International Journal of Engineering Applied Sciences and Technology, 2023
Vol. 8, Issue 06, ISSN No. 2455-2143, Pages 94-99
Published Online October 2023 in IJEAST (http://www.ijeast.com)

NATURAL LANGUAGE PROCESSING BASED THREAT ASSESSMENT

Mayank Raghav
Faculty of Communication Engineering,
Military College of Telecommunication Engineering
Mhow, Indore, Madhya Pradesh, India

Abstract— **Military intelligence agencies are continually faced with the formidable task of evaluating and mitigating diverse risks in an increasingly complex and dynamic global landscape. NLP has emerged as a powerful tool for enhancing the capabilities of Military Intelligence operations. NLP techniques enable the automatic extraction, analysis, and interpretation of information from vast volumes of unstructured textual data, such as open-source intelligence reports, social media, intercepted communications. By processing and understanding these textual sources, NLP systems can assist Military Intelligence analysts in identifying potential threats, uncovering hidden patterns, and making timely and informed decisions. Threat assessment is required to safeguard security interests and maintain the sovereignty of nation. By the rapidly expanding fields in emerging technologies India can gain that decisive advantage.**

Keywords— **Military Intelligence, Natural Language Processing, Artificial Intelligence, Qualitative Analysis, Latent Dirichlet Allocation (LDA)**

I. INTRODUCTION

Languages in which humans can communicate with each other are called “Natural Languages” like Hindi, English or French and many more. Earlier, computers were not designed to understand these natural languages, rather they were used to perform complex problems and calculations only. Due to the need of computers to understand natural languages. The need of processing them was developed. It provides computers the ability to understand information gathered in the form of verbal or written language. Once the ability to understand text and spoken words are enhanced in computers. NLP offers a versatile approach to intelligence analysis. Over the years, researchers have established that NLP operates at five different language levels, encompassing aspects like word classifications (e.g., nouns, verbs) and meanings (animate, count) at the word level, the structure of sentences at the syntactic level, the understanding of meaning at the semantic level, and the examination of how context shapes interpretation, known as pragmatics. Furthermore, the field of information retrieval has developed techniques for text clustering that make use of semantic and pragmatic patterns at the document level to perform thematic or topical analysis. The focus of this paper is on the application of NLP in the analysis of intercepted communications, with a particular emphasis on the semantic level. It utilizes techniques like word association and thematic analysis, estimating associations through similarity metrics applied to word vectors and tracking thematic trends over time using topic models. The primary objective of this research is to illustrate how unsupervised machine learning can identify connections between activities, individuals, and organizations over time within thousands of intelligence reports, all while requiring minimal human intervention.

Languages in which humans can communicate with each other are called “Natural Languages” like Hindi, English or French and many more. Earlier, computers were not designed to understand these natural languages, rather they were used to perform complex problems and calculations only. Due to the need of computers to understand natural languages. The need of processing them was developed. It provides computers the ability to understand information gathered in the form of verbal or written language. Once the ability to understand text and spoken words are enhanced in computers. NLP offers a versatile approach to intelligence analysis. Over the years, researchers have established that NLP operates at five different language levels, encompassing aspects like word classifications (e.g., nouns, verbs) and meanings (animate, count) at the word level, the structure of sentences at the syntactic level, the understanding of meaning at the semantic level, and the examination of how context shapes interpretation, known as pragmatics. Furthermore, the field of information retrieval has developed techniques for text clustering that make use of semantic and pragmatic patterns at the document level to perform thematic or topical analysis.
The focus of this paper is on the application of NLP in the analysis of intercepted communications, with a particular emphasis on the semantic level. It utilizes techniques like word association and thematic analysis, estimating associations through similarity metrics applied to word vectors and tracking thematic trends over time using topic models. The primary objective of this research is to illustrate how unsupervised machine learning can identify connections between activities, individuals, and organizations over time

International Journal of Engineering Applied Sciences and Technology, 2023
Vol. 8, Issue 06, ISSN No. 2455-2143, Pages 94-99
Published Online October 2023 in IJEAST (http://www.ijeast.com)

within thousands of intelligence reports, all while requiring minimal human intervention.

