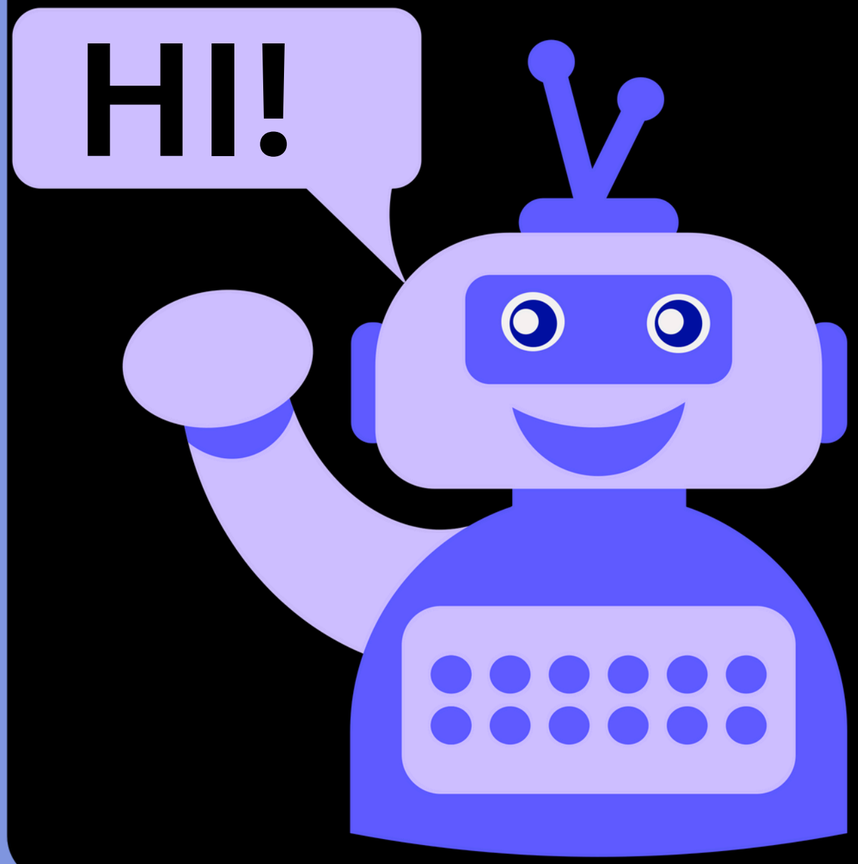
Spam or Not? Building a Robust SMS Classifier



PRESENTED BY:

Sayan Roy

INTRODUCTION

Problems:

- ***Wastes time and inconvenience for users.***
- ***Security risks, including phishing and malware.***
- ***Impact on productivity and system efficiency.***

Need for a Classifier:

- ***Reducing manual filtering.***
- ***Enhancing communication reliability.***

Objective:

To create an end-to-end solution that processes SMS messages, trains a robust classification model, and accurately identifies messages as spam or ham using NLP techniques.

DATASET OVERVIEW:

Dataset Source:

- *The "Spam SMS Classification Using NLP" dataset is sourced from a publicly available repository for research purposes.*

Number of Samples:

- *Total Messages: 5,574*

Class Distribution:

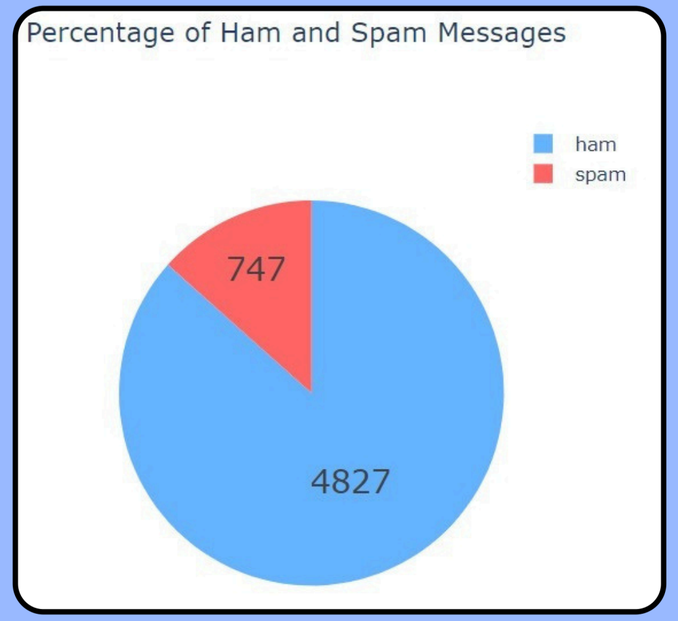
- *Ham (Non-Spam): 4,825 messages (86.6%)*
- *Spam: 747 messages (13.4%)*

