2023 3rd ASIAN CONFERENCE ON INNOVATION IN TECHNOLOGY (ASIANCON 2023)

Social Media Comment Classification and Crime Forecasting

Paper ID: 704

Ram Mohan Reddy, Divya Udayan J, Sai Ganesh, DVS Sravan, Vidyacharan

Department of Computer Science and Engineering

Amrita Vishwa Vidyapeetham, Amritapuri

Contents

- Introduction -- Social Media comments…

- Proposed Methodology

- Pipeline of the proposed methodology.

- Algorithm/Method Details

- Implementation details

- Experimental Results
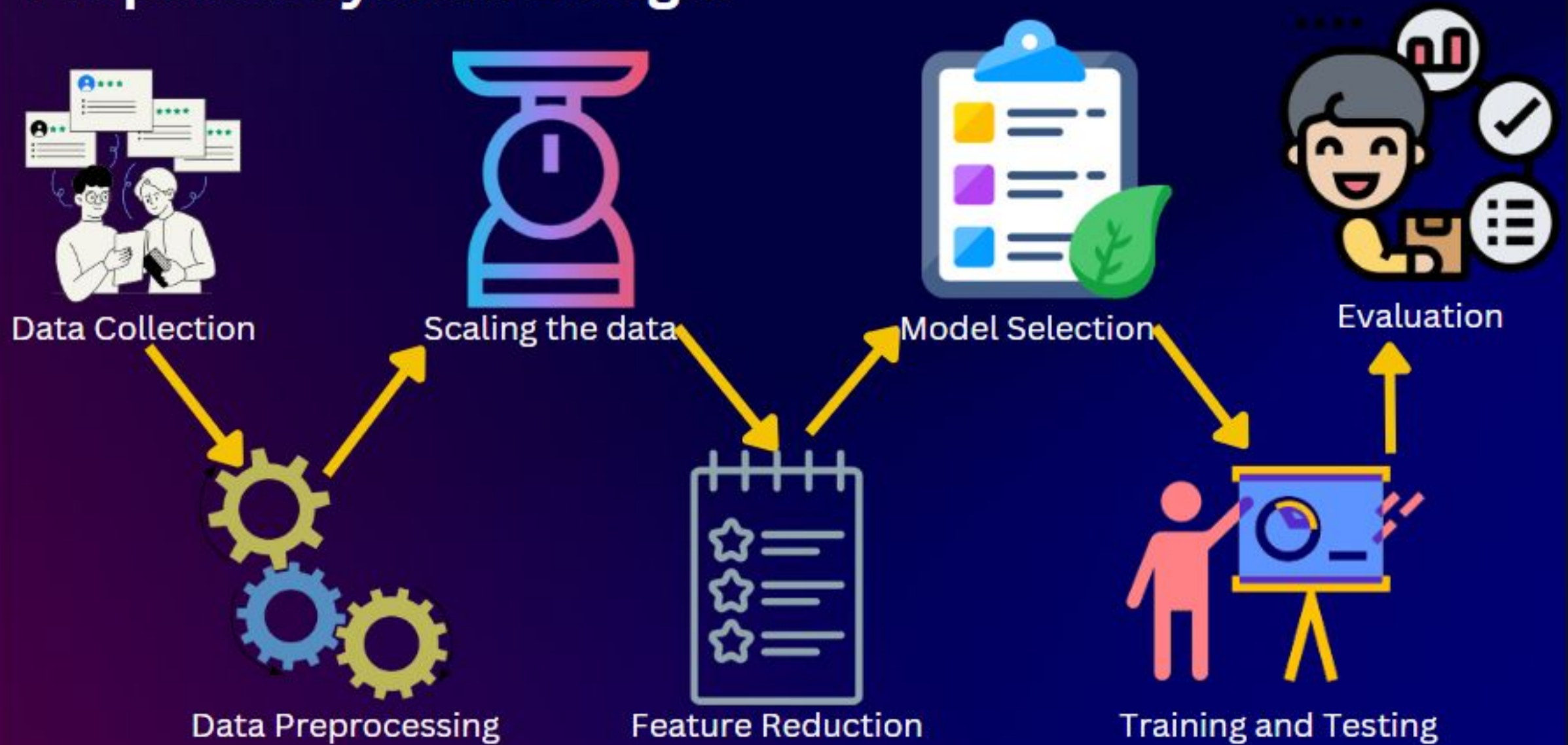
- Results and Discussion