II. PROPOSED ALGORITHM

A. Creation of Database for Intercepts Received –

Preparing a database in MySQL for call intercepts within the sensitive context of the military domain presents unique challenges, primarily stemming from the scarcity of

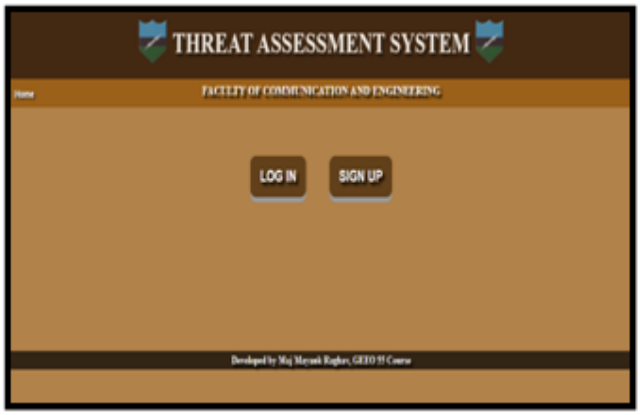

Fig. 1. Centralized Database for receiving call intercepts

openly available datasets. To address this, meticulous planning and data generation are imperative. First, a well-structured schema was designed, accounting for various data types, including call metadata, audio recordings, and user profiles. Security and encryption measures were be implemented to safeguard the classified information. It is crucial to collaborate with military experts to ensure that the database aligns with operational needs and adheres to strict security and privacy.

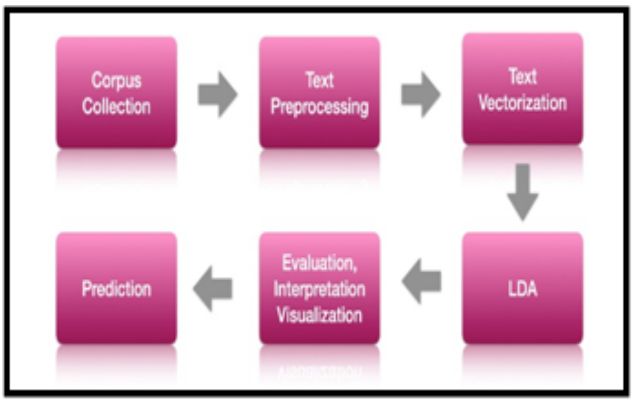

Fig. 2. Topic Modelling Pipeline Block Diagram

regulations. Ongoing maintenance and data enrichment will be crucial to continually improve the database's effectiveness for call intercepts, providing invaluable intelligence to military operations. The intercepts recorded while experiment is mentioned in the table.

Intercept 1	"Sergeant, we need to ensure the smooth mobilization of our troops for the upcoming training exercise. Let us coordinate the logistics and deployment schedules."
Intercept 2	"General, the mobilization of the army troops is crucial for the success of this exercise. What is the status of equipment readiness and troop deployment?"
Intercept 3	"Captain, I need a detailed report on the mobilization process. We should consider topics like transportation, communication, and readiness for the training."
Intercept 4	"Soldiers, remember the importance of speed and precision during mobilization. Our training topics will cover tactics, but first, we must gather our troops efficiently."
Intercept 5	"Lieutenant, please update the training agenda. We will have different mobilization topics for each unit. Ensure everyone is informed."
Intercept 6	"Major, let's focus on optimizing the mobilization routes and timing for the troops. The topics of interest should include road conditions and potential obstacles."
Intercept 7	"Colonel, for our upcoming training, let's create a discussion forum to address mobilization challenges. Invite experts in logistics to share their insights on key topics."
Intercept 8	"Private, you'll be responsible for documenting the troops' experiences during mobilization. Note any issues or notable topics discussed by the soldiers."
Intercept 9	"Brigadier, we must not overlook the mental preparation of our troops. Include psychological topics in the training to ensure their resilience during mobilization."

