TOXICITY Detection in AI Art

Team members: Xun Lei, Chenxi Guo, Jiayi Peng, Yiran Tao

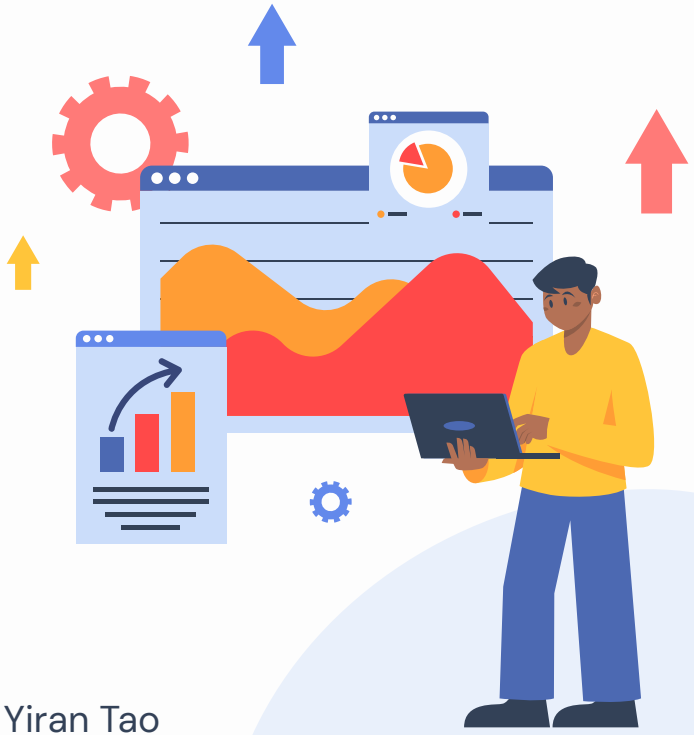
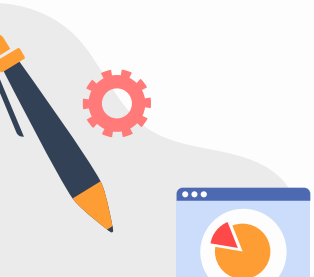


Table of contents



01

INTRODUCION

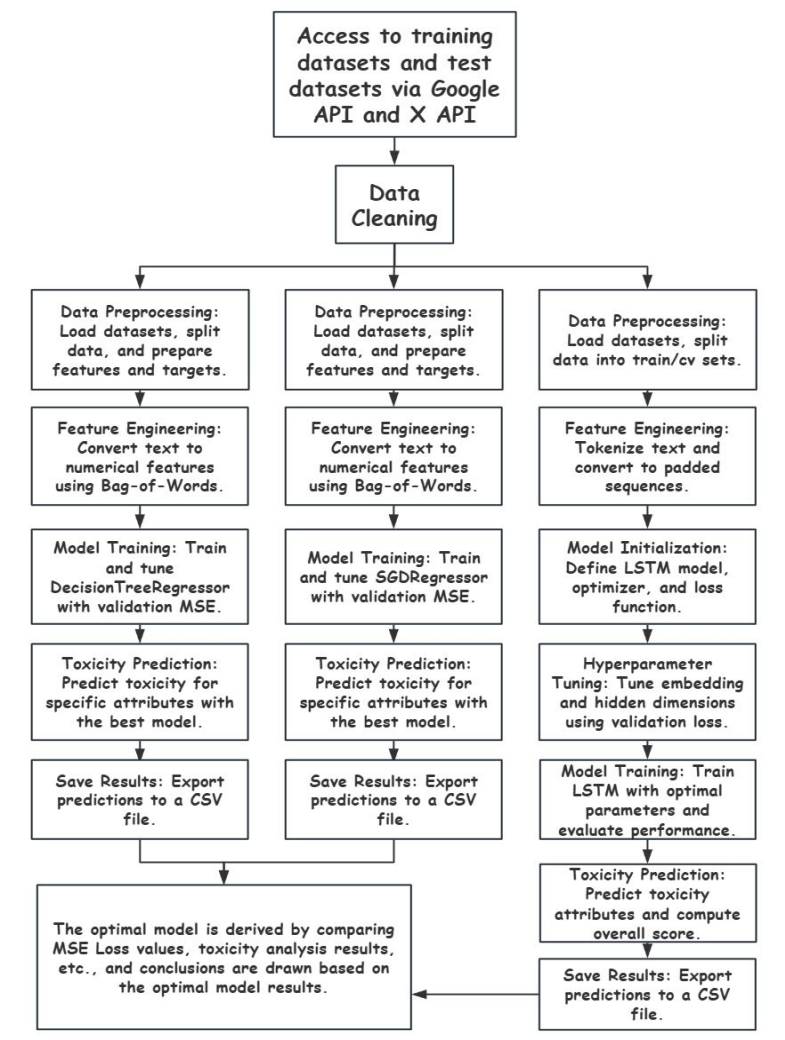
Introduction

AI art, has sparked intense debates in the digital age, with some praising its innovation and others questioning its impact on traditional art.

As more people talk about AI art online, opinions range from admiration to doubt. But sometimes, toxic comments appear, making these discussions less open and welcoming.

Our project aims to explore how people feel about AI-generated art by analyzing online commons. Through this, we hope to gain a deeper understanding of public perceptions and the emotions driving these discussions for AI art.