Characteristics:

- *Spam messages are generally shorter but contain more promotional or suspicious keywords.*
- *Ham messages tend to be more conversational and longer.*

1 to 10 of 5574 entries Filter

Class	Message
ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
ham	Ok lar... Joking wif u oni...
spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's
ham	U dun say so early hor... U c already then say...
ham	Nah I don't think he goes to usf, he lives around here though
spam	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, £1.50 to rcv
ham	Even my brother is not like to speak with me. They treat me like aids patent.
ham	As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to copy your friends Callertune
spam	WINNER!! As a valued network customer you have been selected to receivea £900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only.
spam	Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call The Mobile Update Co FREE on 08002986030

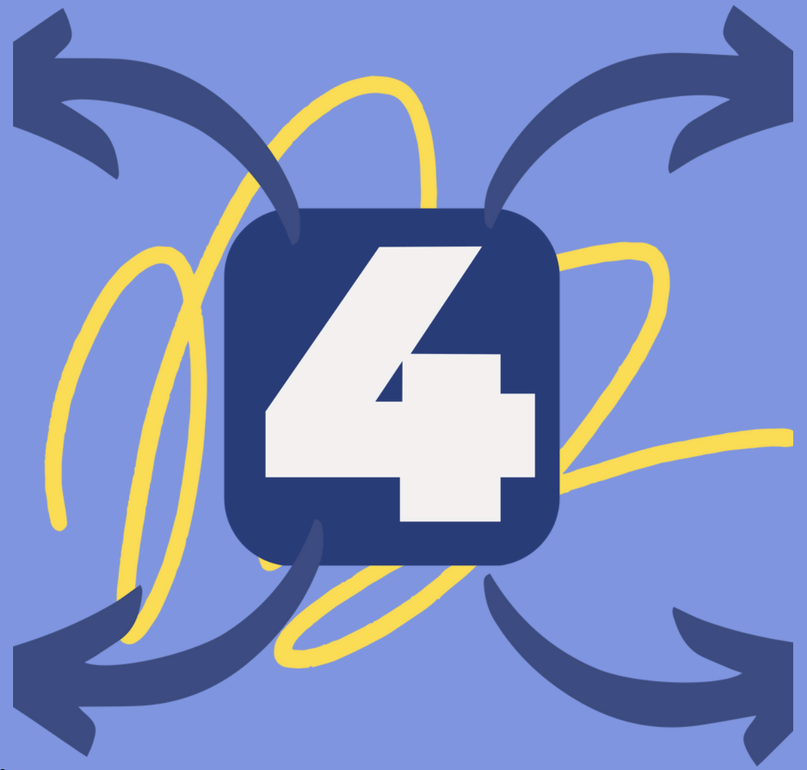


DATASET PREPROCESSING

INPUT DATA

REPLACE URLS

URLs in the text are replaced with <URL>.

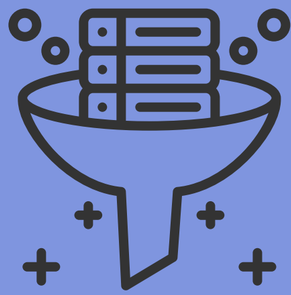


STEMMING

Words like "running" or "checked" are converted to "run" and "check".

