CYBERGUARD

Detection of inappropriate content on Social media indicating Cyberbullying

CSN 382 : Machine Learning

22114058 : Aditya Mundada
22114060 : Nayan Kakade
22114092 : Shubham Kr. Verma

Background

- Social media fosters **cyberbullying** due to its open and anonymous nature.
- Manual reporting not effective.
- Introduce automation through **Machine Learning** and **Deep Learning** models.

Dataset Description

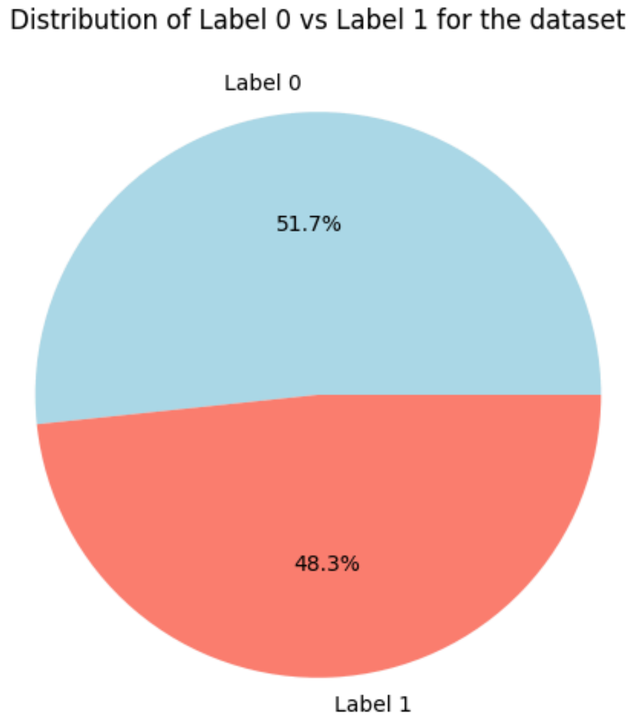
Shape : (93552 , 2)

Columns : Cleaned Text | Label

1 = OFFENSIVE

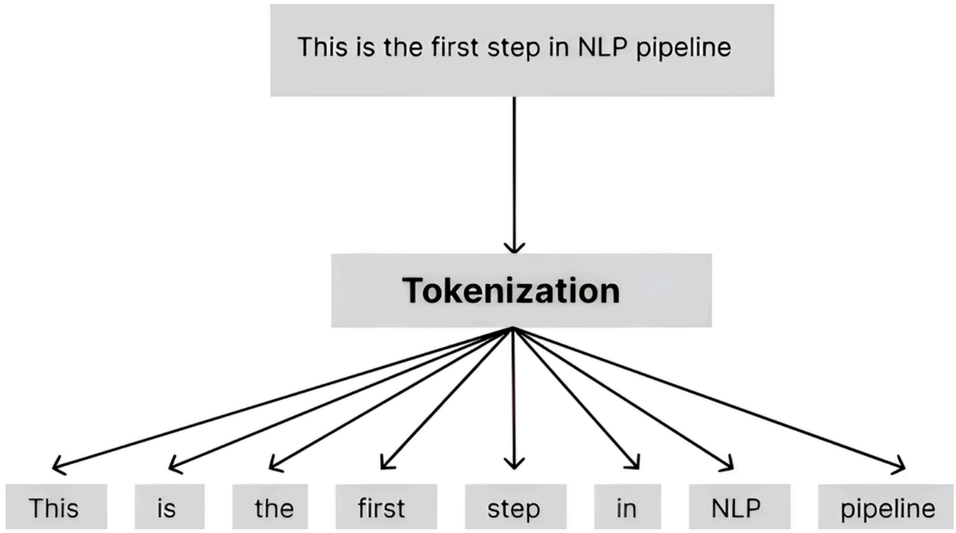
0 = NEUTRAL

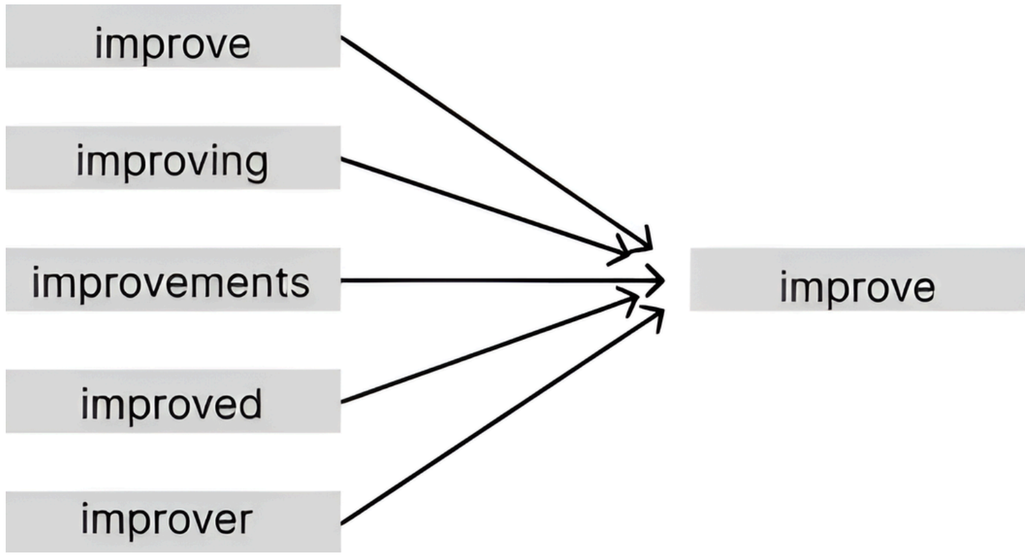
Data Features : Social Media Posts
(includes posts in Hinglish)



