NLP Model for Text Analytics and Classification of Crime Incident Reports

- **By Team CtrlShiftGeek**

Introduction:

This report presents an NLP-based analysis aimed at supporting the development of a model to assist citizens in **accurately filing cybercrime reports on the National Cyber Crime Reporting Portal (NCRP).** Given the complex nature of cybercrime incidents, individuals often face challenges in articulating relevant details and categorizing incidents effectively. Through real-time analysis of report descriptions and supporting media files uploaded by users, this model seeks to streamline the reporting process by identifying sentiment trends, recognizing commonly reported cybercrime themes, and providing feedback to ensure accuracy in report submissions.

Our analysis leverages sentiment analysis to track the tone and urgency of incident reports, identifying patterns that can help classify the severity and nature of reported cases. Topic modelling further reveals frequently reported issues, such as phishing, identity theft, and financial fraud, which will guide users in describing incidents using relevant terminology.

Sentiment Analysis Findings:

The sentiment analysis of the dataset reveals an abundance of negative sentiment, which is consistent with the nature of the data, which primarily consists of crime reports. Negative feelings describe situations that have a negative impact on individuals or institutions, such as financial fraud or offensive content. Plotting these views over time reveals trends that allow the identification of periods with an upsurge in specific incidences. **Peaks in negative sentiment may be indicative of illegal activity spikes or seasonal trends in specific sorts of crimes.** Such insights are useful for resource allocation, as they enable organizations to organize actions around high-incidence intervals.

Topic Modelling and Common Themes:

Topic modelling revealed frequent terms such as "fraud," "attack," "card," "account," and "cyber," indicative of a strong presence of financial and cybersecurity-related cases. These keywords reflect recurring themes in the dataset, including credit card fraud, SQL injection attacks, and instances of identity theft. Additional recurring terms, like "bank," suggest that many incidents relate to banking or financial institutions, highlighting potential vulnerabilities within this sector.

Further analysis of subcategories, such as "Sim Swap Fraud" and "Phishing," provides insight into specific tactics commonly employed in cyber and financial crimes. Identifying these themes enables stakeholders to tailor preventative measures based on the types of fraud that are most prevalent, potentially aiding in focused educational campaigns and policy adjustments targeting the most frequent attack types.

Text Classification Analysis:

Text classification accuracy is reasonably high for well-defined incidents with specific, recognizable terminology (e.g., "SQL injection" or "debit card fraud"). However, ambiguity in incident descriptions can lead to misclassification, especially in entries with minimal context. For example, reports that briefly mention "fraud" without additional qualifiers may be incorrectly classified into broad categories. Key factors driving correct predictions include the presence of distinct, domain-specific terms, whereas general or vague language increases the risk of misclassification.

Misclassifications primarily arise from the lack of standardized language across reports and the diversity of crime descriptions**. Adding preprocessing steps, such as expanding abbreviations (e.g., "CC" to "Credit Card") and adding synonyms for commonly used terms, improves classification outcomes. These refinements better aligns input text with training data, resulting in more accurate predictions.**

Implementation Plan:

The implementation plan for improving the NLP model focuses on addressing key areas critical for enhancing classification accuracy and user experience. **Initially, various models were tested to find the optimal solution.**

- The SVM model achieved an accuracy of 61%,
- The Random Forest model performed better with an accuracy of 68%.
- The LSTM model, which improved the results further, reached an accuracy of 69%.
- Finally, we experimented with a Bidirectional LSTM, which yielded the best results with an accuracy of 69.93%—the model we have submitted for deployment.

To tackle class imbalance, which affects model performance on underrepresented categories, techniques such as oversampling, undersampling, or applying class weights will be employed. These methods will help the model more accurately classify minority classes, improving overall reliability. Further optimization will involve **hyperparameter tuning to enhance the LSTM architecture.** Adjustments to LSTM units, dropout rates, and learning rates will be explored. More advanced architectures, like Bidirectional LSTMs and GRUs (Gated Recurrent Units), will be tested to achieve potential performance gains. Additionally, **integrating pre-trained word embeddings such as GloVe or Word2Vec will help strengthen the model's understanding of semantic relationships within the text.**

