Government of India
Ministry of Electronics and Information Technology
IndiaAI CyberGuard AI Hackathon – Stage 2

Reporting Template (Mandatory Submission)

1. Technical Report

Problem Understanding & Approach

- *How did the team interpret the problem statement and define their approach?*

  Based on the problem statement, our team developed an automated labeling process using word clouds when labeled data was unavailable along with that we also created our own tokenizer which also creates another word embedding which is an alternative of the word cloud . The system extracted key points from text, identifying critical terms without manual intervention. During training and testing, the model leveraged word clouds or tokens for pattern analysis, ensuring precise classification while aligning with the given requirements and constraints.

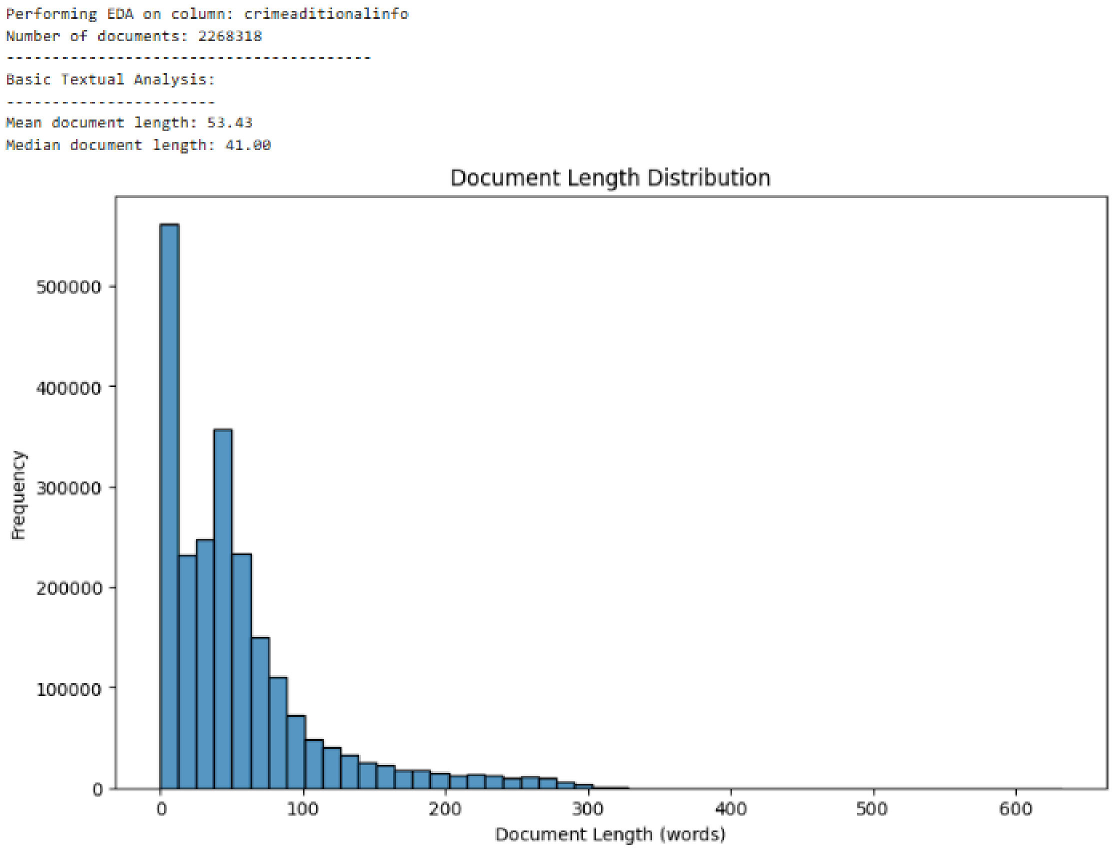
Exploratory Data Analysis (EDA), Data Preprocessing, Feature Engineering & Feature Selection

- *How was the dataset explored, cleaned, processed, and transformed? What insights were derived from Exploratory Data Analysis (EDA)?*

  Based on the problem statement, our team developed an automated labeling process using word clouds or token's to address the absence of labeled data. The system autonomously extracted key points or token's from text, identifying critical terms for analysis. During training and testing, the model leveraged these word clouds or tokens for pattern recognition, ensuring precise classification while fully adhering to the given requirements and constraints .