REMOVE STOPWORDS

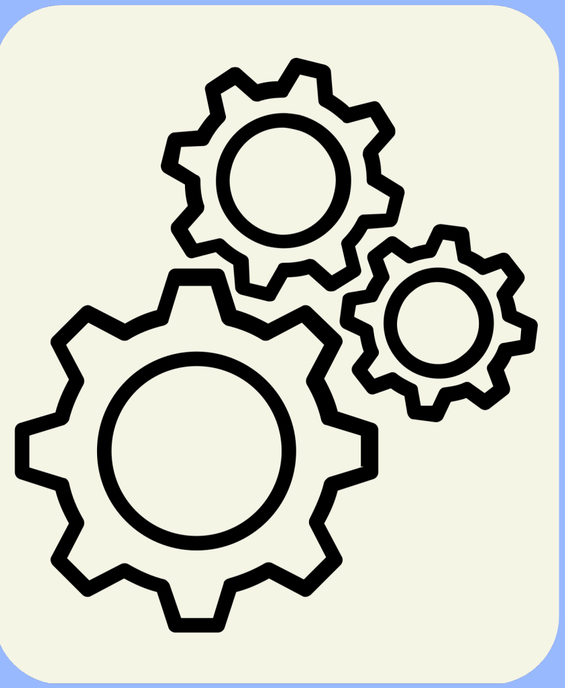
Common stopwords like "the", "and", "is" are eliminated.



CLEANED DATASET

	Class	Message	Message_cleaned	labels
0	ham	Go until jurong point, crazy.. Available only ...	Go jurong point, crazy.. avail bugi n great wo...	0
1	ham	Ok lar... Joking wif u oni...	Ok lar... joke wif u oni...	0
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	free entri 2 wkli comp win FA cup final tkt 21...	1
3	ham	U dun say so early hor... U c already then say...	U dun say earli hor... U c alreadi say...	0
4	ham	Nah I don't think he goes to usf, he lives aro...	nah I think goe usf, live around though	0
...	...	...	...	...
5569	spam	This is the 2nd time we have tried 2 contact u...	thi 2nd time tri 2 contact u. U £750 pound pri...	1
5570	ham	Will ü b going to esplanade fr home?	will ü b go esplanad fr home?	0
5571	ham	Pity, * was in mood for that. So...any other s...	pity, * mood that. so...ani suggestions?	0
5572	ham	The guy did some bitching but I acted like i'd...	the guy bitch I act like i'd interest buy some...	0
5573	ham	Rofl. Its true to its name	rofl. it true name	0