Social media comments…

In social media applications there is no restriction to talk abusive and humiliate people using text. With this benefit the criminals like Drug dealers, Terrorists, and normal users use opportunity to make threats to other people.
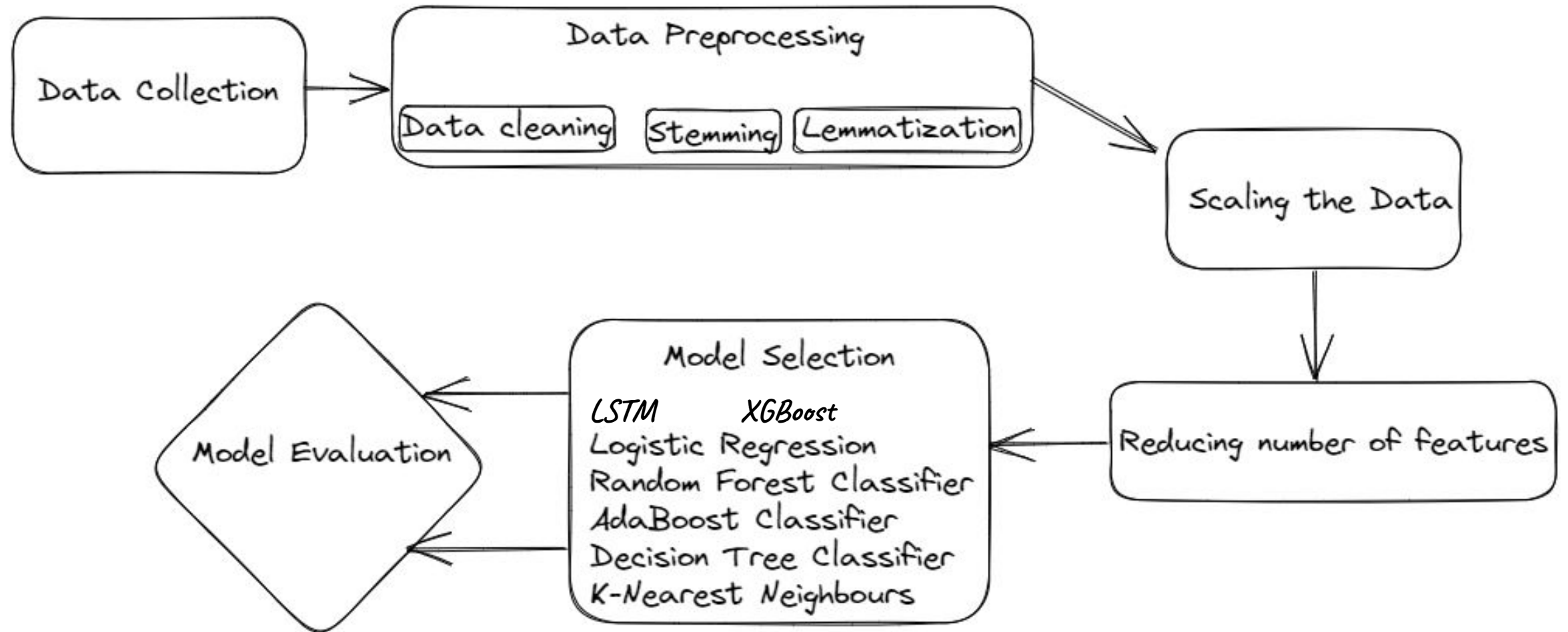
Criminals often take advantage of social media users, leading to a variety of illegal activities and making cybercrime a significant global concern.