Flow Chart


Access to training datasets and test datasets via Google API and X API
Data Cleaning
Data Preprocessing: Load datasets, split data, and prepare features and targets.
Feature Engineering: Convert text to numerical features using Bag-of-Words.
Model Training: Train and tune DecisionTreeRegressor with validation MSE.
Toxicity Prediction: Predict toxicity for specific attributes with the best model.
Save Results: Export predictions to a CSV file.
Data Preprocessing: Load datasets, split data, and prepare features and targets.
Feature Engineering: Convert text to numerical features using Bag-of-Words.
Model Training: Train and tune SGDRegressor with validation MSE.
Toxicity Prediction: Predict toxicity for specific attributes with the best model.
Save Results: Export predictions to a CSV file.
Data Preprocessing: Load datasets, split data into train/cv sets.
Feature Engineering: Tokenize text and convert to padded sequences.
Model Initialization: Define LSTM model, optimizer, and loss function.
Hyperparameter Tuning: Tune embedding and hidden dimensions using validation loss.
Model Training: Train LSTM with optimal parameters and evaluate performance.
Toxicity Prediction: Predict toxicity attributes and compute overall score.
Save Results: Export predictions to a CSV file.
The optimal model is derived by comparing MSE Loss values, toxicity analysis results, etc., and conclusions are drawn based on the optimal model results.

Data Processes

Data Collection

- **Using the X (formerly Twitter) API** to collect public tweets relate AI art
- Retrieving **tweet text** from the collected data.
- **Storing** in a structured format for analysis.

Data Cleaning

- Remove special characters (e.g., punctuation, emojis)
- Remove stop words to reduce noise
- Stemming

Before-and-After Example:

- **Before:** "I love AI art! It's so amazing 😍 #AI #Art"
- **After:** "love AI art amaz"

02 Models

Decision Tree
Regression

Bag of Words (BoW)

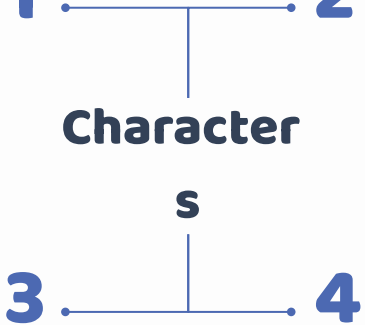
A simple **text representation method** that converts text into **a vector of word frequencies.**

Vocabulary-based

Represents text using a set of unique words.

Order Ignorance

Ignores word order in the text.



Frequency Count

Counts the occurrences of each word in the text.

High Dimensionality

Results in large vectors for large vocabularies.

Introduction of Model

A Decision Tree is an algorithm that splits data into subsets based on feature values, **creating a tree-like model of decisions** and their possible consequences for regression tasks.

Splitting the Data

The tree divides the dataset into subsets based on features, aiming to minimize the prediction error (e.g., MSE).

Choosing the Best Split

The tree selects the best feature and split point that minimizes the prediction error (e.g., MSE).

Tree Growth

The tree recursively splits the data at each node until a stopping criterion is reached.

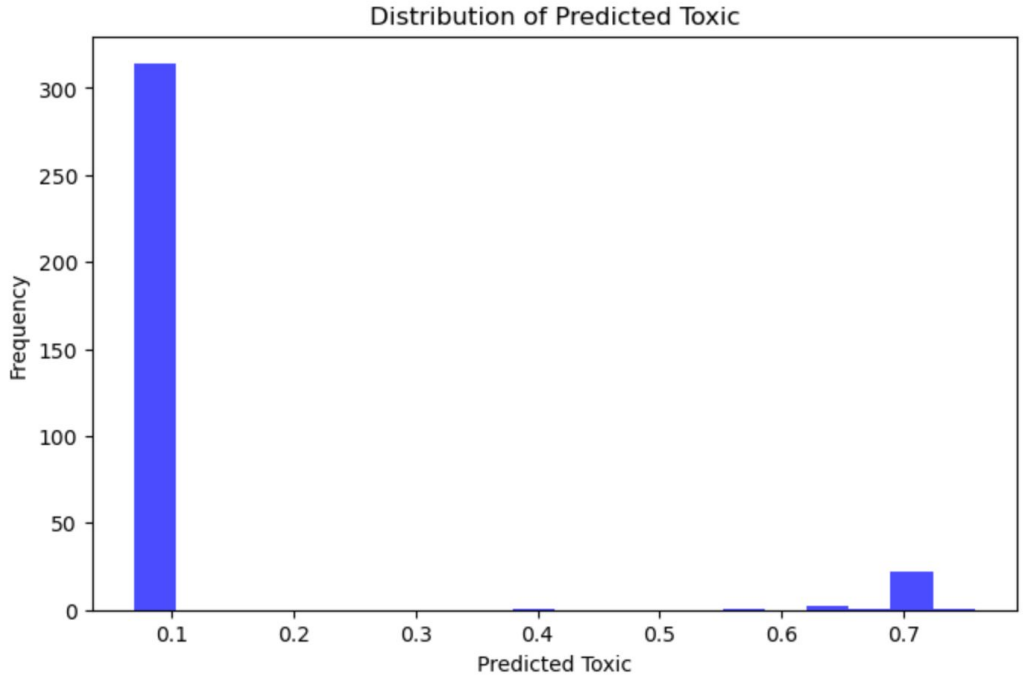
Prediction

After training, the tree predicts new data by averaging the target values (toxicity scores) in the leaf node.