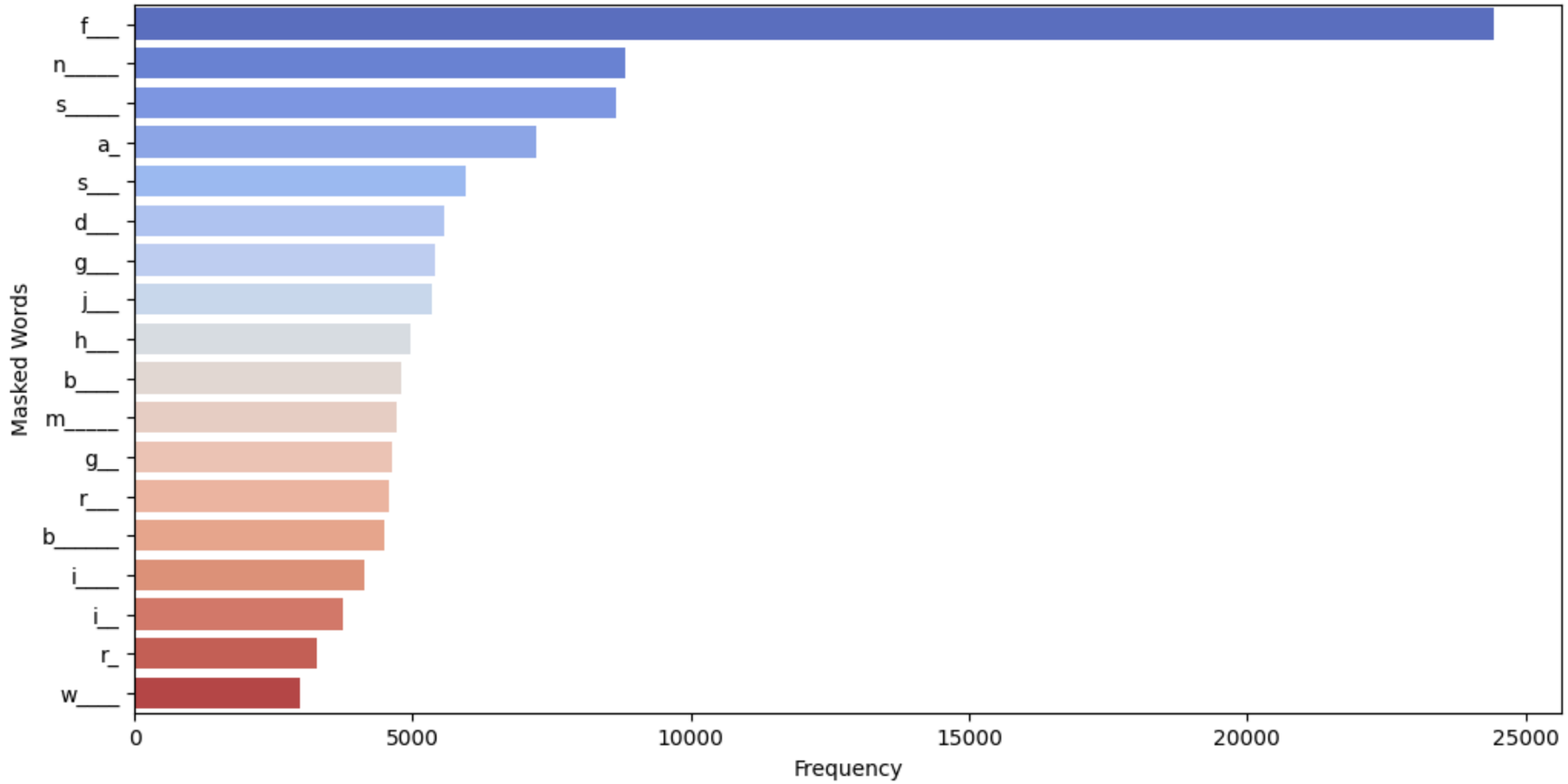
Cleaning & Preprocessing

- Lowercasing
- Removing special characters & punctuation
- Removing stop words
- Removing URLs, hashtags & mentions
- Expanding Contradictions
- Tokenization
- Lemmatization
- Vectorization