- When a comment is about to post, that comment can be analysed and classified.

- Depending on the classification, we can take different actions. If it's a safe comment, it can be posted as-is. If it's potentially harmful, we might give a warning to the person posting it, asking them to reconsider.

- And if it's highly toxic, we might prevent it from being posted altogether to maintain a safe and respectful online environment.

Proposed System Design:
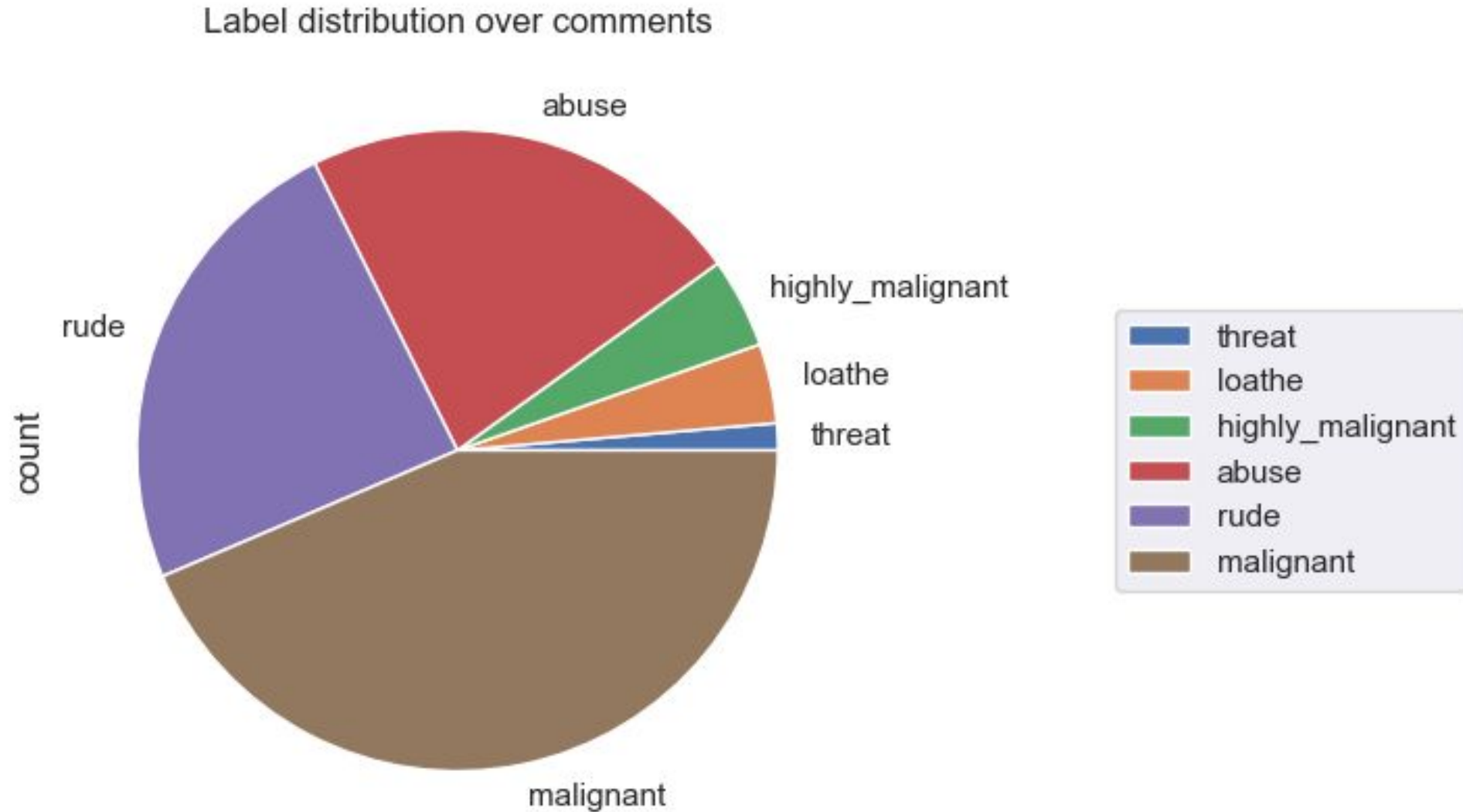
Data Collection

Scaling the data

Model Selection

Evaluation

Data Preprocessing

Feature Reduction

Training and Testing

Pipeline of the proposed methodology:

