

Project Report

India Ai Hackathon 2024 Team Shakti

Rina Mishra(Phd scholar) , Shreya Singh(M.Tech CSE), Khushi Verma(B.Tech CSE)

1. We are provided with the multiclass text classification problem. To solve that problem we used following models:

Model	Benefits	Limitations
SVM [1]	It is possible to apply to unstructured data also such as text, images, and so on; kernel provides strength to the algorithm and can work for high-dimensional data	It needs long training time on large data sets and is difficult to choose a good kernel function, and choosing key parameters varies from problem to problem.
Logistic Regression (LR)[2]	Simple to implement, interpretable, and effective for binary classification tasks; performs well when the relationship between input features and output is linear.	Limited to linear decision boundaries; struggles with multicollinearity or non-linear relationships unless feature engineering is applied.
Naive Bayes (NB)[3]	Simple and fast; works well for text data and problems with categorical features due to its probabilistic nature.	Assumes feature independence, which may not hold in real-world scenarios, reducing accuracy.
Random Forest (RF)[4]	Handles missing values well, robust to overfitting; works for both classification and regression while providing feature importance insights.	Computationally expensive for large datasets; can be slower to predict due to ensemble nature.
BERT[5]	Excels at understanding context in text due to its transformer architecture; achieves state-of-the-art performance in NLP tasks.	Requires substantial computational resources for fine-tuning; can be overkill for simpler tasks or smaller datasets.

XLM-RoBERTA[6]	Multilingual, pre-trained model that excels at understanding text in multiple languages; effective for cross-lingual transfer and large-scale NLP tasks.	High computational cost for training and inference; fine-tuning requires significant resources and expertise.
----------------	--	---

Loopholes :

1. Unbalanced data
2. Data present in both english and hinglish
3. Abbreviations present
4. Noisy data
5. Test data wrongly labeled (We had sent an Email to inform your team regarding the same)
6. Nonsense data present (NULL values, Zibris)

Identified Data Challenges

1. Unbalanced Data

- **Issue:** The dataset has a significant class imbalance, resulting in biased model training and reduced generalization.
- **Potential Impact:** Models may overfit to the dominant class and underperform on minority classes.

Example :

Category	Count
Online Financial Fraud	57434
Online and Social Media Related Crime	12140
Any Other Cyber Crime	10878
Cyber Attack/ Dependent Crimes	3608
RapeGang Rape RGRSexually Abusive Content	2822
Sexually Obscene material	1838
Hacking Damage to computer system etc	1710
Sexually Explicit Act	1552

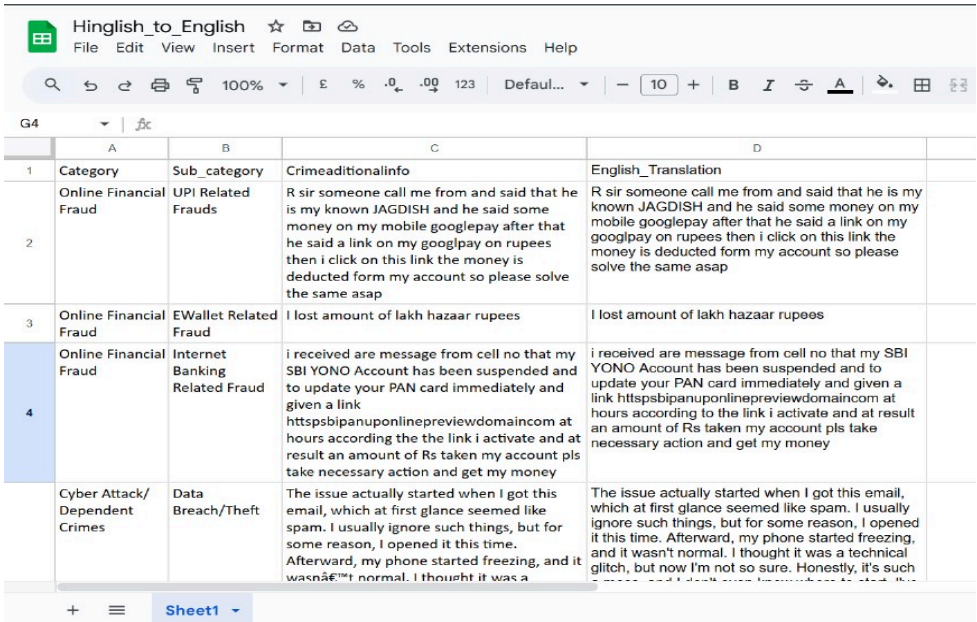
Cryptocurrency Crime	480
Online Gambling Betting	444
Child Pornography CPChild Sexual Abuse Materia...	379
Online Cyber Trafficking	183
Cyber Terrorism	161
Ransomware	56
Report Unlawful Content	1

