AI System For MITRE ATT&CK Threat Classification And Organizational Impact Analysis (Progress Update)

Presented by
Prapatsorn Alongkornpradub - st124846
Vorameth Reantongcome - st124903
Ekkarat Techanawakarnkul - st124945

AT 82.05 AI: Natural Language Understanding
Asian Institute of Technology

Agenda

1. Introduction
2. Research Questions
3. Related Works
4. Methodology
5. Preliminary Results
6. Discussion
7. Conclusion
8. Summary of Progress
9. Limitations and Challenges

Introduction

Cyber security is the practice of protecting networks, devices, and data from unauthorized access or malicious attacks. This field encompasses a variety of technologies, processes, and practices designed to ensure the integrity, confidentiality, and availability of information in the digital age.

Research Questions

1. How effectively can a fine-tuned BERT model classify cybersecurity news into ATT&CK techniques?

2. How does the performance of the BERT model compare to manual classification approaches in cybersecurity threat analysis?

3. What impact does fine-tuning a transformer model like BERT on a cybersecurity-specific dataset have on the accuracy and relevance of threat classification?

Related Works

1. Noise Contrastive Estimation-based Matching Framework for Low-Resource Security Attack Pattern Recognition

- **Summary:** With the complexity, long-tailedness and huge number of multi-label classification of Tactics, Techniques and Procedures (TTPs), this could hinder the learning ability of the model. From following problems the authors propose a new learning paradigm on mapping the TTPs in classification problem which consist of introducing ranking-based Noise Contrastive Estimation (NCE), curating and publicizing an export-annotation dataset, and conducting extensive experiments in their learning methods.
- **Gaps:** The training process took a very long due to a very large datasets of frameworks. This result in the coverage of TTPs which could provide a limit size in the prediction.

2. Automatic Mapping of Unstructured Cyber Threat Intelligence: An Experimental Study

- **Summary:** To support cybersecurity proactively, machine learning will be used in order to classify the CTI into the MITRE ATT&CK category. General machine learning and deep learning were being experimented. Eventually, deep learning could outperform general machine learning. In addition, the sentence-level model could outperform the document-level model. From the result, the author concluded that due to the complexity of natural language making such an accurate model could be challenging.
- **Gaps:** Due to the multi-class and ambiguity of MITRE ATT&CK and NLP model, this provide complex on evaluation process which make the accuracy might be correctly evaluate.

Related Works

3. A Pretrained Language Model for Cyber Threat Intelligence

- **Summary:** To digest the information from many Cyber Threat Intelligence reports within a time limit, the authors develop a pre-trained BERT model tailoring for cybersecurity domain for performing extensive experiments on a wide range of tasks and benchmark datasets for the security domain in order to curate a large amount of high quality cybersecurity datasets specifically designed for cyber-threat intelligence analysis.
- **Gaps:** The model will works only with the English language which in real-world of cybersecurity many news are distributed in many language. In addition, the model was trained with very little datasets which could cause some problem to unknown cybersecurity news or intelligence.

4. AnnoCTR: A Dataset for Detecting and Linking Entities, Tactics, and Techniques in Cyber Threat Reports

- **Summary:** This paper explores NLP techniques for cybersecurity threat intelligence, focusing on NER, entity linking, and text classification. It evaluates transformer models like BERT and RoBERTa, showing they outperform traditional methods. The study also examines entity linking to MITRE ATT&CK and adapts temporal tagging models for cyber threat reports.
- **Gaps:** Existing approaches struggle with disambiguating ATT&CK techniques due to complex terminology and limited data.

Related Works

5. Introducing a New Dataset for Event Detection in Cybersecurity Texts

- **Summary:** This paper introduces CySecED, a dataset for cybersecurity event detection with 30 annotated event types. It improves on previous datasets by incorporating document-level context for better event trigger identification.
- **Gaps:** Existing ED datasets focus on event trigger detection but do not address cyber threat intelligence (CTI) retrieval. Traditional methods struggle with capturing nuanced attack relationships, and graph-based ED models remain underexplored in CTI.

