All the datasets were combined into ScamDataset.csv and shuffled.

# Source of Dataset

https://www.kaggle.com/datasets/narayanyadav/fraud-call-india-dataset

https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset

https://github.com/wspr-ncsu/robocall-audio-dataset/blob/main/metadata.csv

<https://huggingface.co/datasets/alissonpadua/ham-spam-scam-toxic>

## Generated Data with GPT

Prompts used:

* I am creating a machine-learning app to detect Phone call scams. Please produce 50-record samples dataset in CSV format that contains phone conversations and label indicating whether the message is 'fraud' or 'normal'.
* Great, produce 100 more unique record samples dataset in CSV format that contains phone conversations and label indicating whether the message is 'fraud' or 'normal'. Add some conversations with more Singaporean accent. Make sure they have a ration of 60:40 of 'normal' and 'fraud' respectively. shuffle the rows randomly.

A screenshot of a chat window

Description automatically generated