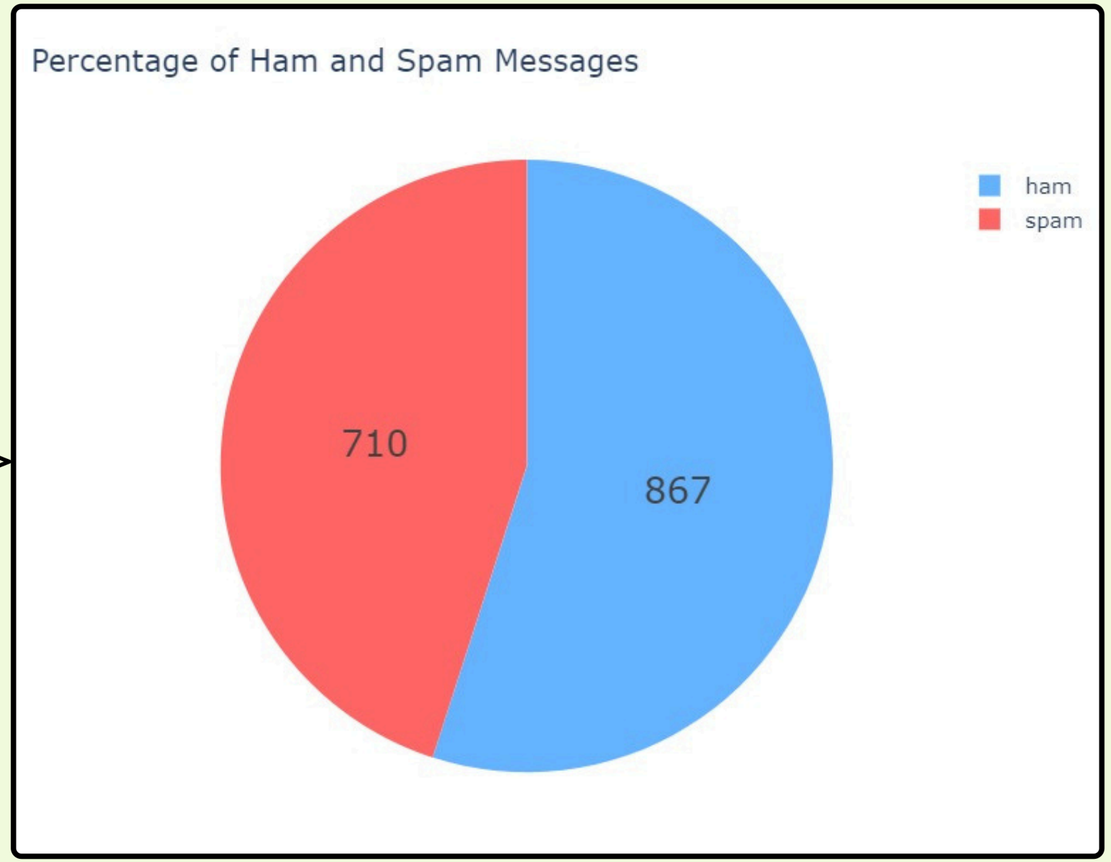
5574 rows × 4 columns



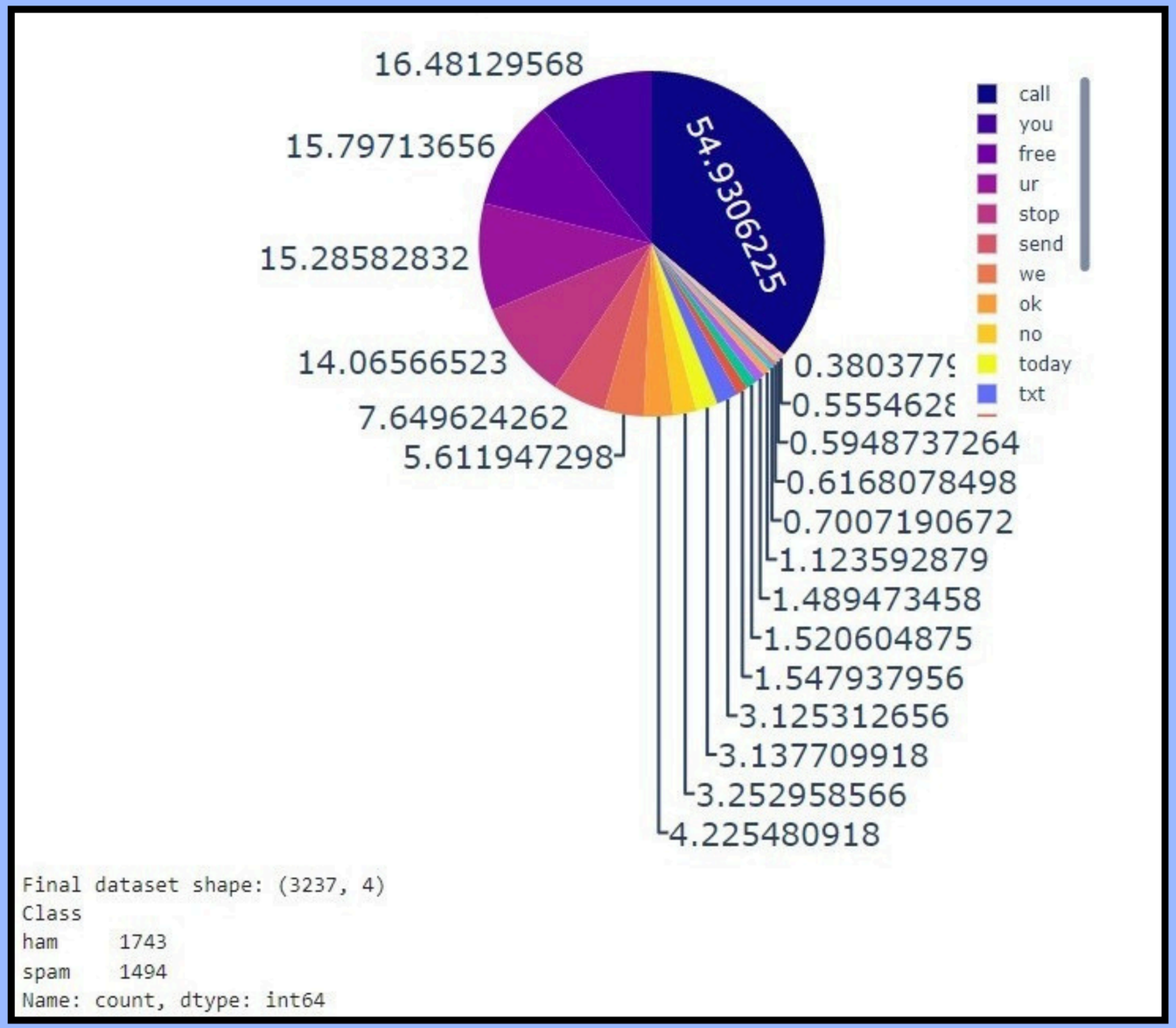
FEATURE ENGINEERING AND BALANCING THE DATASET

1. *Train Word2Vec Model*
2. *Message Vectorization*
3. *Synthetic Spam Generation*
4. *Balanced Dataset Creation*

After Balancing



MOST COMMON PHRASES IN SPAM MESSAGES




```
Final dataset shape: (3237, 4)
Class
ham     1743
spam    1494
Name: count, dtype: int64
```

MODEL DEVELOPMENT

Algorithm and Approach:

- Base Model: ***LSTM*** (Long Short-Term Memory)
- Embedding Layer: ***Pre-trained Word2Vec***

Layers:

- Input: ***Embedding vector*** (size equal to Word2Vec dimensions).
- ***LSTM Layer:*** Captures sequential dependencies in the text.