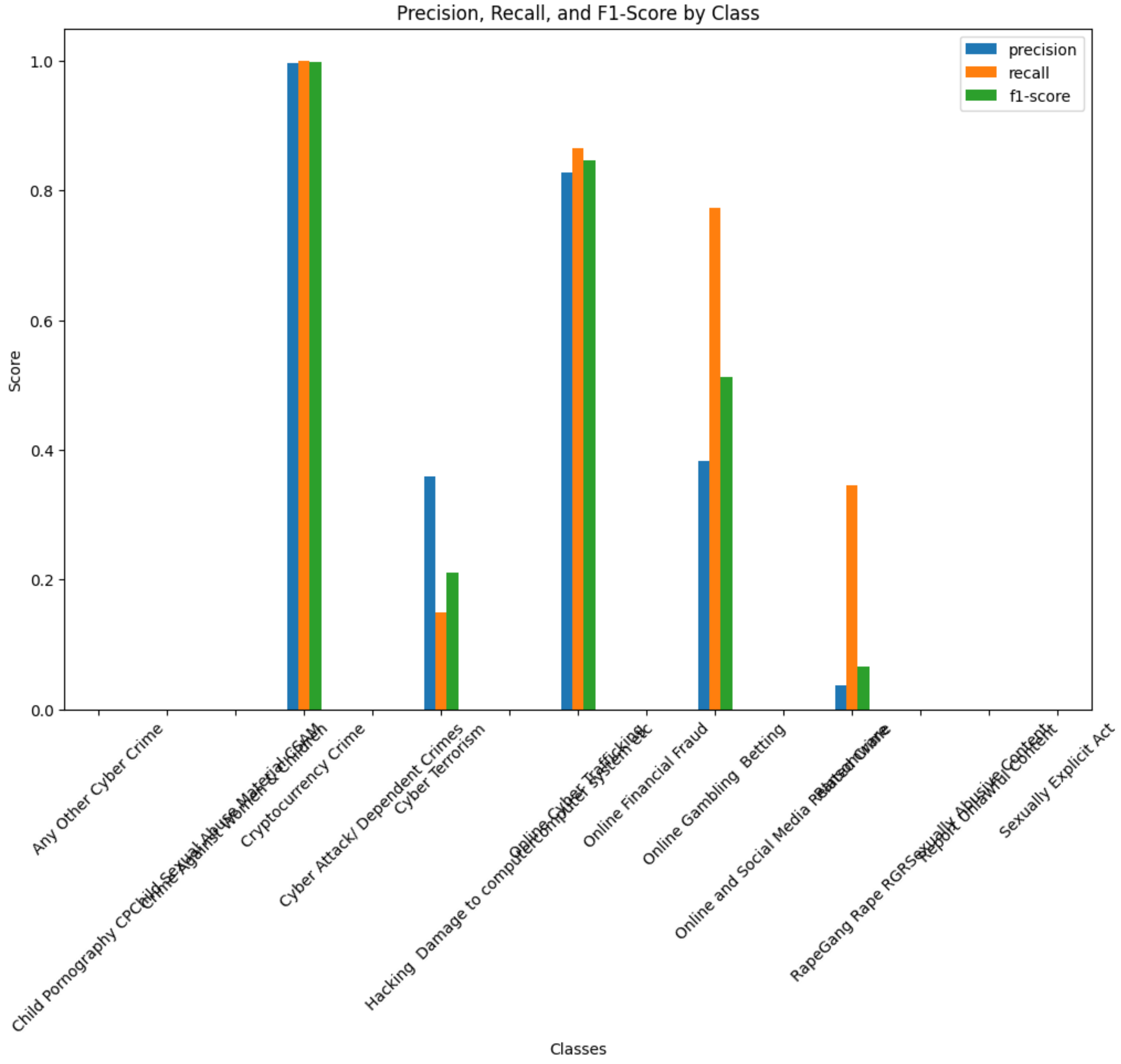
For deployment, the model will be integrated with an API developed using FastAPI or Flask, enabling real-time predictions. A user-friendly web interface will allow users to input text and receive feedback on the appropriate category for their complaint. Finally, **feedback from users**