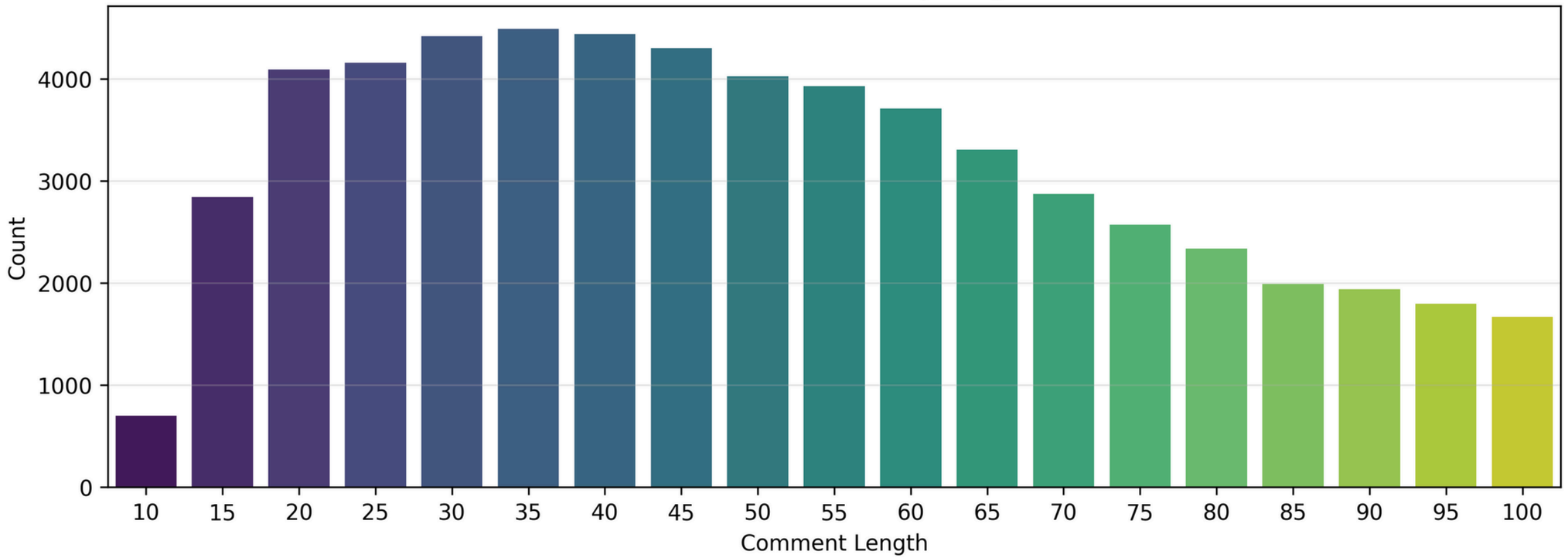


Exploratory Data Analysis



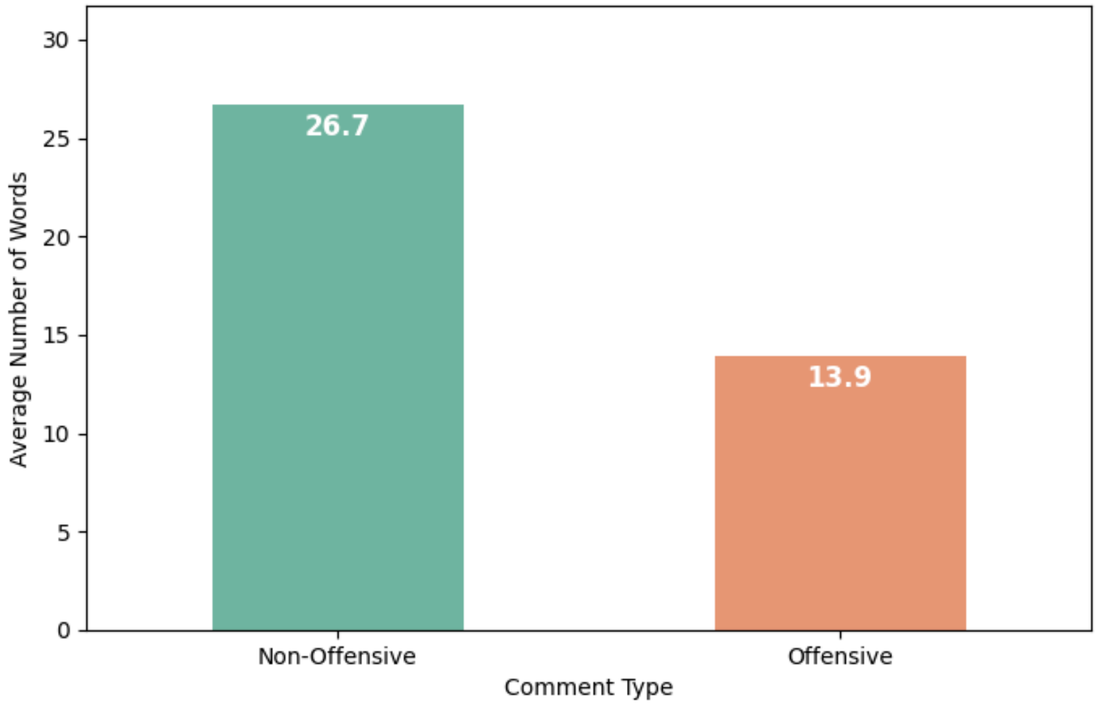
MOST COMMON WORDS

Exploratory Data Analysis



COUNT OF COMMENTS BY LENGTH

Exploratory Data Analysis



AVERAGE COMMENT LENGTH

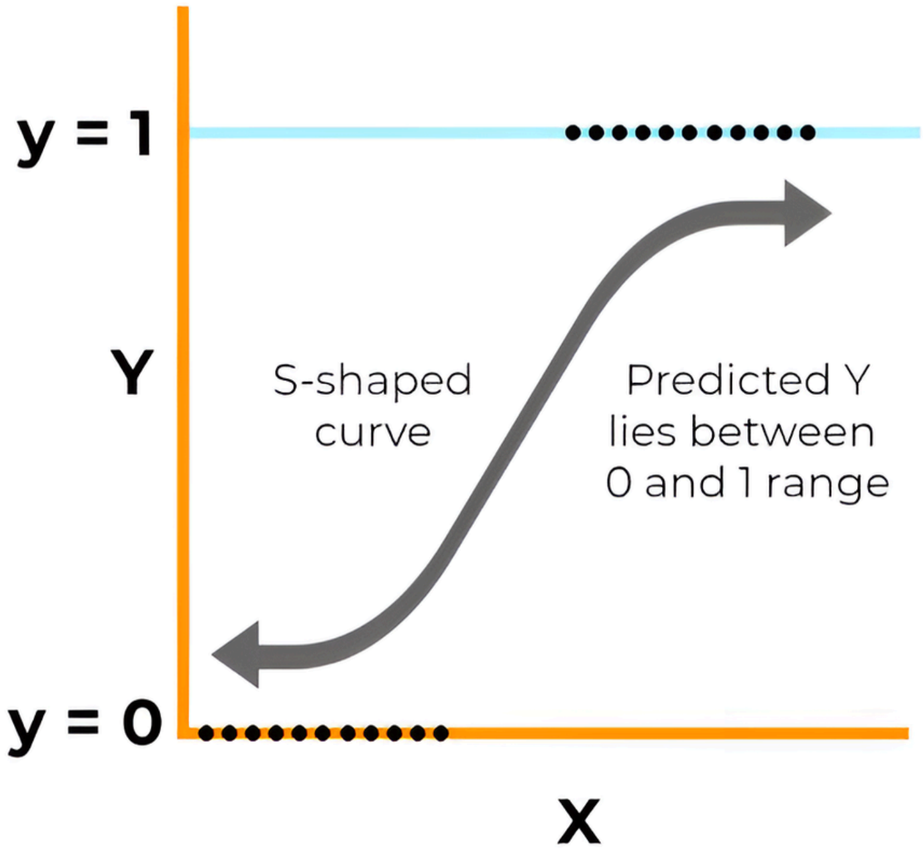