6. Full-Stack Information Extraction System for Cybersecurity Intelligence

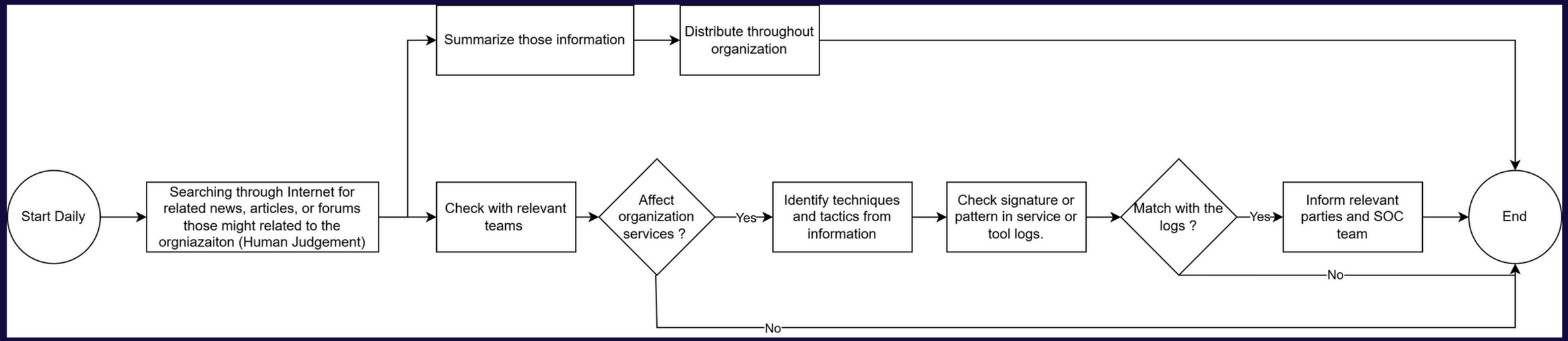
- **Summary:** This system automates the extraction and categorization of cybersecurity-related entities (e.g., malware names, attack vectors) from unstructured sources like news articles and threat reports, using NLP techniques like Named Entity Recognition (NER).
- **Gaps:** The system doesn't use a standard framework like MITRE ATT&CK, which classifies different types of cyberattacks. Without this, the data can be disorganized and hard to use for making quick decisions.

Methodology

Old Workflow

1. Manual Search for Threat Information
2. Human Judgment & Filtering
3. Split into Two Main Tasks
    a. Task1: Security Awareness
    - summarized readable content