Results

Type	Toxicity	Obscene	Identity Attack	Insult	threat
Range	[0.069632 – 0.758003]	[0.009150 – 0.648587]	[0.011833 – 0.086471]	[0.050236 – 0.714464]	[0.005053 – 0.088155]
Mean	0.11970906	0.056506	0.012269	0.07125	0.011625
Std	0.168899	0.163939	0.005699	0.097704	0.021393
MSE	0.0253	0.0027	0.0037	0.0185	0.0022

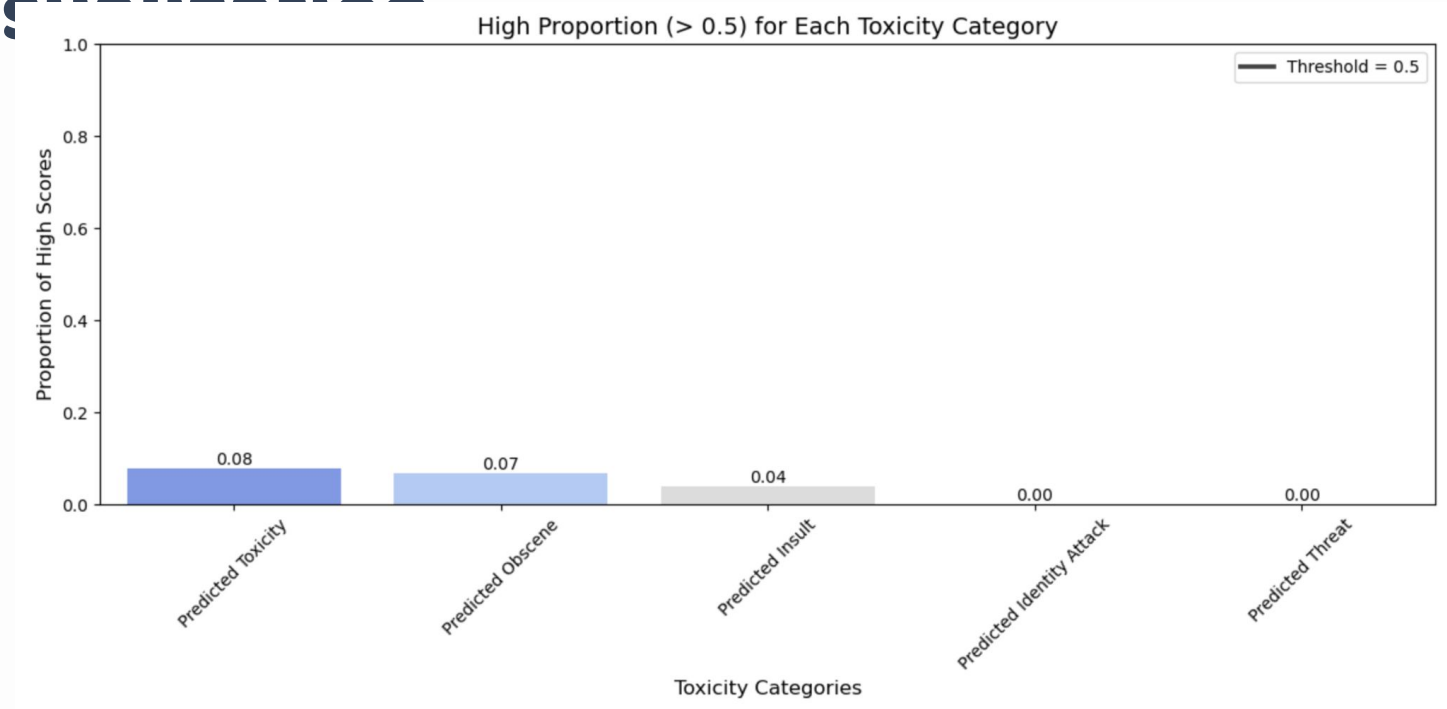
Toxicity Visualization



Although ai art is controversial, the histogram shows that most comments in AI art discussions are non-toxic or mildly toxic, with only a few comments exhibiting higher toxicity.We can also see that some and some of the comments focus on the **very high toxicity range**, which indicates that **a few people are very aggressive on this topic.**

Toxicity Visualization

Toxicity Visualization



03 Models

SGD Regression

Introduction of Model

Stochastic Gradient Descent (SGD) Regression is a computationally efficient optimization algorithm that **minimizes loss functions** by iteratively **updating model parameters.**