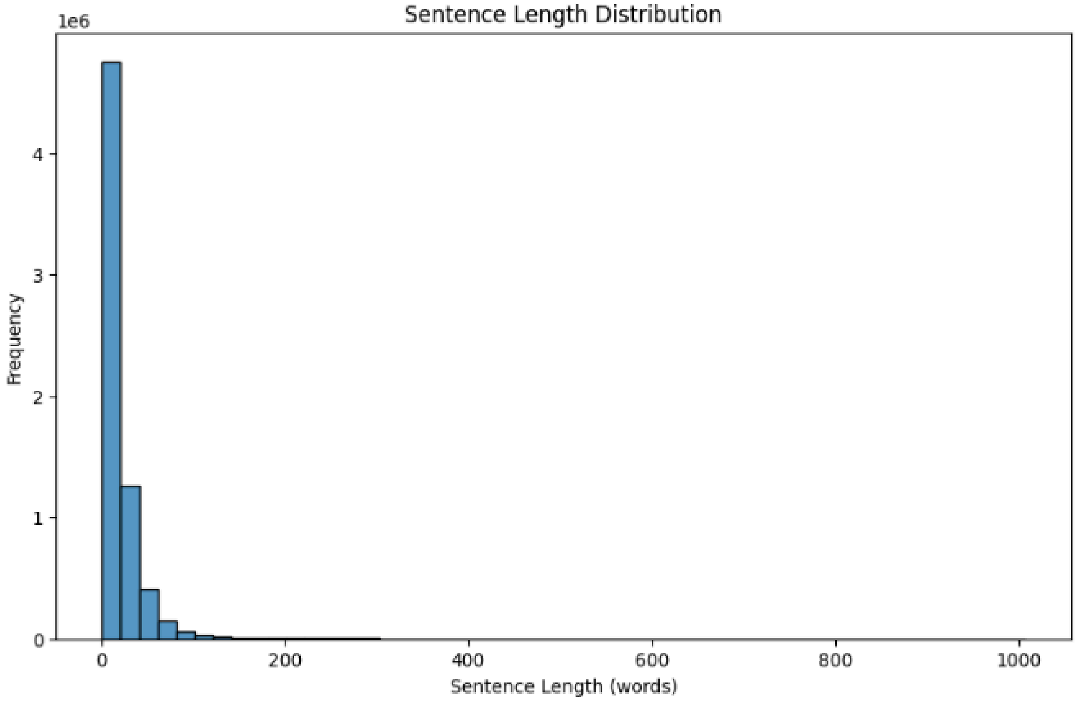
- *What statistical or visualization techniques were used during EDA? What methods were used for feature extraction and optimal feature selection for analysis?*

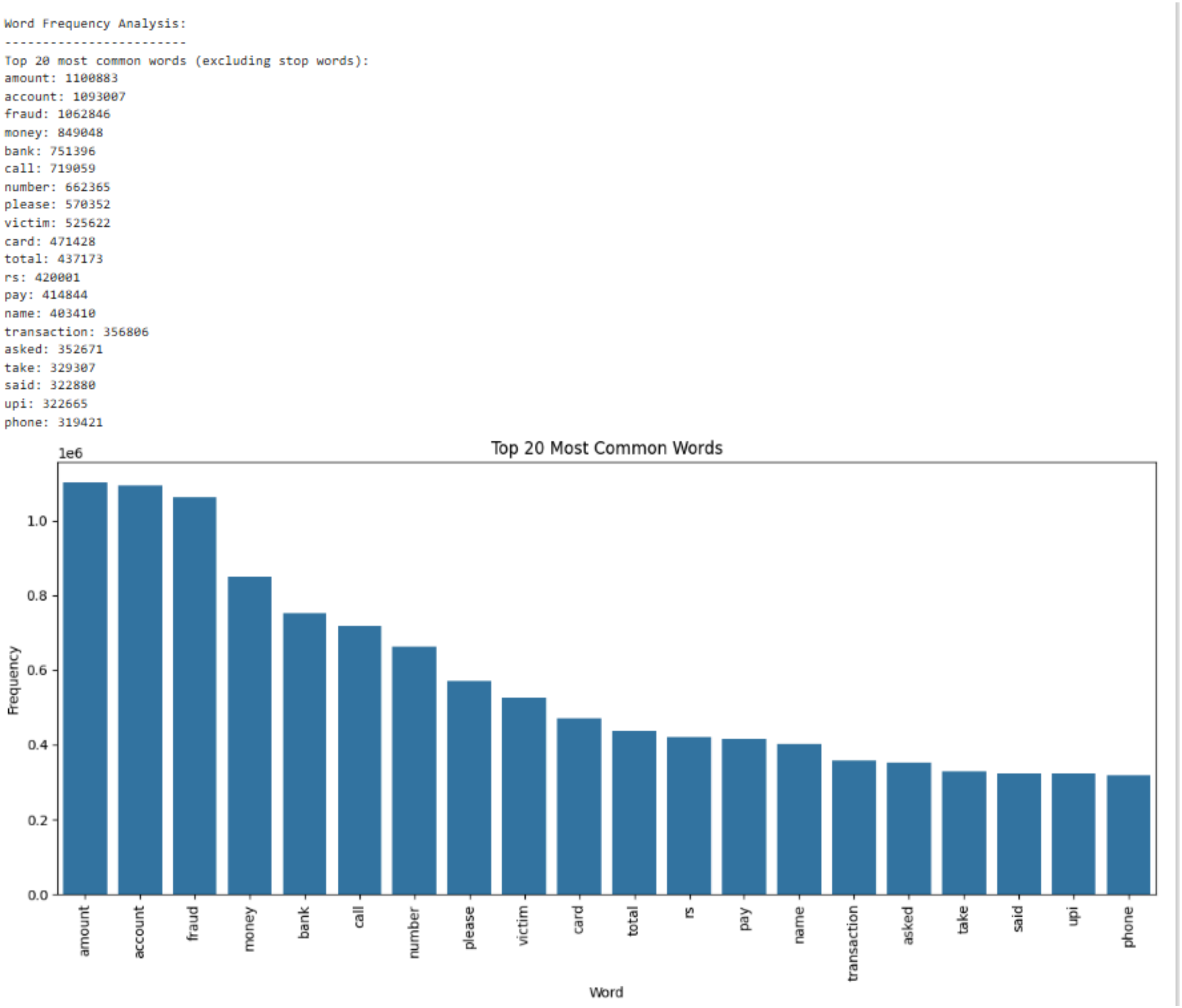
  Apart from all of this we while doing EDA we found the dataset contains Roman Letters , emojis , and some confidential information like phone number and mail which might be unnecessary for *our model so we cleaned the data as shown in the fig below*