International Journal of Engineering Applied Sciences and Technology, 2023
Vol. 8, Issue 06, ISSN No. 2455-2143, Pages 94-99
Published Online October 2023 in IJEAST (http://www.ijeast.com)

B. Pre-processing of Textual Data –
Preprocessing of textual data was a crucial step in processing that involves several techniques to clean and prepare text for analysis. Tokenization, the first step in this process, breaks down the text into individual words or tokens, making it easier to work with. Lemmatization follows, reducing words to their base or root form, which helps in grouping together variations of a word and enhancing the efficiency of subsequent analysis. Lastly, stop word removal eliminates common and uninformative words like "the," "is," and "in," as they add little value and can be a source of noise in text analysis. This combination of tokenization, lemmatization, and stop word removal streamlines textual data, making it more amenable to tasks such as sentiment analysis, information retrieval, and machine learning, ultimately improving the accuracy and effectiveness of natural language processing applications.

C. Text Analysis using LDA –
In this paper, latent Dirichlet allocation (LDA), a form of topic modelling, was employed for thematic analysis. Topic models, such as LDA, deduce latent themes within documents by calculating how words group together to create topics and how these topics combine to form documents. The distributions of words within topics are valuable for understanding the nature of these themes, while the distributions of topics within documents are helpful for gauging thematic prevalence, trends, or pinpointing documents that encompass themes. The flow chart of topic modelling with LDA is explained in Figure 3.

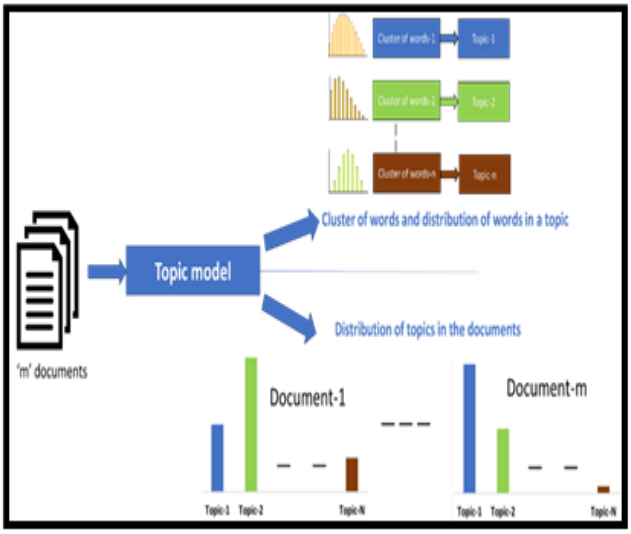

Fig. 3. Latent Dirichlet allocation(LDA) Block Diagram

III. EXPERIMENT AND RESULT

A. Prediction in Mobilization of Troops –
Extraction of features was able to provide us the pattern of troops ready to be mobilized and ensuring readiness as shown in the results found after analysis given in figure 4.

```
[ ] cv.get_feature_names_out()[10]

    'topics'

[ ] import random

    for i in range(10):
        random_word_id = random.randint(0,10)
        print(cv.get_feature_names_out()[random_word_id])

    soldiers
    ensure
    ensure
    ll
    readiness
    ensure
    exercise
    need
    readiness
    soldiers
```

Fig. 4. Features Extracted tells the trend of conversation

B . Challenges in Data Intercepted –
Categorical distributions were a challenging endeavor. Categorical data represented distinct categories or labels, and it differs from continuous numerical data. The presence of categorical data can complicate visualization and statistical testing, necessitating specialized techniques such as chi-squared tests, contingency tables, or logistic regression models for modeling and inference. Consequently, a robust understanding of these categorical distributions and the ability to choose appropriate analytical tools are crucial for accurate and meaningful data analysis when dealing with such data as shown in figure 5 the probability of co-occurrence of words is same when seen from single parameters while it gives different results when other parameters are also being considered as shown in figure 6.

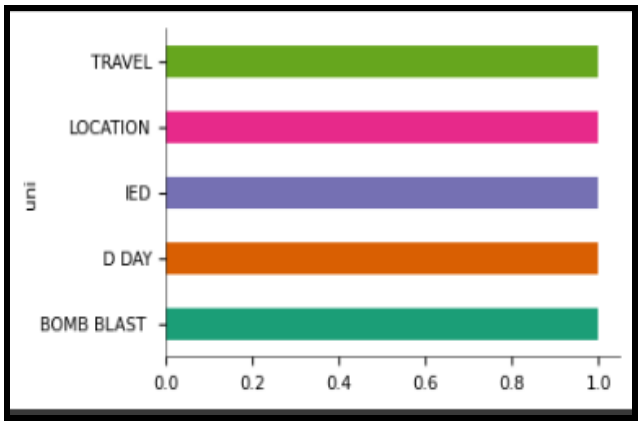

Fig. 5. Words in the Database

International Journal of Engineering Applied Sciences and Technology, 2023
Vol. 8, Issue 06, ISSN No. 2455-2143, Pages 94-99
Published Online October 2023 in IJEAST (http://www.ijeast.com)