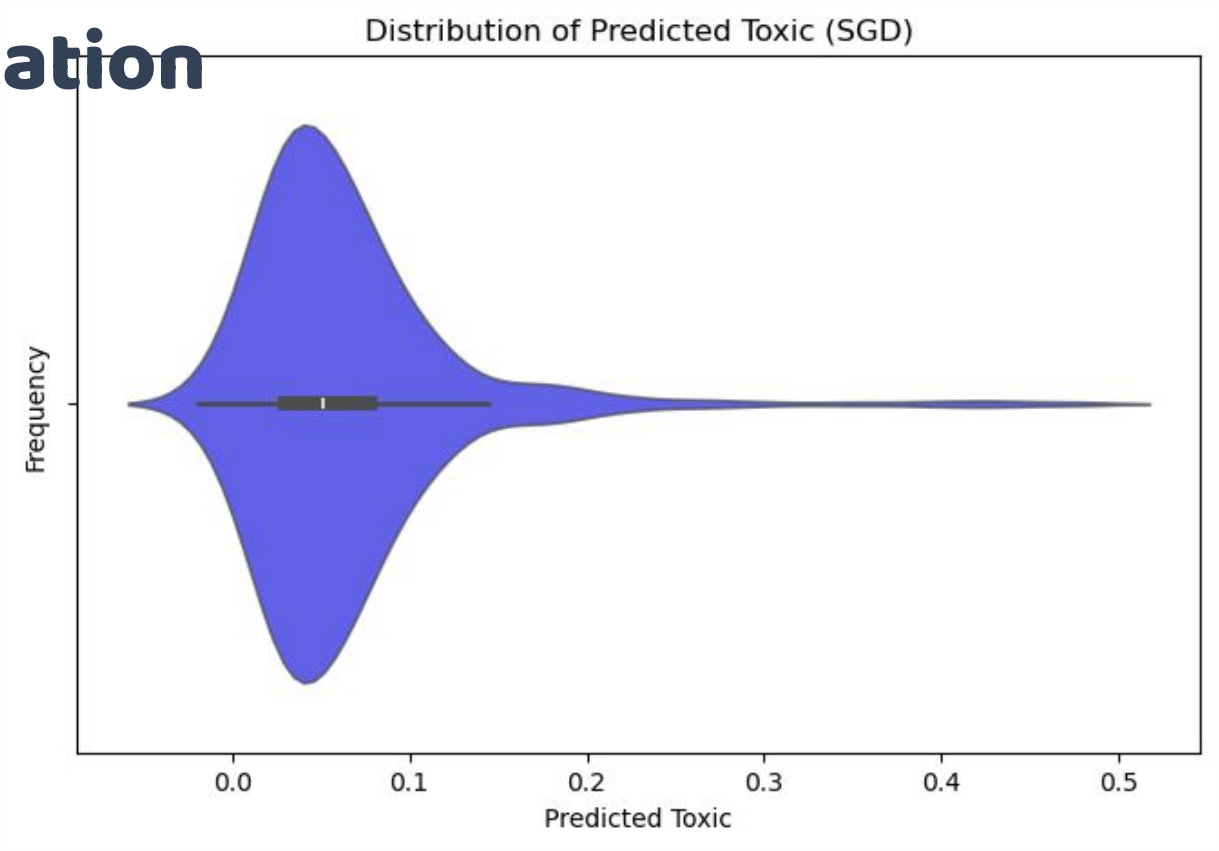
How SGD Regression Works:

- **initialize** the model **parameters (weights)** randomly
- compute the **gradient of the loss function** for each data point
- **update the parameters** incrementally using the learning rate
- **repeat** the process until the model converges to an **optimal solution**

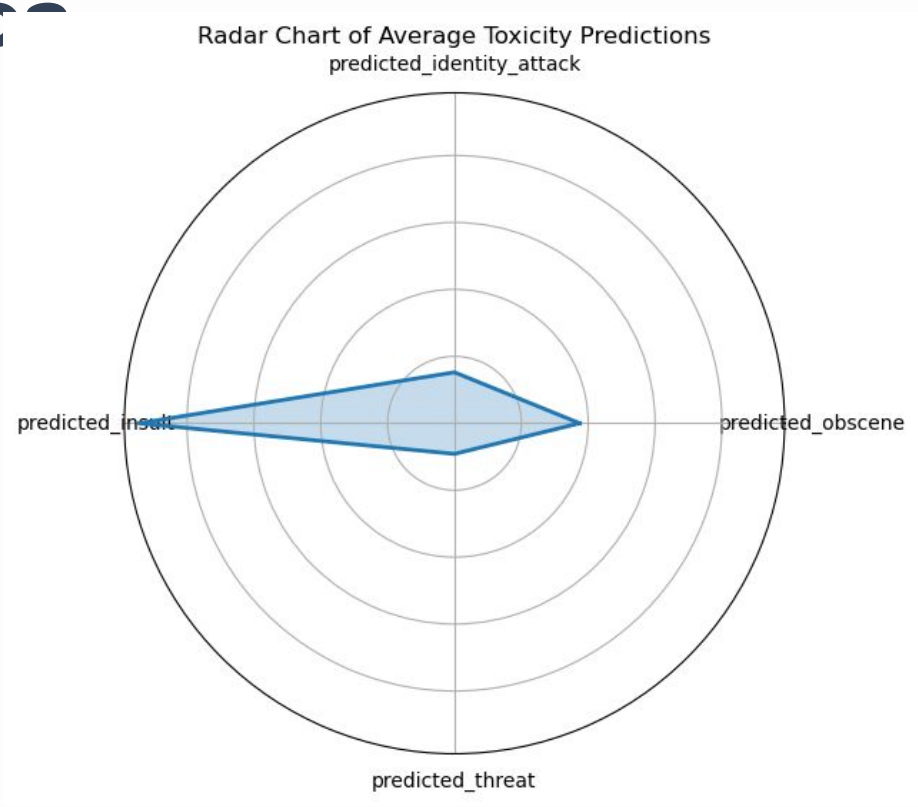
Results

Type	Toxicity	Obscene	Identity Attack	Insult	Threat
Range	[-0.018829, 0.477470]	[-0.007641, 0.198449]	[-0.005281, 0.142287]	[-0.022377, 0.451761]	[-0.008939, 0.074686]
Mean	0.065519	0.018723	0.007620	0.047238	0.004573
Std	0.063688	0.028438	0.010359	0.052428	0.011291
MSE	0.0238370	0.003693	0.003691	0.017609	0.002278