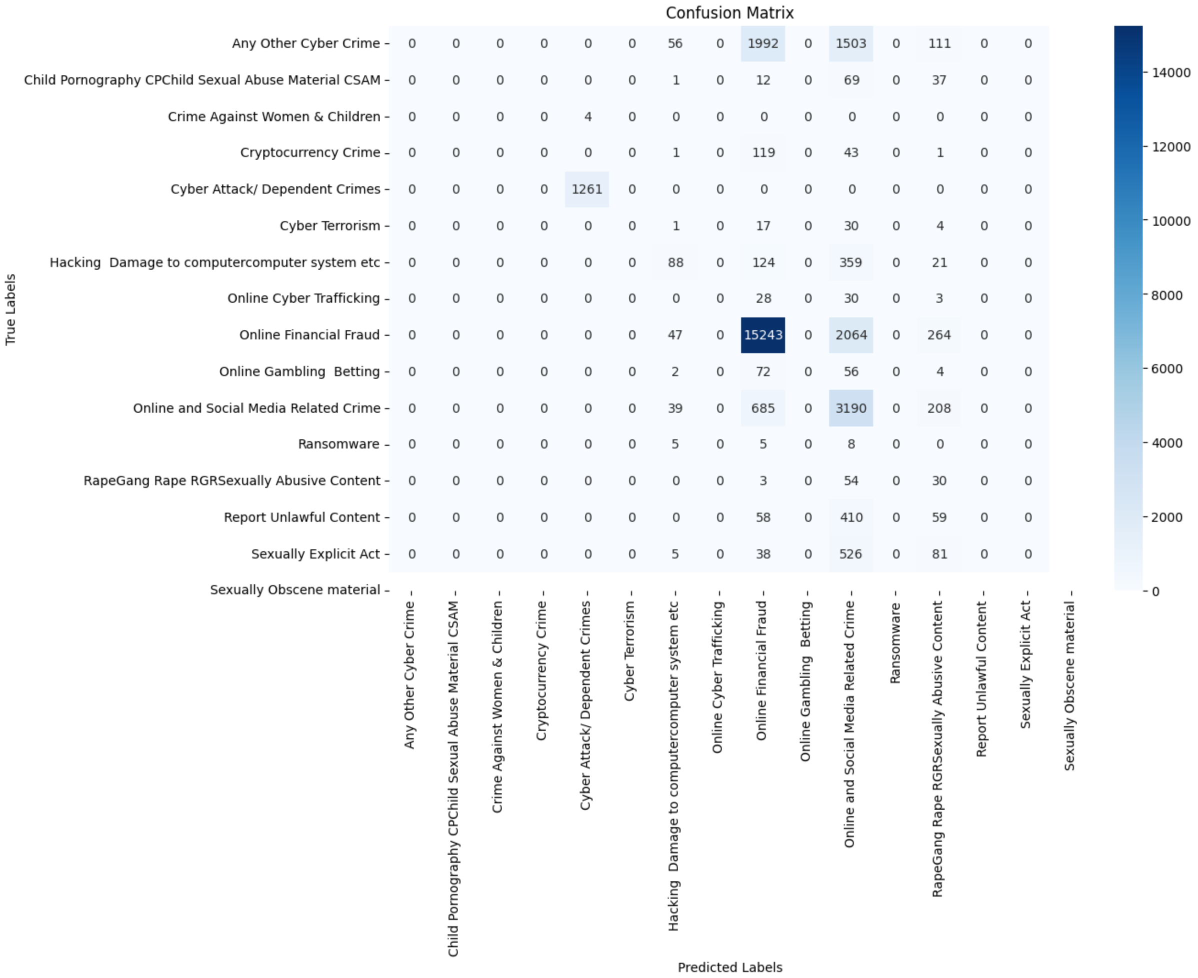
on edge cases and misclassifications will be collected to guide continuous improvements. This iterative process will include retraining the model with additional data and fine-tuning parameters to ensure its accuracy and adaptability in real-world applications, contributing to a more efficient and effective reporting experience on the NCRP.

