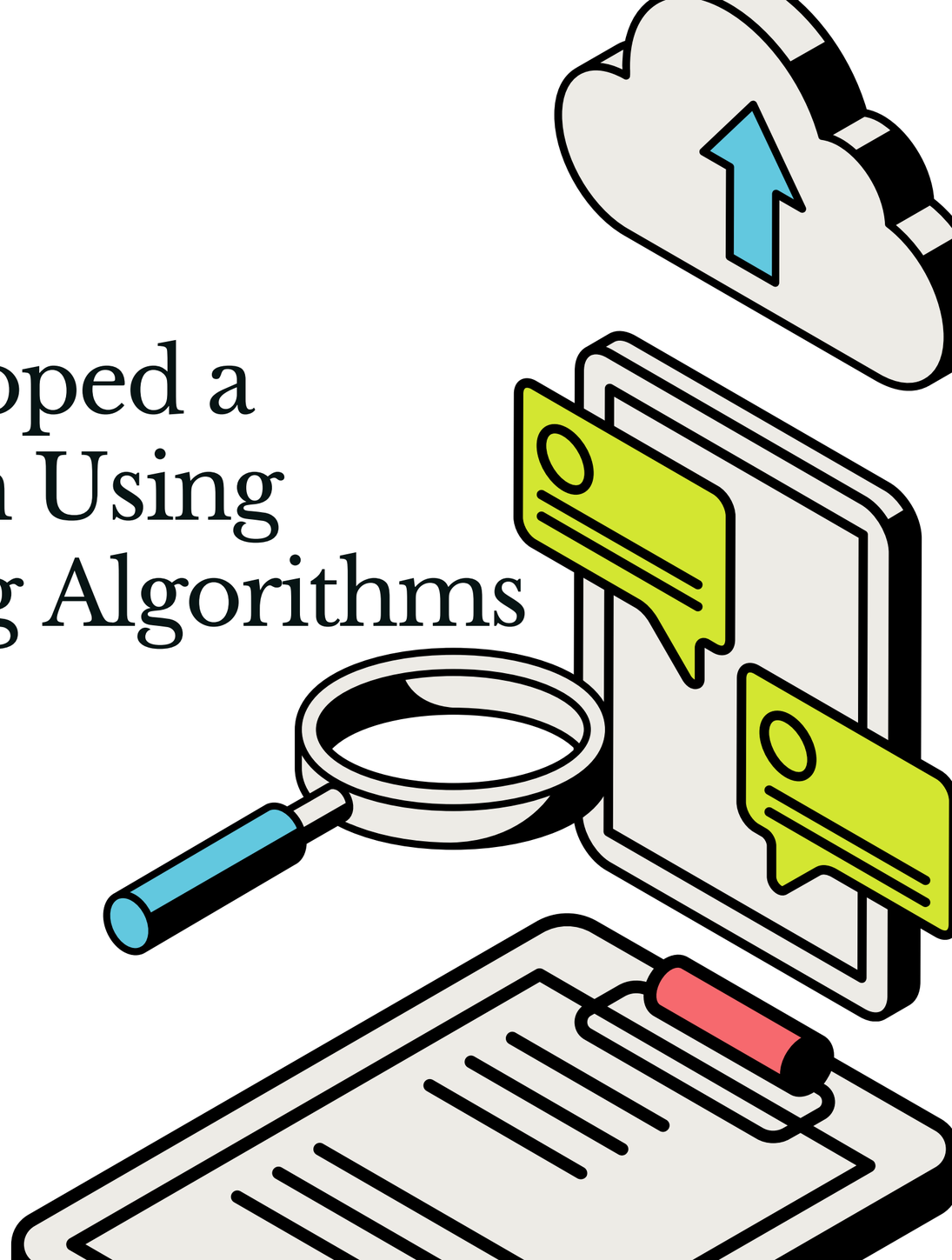
NORTH SOUTH UNIVERSITY
CSE 440, GROUP 2

Sentiment Analysis: Developed a Sentiment Analysis System Using Classical Machine Learning Algorithms



Group Members:

Shamlima Deb Trena - 2011691042

Md Rezwanul Alam Sayem - 1931809042

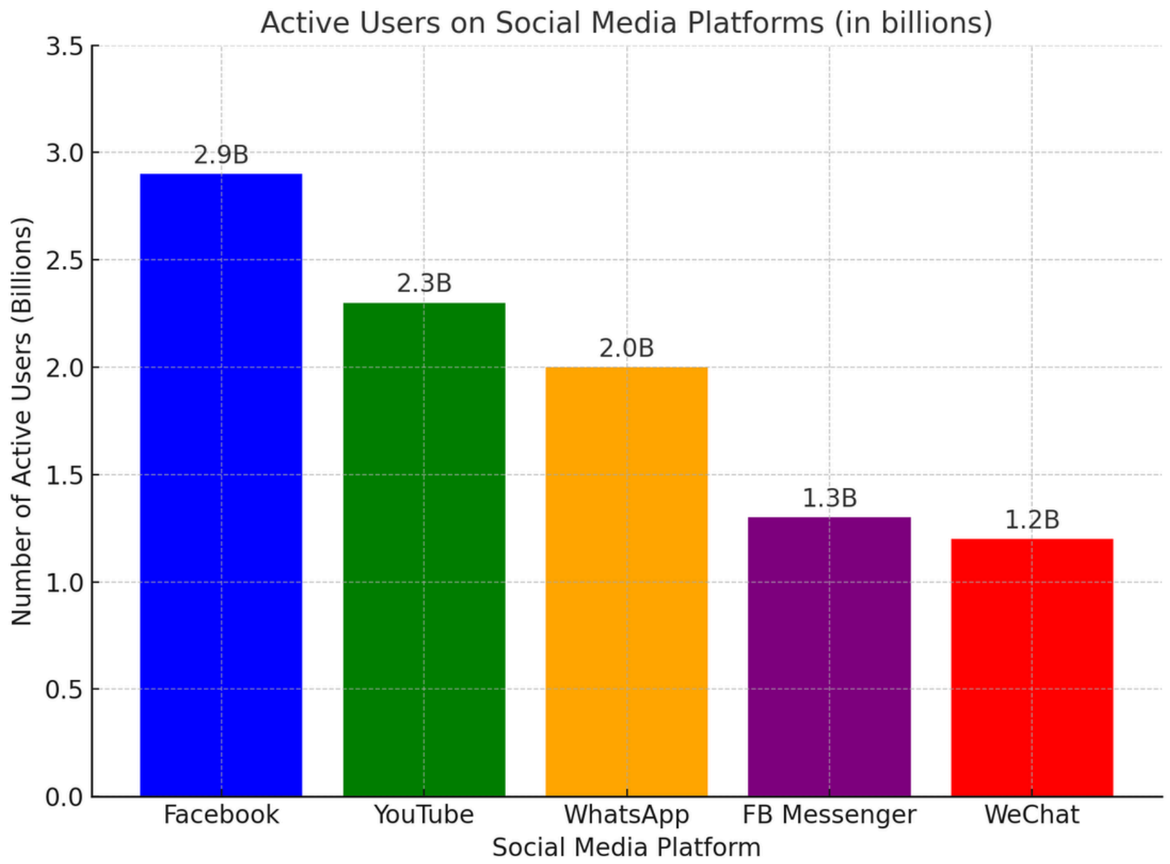
Jannatul Ferdous - 2013252042

Md. Arifur Rahman Akib - 2121359642

Introduction

The Issue

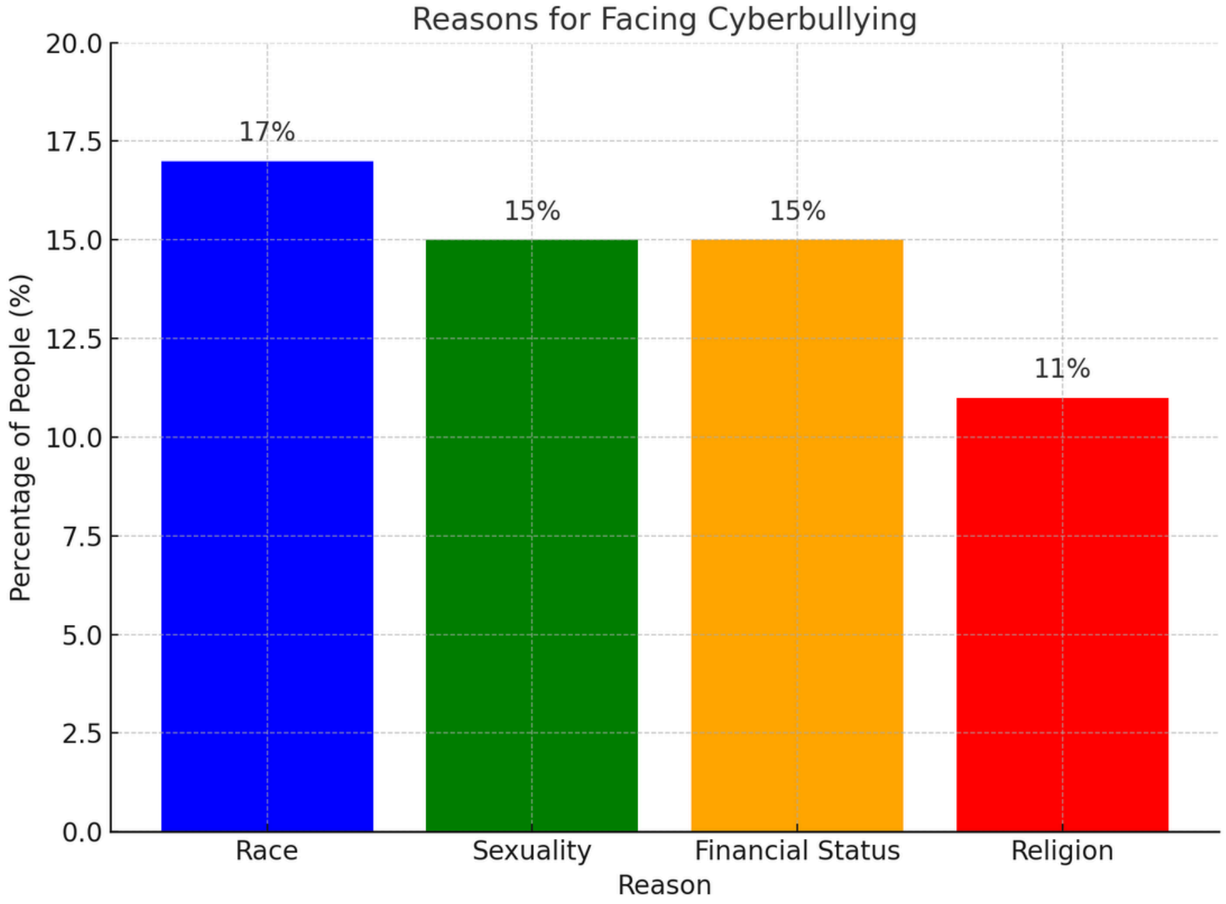
- 4.48 billion active social media users in 2021, a 13.13% increase from 2020.
- Facebook leads with 2.9 billion monthly active users, followed by YouTube, WhatsApp, FB Messenger, and WeChat.
- Social media aims to create connected and secure spaces.
- Increasing user numbers make maintaining community standards challenging.
- Issues like cyberbullying and hate speech are on the rise.



Motivation

The Aim

- 37% of users face hate speech or cyberbullying.
- 15% of users admit to engaging in such behavior.
- Targeting reasons include:
- Race: 17%
- Sexuality: 15%
- Financial status: 15%
- Religion: 11%
- In Bangladesh, 46 million Facebook users, with 60% experiencing harassment.
- Research primarily focuses on high-resource languages like English.
- Bangla lacks necessary computational linguistic resources for NLP tasks.



Previous Works

01 **TF-IDF Feature Extraction:** TF-IDF feature extraction is a common method for processing Bangla text.