Toxicity Visualization



Toxicity Visualizatic

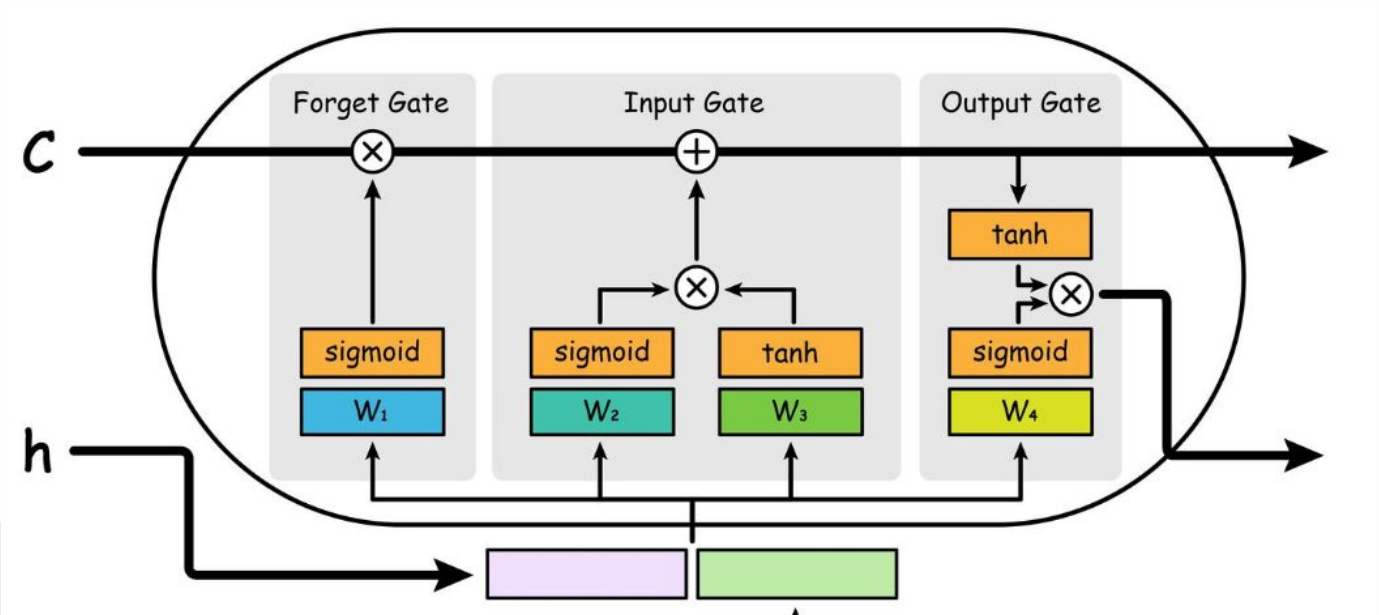


04 Models

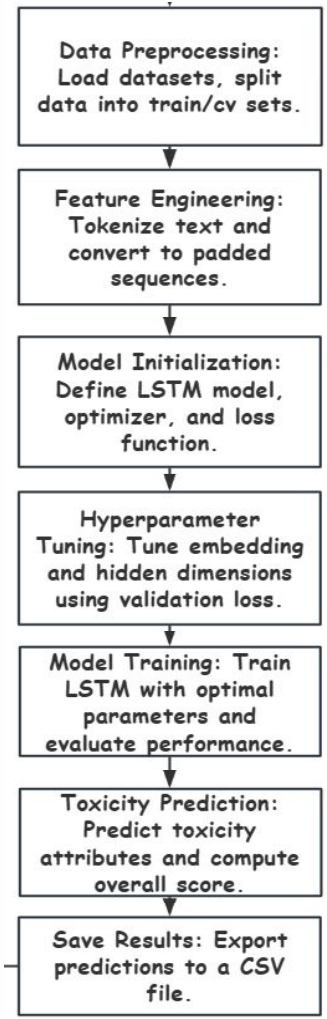
LONG SHORT TERM MEMORY

Introduction of Model

LSTM is a deep learning model commonly used to process sequence data. Compared with traditional **RNN** (Recurrent Neural Network), LSTM introduces three gates (**input gate**, **forget gate**, **output gate**, as shown in the figure below) and a cell state, which are mechanisms that allow LSTM to better handle long-term dependencies in sequences.