Solution: Generated Synthetic Data of undersampled categories.

2. Mixed Language Content (English and Hinglish)

- **Issue:** Data contains a mix of English and Hinglish, making it challenging for models to interpret and process effectively.
- **Potential Impact:** May require language normalization or separate pipelines for handling mixed-language text.

Example Placeholder



Hinglish_to_English			
File Edit View Insert Format Data Tools Extensions Help			
G4			
	A	B	C
1	Category	Sub_category	CrimeadditionalInfo
	Online Financial Fraud	UPI Related Frauds	R sir someone call me from and said that he is my known JAGDISH and he said some money on my mobile googlepay after that he said a link on my googlpay on rupees then i click on this link the money is deducted form my account so please solve the same asap
2			
3	Online Financial Fraud	EWallet Related Fraud	I lost amount of lakh hazaar rupees
4	Online Financial Fraud	Internet Banking Related Fraud	i received are message from cell no that my SBI YONO Account has been suspended and to update your PAN card immediately and given a link httpsbipanuponlinepreviewdomaincom at hours according the the link i activate and at result an amount of Rs taken my account pls take necessary action and get my money
	Cyber Attack/Dependent Crimes	Data Breach/Theft	The issue actually started when I got this email, which at first glance seemed like spam. I usually ignore such things, but for some reason, I opened it this time. Afterward, my phone started freezing, and it wasn't normal. I thought it was a technical glitch, but now I'm not so sure. Honestly, it's such

Solution: We tried many language classification[7] and then translation options[8], but none of the methods was performing well because lack of support for the Hinglish Language[9].

Then we had written a python script and used google translate api for hinglish to english conversion. But it was failed to generate the correct results for the entire dataset.

Hence, finally we decided to use an advanced application of google translate, which was available in google sheet. Consider formula used for language translation:

```
D2      ▼ | fx =Googletranslate(C2, "hi", "en")
```

3. Use of Abbreviations

- **Issue:** Abbreviations and acronyms in the text introduce ambiguity, as they can have multiple meanings depending on the context.
- **Potential Impact:** Requires context-aware disambiguation or abbreviation expansion techniques.
- **Example Placeholder:** ASAP (As soon as possible), SMS(short message service) etc.

Solution: In our dataset, I performed the removal and replacement of abbreviations using a Python script. The process involved defining a dictionary of common abbreviations and their corresponding full forms. Then, I implemented a function to replace these abbreviations in the text using regular expressions for case-insensitive matching.

The transformation was applied to the "English Sentences" column of the dataset. Here's a summary of the steps:

- Created a dictionary mapping abbreviations (e.g., "lol" to "Laugh Out Loud").
- Developed a function, `replace_abbreviations`, to process the text by replacing abbreviations with their full forms.
- Applied this function to the specified column in the dataset to generate a modified version with expanded abbreviations.

This approach ensured that abbreviations in our textual data were expanded into their full forms, making the data cleaner and more interpretable for further analysis.

