REPORT

The objective of the **Cybercrime Classifier** is to automatically categorize crime-related descriptions into specific categories using a transformer-based model (DistilBERT). Our approach involves fine-tuning the model on a domain-specific dataset, optimizing it for text classification through efficient preprocessing, batch processing, and mixed precision training. This solution aims to improve accuracy and scalability for classifying large volumes of cybercrime data.

Preprocessing:

To prepare the crime-related text data for classification, we perform several essential preprocessing steps.

1. **tokenize** the text, splitting it into smaller units (tokens) that are compatible with transformer models like DistilBERT. This allows the model to understand word relationships and handle long descriptions efficiently.
2. remove **stop words** (e.g., "the," "is") as they add little value to the classification task. By eliminating these common words, we reduce noise and allow the model to focus on more meaningful content, improving processing speed and efficiency.
3. **Lemmatization** is applied next to standardize words into their root forms (e.g., "running" becomes "run"), which reduces vocabulary size and helps the model recognize variations of the same word, crucial for crime-related descriptions.
4. **Space removal** eliminates unnecessary spaces that could interfere with tokenization, ensuring cleaner input.
5. Similarly, we remove **special characters** (like punctuation) that don't contribute meaningfully to the text's meaning, though we are cautious not to remove symbols that might be important in certain contexts (e.g., hashtags).

Additionally, we may consider **lowercasing** the text to maintain consistency, handle **negations** (e.g., "not a fraud"), and address **class imbalances** through techniques like oversampling or class weighting. We also plan to handle **negations** (e.g., "not a fraud"), as they can change the meaning of the text and impact classification.

Model:

The **Cybercrime Classifier** is a transformer-based model built by fine tuning **DistilBERT**, a lightweight version of BERT, optimized for text classification tasks. It is designed to categorize crime-related descriptions into specific categories such as **Cyberbullying**, **Fraud**, and **Online Gambling**. This model leverages PyTorch for implementation and uses mixed precision training

for efficiency, especially on large datasets. Given that the size of the data is moderate and there are data imbalances, the model still performs pretty well.

The model is a **sequence classification model** that fine-tunes **DistilBERT** for a domain-specific task: categorizing crime-related text data. DistilBERT's efficiency allows it to capture contextual language features while being computationally lighter than full BERT, making it ideal for processing large crime data quickly and accurately. The model utilizes preprocessing steps like tokenization and label encoding, enabling effective handling of raw text.

Its structure is ideal for the problem at hand as it efficiently classifies multi-class crime categories and handles large datasets, ensuring scalability and robustness in a real-world setting. Mixed precision training and batch processing further optimize performance, reducing memory usage and speeding up training.

Key Features

- **Transformer-based**: DistilBERT captures contextual meaning from text, critical for classifying nuanced crime descriptions.
- **Fine-tuning**: The model is specifically fine-tuned for the crime dataset, enhancing its accuracy for this task.
- **Efficient Training**: Mixed precision training and batch processing improve memory efficiency and speed, allowing for faster model training.
- **Real-time Evaluation**: The model evaluates its performance after each epoch, providing real-time insights into training progress via classification metrics.

Potential Upgrades

1. **Multi-Task Learning**: Predict both category and subcategory labels simultaneously for more granular classification.
2. **Larger Models**: Experiment with more powerful models like **RoBERTa** or **BERT-large** for improved accuracy. However this shall work only with higher amounts of data.
3. **Data Augmentation**: Use techniques like paraphrasing to increase dataset diversity and address class imbalance.
4. **Ensemble Methods**: Combine multiple models to improve performance and handle edge cases. We were planning on making an ensemble of english and hinglish based fine tuned BERT models

Output:

Categories:

	precision	recall	f1-score	support
Any Other Cyber Crime	0.481	0.247	0.326	3670
Child Pornography CPChild Sexual Abuse Material CSAM	0.583	0.171	0.264	123
Crime Against Women & Children	0.000	0.000	0.000	4
Cryptocurrency Crime	0.557	0.410	0.472	166
Cyber Attack/ Dependent Crimes	0.997	1.000	0.998	1261
Cyber Terrorism	0.000	0.000	0.000	52
Hacking Damage to computercomputer system etc	0.367	0.334	0.350	592
Online Cyber Trafficking	0.000	0.000	0.000	61
Online Financial Fraud	0.820	0.948	0.880	18896
Online Gambling Betting	0.000	0.000	0.000	134
Online and Social Media Related Crime	0.565	0.612	0.587	4139
Ransomware	0.000	0.000	0.000	18
RapeGang Rape RGRSexually Abusive Content	0.999	0.906	0.950	912
Sexually Explicit Act	0.000	0.000	0.000	535
Sexually Obscene material	0.372	0.120	0.182	666
accuracy			0.763	31229
macro avg	0.383	0.317	0.334	31229
weighted avg	0.717	0.763	0.731	31229

Subcategories:

	precision	recall	f1-score	support
Business Email CompromiseEmail Takeover	0.000	0.000	0.000	90
Cheating by Impersonation	0.222	0.006	0.011	719
Computer Generated CSAM/CSEM	0.000	0.000	0.000	2
Cryptocurrency Fraud	0.541	0.518	0.529	166
Cyber Blackmailing & Threatening	0.000	0.000	0.000	1
Cyber Bullying Stalking Sexting	0.471	0.534	0.500	1366
Cyber Terrorism	0.000	0.000	0.000	52
Damage to computer computer systems etc	0.000	0.000	0.000	39
Data Breach/Theft	0.000	0.000	0.000	171
DebitCredit Card FraudSim Swap Fraud	0.699	0.700	0.700	3556
DematDepository Fraud	0.000	0.000	0.000	222
Denial of Service (DoS)/Distributed Denial of Service (DDOS) attacks	0.000	0.000	0.000	187
EMail Phishing	0.000	0.000	0.000	54
EWallet Related Fraud	0.661	0.399	0.498	1338
Email Hacking	0.357	0.269	0.307	130
FakeImpersonating Profile	0.458	0.425	0.441	763
Fraud CallVishing	0.314	0.269	0.290	1827
Hacking/Defacement	0.155	0.595	0.245	200
Impersonating Email	0.000	0.000	0.000	13
Internet Banking Related Fraud	0.715	0.560	0.628	2973
Intimidating Email	0.000	0.000	0.000	11
Malware Attack	0.000	0.000	0.000	170
Online Gambling Betting	0.000	0.000	0.000	134
Online Job Fraud	0.292	0.272	0.282	294
Online Matrimonial Fraud	0.000	0.000	0.000	38
Online Trafficking	0.000	0.000	0.000	61
Other	0.381	0.396	0.388	3670
Profile Hacking Identity Theft	0.481	0.463	0.472	751
Provocative Speech for unlawful acts	0.000	0.000	0.000	130
Ransomware	0.000	0.000	0.000	18
Ransomware Attack	0.129	0.344	0.188	186
SQL Injection	0.000	0.000	0.000	167
Sexual Harassment	0.000	0.000	0.000	1
Tampering with computer source documents	0.000	0.000	0.000	194
UPI Related Frauds	0.634	0.859	0.729	8890
Unauthorised AccessData Breach	0.286	0.205	0.239	370
Website DefacementHacking	0.000	0.000	0.000	39
nan	0.683	0.623	0.651	2236
accuracy			0.561	31229
macro avg	0.197	0.196	0.187	31229
weighted avg	0.529	0.561	0.534	31229