International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307 OPEN ACCESS

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/CSEIT2410118

Enhanced Audio-Based Open-Source Intelligence Insights using Machine Learning

Muhammad Ayub*, Sidra Irum, Dr. Zunera Jalil

Department of Cyber Security, Air University, Islamabad, ICT, Pakistan

ARTICLEINFO

Article History:

Accepted: 15 Jan 2024

Published: 31 Jan 2024

Publication Issue

Volume 10, Issue 1

January-February-2024

Page Number

141-149

ABSTRACT

Nowadays, data collection methods and techniques are increasingly used to address intelligence needs in the sense of training models to predict correct information. Open-source intelligence (OSINT) could now incorporate Machine Learning (ML) by correlating diverse data types, such as text, images, audio, and video. In this research, we focused on an essential yet underdeveloped aspect of OSINT, extracting insights from audio data for military intelligence, especially in Pakistan's defence and focused on developing advanced tools for analyzing the expanding audio data, proposing a novel method to extract perfect information for intelligence purposes, specifically targeting key entities like Location, Rank, Operation, Date, and Weapon in military contexts. First, we developed a unique dataset containing 2000 transcribed sentences with annotations for the mentioned entities using an open-source NER annotator. Then, we trained four customized models using advanced NLP frameworks such as Hugging Face's Transformers (DistilBERT), spaCy, NLTK and Stanford CoreNLP, which are subject of assessment to determine their practical use in intelligence contexts. The selected models were evaluated, which proved that AI-based techniques are crucial for enhancing intelligence gathering in the dynamic OSINT landscape. The results also demonstrated the potential of AI integration in OSINT for audio data processing in military intelligence.

Keywords: Audio based OSINT, Audio NER, AI based OSINT, Machine Learning and OSINT

I. INTRODUCTION

In the evolving landscape of intelligence and data analysis, the emergence of sophisticated data collection methods and techniques has revolutionized the way information is gathered and interpreted. Open-source intelligence (OSINT), a field historically reliant on human analysis of publicly available data, is undergoing a transformative shift with the integration of Machine Learning (ML) technologies. This fusion has opened new avenues for comprehensively analyzing various data types, including text, images, audio, and video, to derive actionable intelligence.

The present research delves into a significant yet relatively underdeveloped aspect of OSINT: the extraction of insights from audio data for military intelligence purposes. With a particular focus on Pakistan's defence mechanisms, this study aims to contribute to the development of advanced tools for

analyzing the expanding corpus of audio data. This is of paramount importance given the increasing volume and complexity of audio information in military contexts. The core of this research involves proposing a novel methodology for the extraction of precise intelligence from audio data. This methodology specifically targets the identification of key entities such as location, rank, operation, date, and weapon, which are crucial in military intelligence. To achieve this, the study first involved the creation of a unique dataset containing 2000 transcribed text, each meticulously annotated for the mentioned entities using an open source Named Entity Recognition (NER) annotator. Following this, the research progressed to the training of four customized models employing state-of-the-art Natural Language Processing (NLP) frameworks. These frameworks include Hugging Face's Transformers (DistilBERT), spaCy, NLTK, and Stanford CoreNLP. The selection of these diverse and advanced frameworks was driven by the need to assess their practical applicability and effectiveness in real-world intelligence scenarios. The evaluation of these models is a crucial part of this study, as it aims to establish the efficacy of AI-based techniques in enhancing intelligence gathering processes within the dynamic realm of OSINT. This research is not only about developing a methodology but also about critically assessing the potential and limitations of AI integration in OSINT, particularly for processing audio data in military intelligence.

