Analytical Study of Transliterated Bengali-English Social Media Comments using Machine Learning Models

Ahsanullah University of Science and Technology
Course No: CSE4214
Course Name: Pattern Recognition Lab

Group : 06
Date of Submission: 06/02/2024

Name: Tonmoy Banik
ID : 190204059

Name: Piyal Datta
ID : 190204080

Name: Sanjida Akter
ID : 190204076

Name: Dipanwita Bala
ID : 190204087

Table of Contents

- Problem Statement
- Literature Review
- Dataset
- Methodology
- Model Architecture
- Experimental Result
- Result Analysis

Introduction

- Challenge of Multilingual User-Generated Content
- Emphasis on Transliterated Bengali to English
- Significance of Opinion Understanding
- Importance of Bengali in English-Dominated Digital Space
- Objective: Evaluation of Machine Learning Models
- Advancements and Applications

Problem Statement

- Linguistic Complexity in Social Media
- Limited Research on Transliterated Bengali-English Content
- Model Generalization Across Transliterated Languages
- Implications for User Engagement and Platform Moderation

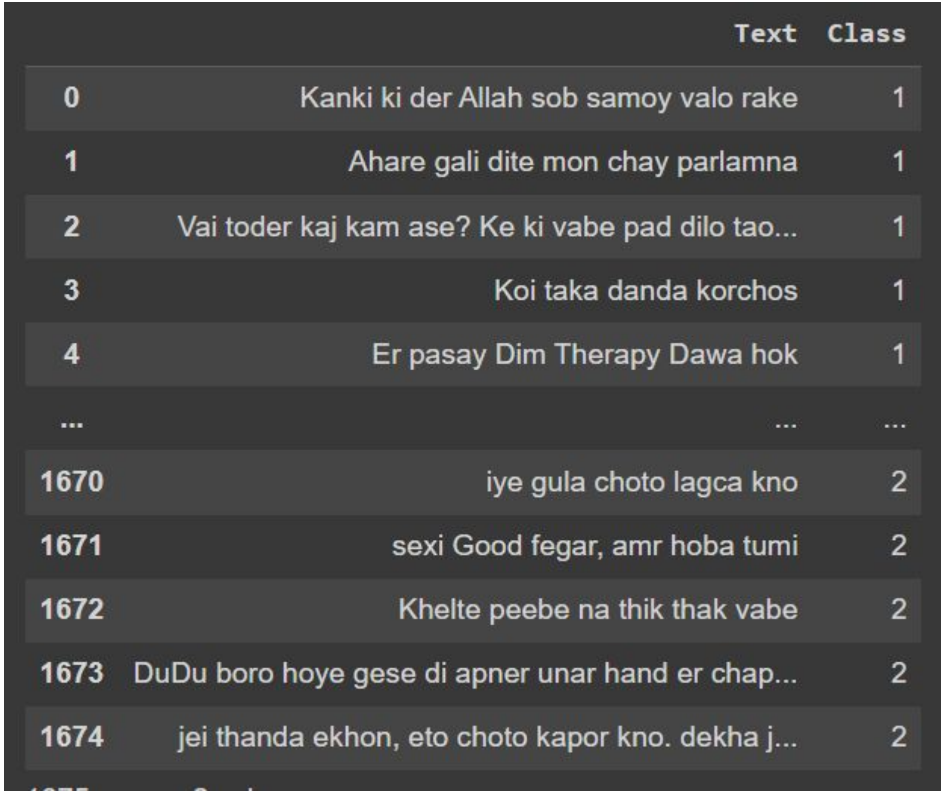
Literature Review

Name	Dataset	Methodology	Accuracy
A Research on Hinglish Sentiments of YouTube Cookery Channels Using Deep learning.	The dataset was divided into seven categories.	MLP	Tf-idf: 98.22 and count : 98.48
Multi-class sentiment classification on Bengali social media comments using machine learning	Dataset consists of 42,036 Facebook comments labeled into four classes: sexual, religious, acceptable, and political	CLSTM	85.8%.
Current State of Hinglish Text Sentiment Analysis	3 class dataset classified as positive, negative or neutral	NLP, SVM, Naïve Bayes, Network, Hybrid	SVM, hybrid: 96.8% maximum