The Dataset…and the distribution:

Label distribution over comments

Algorithm/Model Details

Logistic Regression — 89.716%

Random Forest — 91.591%

AdaBoost — 90.489%

KNN — 90.752%

Decision Tree — 86.206%

XGBoost — 91.617%

LSTM — 99.425%
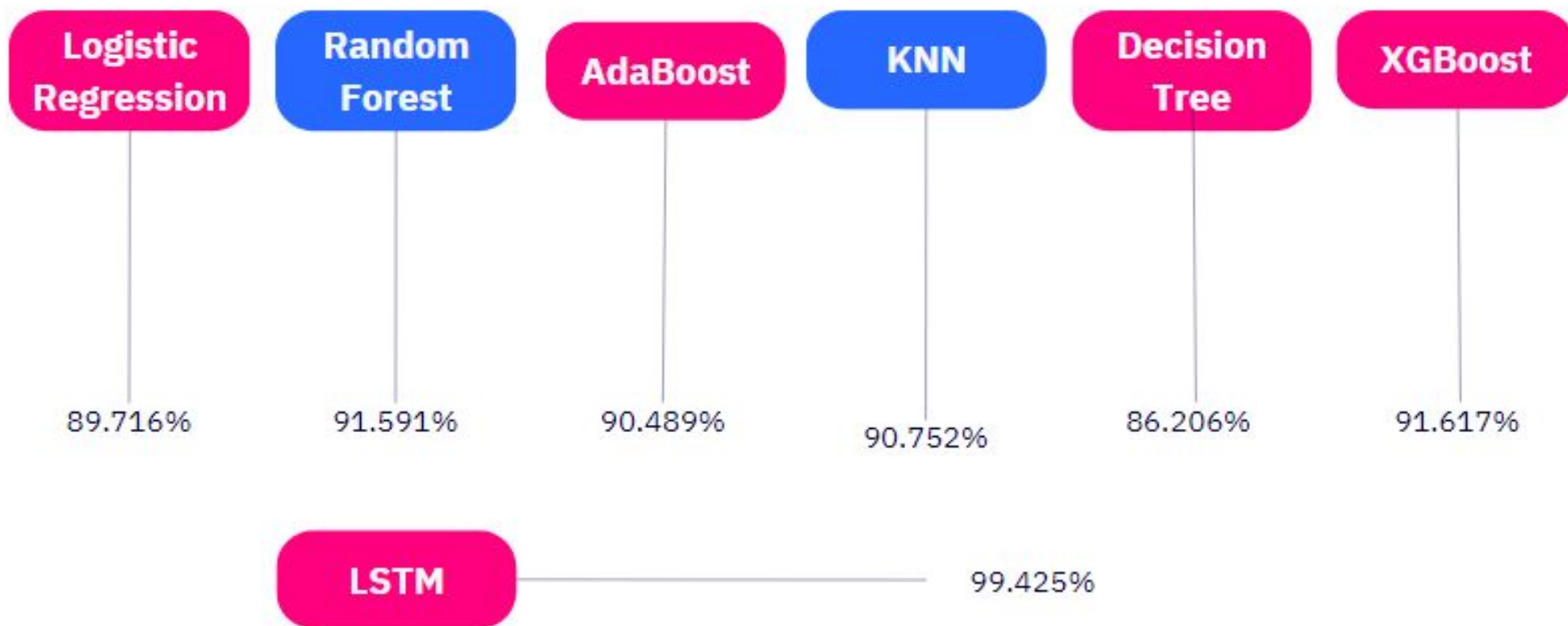
Implementation Details

- The idea is to use the training dataset to train the models with the words in the comments as predictor variables and to predict the probability of the toxicity level of a comment.
- Our project is focused on developing a series of machine learning and deep learning models. The goal is to use different ML and Deep Learning models to classify the comments and identify toxicity levels.