Traditional Models : Logistic Regression

Estimates the probability of an outcome, event or observation

$$f(z) = \frac{1}{1 + e^{-z}} \qquad 0 \leq f(z) \leq 1$$

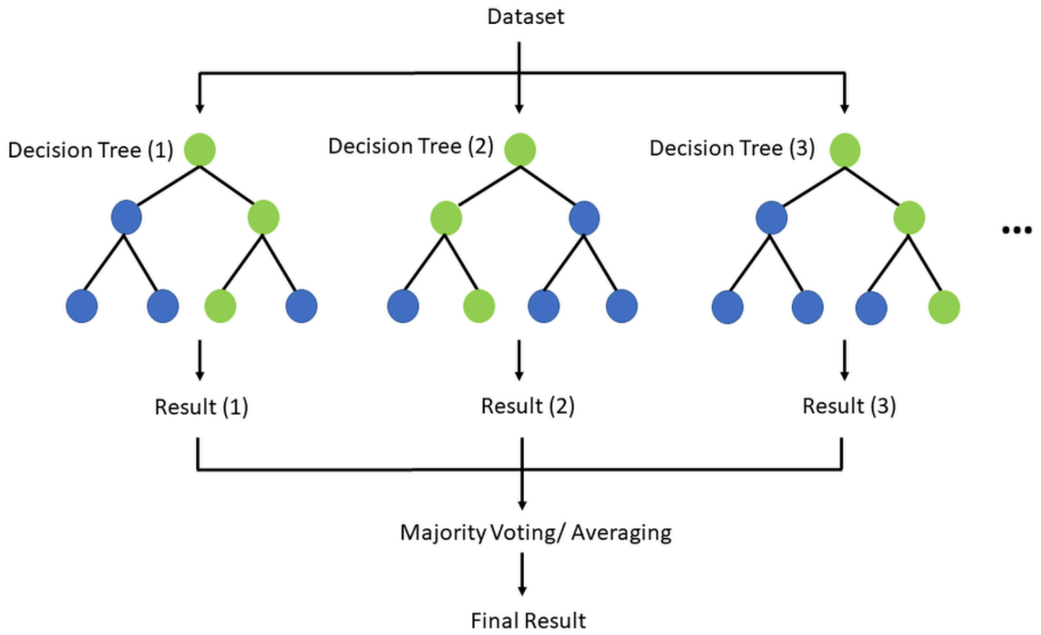
$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