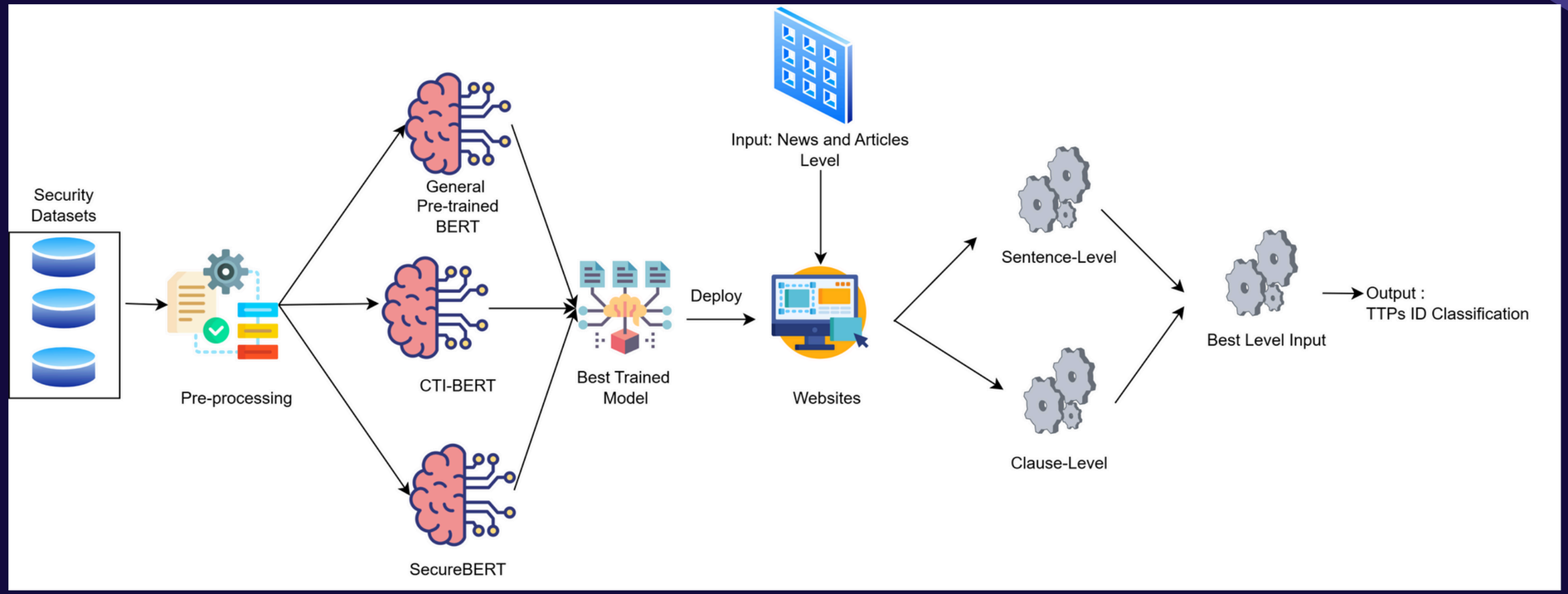
b. Task2: Threat Identification
- Technique & Tactic Extraction: aligned with frameworks like MITRE ATT&CK
- Pattern Matching
- Notify Security Teams
- Mitigation Planning



Methodology

New Workflow

New Workflow Improves the Old Workflow in Identification of Techniques and Tactics part
The old method required manual extraction and mapping of techniques and tactics. With the new workflow, the model classifies the content into MITRE ATT&CK techniques automatically, speeding up threat identification.

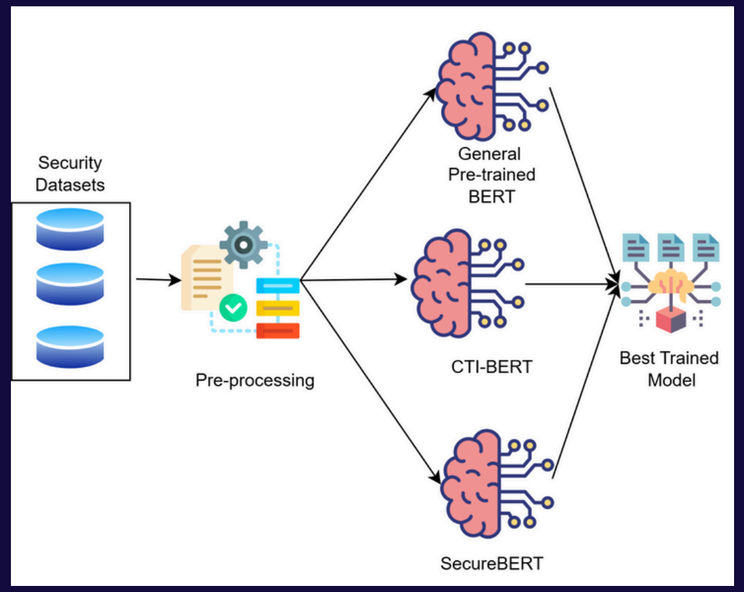


Methodology

Overview Experiment

From new workflow, there are two experiments including Model Experiment and Input Pre-processing Experiment