Logistic Regression:

   It predicts binary outcomes and models the relationship between predictor variables and the outcome.

Decision Tree:

   used for both classification and regression tasks. It creates a tree-like model of decisions and their possible consequences, based on the features of the data.

Random Forest Classifier:

   A popular ensemble learning algorithm that combines multiple decision trees to improve the accuracy and robustness of the model.

AdaBoost Classifier:

   Adaboost classifier is a popular boosting algorithm that combines multiple weak learners to create a strong classifier.

XGBoost Classifier:

It is an implementation of the gradient boosting decision tree algorithm. It uses a gradient descent algorithm to optimize the loss function by adding decision trees sequentially.

KNN Classifier:

It is a simple and popular machine learning algorithm for both classification and regression tasks.

LSTM(Long Short-Term Memory) :

LSTM is a type of recurrent neural network (RNN) that can handle long-term dependencies in sequential data. It includes specialized memory cells that can store information over an extended period and gates that control the flow of information into and out of the cell.
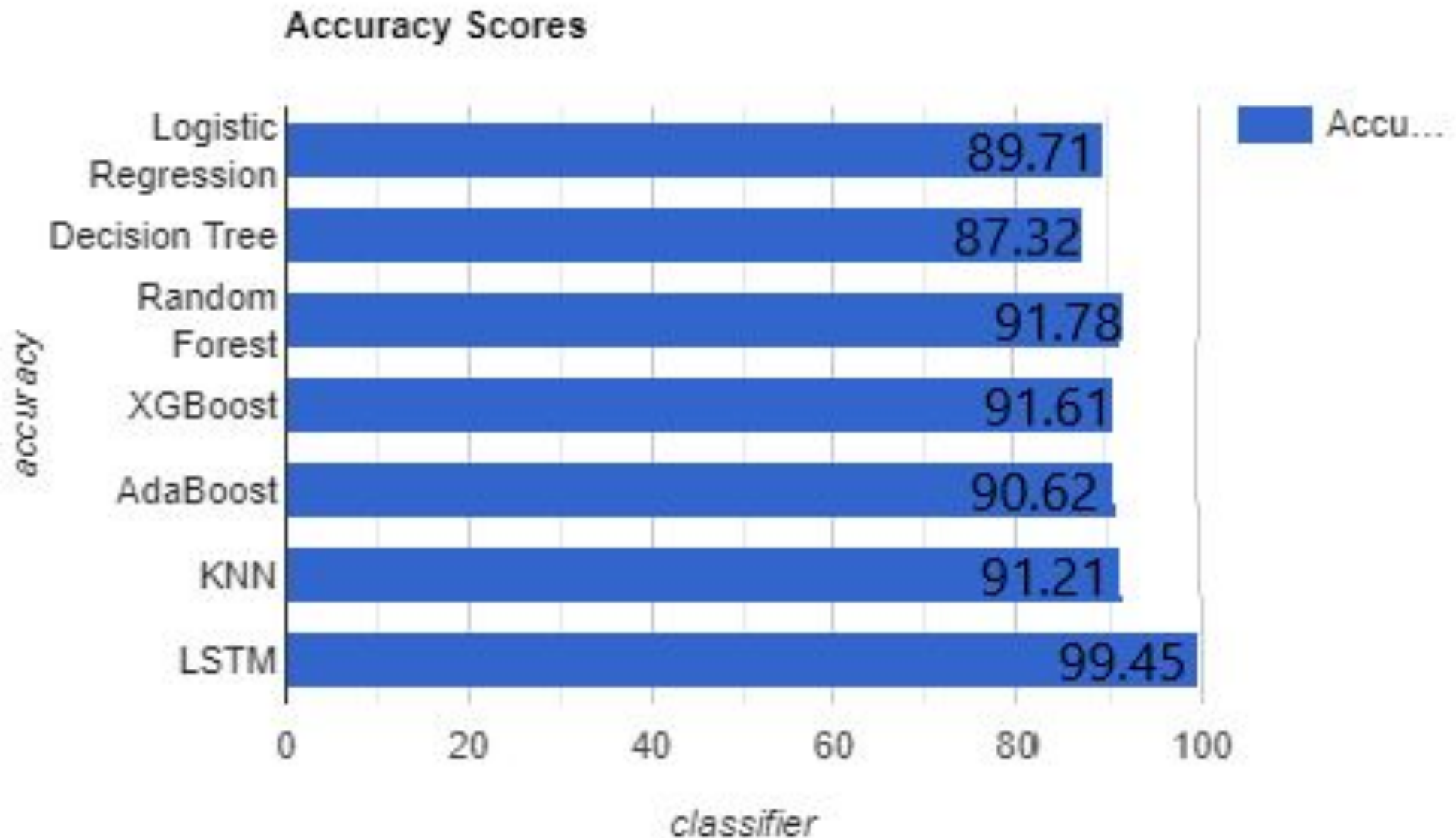
Experimental Results