β_0 is the **intercept** and
$\beta_1, \beta_2, \ldots, \beta_k$ are slopes against
independent variables $\mathrm{x}_1, x_2, x_3, \ldots, x_k$



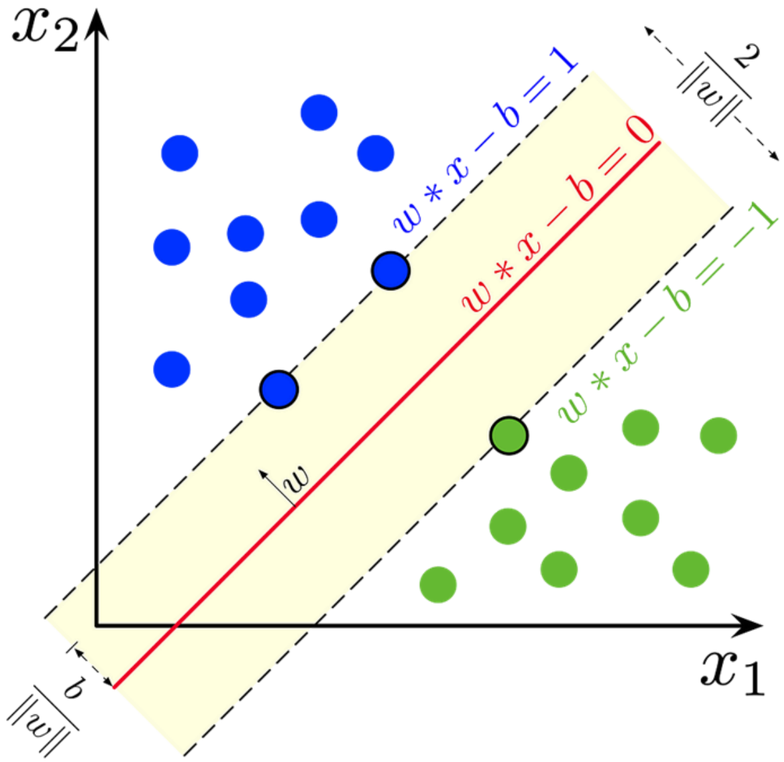
Traditional Models : Random Forests

- Ensemble technique to combine multiple decision trees.
- Trained on random sample & random subset of features.
- Prediction by majority voting among trees for classification



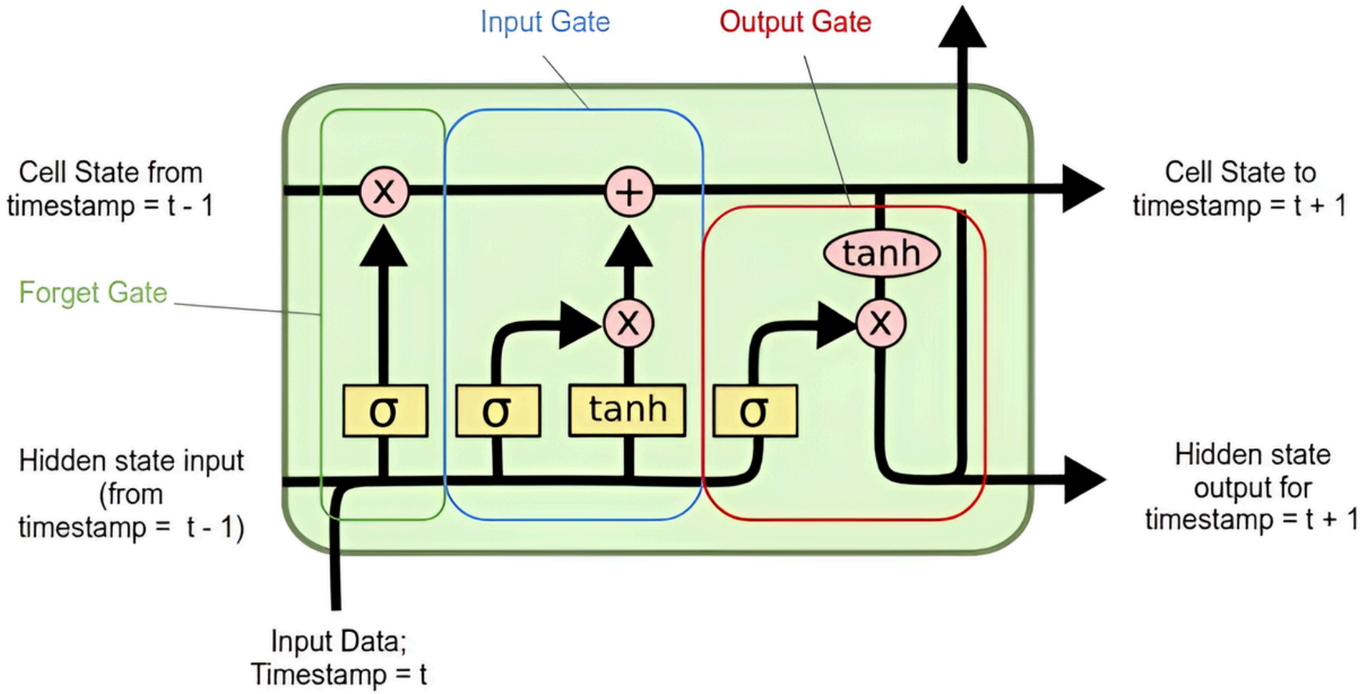
Traditional Models : Support Vector Classifier