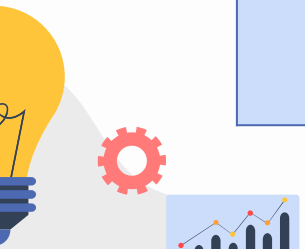


Concrete Implementation

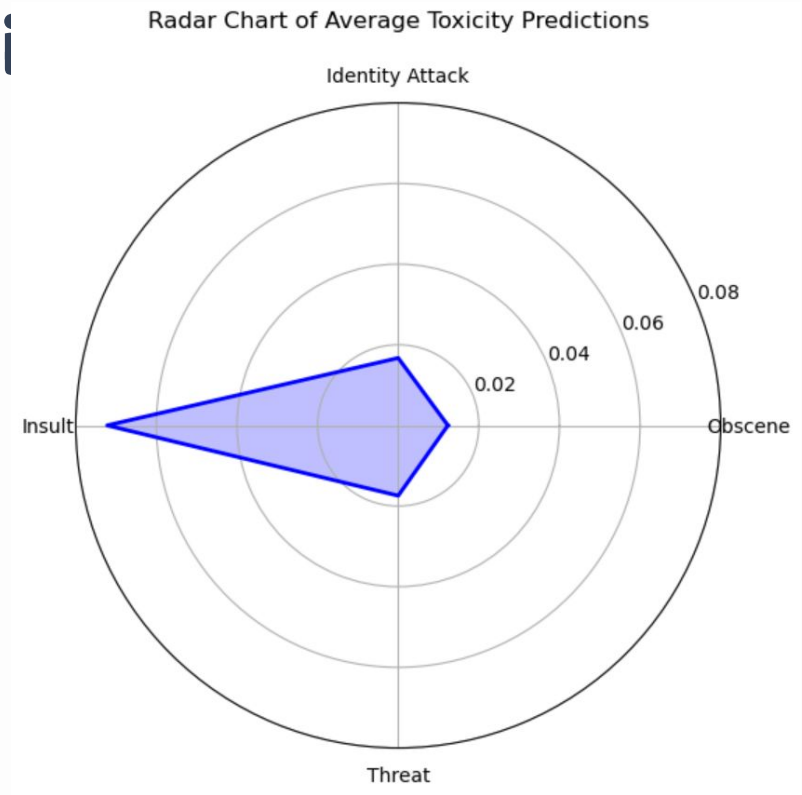


Results

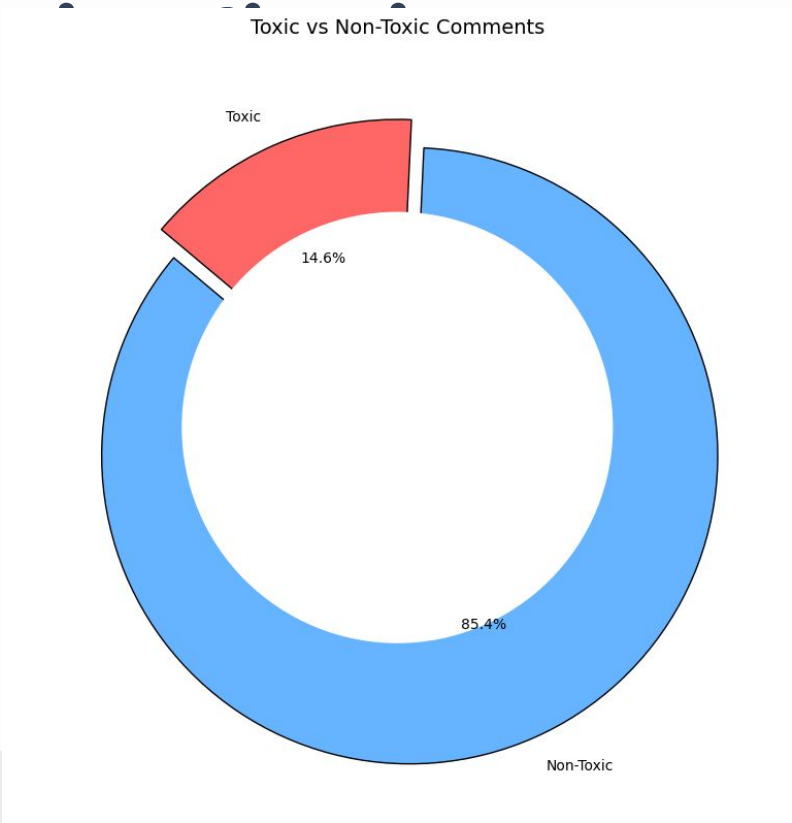
Type	Toxicity	Obscene	Identity Attack	Insult	threat
Range	[0.069317, 0.499014]	[0.005235, 0.158027]	[0.007971, 0.019993]	[0.039182, 0.197283]	[0.005193, 0.022081]
Mean	0.118710	0.012328	0.016725	0.072225	0.017430
Std	0.107424	0.138198	0.003005	0.039525	0.010722
MSE	0.0328	0.0054	0.0044	0.0240	0.0019