Literature Review

Name	Dataset	Methodology	Accuracy
Sentiment Analysis of Code-Mixed Social Media Text (Hinglish)	The data consisted of Code-Mixed tweets containing Hindi and English words written in English script. The tweets were classified among the Negative, Neutral or Positive sentiment polarity.	SVM, KNN, Decision Trees, Random Forests, Naïve Bayes, Logistic Regression	Logistic Regression: 68% maximum for tf-idf
Sentiment Analysis on Bangla Text Using Extended Lexicon Dictionary and Deep Learning Algorithms.	categorical aspect-based dataset: Positive, Negative, Neutral	LSTM models such as HAN-LSTM, Bi-LSTM, BERT-LSTM	78.52%, 80.82%, 84.18% respectively.

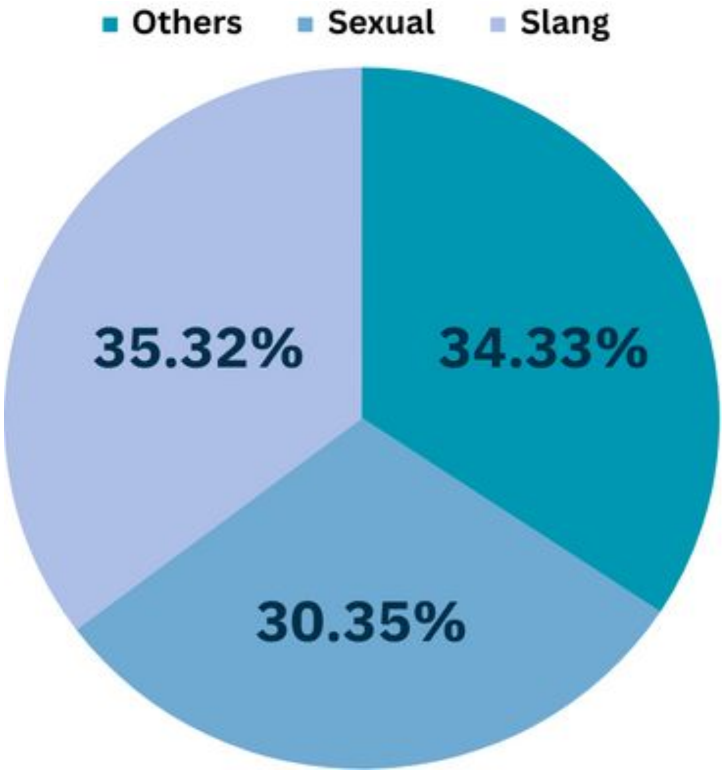
Dataset



	Text	Class
0	Kanki ki der Allah sob samoy valo rake	1
1	Ahare gali dite mon chay parlamna	1
2	Vai toder kaj kam ase? Ke ki vabe pad dilo tao...	1
3	Koi taka danda korchos	1
4	Er pasay Dim Therapy Dawa hok	1
...	...	...
1670	iye gula choto lagca kno	2
1671	sexi Good fegar, amr hoba tumi	2
1672	Khelte peebe na thik thak vabe	2
1673	DuDu boro hoye gese di apner unar hand er chap...	2
1674	jei thanda ekhon, eto choto kapor kno. dekha j...	2