4. Noisy Data

- **Issue:** Dataset includes irrelevant or redundant content, such as random punctuation, special characters, and repeated text.
- **Potential Impact:** Increases preprocessing complexity and reduces data quality for model training.
- **Example Placeholder:**

The issue actually started when I got this email, which at first glance seemed like spam. I usually ignore such things, but for some reason, I opened it this time. Afterward, my phone started freezing, and it wasn't normal. I thought it was a technical glitch, but now I'm not so sure. Honestly, it's such a mess, and I don't even know where to start. I've contacted support, but they keep giving me the runaround. The more I tried to fix it, the more problems came up. My bank account was notification, and then I got locked out of everything. The worst part is, I don't even know who to trust anymore. Friends, family, everyone's acting strange. It's affecting my relationships and my mental health. Even my laptop isn't working right anymore. It's so stange how everything just fell apart after that one email.

Solution:

We performed rigorous Text preprocessing Approaches like:

Conversion from uppercase to lowercase, removed punctuations, numbers, special symbols, stop words removal and sentences less than 3 words removal, samples with values in CrimeadditionalInfo column are also removed, tokenization and lemmatization.

5. Mislabeled Test Data

- **Issue:** Some test samples are incorrectly labeled, leading to misleading performance metrics during evaluation.
- **Potential Impact:** Skewed accuracy and flawed conclusions about model performance.
- **Example Placeholder:**

category	crimeadditionalinfo
RapeGang Rape RGRSexually Abusive Content	Sir namaskar mein Ranjit Kumar PatraPaise nehi the tho sir kuch din pehele online loan aap Credit pearl loan aap se money loan kiya thalekin sir loan bolke jub loan diye tho mein turant return kar diya thalekin din baad whats app pe message aya payment karomein bola diye the aap

	mein wo de diyawo gali diye tho v return kar diyafir v message karke bolte hai full payment karo half payment nehi chalegarape case mein daldenge etcFake or illigal se contact number v hack kar dete haibol rahehai sab ko message karenge ye rapist hai bolke sirpls sir small ammount ke liye goggle play store se loan apply kiya thaFake loan aap v hai socha nehi thapls sir request kar rahahun action lo Sir mera number hai jo v proof chahiye dunga Sir
Sexually Explicit Act	we got call from this number and they ask me for gmail id and then we given code and they acces my gmail account then i got msg from kotak bank transaction of INR has been made of your kOTAK aacount so we request you to please check and resolve this issue at the erliest
Online Financial Fraud	In january month i login in loan application by mistake i allow contact info after that they are started harrashing memy reletivesmy friends now they sending maseges to them and calling them they are calling back to back every five minuts and send maseges to my whole reletive and friend circle

- Solution: The categories of the test dataset should be modified. But we can not perform this on our own, hence we sent an email to organizers to modify it or suggest some alternate solution. If this could be corrected the test accuracy could also be improved up to a great extent. However till now we had performed testing on the provided dataset only.
We performed text preprocessing and all steps which we performed with the training set also.

6. Presence of Nonsense Data

- **Issue:** Dataset includes NULL values, gibberish (e.g., random characters), or nonsensical text that holds no meaningful information.
- **Potential Impact:** Wastes computational resources and negatively affects model performance.
- **Example :**

Null Values : Line 25 in Training dataset

Online and Social Media Related Crime	Cheating by Impersonation	
--	--------------------------------------	--

Gibberish or Nonsensical text :

Example : “ņkhņksanckwskcnwsk” etc.