- Constructs a hyperplane or set of hyperplanes in a high-dimensional space that separates the different classes
- Find best margin that separates the classes.



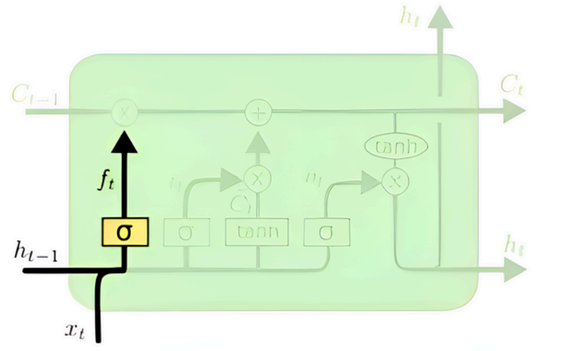
Long Short-Term Memory Network (LSTM)

- An LSTM cell maintains a cell state that runs through the entire sequence chain with only minor linear interactions.
- Capable of learning long-term dependencies.



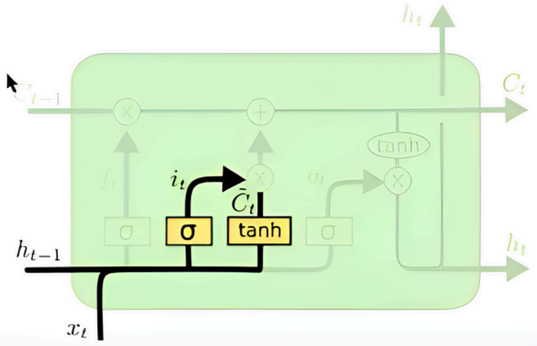
Long Short-Term Memory Network (LSTM)

Forget gate: controls what is kept vs what is forgotten, from previous cell state



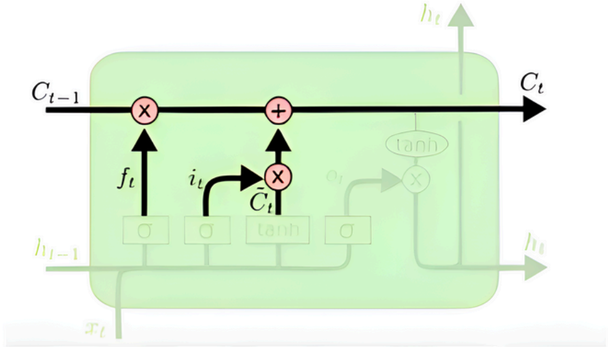
$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

Input gate: decides what information to throw away from the cell state
Cell content: new content to be written to cell



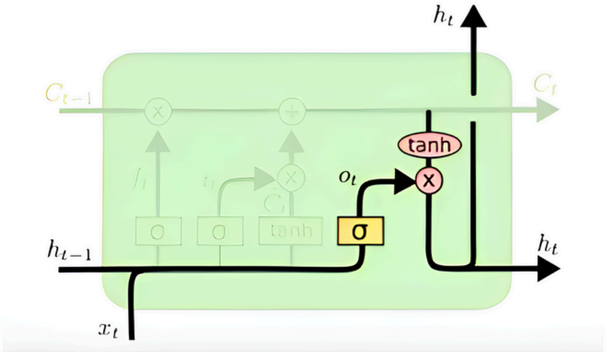
$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Output gate: controls what parts of cell are output to hidden state
Hidden state: read ("output") some content from cell



