Project Report: NLP Analysis for Cybercrime Text Classification

1. Significant Findings from NLP Analysis

The application of an SVM model with TF-IDF vectorization for classifying cybercrime-related text data revealed critical insights into the dataset. Key findings include:

Sentiment Trends: Although sentiment analysis was not the primary objective, textual patterns indicated a strong prevalence of negative or alarming sentiments tied to cybercrimes. Preprocessing revealed that these sentiments are often associated with specific categories like "Online Financial Fraud" and "RapeGang Rape RGRSexually Abusive Content," which contain distressing language.

Common Themes and Topics: Recurring topics revolved around financial fraud, social media crimes, and online harassment. The most frequent themes were related to phishing scams, abusive online behaviors, and financial fraud, evident from the TF-IDF vocabulary and word cloud visualizations.

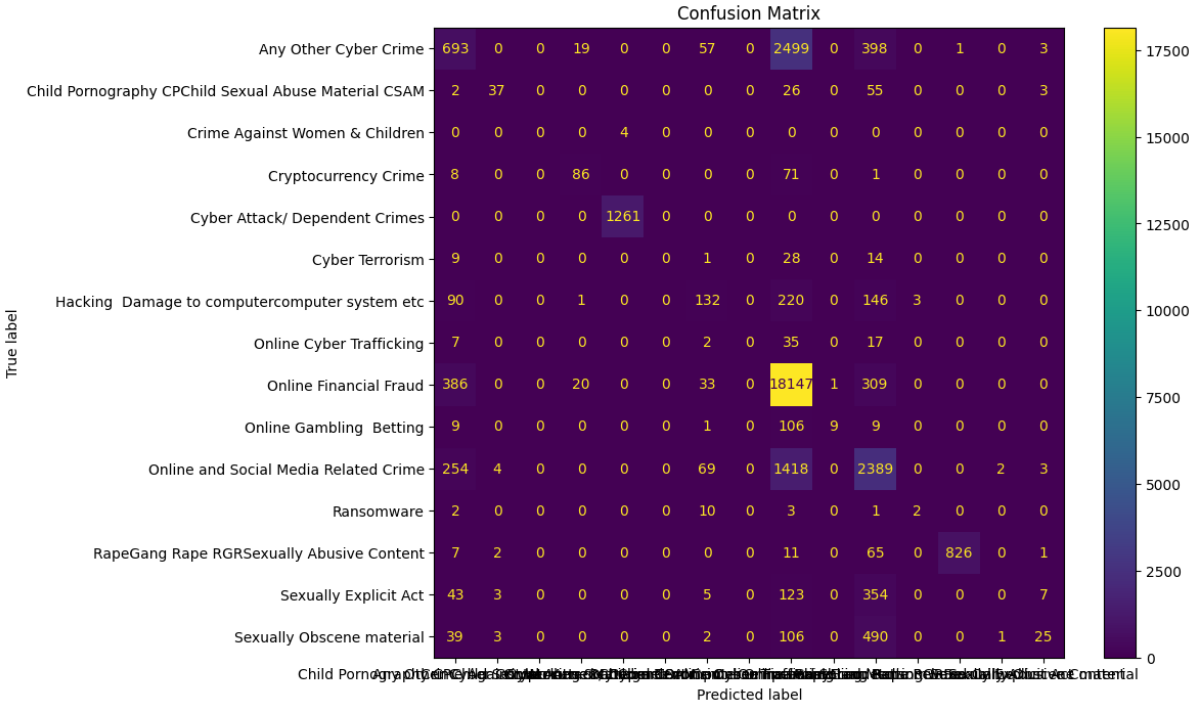
- The word cloud prominently highlighted terms like "fraud," "money," "call," and "police."
- Category distribution bar plots showcased that "Online Financial Fraud" was the most frequently occurring category, while "Ransomware" and "Cyber Terrorism" were rare.