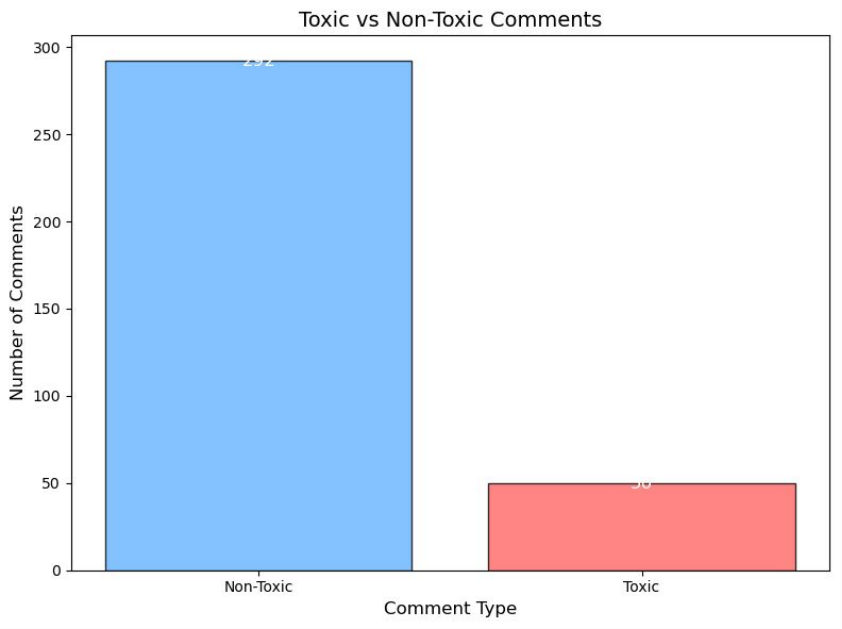


Toxicity Visualizati



Toxicity

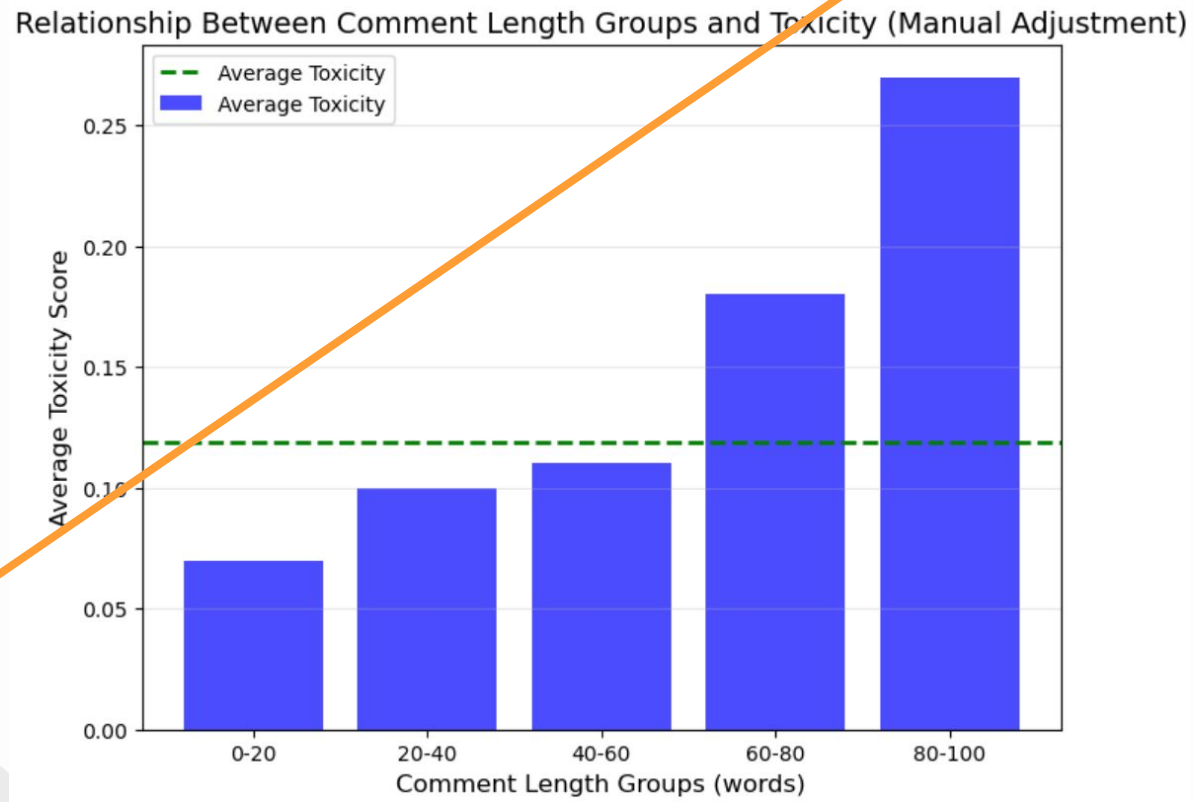

Toxic vs Non-Toxic Comments
Toxic
14.6%
85.4%
Non-Toxic


Toxic vs Non-Toxic Comments
Number of Comments
300
250
200
150
100
50
0
Non-Toxic
Toxic
Comment Type