02 **Deep Learning Models:** TF-IDF feature extraction is a common method for processing Bangla text.

03 **Encoder-Decoder Models:** Encoder-decoder models utilize attention-based LSTM and GRU for categorizing text.

04 **Counter Speech Analysis:** Counter speech analysis examines psycholinguistic impacts on targeted communities.

05 **SVM-based Models:** SVM-based models are applied for hate speech and opinion polarity classification.

06 **Dataset Annotation and Categorization:** Comments are categorized into hateful, non-hateful, and sentiment-based labels through manual annotation and crowdsourcing.

DATASET DETAILS

1. DATASET OVERVIEW:

- CONTAINS 11,006 BANGLA COMMENTS FROM FACEBOOK PUBLIC POSTS.
- COMMENTS ARE CATEGORIZED INTO POSITIVE (17.3%), NEUTRAL (35.1%), AND NEGATIVE (47.6%) SENTIMENTS.

2. ADDITIONAL ATTRIBUTES:

- INCLUDES REACTION COUNTS AND REPLY COUNTS FOR EACH COMMENT.

3. KEY OBSERVATIONS:

- GENDER-BASED INSIGHTS:

MALE PAGES: HIGHER PERCENTAGE OF NEUTRAL COMMENTS.
FEMALE PAGES: DOMINANTLY NEGATIVE COMMENTS, WITH SIGNIFICANT GENDER-BASED AND RELIGIOUS HATE.

- SUBCATEGORY ANALYSIS:

NEGATIVE COMMENTS OFTEN INCLUDE PERSONAL AND GENDER-BASED HATE.

- REACTION TRENDS:

NEUTRAL COMMENTS RECEIVE THE HIGHEST REACTIONS ON AVERAGE.

Dataset Details(continued)

Category	Sub category	# of Comments	Percentage	Ann. Index
Positive	Wishful thinking	967	8.8%	1
	Appreciation	942	8.6%	2
Negative	Gender-based hate	525	4.8%	3
	Religious hate	731	6.6%	4
	Political hate	572	5.2%	5
	Personal hate	1995	18.1%	6
	Sarcasm	1414	12.8%	8
Neutral	N/A	3860	35.1%	7

METHODOLOGY

1.DATA PREPROCESSING:

USED BNLP AND BLTK TOOLKITS FOR:

A) REMOVING BANGLA STOP WORDS AND PUNCTUATION.

B) TOKENIZATION AND STEMMING WITH FATICK STEMMER.

2. FEATURE EXTRACTION:

APPLIED TF-IDF VECTORIZATION (UNI-GRAM, BI-GRAM, TRI-GRAM FEATURES).

3.CLASSIFICATION MODELS:

MACHINE LEARNING:

A) LOGISTIC REGRESSION, DECISION TREES, RANDOM FOREST, MULTINOMIAL NAIVE BAYES (MNB), SVM, KNN.

B) BEST PERFORMANCE: MNB (ACCURACY: 82.6% WITH UNIGRAM).