In this research, we address the significant challenge of efficiently processing and extracting actionable intelligence from extensive audio data, particularly in the crucial field of military intelligence where accuracy, speed, and reliability are paramount. The study introduces an innovative methodology for audio-based Open-Source Intelligence (OSINT) to enhance audio data processing and interpretation for strategic military applications. Key research questions explore how AI methodologies can improve OSINT audio data processing, the challenges of integrating Machine Learning with OSINT in Pakistan's military intelligence, and the accuracy of customized NLP models in extracting military entities from audio data. Objectives include developing a novel OSINT approach, leveraging AI for military intelligence, and enhancing existing OSINT methods. Major contributions encompass creating a customized dataset, comparing various NER frameworks, and exploring AI and ML applications in military OSINT. This research focuses on developing AI-driven methods for audio-based OSINT in Pakistan's military context, involving detailed annotations and training of advanced NER models, and is geared towards providing insightful, detailed information for future research in military intelligence.

The article is organized into four main sections. Section 2 (**Literature Review**) provides a thorough examination of existing research and developments in the fields of AI, NLP, and OSINT, particularly in the context of audio data analysis in military intelligence. Section 3 (**Methods and Material**) details the methodologies, including transcription processes, annotation techniques, and any special equipment or modifications, complete with relevant illustrations. Section 4 (**Results and Discussion**) presents the research findings and provides a comprehensive analysis of these results, focusing on the effectiveness of the AI models in military intelligence and addressing encountered challenges. Finally, Section 5 (**Conclusion**) offers a summary of the key insights, underlining the impact and potential future implications of AI in the realm of military intelligence and OSINT.

II. Literature Review

In terms of OSINT, it means gathering information through open sources to conduct an intelligence operation. It is the information that is collected through numerous sources among them being printed news (newspapers, televisions, radio, and digital

publications), the internet like websites, forums, or social media platforms the government bodies, and finally the academic and professional journals or books. Unlike most intelligent information that comes covertly, OSINT depends on openly available data only and is generally known among people. Therefore, there is increased awareness of how important OSINT is in areas such as national security, law enforcement, business intelligence, and cyber security. Today digital world with an abundance of online information available, OSINT technique is particularly pertinent here [1]. Analysts can obtain vital information without being involved in the complexities associated with legal and ethical dimensions surrounding other concealed forms of intelligence acquisition. Besides, the growth of social media and the presence of other digital platforms have contributed to the expansion and convenience of OSINT which are crucial in current information management concepts.

Open-Source Intelligence (OSINT) is a rapidly evolving field that has gained significant traction in both government and private sectors due to the explosion of publicly accessible information available through the internet, social media, and other digital platforms [2]. OSINT refers to the process of collecting, analyzing, and disseminating intelligence from publicly available sources [3]. This review explores the literature surrounding OSINT, focusing on its evolution, methodologies, applications, and challenges.

Historically, intelligence gathering was largely confined to covert methods. However, with the advent of the digital era, there has been a paradigm shift. Johnson and Wirtz in their works note that OSINT began to gain prominence in the late 20th century as the volume of open-source information exploded with the advent of the internet. The proliferation of smartphones and social media platforms has further accelerated this growth [2]. The methodologies employed in OSINT are diverse, as highlighted by Bazzell and various other scholars. They range from simple manual searches to complex algorithms and AI-driven data analytics. The process typically involves data collection, filtration, analysis, and dissemination. Tools and technologies used in OSINT have also evolved, with advanced software and applications facilitating the automation of data collection and analysis [3]. This includes web scraping tools, social media analytics, and sentiment analysis software. The applications of OSINT are vast and varied. In the context of national security, as described by Mates, OSINT is used for counterterrorism, border security, and crisis management. In the corporate world, as noted by Pyrounakis and Kotsiantis, businesses leverage OSINT for market analysis, competitive intelligence, and risk management. Furthermore, in the realm of journalism, OSINT plays a crucial role in investigative reporting and fact-checking, as emphasized by scholars like Isabelle Böhm & Samuel Lolagar [4].