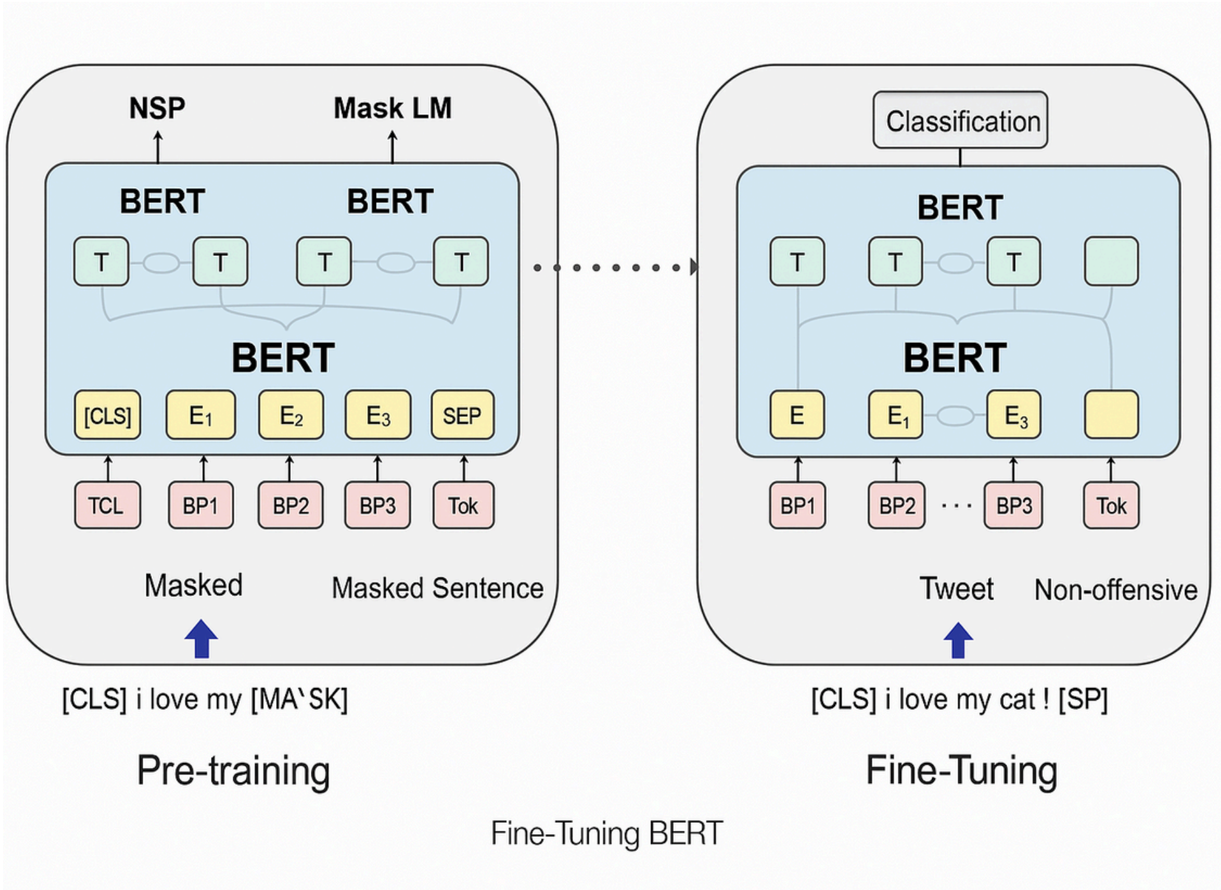
$$o_t = \sigma\left(W_o \; [h_{t-1}, x_t] \; + \; b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

BERT

- Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based model introduced by Google.
- It uses **Attention Mechanism** in order to grasp context.
- It looks at context in both **forward** and **backward** directions and makes **positional embeddings**.

We have **fine tuned** BERT for our classification task.