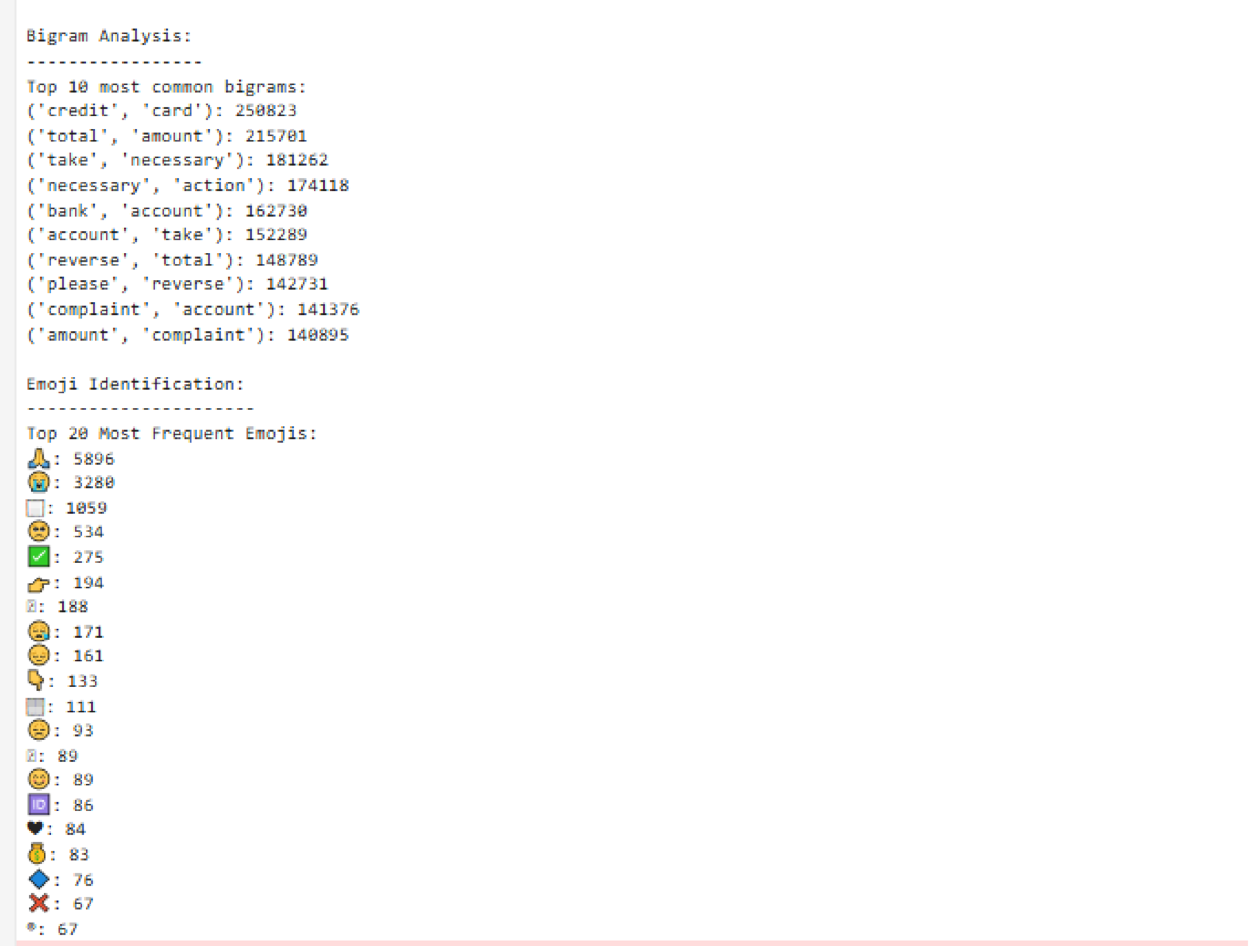
```
Performing EDA on column: crimeaditionalinfo
Number of documents: 2268318
----------------------------------------
Basic Textual Analysis:
-----------------------
Mean document length: 53.43
Median document length: 41.00
```

```
Sentence Length Distribution:
----------------------------
Mean sentence length: 19.08
```