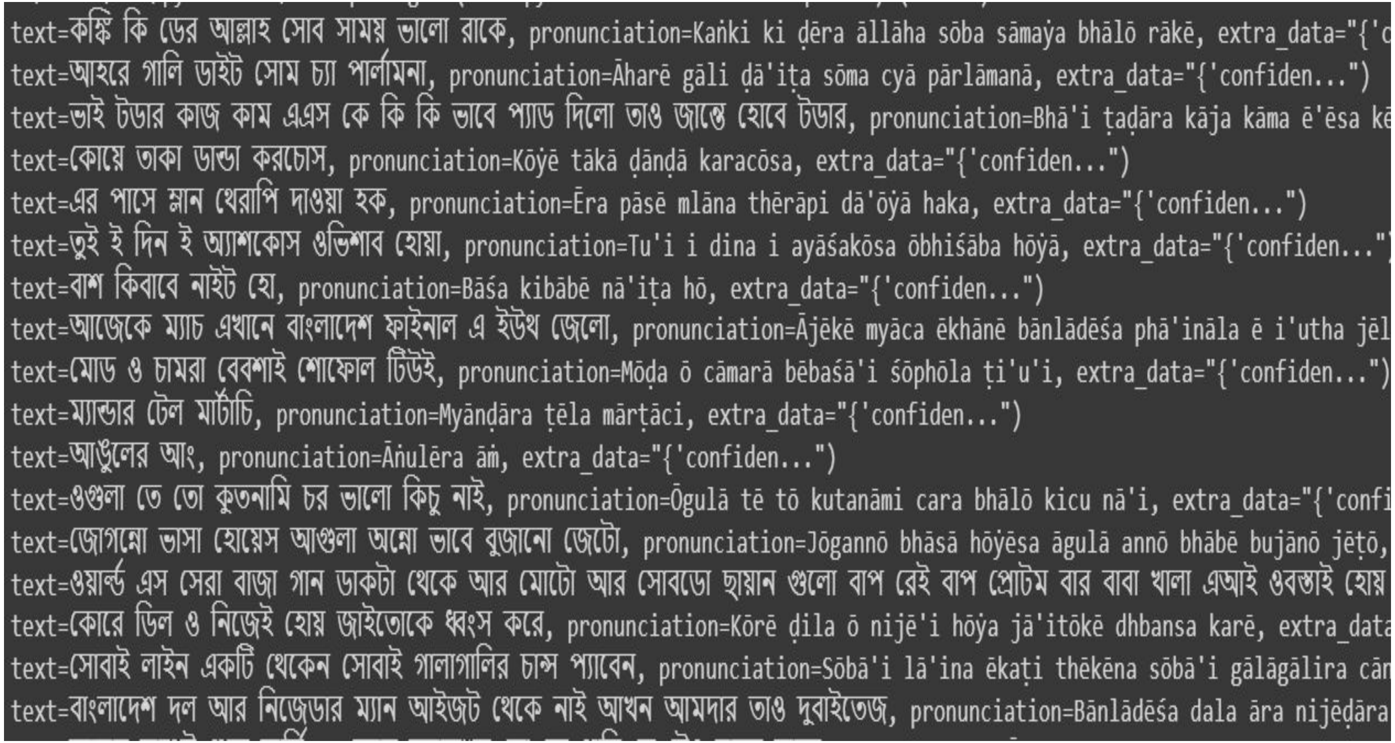
Three classes
Others -1
Sexual - 2
Slang-3

Dataset



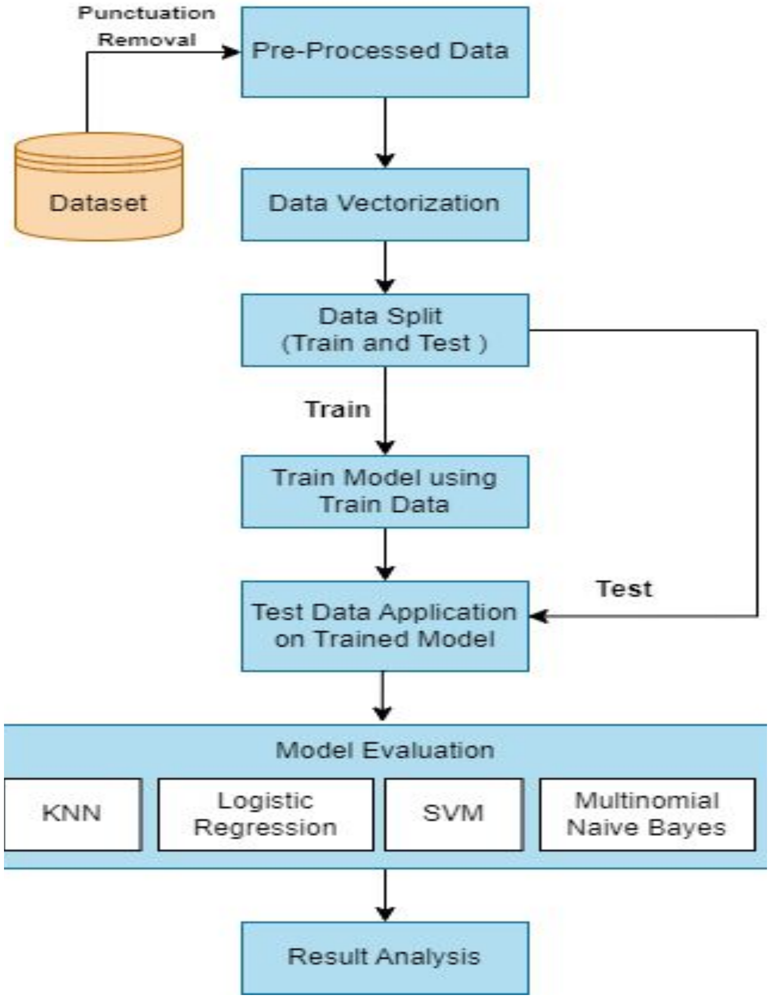
510 Sexual texts,
575 Slang texts
590 Others texts