There are concerns regarding privacy, data protection, and the potential for misuse of information, particularly when personal data is involved. This necessitates a careful and ethical approach to OSINT operations, balancing the need for intelligence with respect for individual rights and legal frameworks. Open-Source Intelligence (OSINT) fundamentally revolves around the collection and analysis of information that is publicly available and accessible. As defined by scholars like Landon-Murray, OSINT includes data gathered from the internet, traditional media, public government records, professional and academic publications, and other publicly accessible sources [5]. The fundamentals of OSINT involve identifying relevant sources, collecting data efficiently, and analyzing this data to produce actionable intelligence [6]. Tomislav and others have emphasized the importance of OSINT in various sectors, including national security, business intelligence, and journalism [7].

Current practices in OSINT involve advanced techniques like data mining, web scraping, sentiment analysis, and network analysis [8, 9]. Tools ranging from simple search engines to complex analytical software are employed for OSINT tasks. As noted by Bean, challenges in OSINT include dealing with the vast amount of data, ensuring data reliability, maintaining privacy and ethical standards, and managing the dynamic nature of open sources [10]. The increasing volume of data, particularly from social media, and the prevalence of misinformation and disinformation are significant challenges for OSINT practitioners.

III. METHODS AND MATERIAL

This research is based on a methodological framework that focuses on tackling complicacies regarding processing and extracting actionable intelligence from audio data in an open-source intelligence (OSINT) context within Pakistan Military information. This section outlines the methodology and represents a particular methodological approach based on the distinct nature of such material regarding OSINT. However, audio data differs from plain text since it involves different speaker's accents and ambient audio, and that interpretation should occur within a meaningful context. Therefore, the approach adopted in this research involves different processes, including audio data collection, pre-processing, transcription, annotation, training-specific models, and NER recognition, which are then utilized for the OSINT process.

Figure 1 showing the proposed methodology diagram divided into five parts, four sections and eight steps followed in the proposed methodology of Audio-based OSINT numbered and written in a square box with a solid border line. This whole methodology is divided into four sections with mentioned specific colours. The first section has a blue dotted line border, and the square section has Part 1: Audio to Text; this is Audio to Text transcription with three steps of audio to text process. The first step in this section is 1. Audio Collection second is 2. Audio pre-processing, and the third is 3. Audio Transcription. In the second section, the colour is purple, and the dotted line border of the square there are two parts. Part 2: Annotation has 4. Annotation Process, Part 3: NER Models Training, and Recognition have step 5. Training NER Models, 6. Entity recognition. In the 3rd section, a red colour dotted square box has Part 4: OSINT Process and Step 7. Intelligence Utilization: In the 4th section, part 5, The feedback and iteration process have step 8. Feedback and Iteration.

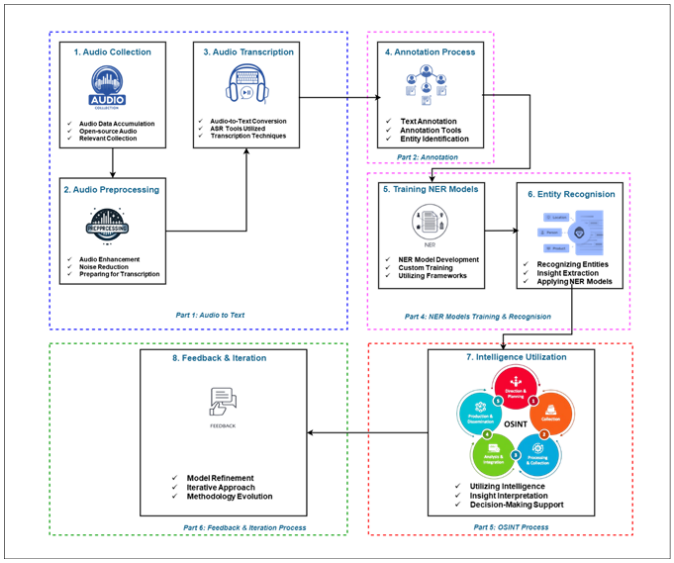

Figure 1. Proposed Methodology

This research outlines a comprehensive methodology for enhancing military intelligence through audio-based Open-Source Intelligence (OSINT). The process begins with the collection of audio data from diverse sources like podcasts, radio broadcasts, and online platforms, ensuring rich content and quality control. In the audio pre-processing stage, techniques like noise reduction and audio enhancement are applied for clearer data. This is followed by audio transcription using Automatic Speech Recognition (ASR) tools, such as Google Cloud Speech-to-Text and IBM Watson Speech to Text, which are tailored to handle military-specific language and terminologies. The next crucial