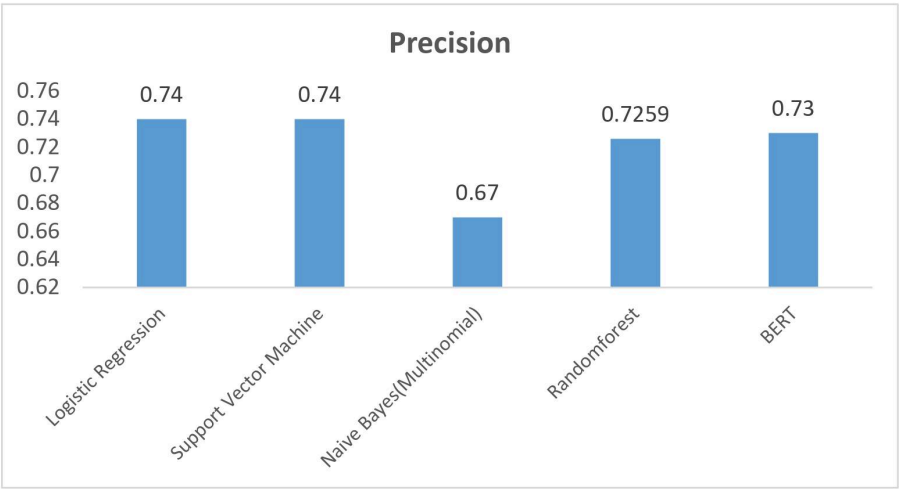
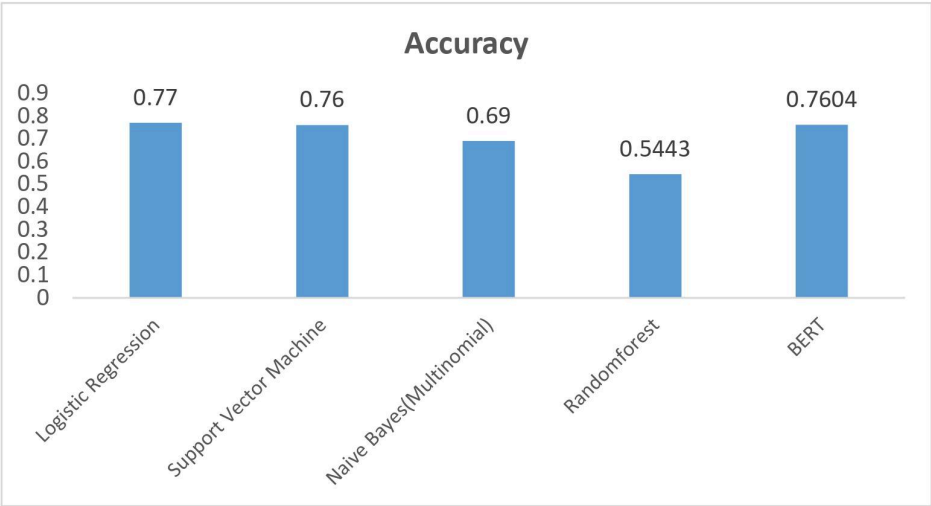
7. Note:

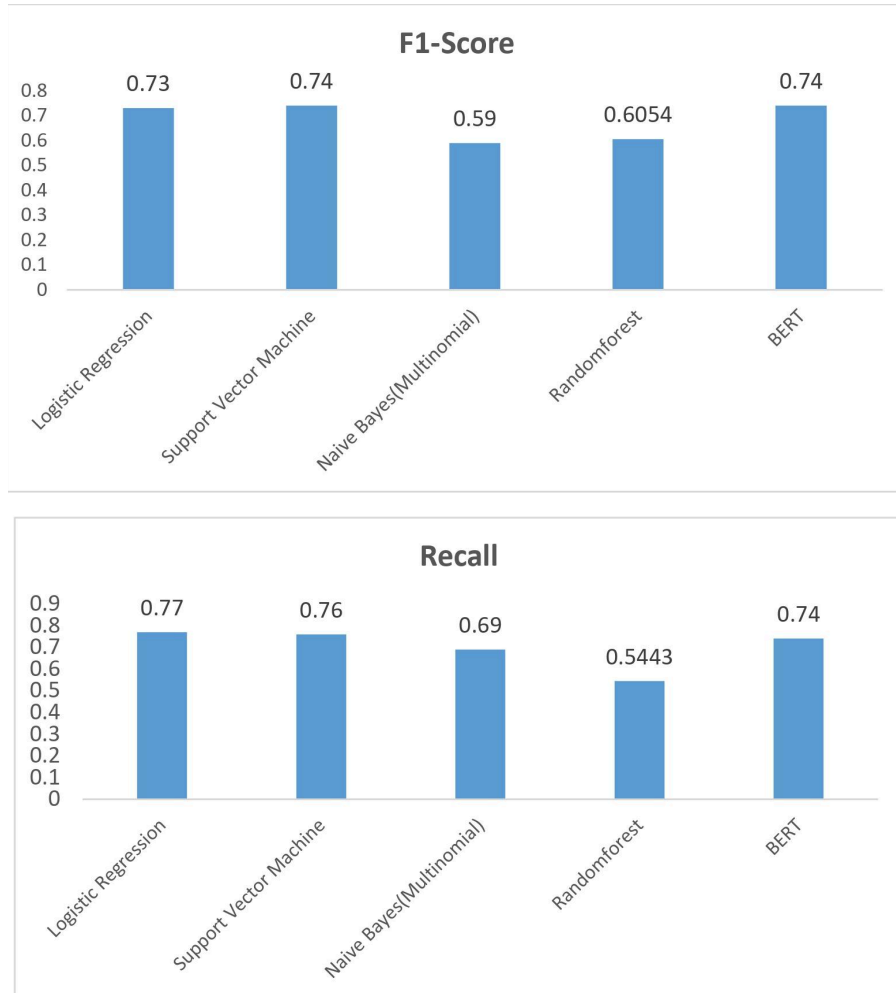
The reason to get the achieved 78% accuracy is due to wrongly labeled data and also Cyber Attack Dependent Crime category has monotonic samples. Hence it was not performing well when we were performing subcategory wise classification.

8.Final Results:

Models	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.77	0.74	0.77	0.73
Support Vector Machine	0.76	0.74	0.76	0.74
Naive Bayes(Multinomial)	0.69	0.67	0.69	0.59
Randomforest	0.5443	0.7259	0.5443	0.6054
BERT	0.7604	0.73	0.74	0.74

Visualization of the Results:





9. Further Improvements:

Since various modern models are available now, which could understand the context of the language written like BERT, XLM-Roberta etc. But they are not able to perform well due to wrong categorization.

10. Plagiarism Declaration

We, the team "Shakti" participating in the India AI Hackathon 2024, hereby declare that the content presented in this Project Report is entirely our own work, except where otherwise indicated. We have used various models like SVM, LR, Naive Bayes, Random Forest, and BERT to address the problem statement provided in the hackathon. All ideas, methods, and findings included in this report, unless explicitly cited, are the results of our own research and work.

Where external sources or references have been used, proper citations have been provided in the report. We acknowledge the contributions of these sources and have made every effort to avoid

any form of plagiarism. We have not copied or reproduced any material without proper acknowledgment, and all external content is appropriately referenced.

We confirm that this work has not been submitted previously for any other academic or professional competition and is not plagiarized in any form.

References:

1. A Complete Process of Text Classification System Using State-of-the-Art NLP Models
Varun Dogra, 2022.
2. Cortes, C., & Vapnik, V. (1995). "Support-Vector Networks." *Machine Learning*, 20, 273–297.
3. Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.
4. Breiman, L. (2001). "Random Forests." *Machine Learning*, 45(1), 5–32.
5. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *NAACL-HLT 2019*.
6. Conneau, A., et al. (2020). "Unsupervised Cross-lingual Representation Learning at Scale." *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
7. Language Classification: <https://github.com/pemistahl/lingua-go>
8. Language Translation: <https://github.com/hazzillrodriguez/flask-multi-language>
9. Gupta, P., & Joshi, S. (2018). "Handling Code-Mixed Hindi-English Social Media Data: A Language Identification Perspective." *Proceedings of the ACL*.