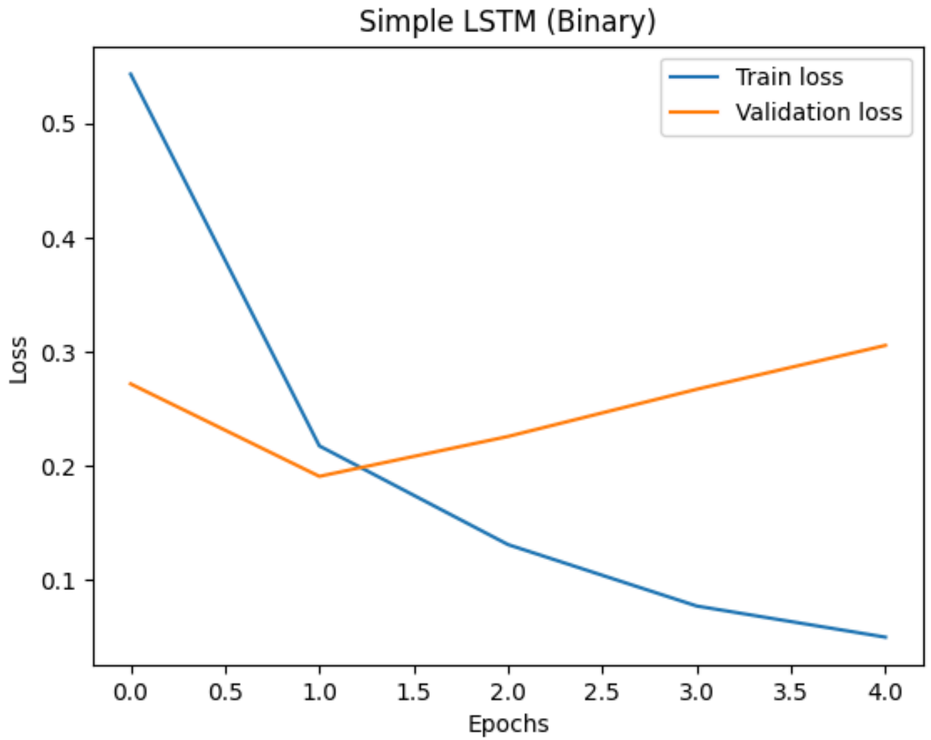


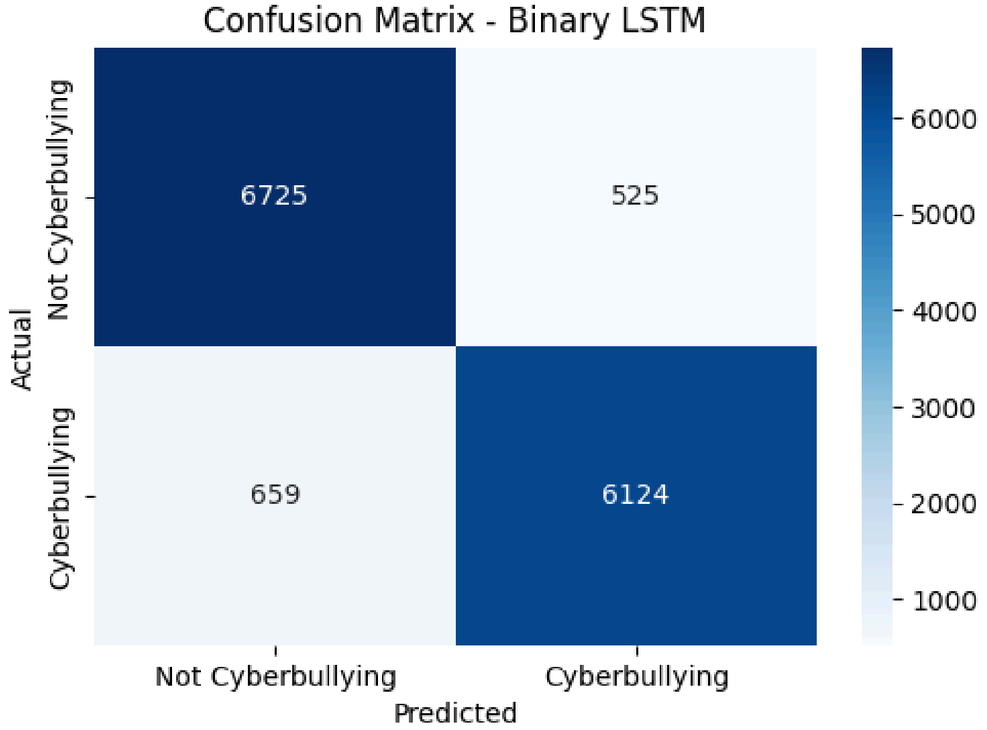
Fine-Tuning BERT

RESULTS : Traditional Models

Model	Class	Accuracy	Precision	Recall	F1-Score
Logistic Regression	Absent	0.8554	0.82	0.93	0.87
	Present		0.91	0.78	0.84
Random Forest	Absent	0.8619	0.82	0.94	0.88
	Present		0.92	0.78	0.85
Support Vector Classifier	Absent	0.8606	0.82	0.94	0.87
	Present		0.92	0.78	0.84

RESULTS : LSTM & BERT





RESULTS : LSTM & BERT

Model	Class	Accuracy	Precision	Recall	F1-Score
LSTM	Absent	0.916	0.91	0.93	0.92
	Present		0.92	0.9	0.91
BERT	Absent	0.918	0.9	0.94	0.92
	Present		0.94	0.89	0.91

CONCLUSION

- The source code of CyberGuard is available at **github.**
- Deep learning techniques like **LSTM** and **BERT** perform better in classification of tweets based on sentiment: they can **capture context efficiently** while traditional models cannot.
- Leveraging BERT helped us understand how pre-trained models can be used for a wide range of different tasks.
- We were able to achieve good metrics on dataset that had **both English** and **Hinglish** samples, thus we solved the problem in a more **Indian context.**

References

- S. Shahane, Cyberbullying Dataset, Kaggle. Available at: https://www.kaggle.com/datasets/saurabhshahane/cyberbullying-dataset.
- Karan Shah, Chaitaniya Phadtare, and Keval Rajpara, “Cyber-Bullying Detection in Hinglish Languages Using Machine Learning,” International Journal of Engineering Research & Technology (IJERT), vol. 11, no. 5, pp. 318–323, May 2022. Available: https://www.ijert.org/cyber-bullying-detection-in-hinglish-languages-using-machine-learning
- Manish Joshi, Dhirendra Pandey, Vandana Pandey, and Mohd Waris Khan, “A fusion framework for Hinglish cyberbullying detection using mBERT and FastText,” Inter-national Journal of Engineering in Computer Science, vol. 7, no. 1, pp. 7–14, 2025. Available: https://www.computersciencejournals.com/ijecs/article/view/149/7-1-3

THANK YOU!