After Translated with Google Translate



```
text=কঙ্কি কি ডের আল্লাহ সোব সাময় ভালো রাকে, pronunciation=Kaṅki ki ḍēra āllāha sōba sāmaẏa bhālō rākē, extra_data="{'c
text=আহরে গালি ডাইট সোম চ্যা পার্লামনা, pronunciation=Āharē gāli ḍā'iṭa sōma cyā pārlāmanā, extra_data="{'confiden...")
text=ভাই টডার কাজ কাম এএস কে কি কি ভাবে প্যাড দিলো তাও জান্তে হোবে টডার, pronunciation=Bhā'i ṭaḍāra kāja kāma ē'ēsa kē
text=কোয়ে তাকা ডান্ডা করচোস, pronunciation=Kōẏē tākā ḍānḍā karacōsa, extra_data="{'confiden...")
text=এর পাসে ম্লান থেরাপি দাওয়া হক, pronunciation=Ēra pāsē mlāna thērāpi dā'ōẏā haka, extra_data="{'confiden...")
text=তুই ই দিন ই অ্যাশকোস ওভিশাব হোয়া, pronunciation=Tu'i i dina i ayāśakōsa ōbhiśāba hōẏā, extra_data="{'confiden..."
text=বাশ কিবাবে নাইট হো, pronunciation=Bāśa kibābē nā'iṭa hō, extra_data="{'confiden...")
text=আজেকে ম্যাচ এখানে বাংলাদেশ ফাইনাল এ ইউথ জেলো, pronunciation=Ājēkē myāca ēkhānē bānlādēśa phā'ināla ē i'utha jēl
text=মোড ও চামরা বেবশাই শোফোল টিউই, pronunciation=Mōḍa ō cāmarā bēbaśā'i śōphōla ṭi'u'i, extra_data="{'confiden...")
text=ম্যান্ডার টেল মার্টাচি, pronunciation=Myānḍāra ṭēla mārṭāci, extra_data="{'confiden...")
text=আঙুলের আং, pronunciation=Āṅulēra āṁ, extra_data="{'confiden...")
text=ওগুলা তে তো কুতনামি চর ভালো কিছু নাই, pronunciation=Ōgulā tē tō kutanāmi cara bhālō kicu nā'i, extra_data="{'confi
text=জোগন্নো ভাসা হোয়েস আগুলা অন্নো ভাবে বুজানো জেটো, pronunciation=Jōgannō bhāsā hōẏēsa āgulā annō bhābē bujānō jēṭō,
text=ওয়ার্ল্ড এস সেরা বাজা গান ডাকটা থেকে আর মোটো আর সোবডো ছায়ান গুলো বাপ রেই বাপ প্রোটম বার বাবা খালা এআই ওবস্তাই হোয়
text=কোরে ডিল ও নিজেই হোয় জাইতোকে ধ্বংস করে, pronunciation=Kōrē ḍila ō nijē'i hōẏa jā'itōkē dhbansa karē, extra_data
text=সোবাই লাইন একটি থেকেন সোবাই গালাগালির চান্স প্যাবেন, pronunciation=Sōbā'i lā'ina ēkaṭi thēkēna sōbā'i gālāgālira cān
text=বাংলাদেশ দল আর নিজেডার ম্যান আইজট থেকে নাই আখন আমদার তাও দুবাইতেজ, pronunciation=Bānlādēśa dala āra nijēḍāra
```

Methodology



Model Architecture

- Logistic Regression
- K Nearest Neighbors
- Support Vector Machine
- Multinomial Naive Bayes