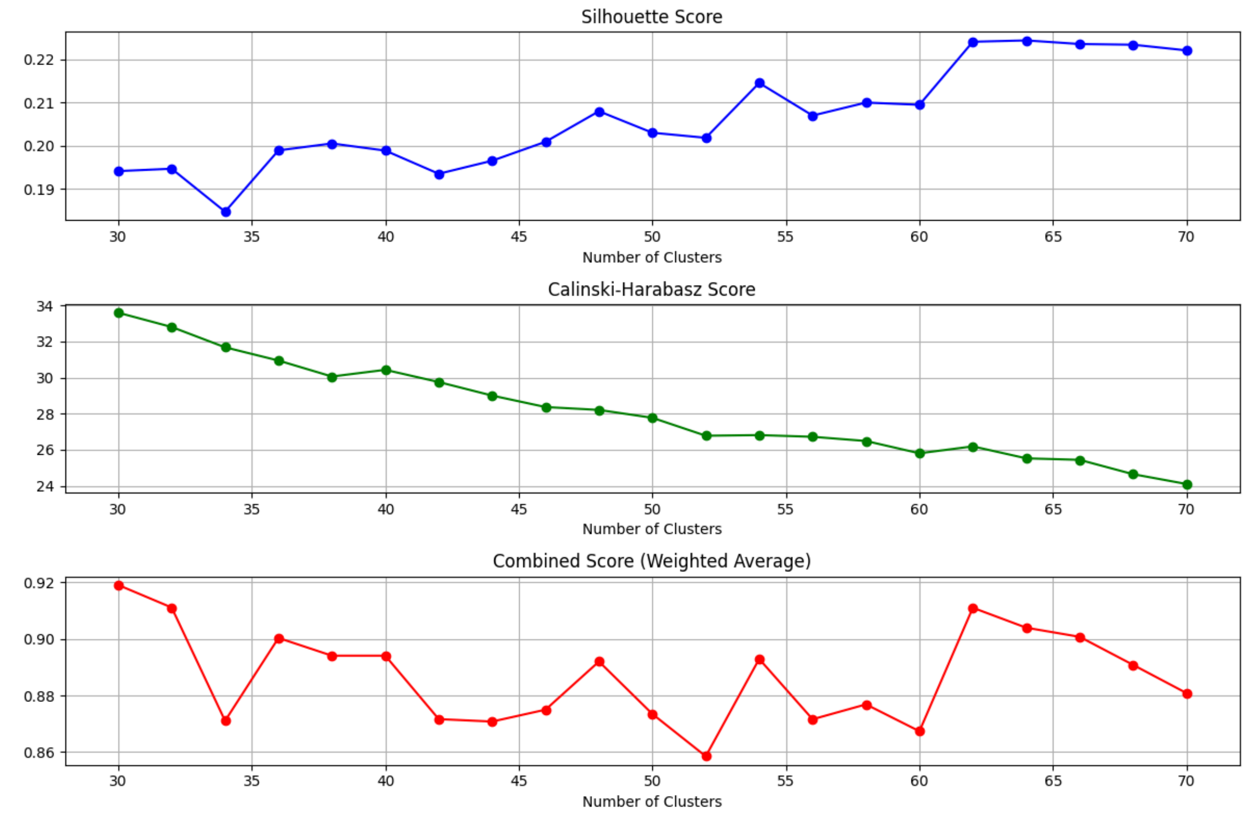
```
Word Frequency Analysis:
------------------------
Top 20 most common words (excluding stop words):
amount: 1100883
account: 1093007
fraud: 1062846
money: 849048
bank: 751396
call: 719059
number: 662365
please: 570352
victim: 525622
card: 471428
total: 437173
rs: 420001
pay: 414844
name: 403410
transaction: 356806
asked: 352671
take: 329307
said: 322880
upi: 322665
phone: 319421
```



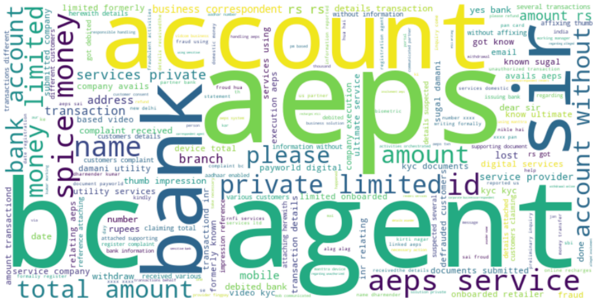
```
Bigram Analysis:
-----------------
Top 10 most common bigrams:
('credit', 'card'): 250823
('total', 'amount'): 215701
('take', 'necessary'): 181262
('necessary', 'action'): 174118
('bank', 'account'): 162730
('account', 'take'): 152289
('reverse', 'total'): 148789
('please', 'reverse'): 142731
('complaint', 'account'): 141376
('amount', 'complaint'): 140895

Emoji Identification:
----------------------
Top 20 Most Frequent Emojis:
🙏: 5896
😭: 3280
□: 1059
😞: 534
✅: 275
👉: 194
□: 188
😢: 171
😔: 161
👎: 133
□: 111
😑: 93
□: 89
😊: 89
🆔: 86
🖤: 84
💰: 83
🔷: 76
❌: 67
®: 67
```