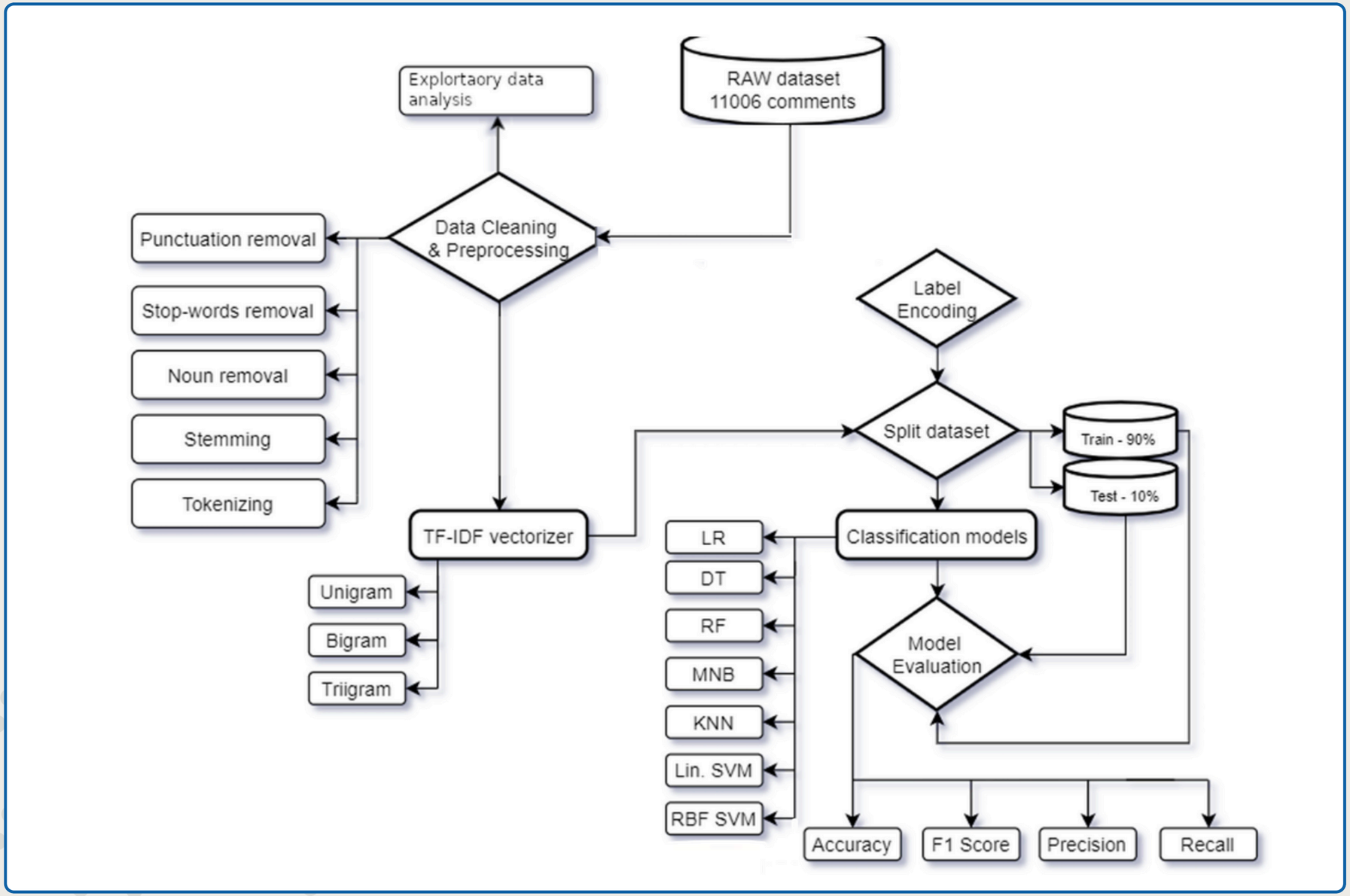
DEEP LEARNING:

A) CNN ACHIEVED 81% ACCURACY.

4.TRAINING SETUP:

USED GOOGLE COLAB FOR GPU RUNTIME AND GOOGLE DRIVE FOR STORAGE.

PROCESS FLOW-CHART



Google Colab Environment

Why Colab?

- Cloud-based environment
- Real-time collaboration
- Pre-installed ML and DL libraries
- GPU and TPU support
- Seamless integration with TensorFlow, PyTorch, etc.
- Easy sharing through Google Drive
- No local setup required

How is it done?

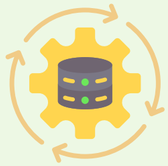
- Share notebook via link/email
- Edit in real-time collaboratively
- Add comments for feedback
- Divide tasks in code cells
- Track changes via version history
- Share outputs and results
- Assign sections for documentation

Data **C**leaning & **P**reprocessing

Cleaning

- Used BNLP and BLTK toolkits for data cleaning
- Removed Bangla stop words and punctuation using BNLP Bangla Corpus
- Tokenized data to word form using BLTK word tokenizer
- Removed names of sample pages from tokens
- Applied Fatick Stemmer for stemming
- Converted tokenized words to root form
- Addressed complexity of Bangla stemming with a lightweight stemmer