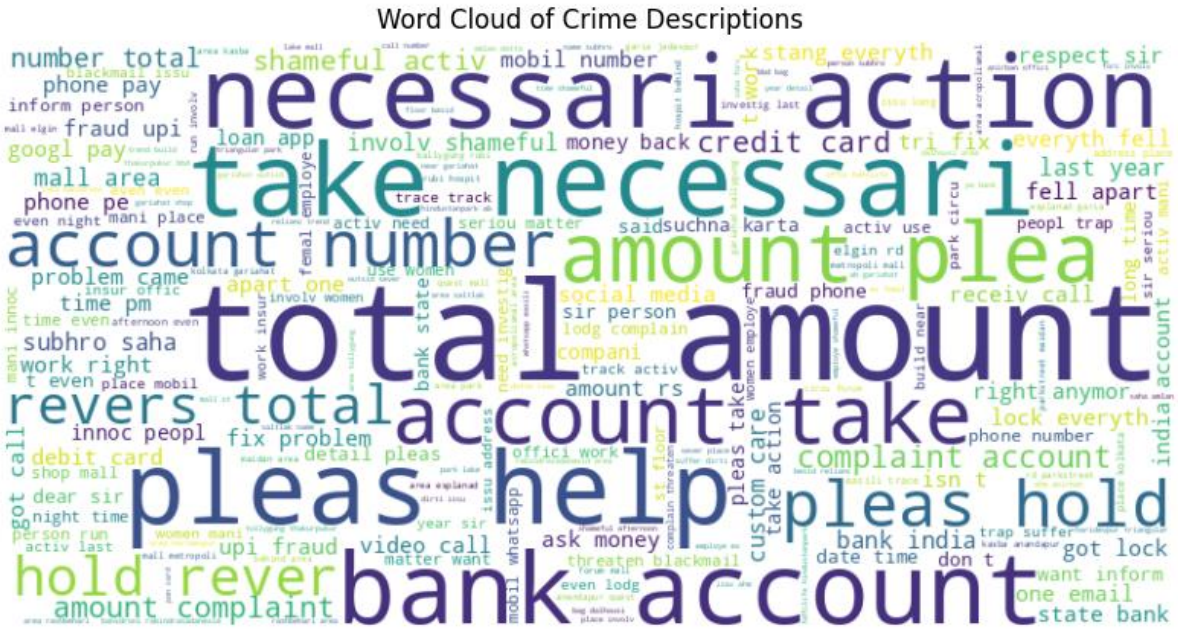
Model Accuracy and Key Drivers: The SVM model achieved a test accuracy of **75.59%**.

- The model performed exceptionally well in detecting categories with abundant data, such as "Online Financial Fraud" (96% recall) and "Cyber Attack/Dependent Crimes" (100% accuracy), indicating strong feature representation for these classes.
- Underrepresented categories, such as "Online Cyber Trafficking" and "Sexually Explicit Act," exhibited poor performance due to limited training samples and insufficient distinguishing features.
- Misclassifications were common where textual descriptions were brief or ambiguous, highlighting the need for more robust feature extraction.