- ***Fully Connected Layer:*** Outputs probabilities for "spam" or "ham."

Model Tuning:

Used techniques like:

- *Hyperparameter tuning: Number of LSTM units, dropout rates, and learning rates.*
- *Early stopping to prevent overfitting.*

Training Process:

- ***Data split:*** *80% training, 10% validation, 10% testing with stratification to maintain class distribution.*
- ***Batch size:*** *32.*
- ***Optimizer:*** *Adam.*
- ***Loss function:*** *Binary Cross-Entropy Loss.*

IMPLEMENTATION DETAILS:

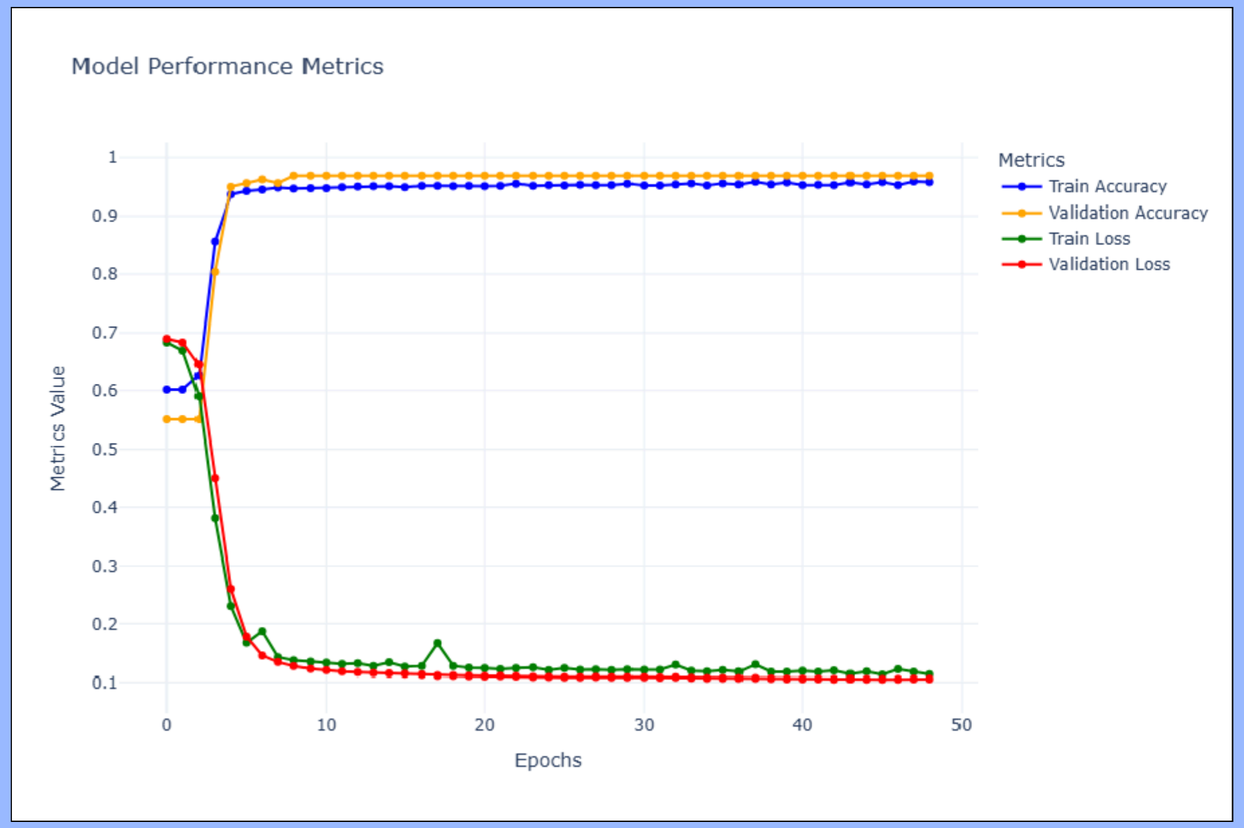
- ***Frameworks/Libraries:*** *PyTorch, Gensim , Scikit-learn.*
- ***Runtime:*** *Model trained for 300 epochs,*
  achieving convergence at high accuracy levels.