Preprocessing

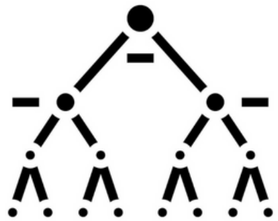


- Remove punctuation and Bangla stop-words
- Exclude names to protect privacy
- Drop data smaller than length 2
- Pass cleaned data to tokenizer and vectorizer
- Encode targets using label encoding (0 to n_classes-1)
- Vectorize data using Tfidf vectorizer
- Use Tfidf to create a feature matrix
- Apply unigram (1,1), bigram (1,2), and trigram (1,3) n-gram ranges
- Feature sizes: unigram (22,873), bigram (124,897), trigram (242,286)

Model Implementation

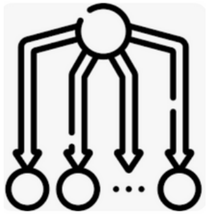
In our project, we have implemented **6 classical machine learning algorithms** to develop a sentiment analysis system.

LR: Logistic Regression

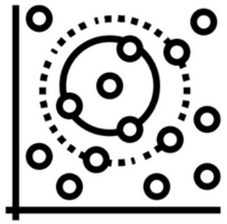


DT: Decision Tree Classifier

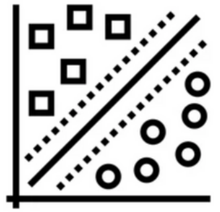
RF: Random Forest Classifier



MNB: Multinomial Naive Bayes



KNN: K-Nearest Neighbor (KNN) classifier



SVM: Support Vector Machine

Model Evaluation

Model Name	Unigram feature			
	Accuracy	Precision	Recall	F1
LR	80.2	81.44	80.2	76.22
DT	76.8	75.7	76.8	76.1
RF	78.93	78.28	78.93	75.18
MNB	**82.6**	**81.74**	**82.6**	**81.56**
KNN	80.48	80.69	80.48	80.58
Linear SVM	76.8	79.51	76.8	69.6
RBF SVM	78.36	80.39	78.36	72.78

Model Evaluation

Model Name	Bigram feature			
	Accuracy	Precision	Recall	F1
LR	77.37	80.93	77.37	70.51
DT	79.07	77.77	79.07	78.01
RF	79.77	80.98	79.77	75.54
MNB	**81.33**	81.4	**81.33**	**81.36**
KNN	80.91	80.81	80.91	80.85
Linear SVM	75.25	81.45	75.25	65.8
RBF SVM	76.52	**82.18**	76.52	68.48

Trigram feature			
Accuracy	Precision	Recall	F1
76.52	**82.18**	76.52	68.48
78.5	77.31	78.5	77.64
78.93	80.32	78.93	74.02
77.79	79.55	77.79	78.43
79.63	80.01	**79.63**	**79.8**
74.82	81.22	74.82	64.86
75.53	81.61	75.53	66.41

References:

[1]. Bltk: The bengali natural language processing toolkit. https://pypi.org/project/bltk/description. Accessed: 2021-08-14.

[2]. Amit Kumar Das, Abdullah Al Asif, Anik Paul, and Md Nur Hossain. Bangla hate speech detection on social media using attention-based recurrent neural network. Journal of Intelligent Systems, 30(1):578–591, 2021.

[3] Md Serajus Salekin Khan, Sanjida Reza Rafa, Amit Ku- mar Das, et al. Sentiment analysis on Bengali facebook comments to predict fan's emotions towards a celebrity. Journal of Engineering Advancements, pages 118–124, 2021.

[4] Muhammad Mahmudun Nabi, Md Tanzir Altaf, and Sabir Ismail. Detecting sentiment from bangla text using machine learning technique and feature anal- ysis. International Journal of Computer Applications, 153(11):28–34, 2016.

[5]. Harikrishnan N B. Confusion matrix, accuracy, precision, recall, f1 score. https://medium.com/analytics-vidhya/confusion-matrix-accuracyprecision-recall-f1- score-ade299cf63cd.

Thank You, Everyone!