Model Experiment



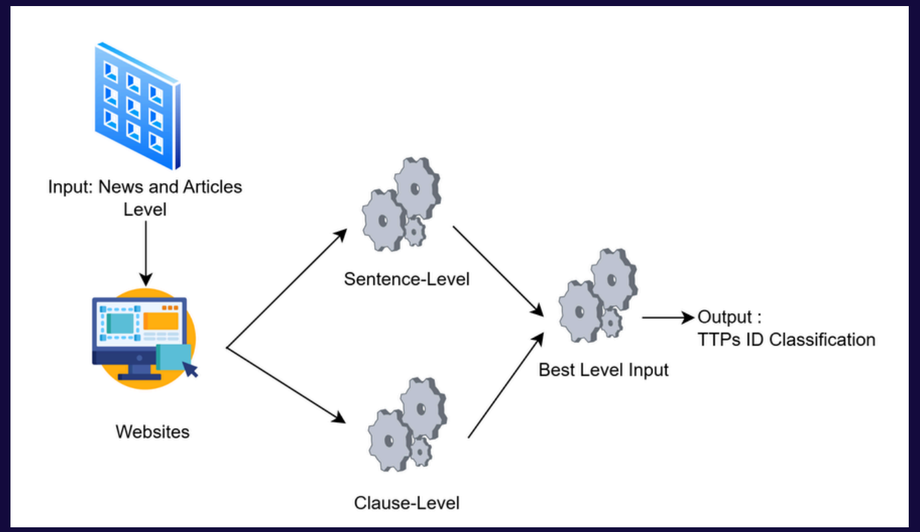
1. Select Models for Comparison
   Three models selected for comparison are:
   BERT-Base-Uncased, CTI-BERT and Secure-BERT
2. Tokenizer and Word Embeddings
3. Training the Models: train 3 models with the same dataset and turn hyperparameters
4. Evaluate Performance
5. Select the Best Model

Methodology

Overview Experiment

From new workflow, there are two experiments including Model Experiment and Input Pre-processing Experiment

Input Pre-processing Experiment

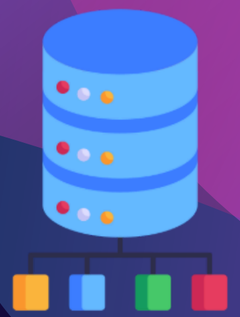


1. Pre-process the News or Articles
   Method 1: Sentence-Level Processing
   - The text will be processed at sentence level, mapping the relation between the sentence and MITRE ATT&CK attack patterns.

   Method 2: Clause-Level Processing
   - The text will be processed at clause level, where each clause is mapped to MITRE ATT&CK attack patterns.
2. Compare Accuracy of These Two Methods
3. Human Evaluation
   - A cybersecurity analyst or engineer will verify the classification results

Methodology

Datasets



There are three datasets that will be used for training the model and using it as a document retrieval.

Tumeteor Dataset (tumeteor/Security-TTP-Mapping

- Maps security-related text to Tactics, Techniques, and Procedures (TTPs).
- Helps identify how hackers might use similar patterns in real-world news.

Vittorio Dataset

- Aggregated dataset of attack pattern objects and relationships in JSON format.
- Includes documents mapped to MITRE ATT&CK techniques and tactics.

Zainabsa99 Dataset (Zainabsa99/mitre_attack)

- Translates the MITRE ATT&CK IDs into the consumable datasets through a Huggingface platform consisting of ID, name, description, and detection.

Methodology

Data Preprocessing

Generalized sentences

- Converting all text to lower case and removing unnecessary characters

Tokenization

- During model training, the tokenizer from the selected model will be used to transform each word into tokens.

Prepare Sentences for the Model

- Padding
- Beginning-of-sentence (BOS) tag
- End-of-sentence (EOS) tag

real-world implementation

- The sentence and clause level will be used in order to evaluate the best accuracy

Methodology

Model

The following BERT-based models will be evaluated for their ability to classify cybersecurity data and identify MITRE ATT&CK techniques:

BERT-Base-Uncased

- Base BERT model available on Huggingface.
- Serves as the baseline for comparison.
- Will be evaluated against CTI-BERT and Secure-BERT.

CTI-BERT

- Specialized in the cybersecurity domain.
- Built from the SecBert model to focus on cyber threat intelligence.
- Primary model for the project.