Final Training Metrics:

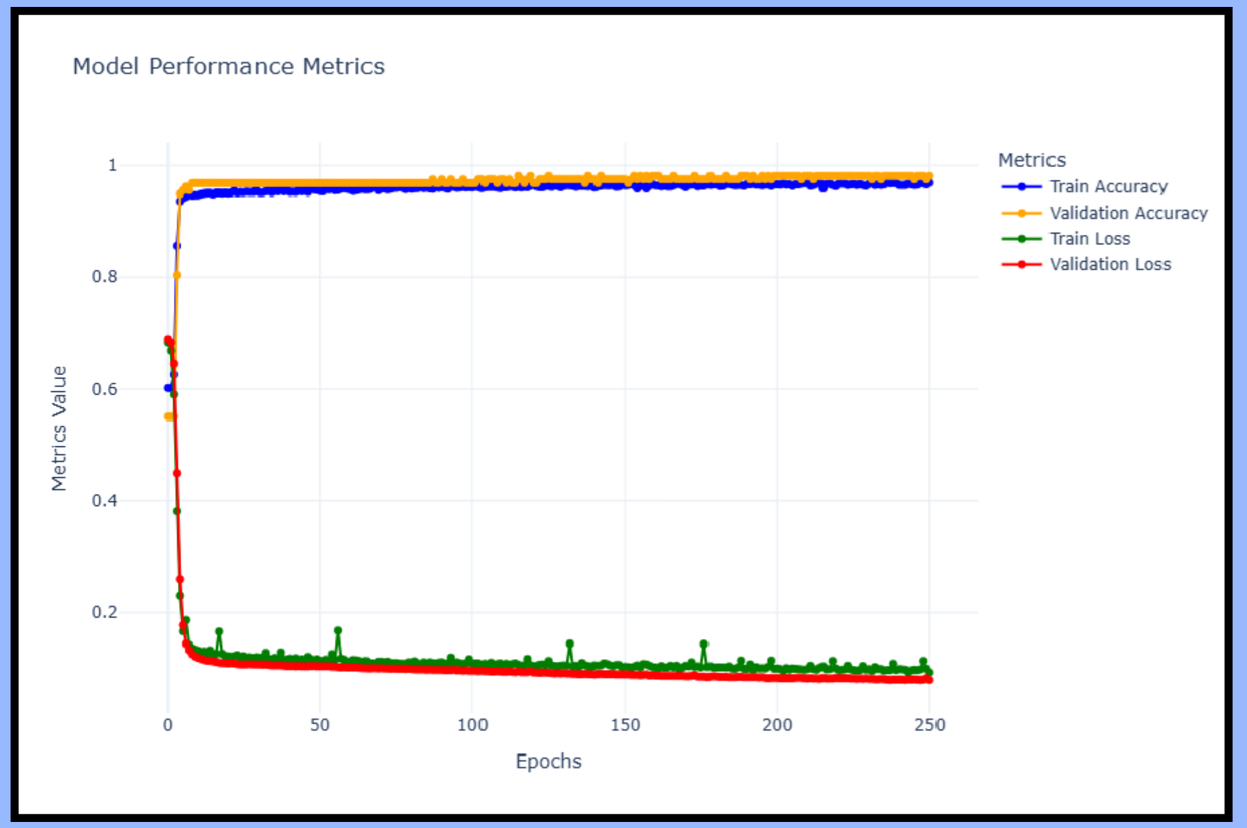
- ***Train Accuracy:*** *96.82%*
- ***Validation Accuracy:*** *97.47%*
- ***Valiation Loss:*** *8.07%*
- ***Train loss:*** *9.09%*
- ***Test Accuracy:*** *98.21 %*
- ***Test F1 Score:*** *0.93*
- ***Test Precision:*** *0.92*
- ***Test Loss:*** *0.0717*