step is annotation, where audio transcripts are meticulously tagged with military-relevant entities by experts, ensuring precision and consistency. Training Named Entity Recognition (NER) models is then undertaken, utilizing frameworks like SpaCy and Hugging Face's Transformers, focusing on entity types crucial for military intelligence. The research also emphasizes the importance of entity recognition, ensuring the NER models align with the output from ASR and accurately process and extract relevant entities. Finally, the methodology includes the utilization of this intelligence in military OSINT investigations, leveraging the extracted entities to enhance efficiency and efficacy, and concludes with a feedback and iteration process to continually refine and improve the intelligence gathering methods.

A. DistilBERT

In this methodology, a sophisticated Named Entity Recognition (NER) model was developed for extracting military intelligence data, utilizing Google Colab, PyTorch, and the Transformers library. The approach began with integrating Google Drive with Google Colab for seamless dataset access, followed by preparing 1700 annotated sentences in the IOB format for precise entity recognition. The DistilBert Tokenizer Fast from Hugging Face's Transformers library tokenized these sentences, and a custom PyTorch-based Dataset class managed the data during training. The model, DistilBert for Token Classification, was chosen for its efficiency and accuracy in NER tasks. Training involved several epochs with the AdamW optimizer, closely monitored for performance optimization. Post-training, the model was evaluated using metrics like precision, recall, and F1-score, indicating high accuracy in identifying key entities like DATE, LOCATION, OPERATION, RANK, and WEAPON. Results were visualized through charts and heatmaps for interpretability, showcasing the model's effectiveness and guiding further refinements. This methodology exemplifies the use of advanced AI and NLP tools in enhancing OSINT capabilities within military intelligence.

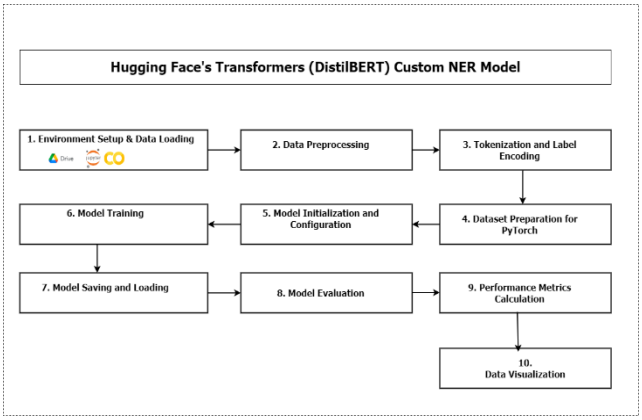

Figure 2. DistilBERT Custom NER model methodology diagram illustrated with steps.

B. NLTK

This methodology, designed for Named Entity Recognition (NER) in military intelligence, combines Python libraries and machine learning within Google Colab's computational environment. The process begins by integrating Google Drive with Colab for direct access to a dataset of 1700 annotated sentences, each richly tagged with military-specific entities. Utilizing Python's Pandas library, the data is efficiently managed and visualized, aiding in understanding its structure. The data is converted into the IOB format, a key step for accurate entity tagging and model training. The methodology further employs the Natural Language Toolkit (NLTK) for tokenization and part-of-speech tagging, essential for feature extraction and providing contextual information to the NER model. Various machine learning models, such as Random Forest, SVM, Logistic Regression, Naive Bayes, and Gradient Boosting, are trained and assessed using scikit-learn tools. This approach allows for comprehensive evaluation through metrics like precision, recall, and F1-score, offering insights into each model's performance in accurately identifying and classifying military-related entities.