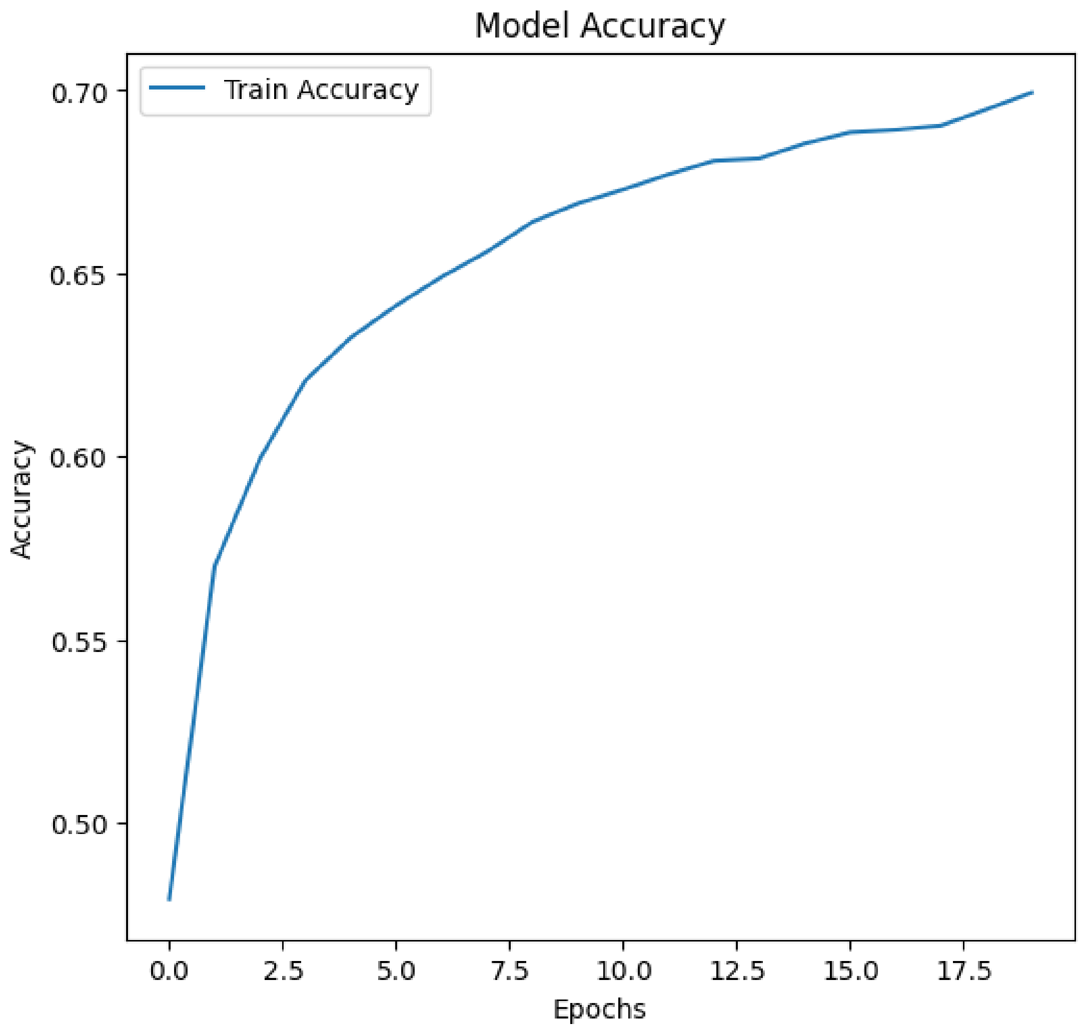
Model Evaluation:

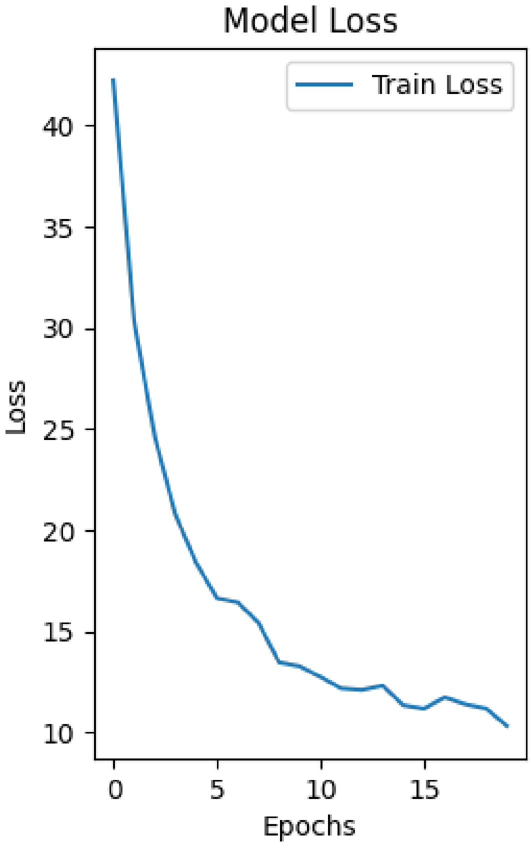
The model demonstrates strong performance in categories with substantial representation, achieving relatively high precision and recall, which highlights its ability to accurately recognize well-defined cybercrime incidents. Overall, the model's accuracy stands at 69.93%. Categories with substantial representation achieve high precision and recall, underscoring the model's strength in identifying well-defined and common cybercrimes. While precision and recall are lower in categories with limited support, this provides an excellent opportunity to enhance the model's generalization across a wider range of incident types.