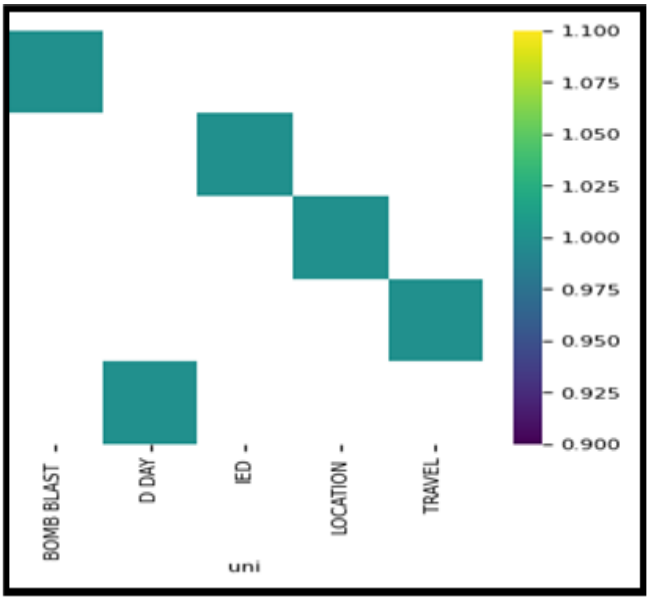

Fig. 6. Words in different topics

C . Extract of Conversation in Intercepts –

While extraction of topics top 5 words from each corresponding topic were extracted which provided gist of conversation of troops for better correlation with environment. As shown in figure 7.

```
for index,topic in enumerate(LDA.components_):
    print(f'THE TOP 5 WORDS FOR TOPIC #{index}')
    print([cv.get_feature_names_out()[i] for i in topic.argsort()[-5:]])
    print('\n')

THE TOP 5 WORDS FOR TOPIC #0
['logistics', 'upcoming', 'topics', 'training', 'let']


THE TOP 5 WORDS FOR TOPIC #1
['deployment', 'exercise', 'need', 'logistics', 'upcoming']


THE TOP 5 WORDS FOR TOPIC #2
['training', 'readiness', 'deployment', 'exercise', 'troops']


THE TOP 5 WORDS FOR TOPIC #3
['soldiers', 'readiness', 'need', 'topics', 'training']


THE TOP 5 WORDS FOR TOPIC #4
['include', 'll', 'topics', 'ensure', 'training']


THE TOP 5 WORDS FOR TOPIC #5
['ensure', 'let', 'topics', 'include', 'troops']
```

Fig. 7. Top words in each topic

D . Results –

Probabilities play a pivotal role in the context of topic modeling, particularly when employing Latent Dirichlet Allocation (LDA) models. In this process, each topic is represented as a probability distribution over a set of words. These probabilities signify the likelihood of a specific word being associated with a particular topic. The core idea behind LDA is to view each document as a mixture of various topics, with each word within the document generated based on these topic-word probabilities. As shown in the graphs below for five topics extracted from the intercepts, word probability is being explained.