Secure-BERT

- Extends from the pre-trained Ro-Bert model.
- Aimed at improving cybersecurity-specific tasks.
- Will be compared with other models for performance.

Methodology

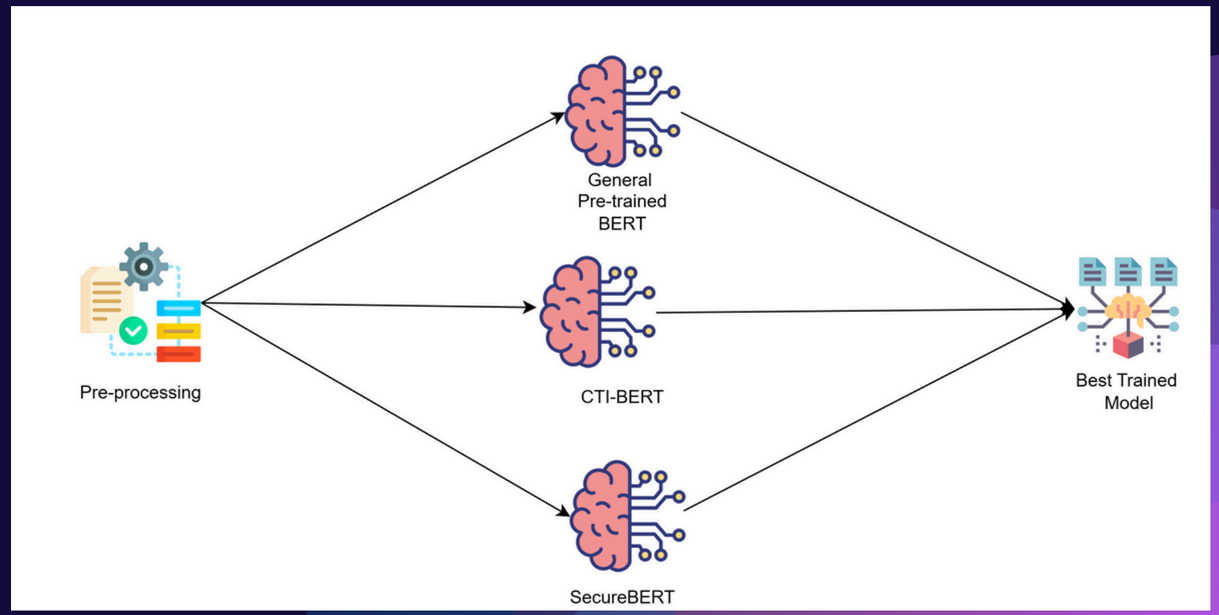
Training Process

The training phase customizes the pre-trained models for cybersecurity classification tasks using specialized datasets. Here's how it works:

1. Customize Pre-trained Models for Cybersecurity
   - Datasets from the cybersecurity context will be applied to the pre-trained models.
   - This process fine-tunes the models to specialize in cybersecurity-related tasks and threat identification.

2. Models Used for Training
   Three pre-trained models will be tested and compared:
   - CTI-BERT (Park et al., 2023)
   - Secure-BERT (Aghaei et al., 2023)
   - BERT-Base-Uncased (Huggingface)

3. Model Comparison for Best Performance
   The goal is to determine which model best handles cybersecurity data and provides accurate classification based on MITRE ATT&CK techniques.



Methodology

New Workflow

Evaluation / Metrics

To evaluate the models, two mechanisms will be used.