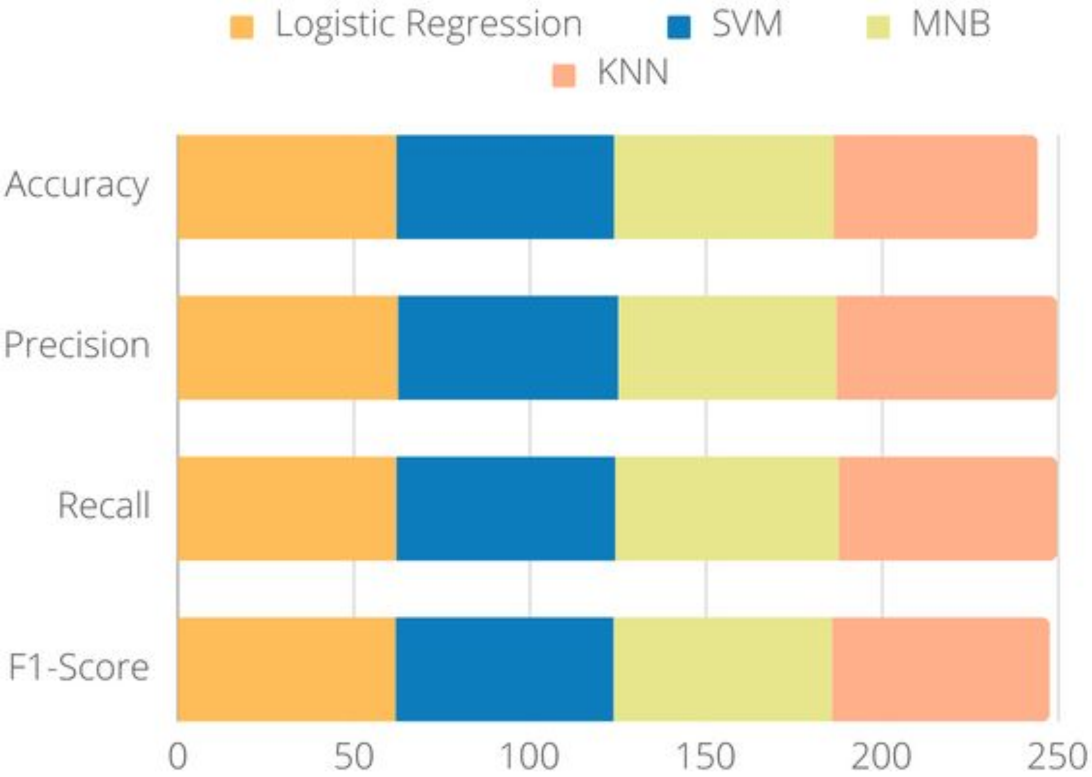
Experimental Result Before Translation

Model	Accuracy	Precision	Recall	F1
Logistic	62%	62%	62%	62%
KNN	58%	62%	62%	62%
SVM	62%	62%	62%	62%
Multinomial Naive Bayes	62%	62%	63%	62%

Experimental Result After Translation

Model	Accuracy	Precision	Recall	F1
Logistic	50%	52%	50%	49%
KNN	30%	52%	50%	49%
SVM	49%	52%	50%	49%
Multinomial Naive Bayes	51%	54%	59%	57%

Result Analysis



References

[1] Suraj Kumar Donthula, Abhishek Kaushik, "A Research on Hinglish Sentiments of YouTube Cookery Channels Using Deep learning", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878,Volume-8, Issue-2S11, September 2019

[2] R. Haque, N. Islam, M. Tasneem, and A. K. Das, "Multi-class sentiment classification on bengali social media comments using machine learning," International Journal of Cognitive Computing in Engineering, vol. 4, pp. 21–35, 2023.

[3] V. Thakur, R. Sahu, and S. Omer, "Current state of hinglish text sentiment analysis," SSRN Electronic Journal, 01 2020.

References

[4] Gaurav Singh, “Sentiment Analysis of Code-Mixed Social Media Text (Hinglish)”,School of Computing, University of Leeds, Leeds, LS29JT, UK

[5] N. R. Bhowmik, M. Arifuzzaman, and M. R. H. Mondal, “Sentiment analysis on bangla text using extended lexicon dictionary and deep learning algorithms,” Array, vol. 13, p. 100123, 2022.

[6] Faisal Bin Ashraf, “Banglish (bengali in english letter) hate speech dataset,” 2022.

Thanks!!

Any questions?