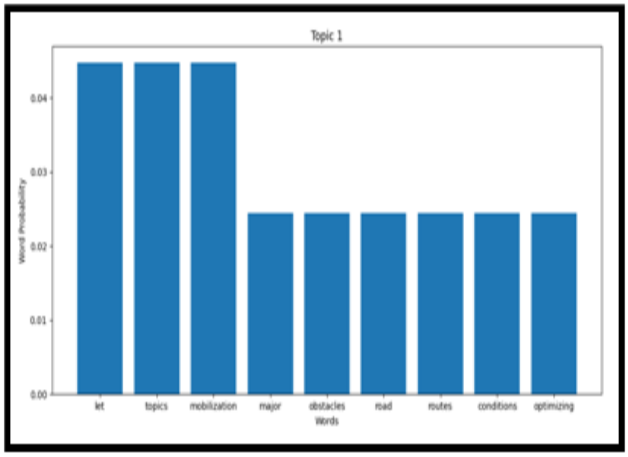

Fig. 8. Topic 1 predicting mobilisation planning

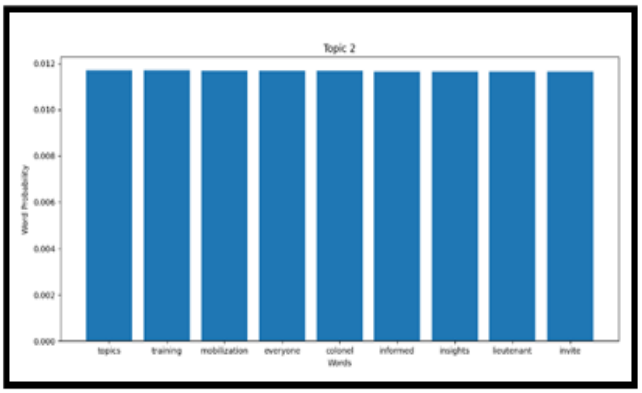

Fig. 9. Topic 2 predicting training related to mobilisation

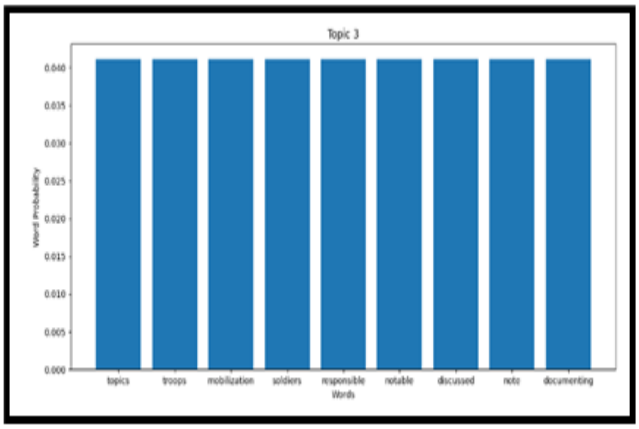

Fig. 10. Top 3 predicting soldiers documented responsible for mobilisation

International Journal of Engineering Applied Sciences and Technology, 2023
Vol. 8, Issue 06, ISSN No. 2455-2143, Pages 94-99
Published Online October 2023 in IJEAST (http://www.ijeast.com)

