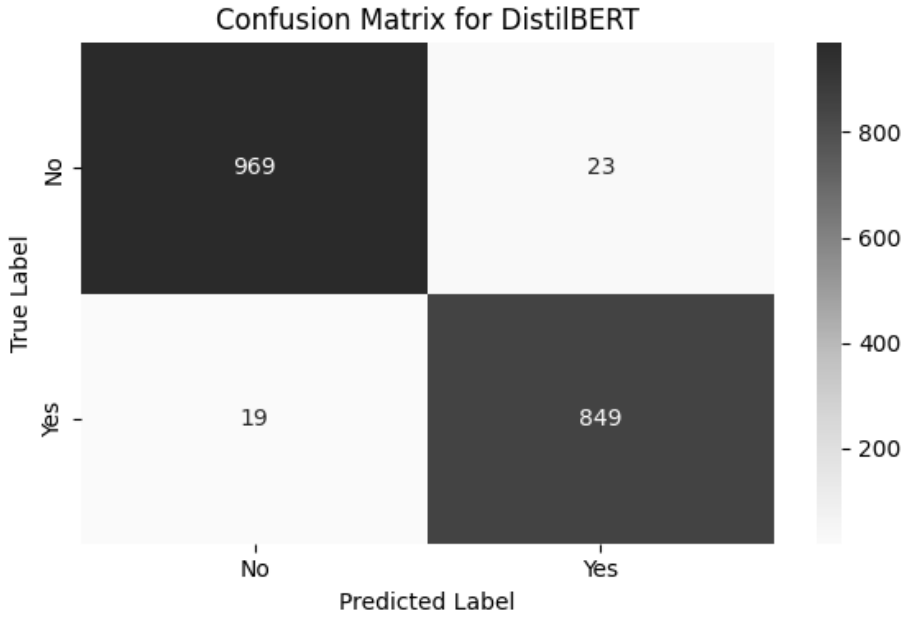

Fig. 4: Confusion Matrix for DistilBERT after Fine-Tuning

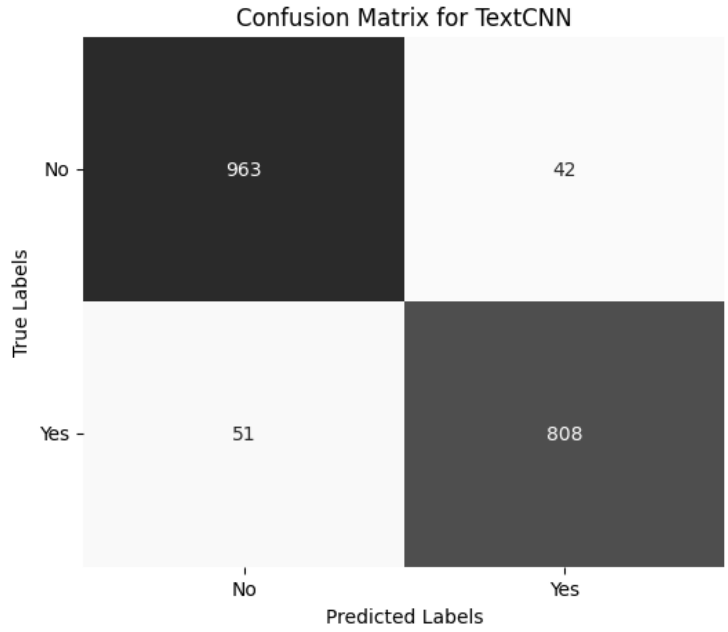

Fig. 5: Confusion Matrix for TextCNN after Fine-Tuning

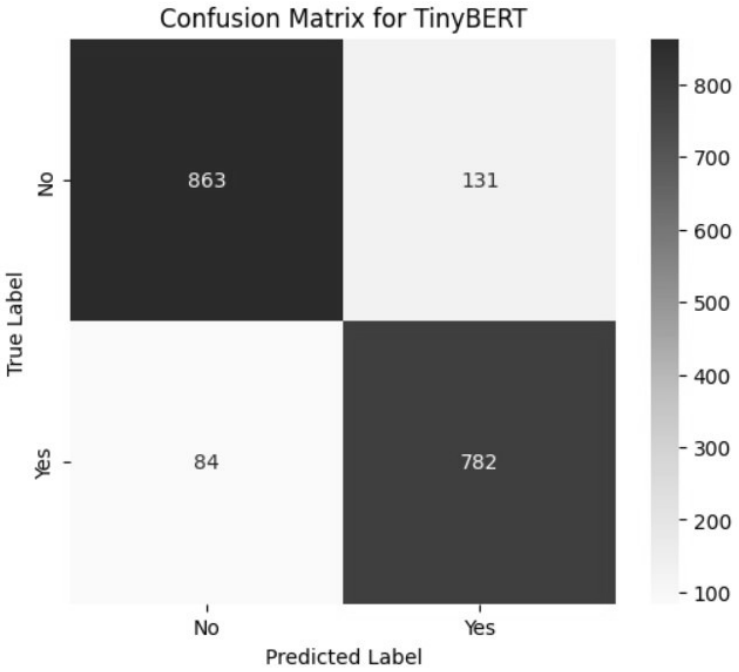

Fig. 6: Confusion Matrix for TinyBERT after Fine-Tuning

The notable improvement in model performance following fine-tuning can be attributed to several factors.