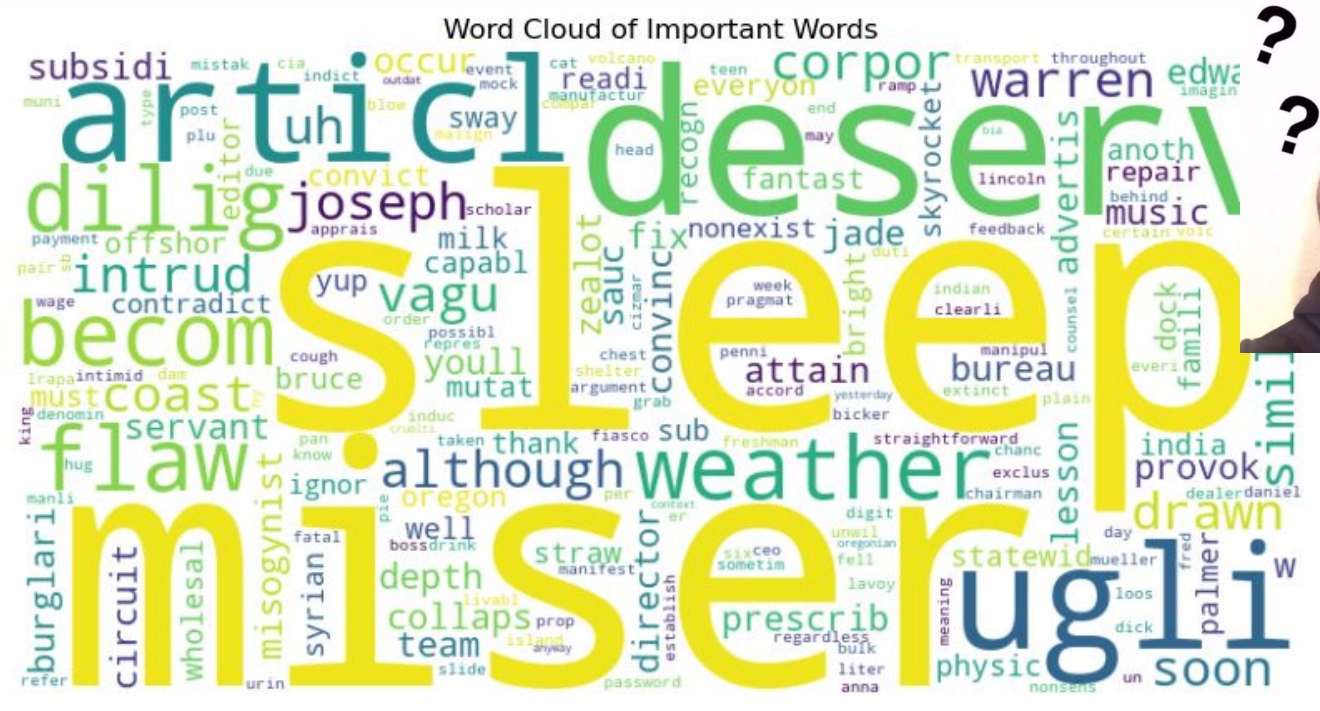
LSTM vs BoW-Based Models

Aspect	LSTM	Bag-of-Words + SGD/Decision Tree
Long Comments	Excels at capturing complex linguistic features.	Lacks context handling, poor performance.
Short Comments	May overfit simple scenarios, still effective.	Efficient, accurately detects explicit toxicity.
Large Data	Adapts well, strong learning capacity.	Handles large data but lacks semantic insights.
Small Data	Prone to underfitting without sufficient data.	Converges quickly, friendly to small datasets.
Deployment	High complexity, hardware-intensive.	Simple, fast deployment.

Toxicity Visualization



Toxicity Visualization （Error Demonstration Version ）



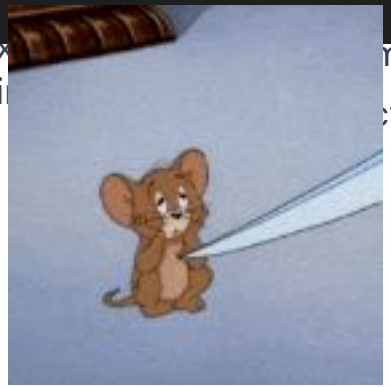
Why not BERT?

```
[17]  6.1s
...  Epoch 1:   0%|          | 1/1244 [00:04<1:43:17,  4.99s/it, loss=0.046]
```

✓ Ideal for complex
requiring deep li
understanding

much longer than

t time budget

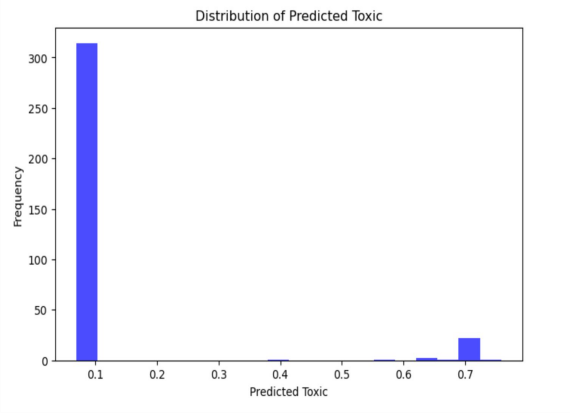


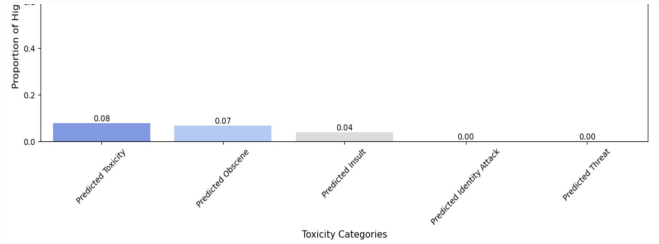
05 Conclusions

Conclusions

	DTR	SGD	LSTM
MSE Loss	0.0253	0.0238370	0.0328
Time-consuming (s)	4.12	5.43	5.94
Toxicity Interval	[0.069632, 0.758003]	[-0.018829, 0.477470]	[0.069317, 0.499014]

Conclusions





Conclusions

As technology advances and artificial intelligence integrates with art, **ethical disputes** are shifting from fear and rejection to **rationality and calm**. This shift reflects human resilience and highlights the growing importance of technological ethics in the modern era. Unlike the panic during the First Industrial Revolution, today's society is approaching change with increasing rationality, though toxicity persists, especially among skeptics. This reminds us that while technological progress drives innovation, it also challenges us to **balance technology with humanity** and **uphold social justice and dignity**. As technological change deepens, we must foster rational dialogue and work together to build an inclusive, understanding, and ethical society.

Thanks!

Q & A

References

[1] Maslej-Krešňáková, V.; Sarnovský, M.; Butka, P.; Machová, K. Comparison of Deep Learning Models and Various Text Pre-Processing Techniques for the Toxic Comments Classification. Appl. Sci. 2020, 10, 8631. https://doi.org/10.3390/app10238631

[2] N. Dughyala, S. Potluri, S. KJ and V. Pavithran, "Automating the Detection of Cyberstalking," 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2021, pp. 887-892, doi: 10.1109/ICESC51422.2021.9532858.

[3] Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. 2024. “HOT” ChatGPT: The Promise of ChatGPT in Detecting and Discriminating Hateful, Offensive, and Toxic Comments on Social Media. ACM Trans. Web 18, 2, Article 30 (May 2024), 36 pages. https://doi.org/10.1145/3643829