Performance Metrics (Automated Evaluation)	• **Accuracy** – Measures overall correctness • **Precision** – Checks how many classified threats are relevant • **Recall** – Ensures the model identifies all actual threats • **F1-Score** – Balances precision & recall for overall performance
Human Evaluation (Expert Assessment)	• Cybersecurity experts review and validate model outputs • Experts answer specific questions to gauge real-world effectiveness

🎯 Goal: Ensure both quantitative performance and practical reliability in cybersecurity contexts

Preliminary Results

Table: Preliminary Results of BERT-based Models

Metrics	BERT-base-uncased	CTI-BERT	SecureBERT
Training loss	3.575500	2.774000	3.579800
Validation loss	3.438241	2.707260	3.307534
Accuracy	0.437643	0.546388	0.444487
Precision	0.292122	0.431662	0.299807
Recall	0.437643	0.546388	0.444487
F1-Score	0.333476	0.466138	0.337116

Discussion

Addressing Core Research Questions

RQ1: Effectiveness of the BERT Model for Threat Classification

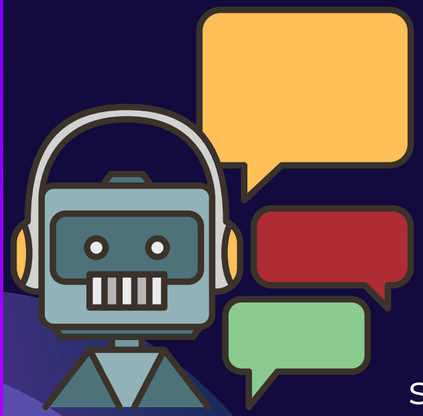
- The preliminary results of this study demonstrate that the BERT-based models, particularly CTI-BERT, has a capability to understand and categorize complex cybersecurity incidents, even with the challenging terminology and writing styles prevalent in cybersecurity news.

RQ2: Comparison to Manual Classification Methods

- While human classifiers are often overburdened with the complexity and volume of cybersecurity reports, the BERT model can process and classify large datasets in a much shorter time

RQ3: Impact of Fine-Tuning a Transformer Model on Accuracy

- Fine-tuning a transformer model like BERT on a cybersecurity-specific dataset improved accuracy. The fine-tuned model showed improvements in training and validation loss, as well as a higher F1-score (46.61%) compared to the base BERT model.



Conclusion

In conclusion, this research demonstrates the potential of AI-driven NLP systems to improve the **efficiency and accuracy of cybersecurity threat classification**. Although the system is still in development, the fine-tuned BERT model has shown promising results so far. Continued work will focus on **refining the model**, **incorporating expert feedback**, and **addressing challenges related to data quality**, **scalability**, and **integration** into real-world environments. Once fully implemented, the system has the potential to significantly **reduce manual effort, minimize human error, and help organizations respond to cybersecurity threats more quickly and effectively**.

Summary of Progress

What Has Been Done:

- **Scope the project and system:** From the first proposed solution, there are multiple redundancy and out-of-scope task. This made the scope to be reduced and complying feasible scope within timeline and knowledge.
- **Corporate more related works:** In order to increase more relatable models, the paper were being reviewed more for ingesting the some potential model or methodology in proposed solutions.
- **Model Comparison:** Model with three comparisons will be evaluated in order to pick the best model for this task.
- **Datasets Exploration:** More datasets had been explored which could be potentially insert as a trained data for providing the most relatable results.
- **Consultant with TA:** For scoping the topic, consultation with TA's NLP had been done for get a clear picture and staying on the right track.

What Still Needs to Be Done:

- **Keyword Extractions:** According to the TA's comments, the main improve of the model is identifying the potential keywords.
- **Experiment on Pre-processing:** In order to extract the context from sentence, pre-processing methods will be experimented for selecting the best method to processing the sentence in order to improve accuracy.
- **System Evaluation:** After finishing all the coding, the evaluation process will be performed through a human and machine judgment as the final validation.
- **Deployment**: The model will be deployed on the website allowing user to access for real-word use case.

Limitations and Challenges

Imbalanced and Limited Ground Truth Datasets

- Difficulty in verifying model accuracy due to limited real-world datasets.
- Data is unbalanced, which makes it harder to train the model properly.

Specialized Domain Limitations

- Datasets come from specific areas, so they may not work well for other types of data.

Complexity of the MITRE ATT&CK Framework

- The framework’s complexity may affect model's ability to fully capture its context.
- Potential risk of inaccurate classification for new use cases.

Thank You