- The models were trained on a domain-specific dataset curated from cybersecurity-related Reddit discussions. Fine-tuning on this task-specific corpus allowed the models to adapt to the unique linguistic patterns, terminologies, and contextual cues present in cybersecurity discourse, something that general-purpose pretraining might not fully capture.
- The manual annotation process ensured high-quality, noise-free labels, which contributed to better learning outcomes.
- Lightweight transformer models like DistilBERT and TinyBERT, although pretrained on general corpora, are efficient at transfer learning when guided by good data despite being small.

The efficiency of our approach in filtering out irrelevant data and structuring threat information stems from the integration of fine-tuned NLP models with a carefully curated dataset. By training on labeled Reddit posts specifically annotated for imminent cybersecurity threats, our models learned to distinguish between general discussions and actionable threat intelligence with high precision and recall.

Additionally, Reddit's subreddit-based structure inherently organizes discussions by topic, which reduces the noise-to-signal ratio compared to other social platforms. This topical coherence further improves the model's ability to isolate relevant content.

Thus, our approach significantly enhances OSINT-based cyber threat detection through effective noise filtration and structured threat information extraction.

Table 1: Performance Metrics Before Fine-Tuning

Model Name	Accuracy	Precision	Recall	F-1 Score
AWD-LSTM	0.52	0.36	0.06	0.10
ULMFiT	0.55	0.62	0.07	0.12
DistilBERT	0.56	0.65	0.14	0.23
TinyBERT	0.54	0.55	0.01	0.03
TextCNN	0.54	0.78	0.01	0.02

5 Conclusion

By extending the system to multiple data sources, refining LLM capabilities, improving structured data extraction, and integrating real-world deployment

Table 2: Performance Metrics After Fine-Tuning

Model Name	Accuracy	Precision	Recall	F-1 Score
AWD-LSTM	0.79	0.75	0.82	0.78
ULMFiT	0.79	0.75	0.82	0.79
DistilBERT	0.98	0.97	0.98	0.98
TinyBERT	0.88	0.86	0.90	0.88
TextCNN	0.95	0.95	0.94	0.95

features, this research can evolve into a *fully automated cybersecurity threat intelligence system*. These future advancements will enable more *accurate, scalable and actionable cybersecurity insights* for researchers and security professionals.

Future work could expand beyond Reddit to include other platforms, even dark web sources, enabling richer, multisource threat intelligence. Enhancing LLM capabilities through domain-specific fine-tuning and integration with cybersecurity knowledge bases can improve detection accuracy. Structured data extraction could be refined with techniques like Named Entity Recognition and graph-based modeling to better map relationships between threat entities. Advanced visualizations, such as interactive dashboards and temporal or geospatial mapping would help analysts monitor trends more effectively. Finally, deploying the system as an API and integrating it with SIEM tools would move it toward real-world applicability.

References

1. Achuthan, K., Khobragade, S., Kowalski, R.: Public sentiment and engagement on cybersecurity: Insights from Reddit discussions. Computers in Human Behavior Reports 17, 100573 (2025)
2. Adewopo, V., Gonen, B., Adewopo, F.: Exploring Open Source Information for Cyber Threat Intelligence. In: 2020 IEEE International Conference on Big Data (Big Data), pp. 2232–2241 (2020). doi:10.1109/BigData50022.2020.9378220
3. Ait Maalem Lahcen, R., Caulkins, B., Mohapatra, R., Kumar, M.: Review and insight on the behavioral aspects of cybersecurity. Cybersecurity 3(1), 10 (2020)
4. Böhm, I., Lolagar, S.: Open source intelligence: Introduction, legal, and ethical considerations. International Cybersecurity Law Review 2 (2021). doi:10.1365/s43439-021-00042-7
5. HoppA, D., Hsieh, K.C.: TwitterOSINT: Automated Open Source Intelligence Collection, Analysis & Visualization Tool. Annual Review of Cybertherapy & Telemedicine (2020)
6. Ju, Y., Li, Q., Liu, H., Cui, X., Wang, Z.: Study on application of open source intelligence from social media in the military. Journal of Physics: Conference Series 1507 (2020)

7. Kolade, T.M., Obioha-Val, O., Balogun, A., Gbadebo, M., Olaniyi, O.: AI-Driven Open Source Intelligence in Cyber Defense: A Double-edged Sword for National Security. Asian Journal of Research in Computer Science 18, 133–153 (2025)
8. Medvedev, A., Lambiotte, R., Delvenne, J.-C.: The Anatomy of Reddit: An Overview of Academic Research. Springer Proceedings in Complexity, pp. 183–204 (2019)
9. Sufi, F.: An innovative GPT-based open-source intelligence using historical cyber incident reports. Natural Language Processing Journal 7, 100074 (2024)
10. Szymoniak, S., Foks, K.: Open Source Intelligence Opportunities and Challenges: a Review. Advances in Science and Technology Research Journal 18, 123–139 (2024)
11. Tabatabaei, F., Wells, D.: OSINT in the Context of Cyber-Security. pp. 213–231 (2016). doi:10.1007/978-3-319-47671-1_14
12. Yang, L., Shami, A.: IDS-ML: An open source code for Intrusion Detection System development using Machine Learning. Software Impacts 14, 100446 (2022). doi:10.1016/j.simpa.2022.100446