Project Result and Analysis Submission
Title: Cyber Bullying Detection on Social Media Platforms Using Machine Learning

Course title: Machine Learning Laboratory
Course code: CSE-458
4th Year 2nd Semester Examination 2023

Date of Submission: 11.02.2024

Submitted to-

Sarnali Basak
Associate Professor
Department of Computer Science and Engineering
Jahangirnagar University
Savar, Dhaka-1342

Submitted by-

SL	ID	Name
01	343	Tanzila Akter
02	361	Nurun Nahar Fiha
03	378	Md. Parvej Haque Palash
04	387	Md. Sakib Mollah
05	2142	Serajum Monira

DATASET DESCRIPTION:

A. Data Collection:

We have used multi label Bengali toxic comments obtained from Kaggle for reaching the final results. As we are detecting cyber bullying using Bengali toxic comments, a complete, reliable dataset was needed. So after tried out of many other open source datasets and finally decided to use this dataset. This dataset was originally for Multi Labeled Bengali Toxic Comments Classification. We processed this dataset to train our models to detect cyber bullying.

Dataset Link - https://www.kaggle.com/datasets/tanveerbelaliut/multi-labeled-bengali-toxic-comments

Here is the detailed description of the dataset

- This dataset is combination of two raw dataset and labeled them into six categories.
- Total instances 16074.

This dataset has 7 attribute- text, vulgar, hate, religious, threat, troll and Insult.

B. Data Cleaning

This dataset used was set in a CSV format. As this dataset is originally for classification we add a new category named bullying and marked them 1/0 based on the 6 attribute text, vulgar, hate, religious, threat, troll and Insult.

Table 1: Instances Description

Total instances	16074
Cyber Bullying instances	8488
Non Cyber Bullying instances	7585

C. Data Preprocessing

- Word Tokenization: This is done to break down the piece of text into individual words or tokens.
- Stopwords Filtering: Stopwords filtering is done using stop = set (stopwords.words('bengali'))

  To fetch a list of stopwords in the Bangla dictionary, after which they are removed. The stopwords are words such as 'হলে', 'দুই', 'কেউ' etc. which are not significant and do not affect the meaning of the data to be interpreted.

- To remove punctuation, we save only the characters that are not punctuation, which can be checked by using string.punctuation.
- Digit removal: We also filtered out any numeric content as it doesn't contribute to cyberbullying
- The next step was to extract the features so that the data could be utilized with machine learning methods. To do this, we used Python's sklearn library and the TF-IDF Transform. A statistical metric known as TF-IDF is used to assess a word's relevance. It is essentially calculated by multiplying the word's inverse document frequency by the number of times it appears in the document. Rather than only counting the frequency of words like CountVectorizer does, TF-IDF utilizes a technique that reduces the weight (importance) of words that appear in many texts in common, viewing them as incapable of differentiating the documents. The top 25 words based on the determined tf-idf score are presented here, demonstrating the manual nature of attribute evaluation. Several of the top terms for [কুত্তা, চোদনা, ফকিন্নি, মাদারটোস্ট, আবালচোদা, সেক্স, নাস্তিক] was the dataset.

D. Dataset Split: The dataset was divided into two parts using an 80:20 split. 80% of the data was allocated for training purposes. The remaining 20% of the dataset was set aside for testing the trained model's performance.

Table 2: Split Dataset

	Test	Training
Total Instance	3215	12858
Cyber Bullying instance	1746	6742
Non Cyber Bullying instance	1469	6116

EXPERIMENT AND RESULTS:

For our supervised learning technique, we've used Logistic Regression, Decision Tree, Random Forest and Support Vector Machine. In our research Decision Tree performed the poorest, whereas Logistic Regression performed best. The Random Forest performed batter then Decision Tree classifier as it is an extension of the Decision Tree Classifier, averaging out results of multiple recursion of the same.

The Metrics used for determining the performance of models are given below

- Accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$
- Precision $= \frac{TP}{TP+FP}$
- Recall $= \frac{TP}{TP+FN}$

Where,
TP = No. of True Positives
TN = No. of True Negatives
FP = No. of False Positives
FN= No. of False Negative

Table 3: Supervised Traditional Method

	Decision Tree	Logistic Regression	Support Vector Machine
Accuracy	0.736547434	0.81150855	0.80808709
Precision	0.752954418	0.79290853	0.79053011
Recall	0.761168385	0.88373425	0.87972509