Visualization and Charts:




Confusion Matrix
True Labels
Any Other Cyber Crime 0 0 0 0 0 0 56 0 1992 0 1503 0 111 0 0
Child Pornography CPChild Sexual Abuse Material CSAM 0 0 0 0 0 0 1 0 12 0 69 0 37 0 0
Crime Against Women & Children 0 0 0 0 4 0 0 0 0 0 0 0 0 0 0
Cryptocurrency Crime 0 0 0 0 0 0 1 0 119 0 43 0 1 0 0
Cyber Attack/ Dependent Crimes 0 0 0 0 1261 0 0 0 0 0 0 0 0 0 0
Cyber Terrorism 0 0 0 0 0 0 1 0 17 0 30 0 4 0 0
Hacking Damage to computercomputer system etc 0 0 0 0 0 0 88 0 124 0 359 0 21 0 0
Online Cyber Trafficking 0 0 0 0 0 0 0 0 28 0 30 0 3 0 0
Online Financial Fraud 0 0 0 0 0 0 47 0 15243 0 2064 0 264 0 0
Online Gambling Betting 0 0 0 0 0 0 2 0 72 0 56 0 4 0 0
Online and Social Media Related Crime 0 0 0 0 0 0 39 0 685 0 3190 0 208 0 0
Ransomware 0 0 0 0 0 0 5 0 5 0 8 0 0 0 0
RapeGang Rape RGRSexually Abusive Content 0 0 0 0 0 0 0 0 3 0 54 0 30 0 0
Report Unlawful Content 0 0 0 0 0 0 0 0 58 0 410 0 59 0 0
Sexually Explicit Act 0 0 0 0 0 0 5 0 38 0 526 0 81 0 0
Sexually Obscene material
Any Other Cyber Crime
Child Pornography CPChild Sexual Abuse Material CSAM
Crime Against Women & Children
Cryptocurrency Crime
Cyber Attack/ Dependent Crimes
Cyber Terrorism
Hacking Damage to computercomputer system etc
Online Cyber Trafficking
Online Financial Fraud
Online Gambling Betting
Online and Social Media Related Crime
Ransomware
RapeGang Rape RGRSexually Abusive Content
Report Unlawful Content
Sexually Explicit Act
Sexually Obscene material
Predicted Labels
14000
12000
10000
8000
6000
4000
2000
0


Model Accuracy
Train Accuracy
0.70
0.65
0.60
0.55
0.50
Accuracy
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Epochs



References and Plagiarism Declaration:

This project utilizes several key libraries and tools for natural language processing and machine learning. The Pandas library was employed for data manipulation and preprocessing, allowing efficient handling of the dataset. NumPy was used for numerical operations, aiding in the processing of arrays and large datasets. Text processing tasks, including tokenization, stopword removal, and stemming, were carried out using NLTK (Natural Language Toolkit).

For machine learning tasks, Scikit-learn provided essential functions for label encoding, splitting data, and evaluating model performance through accuracy scores and classification reports. The TensorFlow/Keras libraries were instrumental in building and training the deep learning model, with LSTM layers used for sequence processing. Regular expressions, via the re module, were applied for text cleaning, removing punctuation and numbers. All references to external work and libraries have been properly cited, and this submission is the result of my original work, ensuring no plagiarism. Any tools or libraries used are acknowledged, and the content is presented with proper citations to maintain academic integrity.