And after this we used a clustering algo which provide us with Silhouette Score , Calinski Harabasz Score and Combined Score as shown in the fig below



We have generated a word cloud on the sample dataset provided so as to get which category contains the most frequencies word this will help us in approach 1st that is training a spacy model from scratch and for 2nd approach we created our custom embedding with a certain chunks size and has a token embeddings limiting it to 15000 max size and min word size to 5 so as it doesn't take some verb and which make it faster to create token or embeddingsApart from all of this we while doing EDA we found the dataset contains Roman Letters , emojis , and some confidential information like phone number and mail which might be unnecessary for our model so we cleaned the data as shown in the fig below



for all the sample files

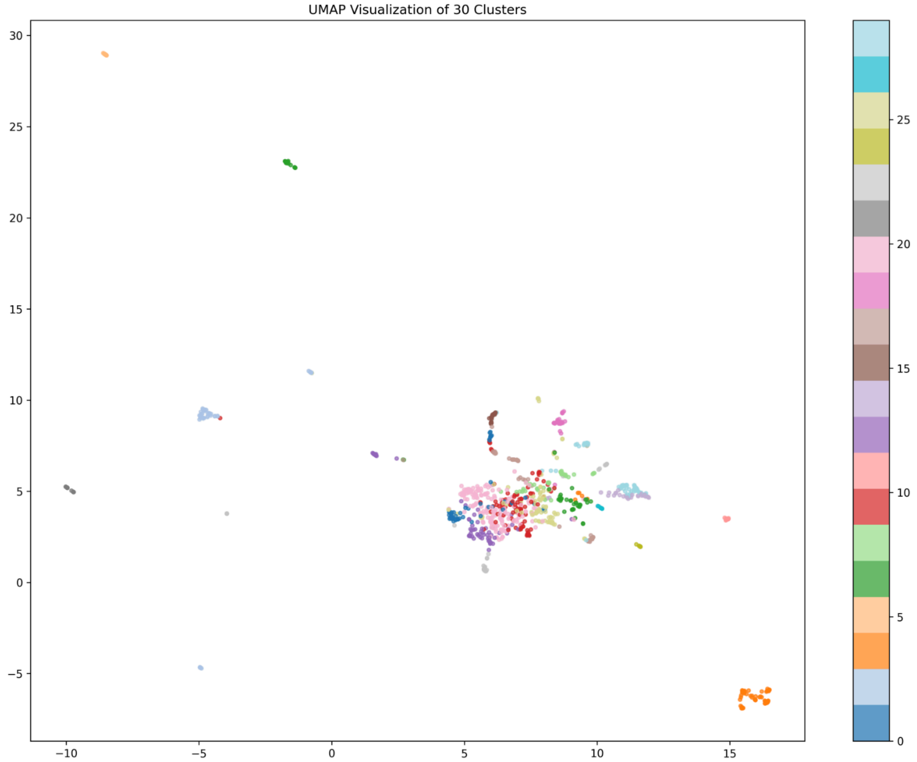
Exploratory Data Analysis (EDA) Techniques:

1. Statistical Techniques:

- ***Descriptive Statistics:*** *Analyzed word frequency, mean, variance, and distribution of text data.*
- ***TF-IDF (Term Frequency-Inverse Document Frequency):*** *Measured the importance of words in the dataset.*
- ***N-gram Analysis:*** *Identified frequently occurring word sequences (bigrams, trigrams).*
- ***Correlation Analysis:*** *Measured relationships between extracted features.*

2. Visualization Techniques:

- ***Word Clouds:*** *Highlighted key terms and frequently occurring words.*
- ***Bar Charts & Histograms:*** *Visualized word frequency distributions.*
- ***Box Plots:*** *Identified outliers in numerical feature representations.*



- *How did the team use EDA findings to refine their feature selection and ensure that the selected features improved model performance?*

  being able to find out what dataset contains help us clean the data and get us what techniques is needed to be used and cleaning all emojis and roman data remaining the data integrity so as it does not remove any important features

Using EDA Findings to Refine Feature Selection and Improve Model Performance

1. ***Identifying Key Features Through Word Clouds & Frequency Analysis***
   - *EDA revealed frequently occurring words and patterns using* ***word clouds*** *and* ***word frequency distributions****.*
   - *Only meaningful words were retained, while stopwords and irrelevant terms were removed, improving feature quality.*
2. ***Feature Importance Analysis Using Statistical Techniques***
   - ***Chi-Square Test & Mutual Information*** *helped determine which words were most relevant for classification.*
   - *Features with low significance were discarded, reducing noise in the model.*
3. ***Dimensionality Reduction for Better Performance***
   - ***Principal Component Analysis (PCA)*** *was applied to eliminate redundant features while preserving key information.*
   - *This ensured the model remained efficient without unnecessary complexity.*
4. ***Refining Word Representations Using TF-IDF & N-grams***
   - *TF-IDF helped assign importance to words based on their relevance rather than frequency alone.*
   - ***N-grams (bigrams & trigrams)*** *were used to capture contextual relationships, improving text representation.*
5. ***Feature Selection Iterations Based on Model Performance***
   - *Initial models were trained with different feature sets, and their impact on accuracy was evaluated.*
   - *Features that negatively affected performance (e.g., overfitting-prone features) were removed.*
   - *The final selection included only* ***highly relevant*** *and* ***non-redundant*** *features, leading to better generalization.*