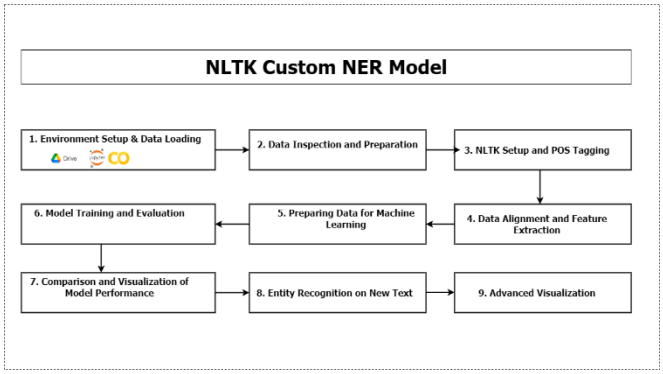

Figure 3. NLTK Custom NER model methodology diagram illustrated with steps.

C. spaCy

This methodology employs advanced data visualization and practical application techniques for Named Entity Recognition (NER) in the context of military intelligence. Utilizing Matplotlib and Seaborn, the model's performance is intuitively visualized through bar charts and heatmaps, highlighting its strengths and weaknesses across different entity types. A crucial phase involves processing new texts to predict and visualize entities, showcasing the model's real-world applicability and accuracy in identifying military-related entities. Additionally, the methodology integrates SpaCy, a sophisticated NLP library, for developing custom NER models. The process, executed in Google Colab, involves mounting Google Drive for direct dataset access, followed by loading and formatting 1700 annotated sentences into SpaCy's training structure. This comprehensive approach highlights the synergy of Python programming, data handling, machine learning, and visualization in enhancing the extraction of valuable insights from complex datasets in specialized domains like military intelligence.

Figure. 4 Named Entity Recognition

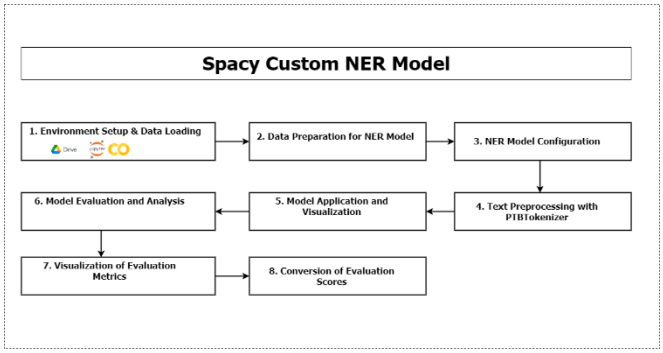

Figure 5. spaCy Custom NER model methodology diagram illustrated with steps.

D. Stanford CoreNLP

This methodology develops a high-performance Named Entity Recognition (NER) system for complex datasets using Google Colab and the Stanford CoreNLP toolkit. The process begins with integrating Google Drive into Colab for easy access to datasets and models. The Stanford CoreNLP package is installed in Colab, enabling advanced natural language processing capabilities.

Text preprocessing uses the Penn Treebank Tokenizer from Stanford CoreNLP to split text into tokens, preparing it for NER. The dataset, comprising 1700 annotated transcribed sentences, is loaded, and visualized for structure and content accuracy, then converted into the IOB format for model training. The training utilizes Stanford CoreNLP's Conditional Random Field (CRF) Classifier, with a custom properties file defining training parameters. This phase ensures the model's accuracy and effectiveness. Rigorous testing and evaluation follow, using separate test data to assess the model's precision, recall, and F1 scores, especially for entities like DATE, LOCATION, OPERATION, RANK, and WEAPON. The final step involves visualizing tagged data, applying the model to new sentences to demonstrate its tagging capabilities and provide an intuitive understanding of its performance. Overall, this methodology leverages advanced tools and frameworks to create a robust solution for extracting insights from complex text data.