RESULTS AND EVALUATION

INITIAL EPOCH (TILL 50) :



FINAL EPOCH (TILL 250) :



MODEL PIPELINING

Dataset

Preprocessing using NLTK

Word2Vec

Training Loop

Vectorised Messages of 100 dimensions

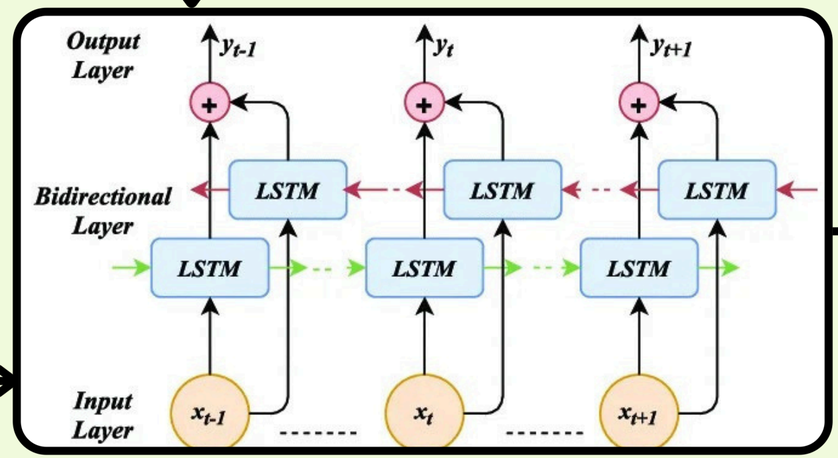
Spam Generation

Ham Downsample

Down Sampler

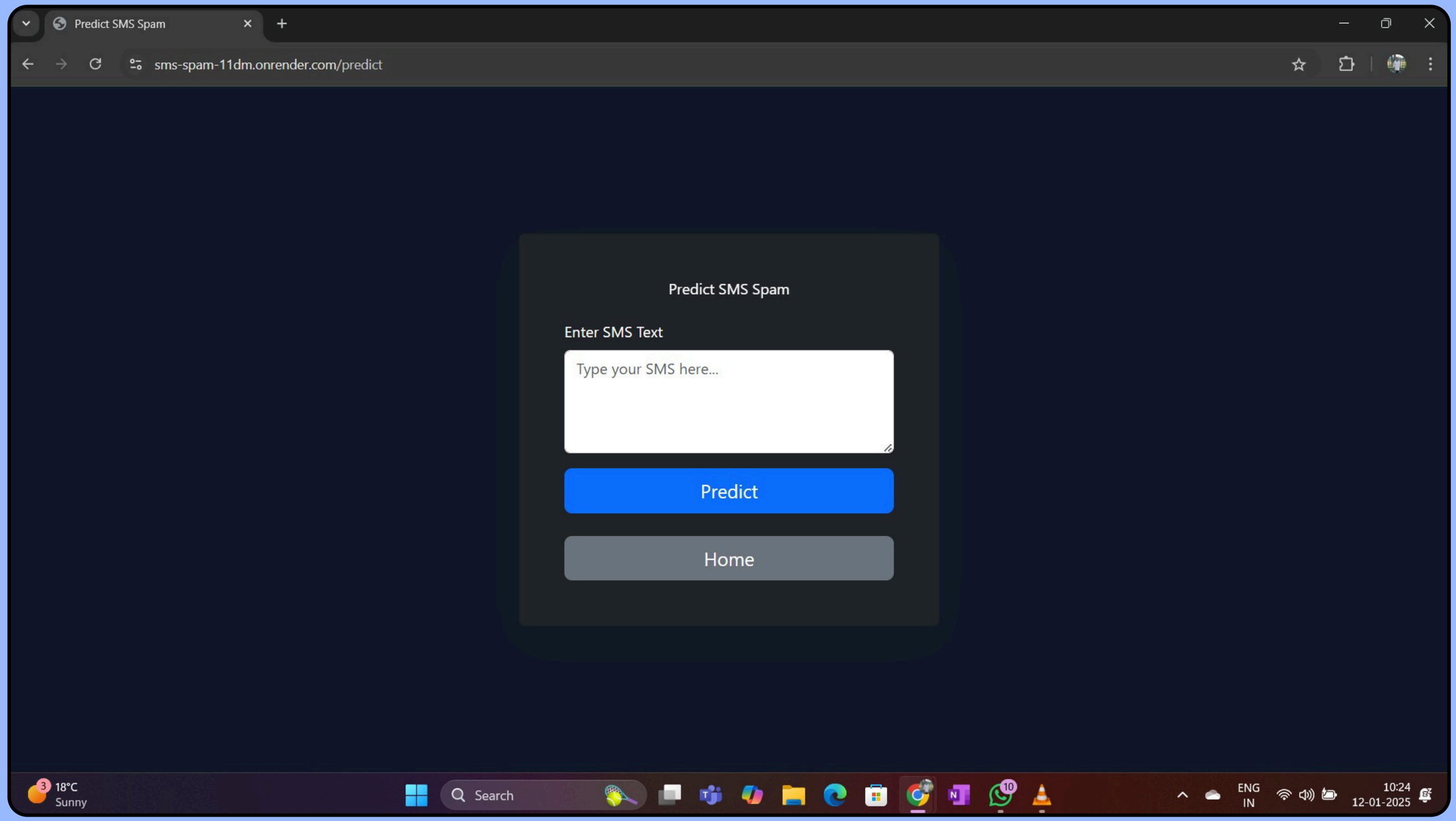
Synthetic Data Generator

Class Balancer



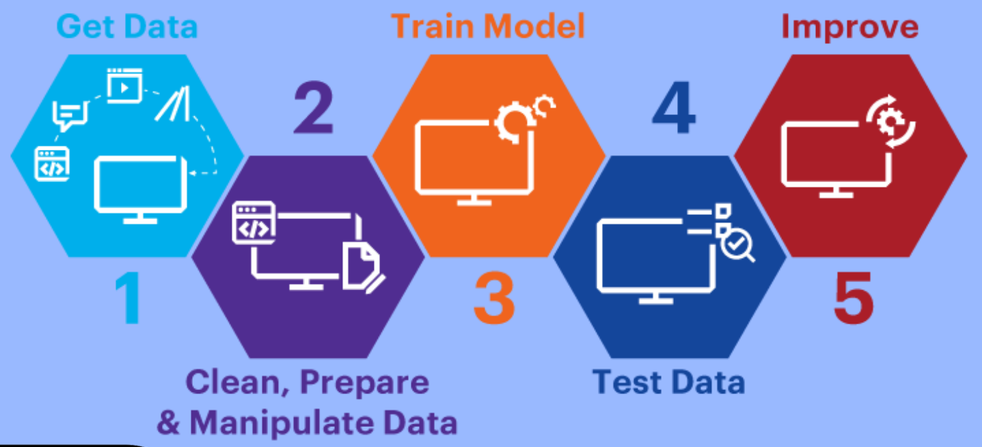
OUTPUT

USER INTERFACE



EVOLUTION OF THE MODEL

PRELIMINARY TENSORFLOW MODEL WITH CLASS IMBALANCING



CLASS BALANCED SIMPLE NEURAL NETWORK MODEL

CLASS IMBALANCED RESISTANT BIDIRECTIONAL LSTM MODEL

LINKS

- Github Repository Link: https://github.com/sortira/sms-spam
- Website Link: https://sms-spam-11dm.onrender.com/

FUTURE INSIGHTS

- A smarter approach to **vectorization of messages** can be implemented
- Other than LSTMs, much **simpler algorithms** like Random Forests Classifier or Support Vector Machine can be **experimented** with.
- The **small dataset** affects prediction accuracy. Future improvements should focus on **increasing data volume**. Given
- enough time, we would love to explore new ways of detecting spams because in a growing AI field, spams like this will go out of hand soon.