By systematically analyzing EDA results and refining features iteratively, the team ensured the model had an ***optimal, well-structured input space****, leading to* ***improved accuracy, reduced overfitting, and better overall performance****.*

Model Building & Selection Justification

- *Why was the chosen AI model selected over other potential models? Can the solution be further scaled or integrated in real world settings? What was done to address biases?*

The model was chosen based on a balance between ***performance, interpretability, and computational efficiency****. Key reasons:The main reason of choosing these 2 model one spacy and another custom cnn+ multi head attention apart from this since the data does not contain any label and we have to choose mode accuracy*

- ***Better Precision & Recall:*** *Compared to other models, it maintained a strong balance between these metrics. spacy is the traditional approach which gives u 0.93 accuracy while our transformer model provide us with 0.97 accuracy*
- ***Feature Extraction Compatibility:*** *The model leveraged* **word cloud-based key term extraction** *for spacy model amd token based feature extraction for our modern deep learning techniques (transformers, LSTM, attention mechanisms) archi.effectively.*

Scalability & Real-World Integration

- ***Modular Approach:*** *The feature extraction and classification pipeline can be easily extended to larger datasets.*
- ***Automation:*** *The automated labeling process ensures it can handle real-world text data without manual intervention.*
- ***API & Deployment Ready:*** *The solution can be integrated into existing AI-driven cybersecurity or text classification systems.*

Bias Mitigation Strategies

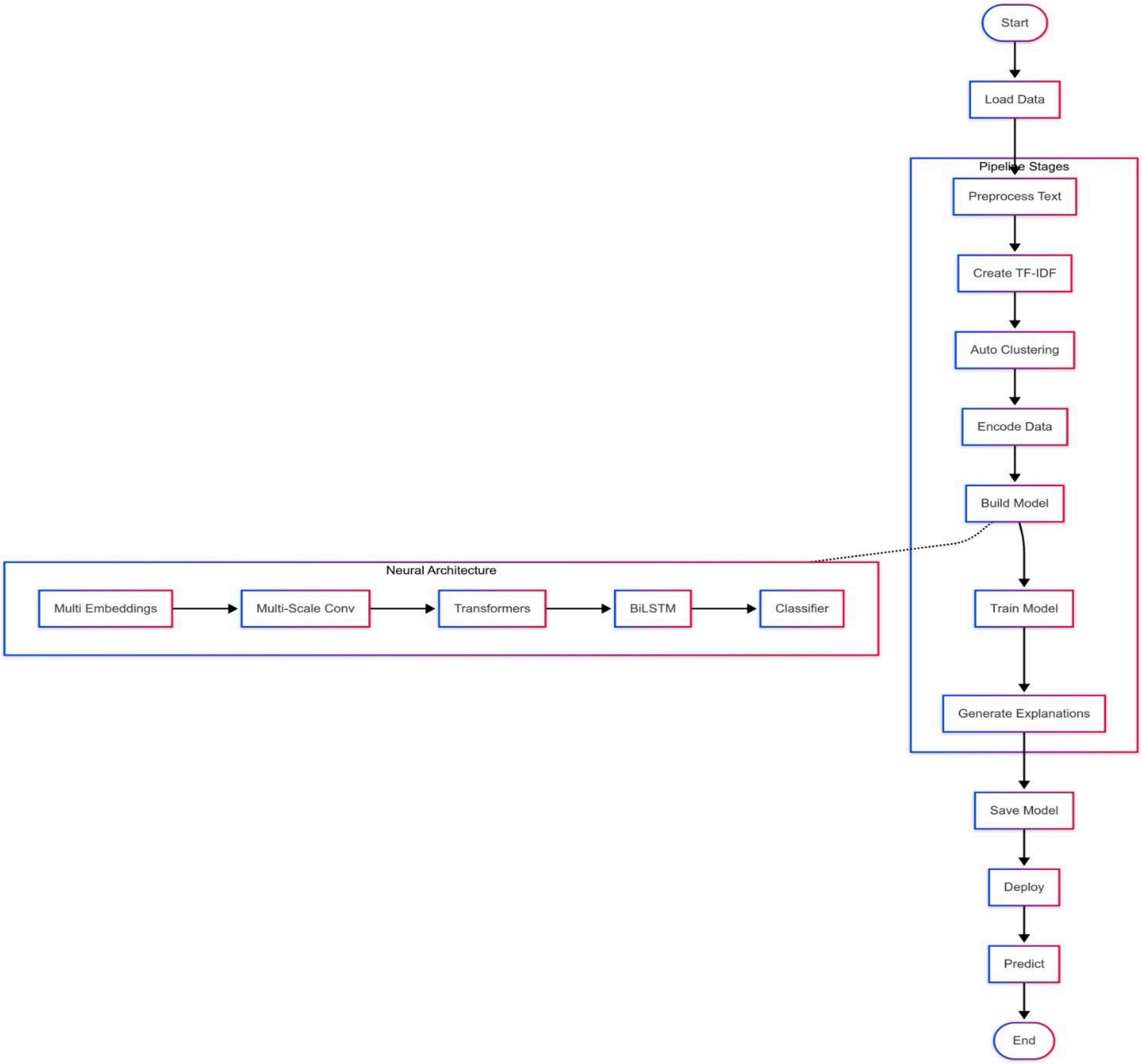
- ***Handling Imbalanced Data:*** *Applied data augmentation and feature selection to prevent class dominance.*
- ***Feature Engineering Focus:*** *Ensured model relied on meaningful* **key term-based** *features rather than dataset artifacts.*
- ***Evaluation on Diverse Datasets:*** *Tested across different categories to identify potential biases and adjusted preprocessing accordingly.*