Table 4: Supervised Ensemble Method

	Random Forest
Accuracy	0.792982549
Precision	0.792982549
Recall	0.791912908

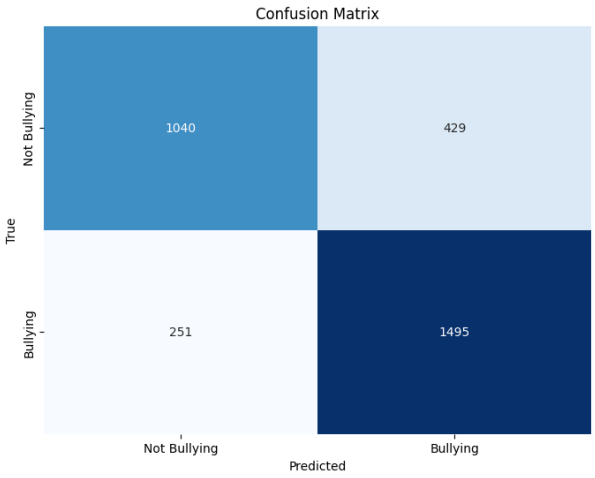

Fig 1: Confusion Matrix of Logistic Regression

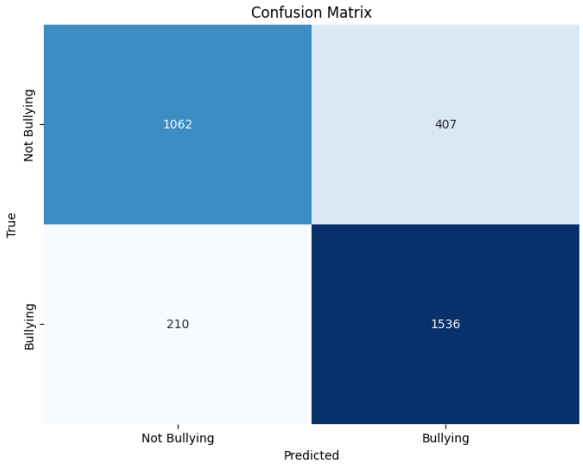

Fig 2: Confusion Matrix of Decision Tree

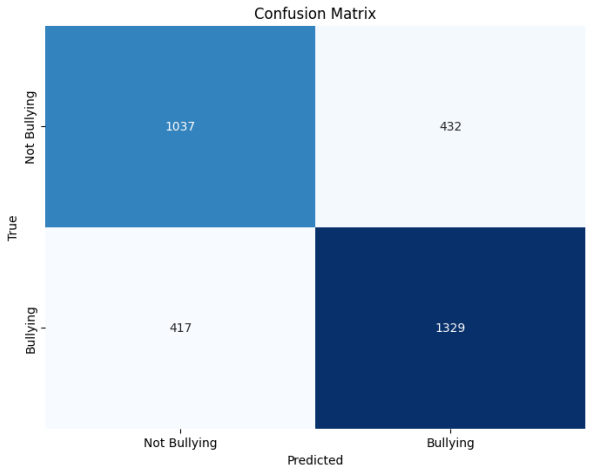

Fig 3: Confusion Matrix of Support Vector Machine

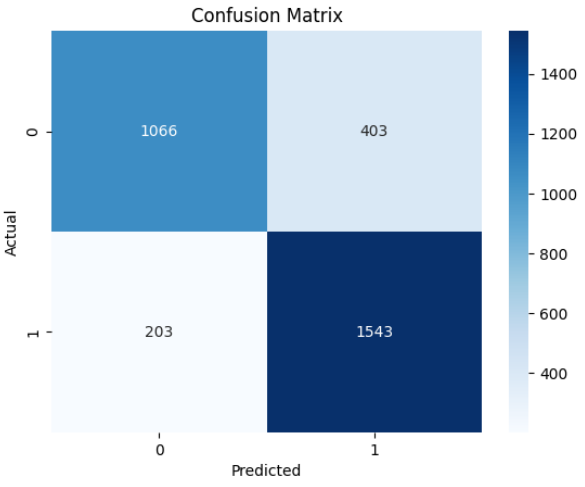

Fig 4: Confusion Matrix of Random Forest

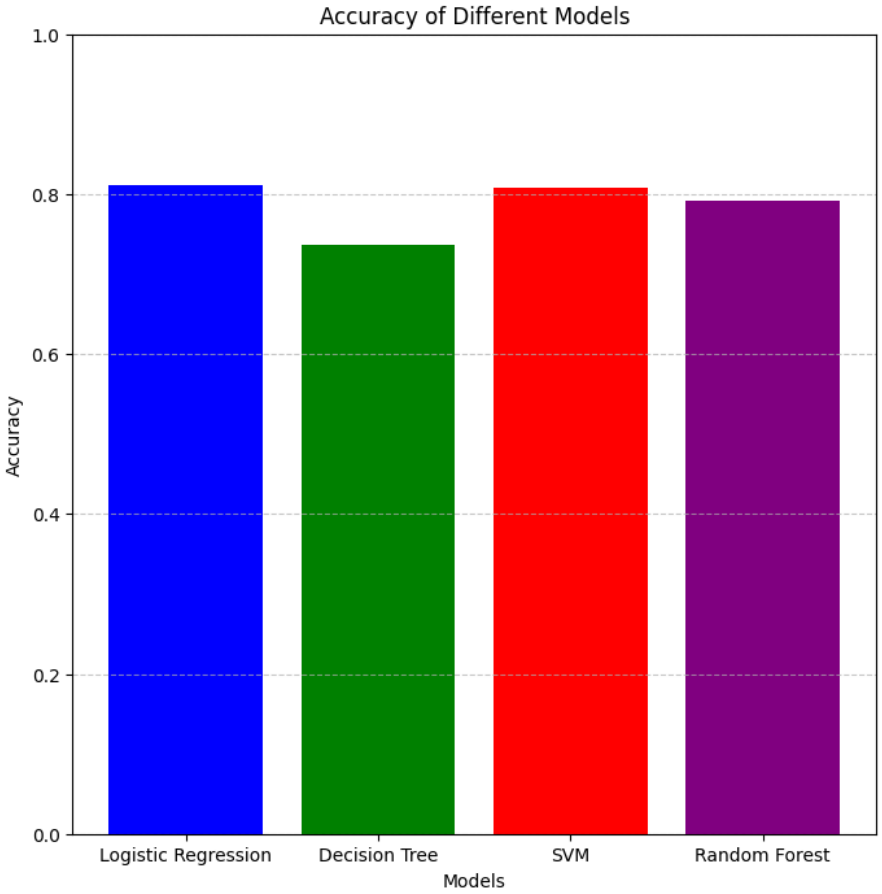

Fig 5: Accuracy

Traditional Supervised Learning used: Logistic Regression, Decision Tree, Support Vector Machine
The Ensemble Learning Method used: Random Forest classifier.
Figure 5 shows a graphical comparison between the aforementioned algorithms.

We have taken cv = 10 for cross validation.

- **Logistic Regression:**
  Cross-validation scores: [0.8028607, 0.81281095, 0.80721393, 0.81331674, 0.81331674, 0.81144991
  0.80460485, 0.8226509, 0.80211574, 0.81829496]
  Mean cross-validation score: 0.8108635416569919

- **Decision tree:**
  Cross-validation scores: [0.75186567, 0.75808458, 0.75248756, 0.75233354, 0.74113255 0.72806472, 0.73242066, 0.75482265, 0.72868699, 0.7423771]
  Mean cross-validation score: 0.744227601878597

- **SVM:**
  Cross-validation scores: [0.79850746, 0.81654229, 0.81218905, 0.82389546, 0.80958307, 0.81953951, 0.8089608, 0.81953951, 0.81082763, 0.81953951]