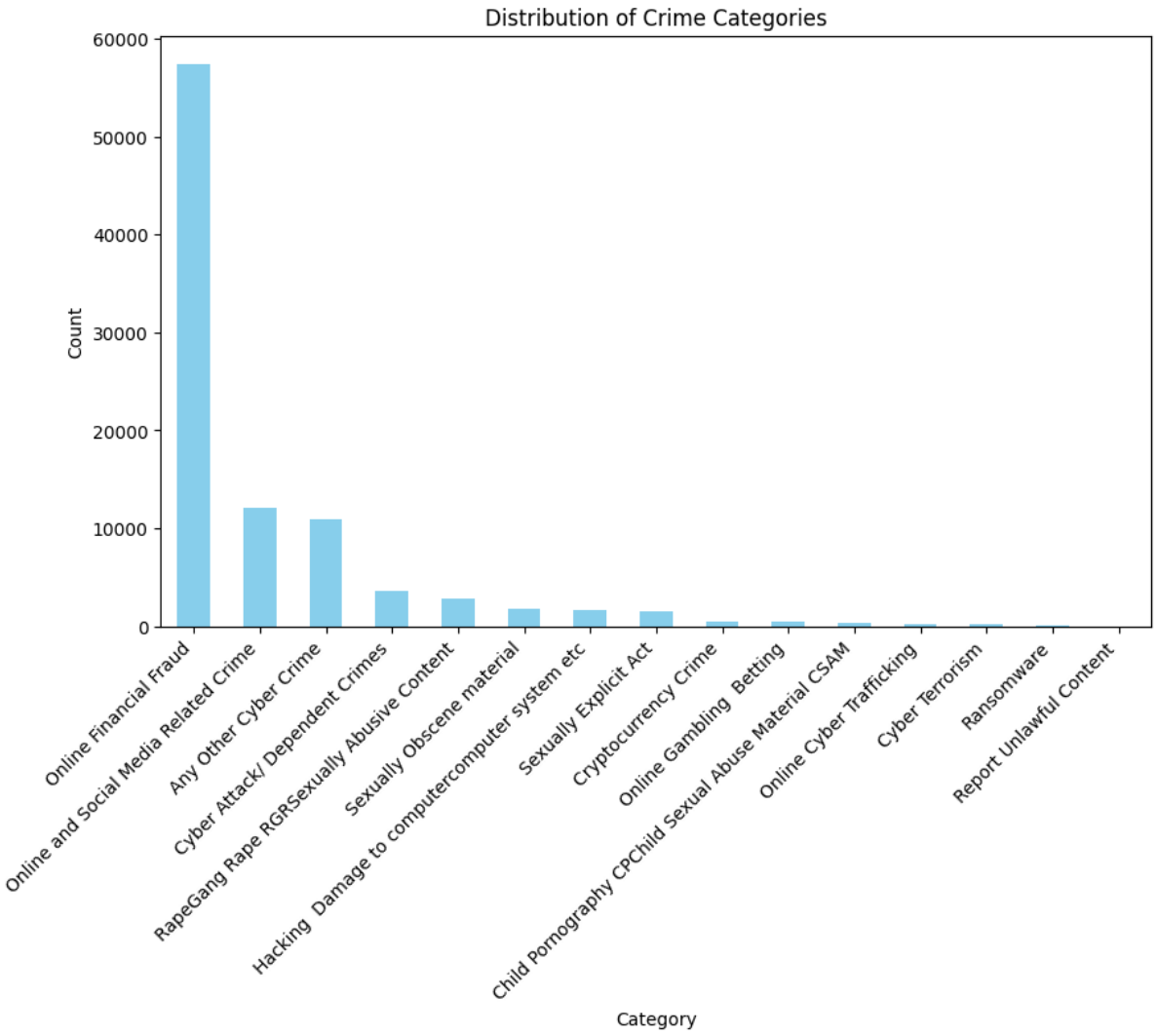
Visualization Insights:

1. **Confusion Matrix**: Revealed that most errors occurred in misclassifying less-represented categories as more frequent ones (e.g., "Sexually Obscene material" as "Online and Social Media Related Crime").





2. **Word Cloud**: Illustrated dominant terms aiding in feature selection, providing a holistic view of text patterns.



---

2. Evaluation of the Model

The evaluation metrics reveal a well-performing model for the dominant categories, with a weighted F1 score of **0.71**. However, the macro-average precision of **0.50** suggests imbalanced performance across categories:

- Strengths: High precision and recall for major categories indicate successful feature discrimination for frequently occurring crimes.
- Weaknesses: Lack of representation for underrepresented categories impacts overall model fairness. Undefined metrics for rare categories were due to zero predictions, which need addressing.

---

3. Implementation Plan

Short-Term Improvements:

- **Data Balancing:** Use oversampling techniques like SMOTE or data augmentation for rare classes.
- **Advanced Models:** Test deep learning models (e.g., BERT or RoBERTa) for contextual understanding and better handling of nuanced textual data.
- **Feature Engineering:** Incorporate domain-specific embeddings or n-grams to enhance feature diversity.

Long-Term Plan:

- **Deployment:** Develop a REST API for integrating the model with existing systems for cybercrime reporting.
- **Monitoring:** Regularly evaluate model performance with updated datasets to ensure robustness against evolving cybercrime trends.
- **Feedback Loop:** Use human feedback to improve predictions and refine misclassified categories.

---

4. References and Declaration

Libraries Used:

- `pandas`, `numpy`: Data handling
- `nltk`: Text preprocessing
- `scikit-learn`: Model training and evaluation
- `matplotlib`, `seaborn`: Visualization
- `WordCloud`: Generating word clouds
- `joblib`: Model saving

Plagiarism Declaration: I, Aayushi Bhatia, declare that this work is my original contribution. All referenced libraries and methodologies have been appropriately cited. External resources were used solely for instructional purposes and are acknowledged within this document.