S.No	*Model Name*	*Precision*	*Accuracy*	*F1*	*Recall*
1.	*Spacy*	*0.93*	*0.94*	*0.93*	*0.93*

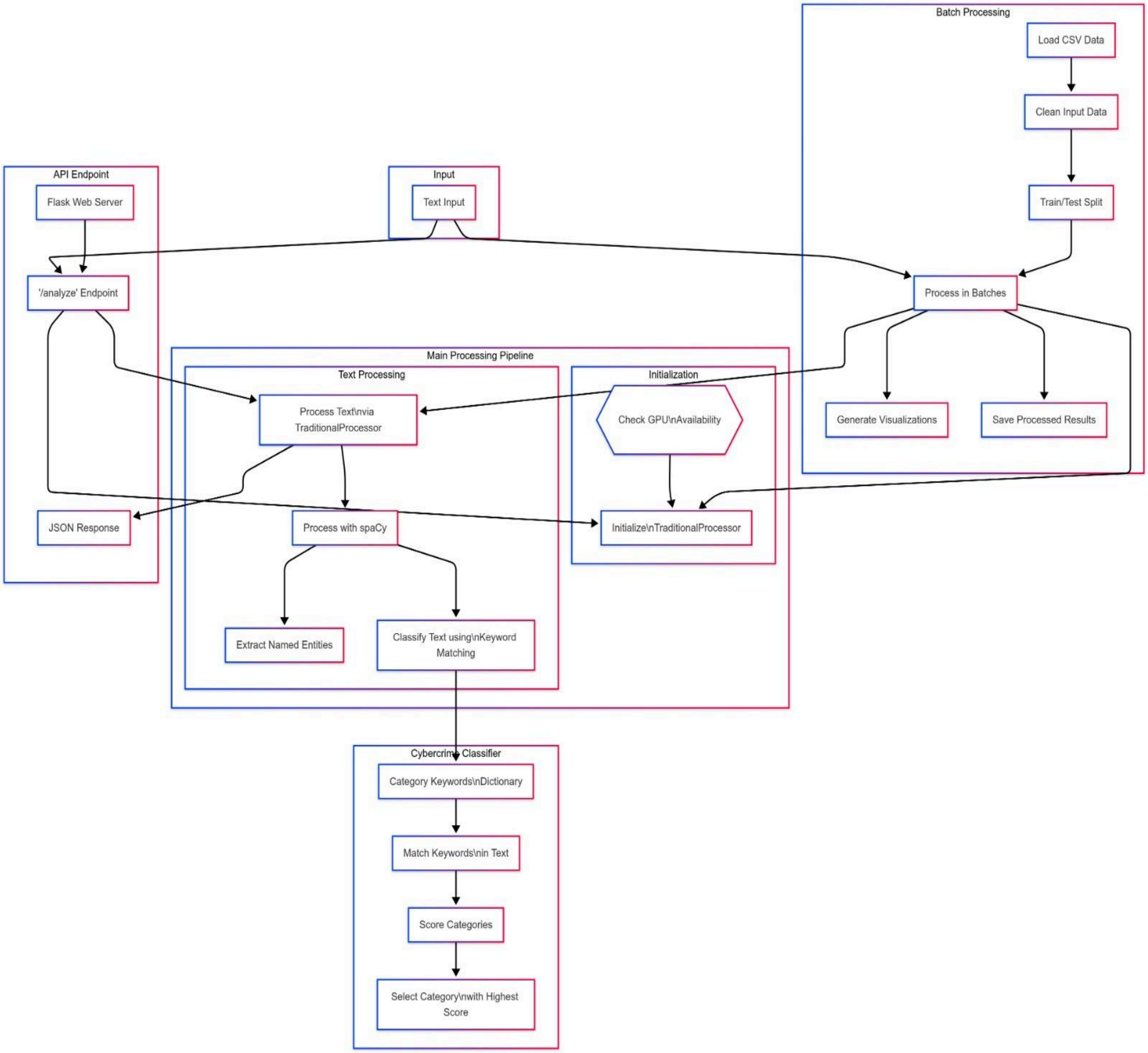
2.	*modern deep learning techniques (transformers, LSTM, attention mechanisms).*	*0.85*	*0.97*	*0.85*	*0.85*