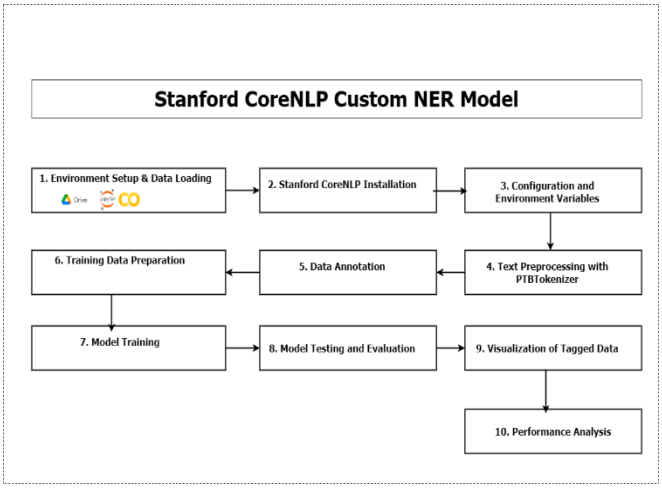

Figure. 6 Stanford CoreNLP Custom NER model methodology diagram illustrated with steps.

IV. RESULTS AND DISCUSSION

This section interprets and discusses the results and their implications, providing a comprehensive understanding of the experimental outcomes and what they suggest about the effectiveness of various Named Entity Recognition (NER) models in the context of military intelligence. The results obtained using Hugging Face's Transformer (DistilBERT) are presented in Table 1, while Table 4 displays the results of the Stanford CoreNLP model. The outcomes of NLTK-based machine learning models are detailed in Table 2, and the results of the SpaCy model are shown in Table 3. Fig. 7 illustrates the performance graph based on five entities: Data, Location, Operation, Rank, and Weapon, focusing on Precision, Recall, and F1 Score. Additionally, Fig. 8 depicts a heatmap based on the performance results of the DistilBERT model. Fig. 9 and 10 showing the performance graph and heatmap.

Table 1. Performance Evaluation of Hugging Face's Transformer (DistilBERT) on Different Entities

Entity	Precision (%)	Recall (%)	F1-Score (%)
DATE	99.0	100.0	99.0
LOCATION	99.0	91.0	95.0
OPERATION	91.0	99.0	94.0
RANK	80.0	72.0	76.0
WEAPON	78.0	86.0	82.0

Table 2. Summary of the NLTK model performance evaluation table.

Model	Precision (%)	Recall (%)	F1-Score (%)
Random Forest	94.00	93.94	93.81
SVM	94.03	93.96	93.82
Logistic Regression	94.03	93.96	93.82
Naive Bayes	93.44	93.25	93.07
Gradient Boosting	93.94	93.86	93.73

Table 3. Performance Evaluation of SpaCy on Different Entities

Entity	Precision (%)	Recall (%)	F1-Score (%)
OPERATION	84.57	89.70	87.06
DATE	95.35	87.23	91.11
WEAPON	38.03	58.06	45.96
RANK	59.21	77.59	67.16

Table 4

Entity	Precision (%)	Recall (%)	F1-Score (%)
LOCATION	82.73	65.34	73.02
OPERATION	88.30	91.52	89.88
RANK	79.55	60.34	68.63
WEAPON	42.74	56.99	48.85