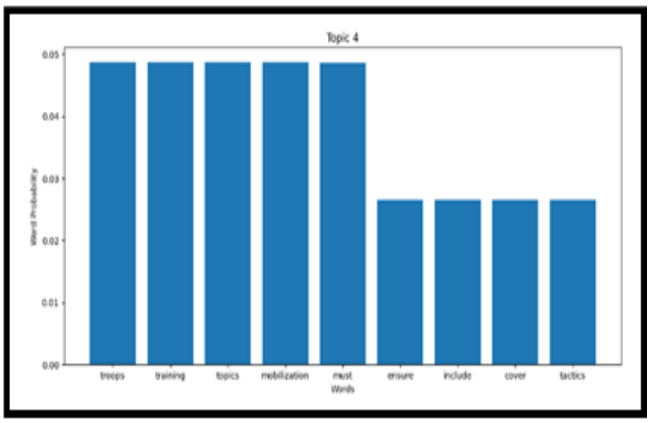

Fig. 11. Topic 4 predicting training related conversation for mobilisation

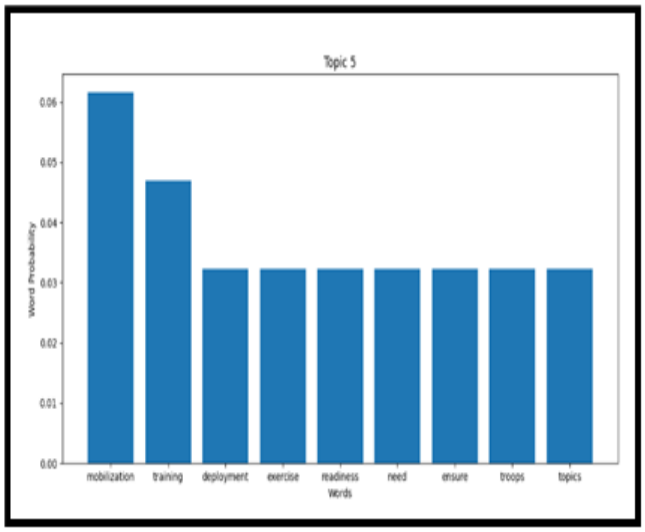

Fig. 12. Topic 5 predicting mobilisation being ensured by troops

In the realm of military applications, the integration of LDA and probabilistic graphs becomes highly relevant. Military intelligence often deals with vast quantities of text-based data, including reports, documents, and communication intercepts. Topic modeling techniques such as LDA allow military analysts to automatically categorize and summarize this information, making it easier to identify critical trends, emerging threats, and areas of interest. The use of probabilistic graphs, which are built on topic-word probabilities, helps in visualizing the relationships between topics and their constituent words. This graphical representation assists military analysts in comprehending the complex information landscape, aiding in decision-making processes, threat assessments, and strategic planning. Moreover, the combination of LDA and probabilistic graphs can enable real-time monitoring and early warning systems in the military domain. By continuously updating topic models and their associated probabilities, defense organizations can track evolving narratives, spot changes in adversary tactics, and identify potential security risks. The visual aspect of probabilistic graphs makes it easier to communicate insights to decision-makers, facilitating a better understanding of the evolving military landscape. In summary, the synergy between LDA models and probabilistic graphs enhances the capabilities of topic modeling in military applications. It assists in text data analysis, topic discovery, and understanding the dynamics of information in a structured and visual manner. This not only aids in extracting actionable intelligence but also plays a crucial role in bolstering national security by providing timely and relevant information for decision-making in a complex and ever-changing geopolitical environment.

IV.CONCLUSION

This article showcases the application of Natural Language Processing (NLP) in enhancing the analysis of intelligence reports for research purposes. It introduces a systematic and objective approach to quantifying connections among activities, individuals, and organizations across a vast corpus of documents over time. It is essential to note that the intent is not to advocate for NLP as a replacement for qualitative analysis but, rather, to present a case study illustrating how NLP can complement qualitative analysis in gaining insights from intelligence reports. In practice, qualitative expertise remains crucial, particularly when dealing with inconsistent spelling and errors that might introduce ambiguity to critical terms. The article underscores the value of NLP as a valuable tool for historians and analysts in comprehending historical reports and contributing to the intelligence cycle's analysis phase. In a broader historical context, this work emphasizes that scholars can harness NLP to scrutinize government documents, providing an alternative avenue to understanding extensive archival resources. Ultimately, the goal of this article is to stimulate dialogue among scholars and intelligence practitioners regarding the utility, constraints, and future potential of NLP in research and analysis, with the overarching aim of advancing machine learning methodologies in the field.

V. REFERENCE

[1] Advancing intelligence analysis: using natural language processing on East Pakistani intelligence documents Ryan Shaffer & Benjamin Shearn DOI: 10.1080/02684527.2023.2170744 Published online: 13 Feb 2023.

[2] Garicano, Luis, and Richard A. Posner. 2005. "Intelligence Failures: An Organizational Economics Perspective." Journal of Economic Perspectives, 19 (4): 151-170.DOI: 10.1257/089533005775196723

[3] Jurafsky, D., and J. H. Martin. “Speech and Language Processing (Draft), 3rd. “2022. https://web.stanford.edu/~jurafsky/slp3/ Katagiri, A.,

and E. Min. "The Credibility of Public and Private Signals: A Document-Based Approach." The American Political Science Review 113, no. 1 (2019): 156–172. doi:10.1017/S0003055418000643

[4] Richelson, J. "The Wizards of Langley: The Cia's Directorate of Science and Technology." Intelligence and National Security 12, no. 1 (1997): 82–103. doi:10.1080/02684529708432400.

[5] Barr, A. "Natural Language Understanding." AI Magazine 1, no. 1 (1980): 5–10. doi:10.1609/aimag. v1i1.85.

[6] Antoniak, M., and D. Mimno. "Evaluating the Stability of Embedding-Based Word Similarities." Transactions of the Association for Computational Linguistics 6 (2018): 107–119. doi:10.1162/tacl_a_00008.