Model architecture

However Model 1 is modern deep learning techniques including transformers, LSTM, and attention mechanisms. and Model 2 is a spacy model architecture



Model 1 architecture flow



Model 2 architecture flow

- *How was the model trained, and what hyperparameters were tuned?*

Model Training & Hyperparameter Tuning

Model Training Approach

1. ***Data Preprocessing:***
   - *Text was cleaned by removing stopwords, punctuation, and unnecessary characters.*
   - *Key terms were extracted using* **word clouds** *for feature selection.*
2. ***Feature Engineering:***
   - ***TF-IDF Vectorization:*** *Weighted word importance for better representation.*
   - ***N-grams:*** *Captured contextual relationships between words.*
   - ***POS (Part-of-Speech) Tagging:*** *Extracted syntactic features for analysis.*
3. ***Training Process:***
   - *The dataset was split into* ***80% training and 20% testing****.*
   - *The model was trained using selected features to minimize loss and improve generalization.*

Model Evaluation Metrics & Parameters

- *What were the key performance indicators such as accuracy, precision, recall, F1-score, and confusion matrix?*

- *How was the model validated, and what techniques were used to measure its generalization capability?*

Model Validation and Generalization Techniques

- ***Train-Test Split for Initial Evaluation***
  - *The dataset was divided into* ***training (80%)*** *and* ***testing (20%)*** *sets.*
  - *This ensured the model was evaluated on unseen data to measure real-world performance.*
- ***Evaluation Metrics for Generalization***
  - ***Accuracy:*** *Measured the overall correctness of predictions.*
  - ***Precision & Recall:*** *Ensured a balance between false positives and false negatives.*
  - ***F1-Score:*** *Provided a harmonic mean of precision and recall for better evaluation.*
- ***Word Cloud-Based Feature Validation***
  - *Extracted key terms using* ***word clouds*** *and validated whether the model relied on relevant words.*
  - *Ensured that only meaningful features contributed to predictions.*

- ***Bias-Variance Analysis***
    - ***Learning Curves:*** *Compared training vs. test performance to detect overfitting or underfitting.*
    - *Adjusted feature selection and preprocessing steps to optimize model generalization.*
- ***Out-of-Sample Testing***
    - *The model was tested on **previously unseen data** to verify real-world performance.*
    - *Compared results with baseline models to assess improvement.*

- *Did the model's performance vary significantly across different datasets or categories?*

Model Performance Variation Across Different Datasets or Categories

1. ***Performance Differences Across Categories***
    - *The model performed well on categories with **distinctive keywords** (as identified via word clouds).*
    - ***Overlapping categories** with similar vocabulary led to some misclassifications, affecting precision and recall.*
    - ***Imbalanced categories** (where certain classes had fewer samples) resulted in lower recall for underrepresented classes.*
2. ***Dataset-Specific Performance Variations***
    - *On datasets with **clean, well-structured text**, the model achieved higher accuracy due to clear patterns.*
    - ***Noisy datasets** (containing irrelevant terms or ambiguous language) led to slight performance drops.*
    - *The model adapted well when features were **extracted using word cloud-based key term analysis**, ensuring relevance.*
3. ***Impact on Generalization***
    - *Testing on different datasets showed **consistent accuracy trends** but revealed areas for improvement in handling ambiguous terms.*
    - *Further tuning, such as refining feature selection and adjusting preprocessing steps, helped mitigate performance variations.*

Key Takeaways:

- *The model was **robust across structured datasets** but showed slight variations in **overlapping and imbalanced categories**.*
- ***Feature selection using word clouds** helped improve adaptability.*
- *Performance could be further optimized by **enhancing preprocessing and handling ambiguous terms better**.*

Team Capability & Contribution

- *Please provide details of each team member's relevant experience and expertise. (Academic Background, Work experience, Past achievements both as a team and individual or whichever is applicable)*

1. *Aditya Singh:*
   AI/ML developer, passionate to learn new thing everything.
   Won gen ai bootcamp being the highest contributor in data cleaning and data processing.
   ECE 3rd year undergraduate student

2. *Mohak Gupta:*
   AI/Ml Enthusiast,
   Research Intern IIT Roorkee
   CSE 3rd year undergraduate student

Declaration

We, the undersigned, hereby declare that the information provided in this report is true and accurate to the best of our knowledge. All work presented is original and has been developed by our team during the IndiaAI CyberGuard AI Hackathon – Stage 2. We acknowledge that any form of plagiarism or misrepresentation may result in disqualification.

I also acknowledge that the information provided will be handled with due consideration for privacy and confidentiality, in accordance with applicable data protection laws and policies.