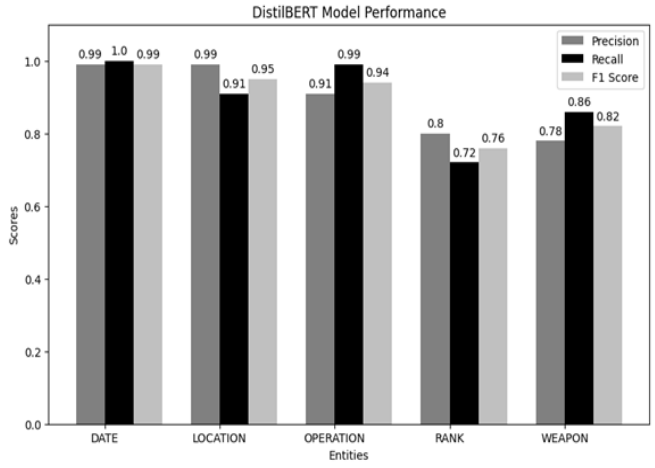

Figure. 7 DistilBERT performance

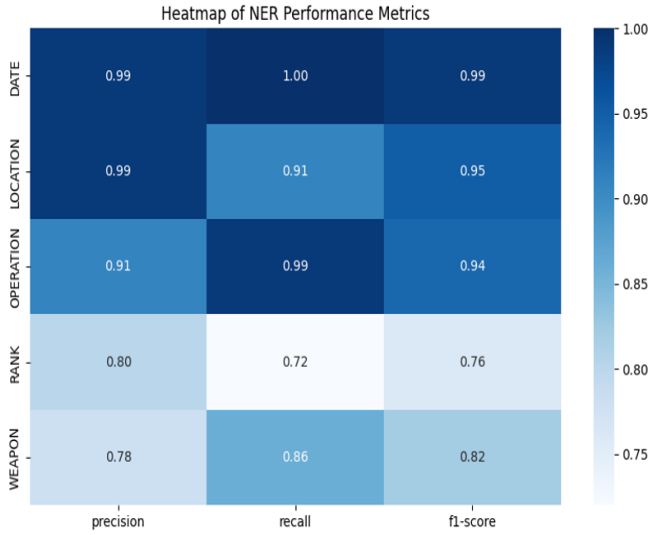

Figure. 8 heatmap of distilBERT

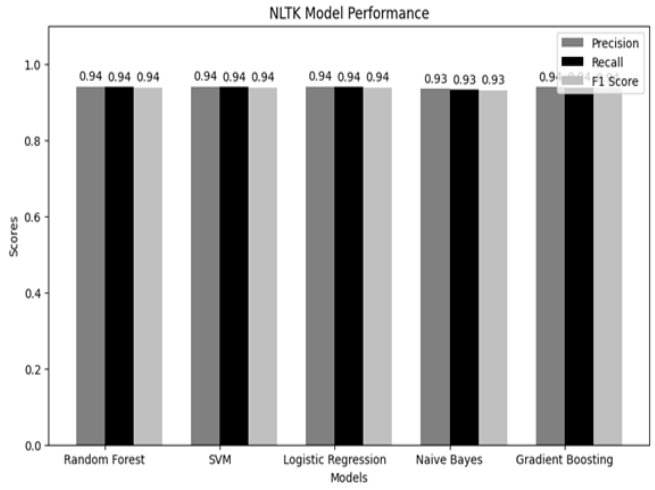

Figure. 9 nltk model performance graph

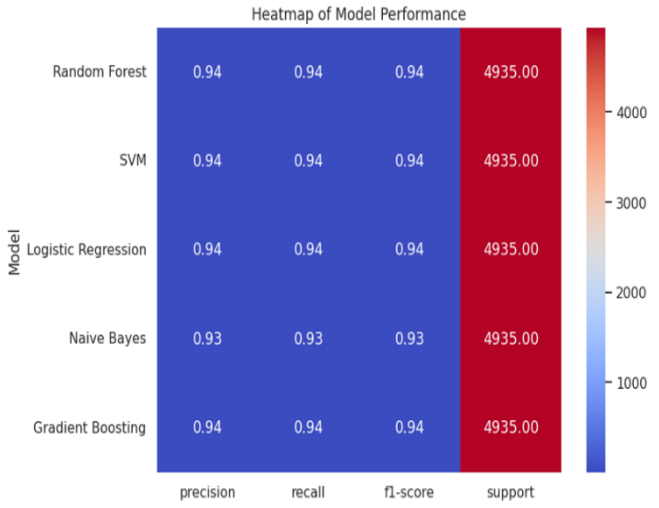

Figure. 10 heatmap of model performance

The discussion covers the practical implications of the findings, particularly in terms of their applicability in real-world intelligence analysis. It also explores the potential impact of these results on future research and development in the field of OSINT and NLP.