  Mean cross-validation score: 0.8139124306903565

- **Random Forest:**
  Cross-validation scores: [0.78420398, 0.8090796, 0.79042289, 0.80647169, 0.80087119, 0.79713752, 0.79838208, 0.80273802, 0.78282514, 0.80149347]
  Mean cross-validation score: 0.797362557158204

SVM has the highest mean cross-validation score (0.8139), followed closely by Logistic Regression (0.8109), Random Forest (0.7974), and Decision Tree (0.7442). SVM and Logistic Regression show relatively consistent performance across folds, with smaller variability compared to Decision Tree and Random Forest. Decision Tree has the highest variability among the models, indicating less stable performance across folds.

Comparison:

1. SVM vs. Logistic Regression: SVM performs slightly better than Logistic Regression in terms of mean cross-validation score, with a slightly narrower range of scores.

2. SVM vs. Random Forest: SVM outperforms Random Forest in terms of mean cross-validation score, with a narrower range of scores.

3. SVM vs. Decision Tree: SVM significantly outperforms Decision Tree in terms of mean cross-validation score and shows less variability.

Final Assessment: Based on the mean cross-validation score and variability, SVM appears to be the best-performing model among the four for this dataset. It consistently achieves higher mean scores and demonstrates relatively stable performance across folds compared to the other models.