- Grid search CV (Cross-Validation), technique used in machine learning to tune the hyperparameters of a model by exhaustively searching over a range of parameter values.

- For some machine learning algorithms, using Grid search cv makes the algorithm to take less time than the usual algorithm.

- From the research of analyzing the toxic comments classification and forecasting, it was said that deep learning algorithms will perform better than the normal machine learning algorithms.

Time taken and Accuracy Scores:

S.No	Algorithm Name	Accuracy
1	Logistic Regression	89.714238
2	Decision Tree Classifier	87.322443
3	Random Forest Classifier	91.788519
4	AdaBoost Classifier	90.921624
5	K Nearest Neighbours	91.205715
6	XGBoost Classifier	91.617229

S.No	Algorithm Name	Accuracy
1	LSTM	99.45

Performance analysis using Accuracy Scores:

Accuracy Scores

Classifier	Accuracy
Logistic Regression	89.71
Decision Tree	87.32
Random Forest	91.78
XGBoost	91.61
AdaBoost	90.62
KNN	91.21
LSTM	99.45

Results and Discussion:

- As concluded in the research, deep learning algorithms will perform better when compared with normal machine learning algorithms.

- After the implementation, LSTM performs very well than the normal machine learning algorithms. And in the machine learning algorithms, XGBoost performs well with a comparatively good accuracy score

- Initially, our model had an MSE of 1.2. To improve its performance, we fine-tuned the parameters using a grid search and cross-validation. With these optimized parameters, the MSE was reduced to around 0.5.

References.

- Shojaee S, Mustapha A, Sidi F, Jabar MA (2013) A study on classification learning algorithms to predict crime status. Int J Digital Content Technol Appl 7(9):361–369

- Obuandike GN, Isah A, Alhasan J (2015) Analytical study of some selected classification algorithms in WEKA using real crime data. Int J Adv Res Artif Intell 4(12):44–48.

- Iqbal R, Murad MAA, Mustapha A, Panahy PHS, Khanahmadliravi N (2013) An experimental study of classification algorithms for crime prediction. Indian J Sci Technol 6(3):4219–4225.

- Jangra M, Kalsi S (2019) Crime analysis for multistate network using naive Bayes classifier. Int J Comput Sci Mob Comput 8(6):134–143

- Wibowo AH, Oesman TI (2020) The comparative analysis on the accuracy of k-NN, naive Bayes, and decision tree algorithms in predicting crimes and criminal actions in Sleman regency. J Phys Conf Ser 1450:012076.