V. CONCLUSION

This research embarked on a pioneering journey to integrate advanced Named Entity Recognition (NER) techniques into the realm of Open-Source Intelligence (OSINT), specifically tailored for military intelligence within the context of Pakistan. The key findings of this study are multifaceted, highlighting the effectiveness and adaptability of various NER models - SpaCy, NLTK, Stanford CoreNLP, and Hugging Face's Transformers (DistilBERT) in extracting critical entities such as locations, ranks, operations, dates, and weapons from textual data. The study revealed that while each model has unique strengths, DistilBERT stood out for its superior performance across precision, recall, and F1 scores. This finding underscores the potential of utilizing advanced deep learning techniques in OSINT for military intelligence. Additionally, the research demonstrated the feasibility and value of creating a specialized dataset, tailored for the specific context of Pakistan Military Intelligence, which significantly enhanced the training and performance of the NER

models. Furthermore, the comparative analysis provided insightful revelations into how different models respond to the nuances of military-related text, offering a guide for future implementations in similar contexts. The research concludes that the incorporation of NER models, particularly those based on advanced deep learning techniques like DistilBERT, can significantly enhance the capabilities of OSINT in the military intelligence domain. This research work highlights the significant potential of AI and NLP advancements in Open-Source Intelligence (OSINT), especially for military intelligence. Key future areas include advanced AI for complex data analysis, real-time NLP, enhanced machine learning for predictive analytics, cross-domain data fusion, ethical AI, and scalable frameworks. The research emphasizes the importance of ongoing innovation in AI and NLP to meet evolving data and technology challenges, marking a new chapter in intelligence and national security.

VI. REFERENCES

[1]. C. Hobbs, M. Moran and D. Salisbury, Open source intelligence in the twenty-first century: new approaches and opportunities, Springer, 2014.

[2]. T. K. a. W.-S. V. A. Shackelford, Encyclopedia of evolutionary psychological science, Springer Cham, 2021.

[3]. R. a. T. M. a. F. L. Ghioni, "Open source intelligence and AI: a systematic review of the GELSI literature," AI & society, pp. 1-16, 2023.

[4]. I. Bohm and S. Lolagar, "Open source intelligence: Introduction, legal, and ethical considerations," International Cybersecurity Law Review, pp. 317-337, 2021.

[5]. J. a. N. P. a. M. F. G. a. P. G. M. Pastor-Galindo, "The not yet exploited goldmine of OSINT: Opportunities, open challenges and future trends," IEEE Access, pp. 10282-10304, 2020.

[6]. B. Akhgar, "Osint as an integral part of the national security apparatus," Open Source Intelligence Investigation: From Strategy to Implementation, pp. 3-9, 2016.

[7]. T. a. B. T. a. K. M.-A. a. R. C. Riebe, "Privacy Concerns and Acceptance Factors of OSINT for Cybersecurity: A Representative Survey," Proceedings on Privacy Enhancing Technologies, vol. 1, pp. 477-493, 2023.

[8]. M. a. M. E. a. N. B. Landon-Murray, "Disinformation in contemporary US foreign policy: Impacts and ethics in an era of fake news, social media, and artificial intelligence," Public Integrity, vol. 22, pp. 512-522, 2019.

[9]. T. a. D. T. Ivanjko, "Open Source Intelligence (OSINT): issues and trends," INFuture 2019: knowledge in the digital age, pp. 191-196.

[10]. M. a. R. A. C. S. a. K. C. Wankhade, "A survey on sentiment analysis methods, applications, and challenges," Artificial Intelligence Review, vol. 55, pp. 5731-5780, 2022.

Cite this article as :

Muhammad Ayub, Sidra Irum, Dr. Zunera Jalil, "Enhanced Audio-Based Open-Source Intelligence Insights using Machine Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 10, Issue 1, pp.141-149, January-February-2024. Available at doi : https://doi.org/10.32628/CSEIT2410118
Journal URL : https://ijsrcseit.com/CSEIT2410118