Ushoshi2023 at BLP-2023 Task 2: A Comparison of Traditional to Advanced Linguistic Models to Analyze Sentiment in Bangla Texts

Sharun Akter Khushbu[1], **Nasheen Nur[2]**, Mohiuddin Ahmed[3], Nashtarin Nur[4]

1. Daffodil International University, 2. Florida Institute of Technology, 3. UNC Charlotte, 4. United International University

Introduction

Our analytical approach is designed for BLP Workshop-2023 Task-2: **Sentiment Analysis**. We achieved a **68%** accuracy and a **68%** F1 micro score with vanilla **LSTM**, where **XLM-RoBERTa-base** model achieved **65%** accuracy. Traditional machine learning models were applied to compare the result where **75%** accuracy was achieved with traditional SVM.

- Our contributions are-
  - Data augmentation using the oversampling method to remove data imbalance
  - Attention masking for data encoding with masked language modeling to capture representations of language semantics effectively, by further demonstrating it with explainable AI
- https://github.com/sharunakter/BLPWorkshop_2023_SentimentAnalysisInBangla

Observations

1. Classifiers with no boosting, oversampling, or under-sampling gave lower recall with a lower false positive rate (FPR).For example, after applying these techniques and masking, we get 66% accuracy for the XLM-RoBERTa-base, which was previously 41.45% on the XLM-RoBERTabase.
2. XAI on XLM-RoBERTabase's output shows how the Masked Language Modeling (MLM) approach captures the nuanced sentiment expressed in Bangla text, even in the presence of codemixing, sarcasm, or subtle linguistic cues.

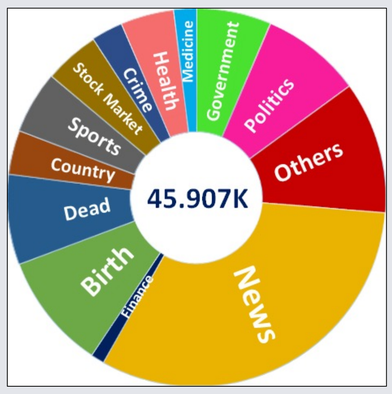

Figure 1.
Data Distribution for Different Categories

System Overview: Experiment and Results

Preprocessing and Data Augmentation

After necessary preprocessing, boosting is applied with oversampling. There is a lack of balance in the class distribution of the Bangla dataset provided. We used oversampling techniques to balance the class distribution. We merged the train and dev-test set to train the model.

We used an 80-20 training-validation split for training all the classifiers: complex deep learning models, pre-trained transformers, and traditional machine learning algorithms.

Traditional Machine Learning Models

Traditional Models	Accuracy	Precision	Recall	F1-Score
LR	71.91	72.54	71.91	71.52
DT	64.81	64.31	64.81	64.18
RF	72.66	73.55	72.66	72.00
MNB	71.22	72.51	71.22	70.83
KNN	53.69	54.79	53.69	53.64
SVM	75.02	75.26	75.02	74.85
SGD	60.40	65.94	60.40	58.69

Table 1.
Evaluation Metrics: Traditional ML models

Deep Learning and MLM

Class Label	Model	Accuracy	Precision	Recall	F1	Micro F1	Macro F1
Negative	LSTM	0.68	0.70	0.64	0.67	0.68	0.62
Neutral			0.70	0.78	0.74		
Positive			0.63	0.63	0.63		
Negative	XLM-RoBERTa-base	0.66	0.71	0.76	0.73	0.65	0.58
Neutral			0.51	0.26	0.34		
Positive			0.62	0.74	0.67		
Negative	BanglaBERT	0.64	0.71	0.72	0.71	0.64	0.59
Neutral			0.44	0.38	0.41		
Positive			0.63	0.67	0.65		
Negative	Multilingual BERT	0.64	0.68	0.77	0.72	0.64	0.57
Neutral			0.46	0.29	0.36		
Positive			0.65	0.66	0.66		
Negative	DistilBERT	0.55	0.54	0.54	0.54	0.55	0.51
Neutral			0.60	0.64	0.61		
Positive			0.20	0.33	0.24		

Table 2.
Evaluation of Top Deep Learning Models based on Individual Class Labels

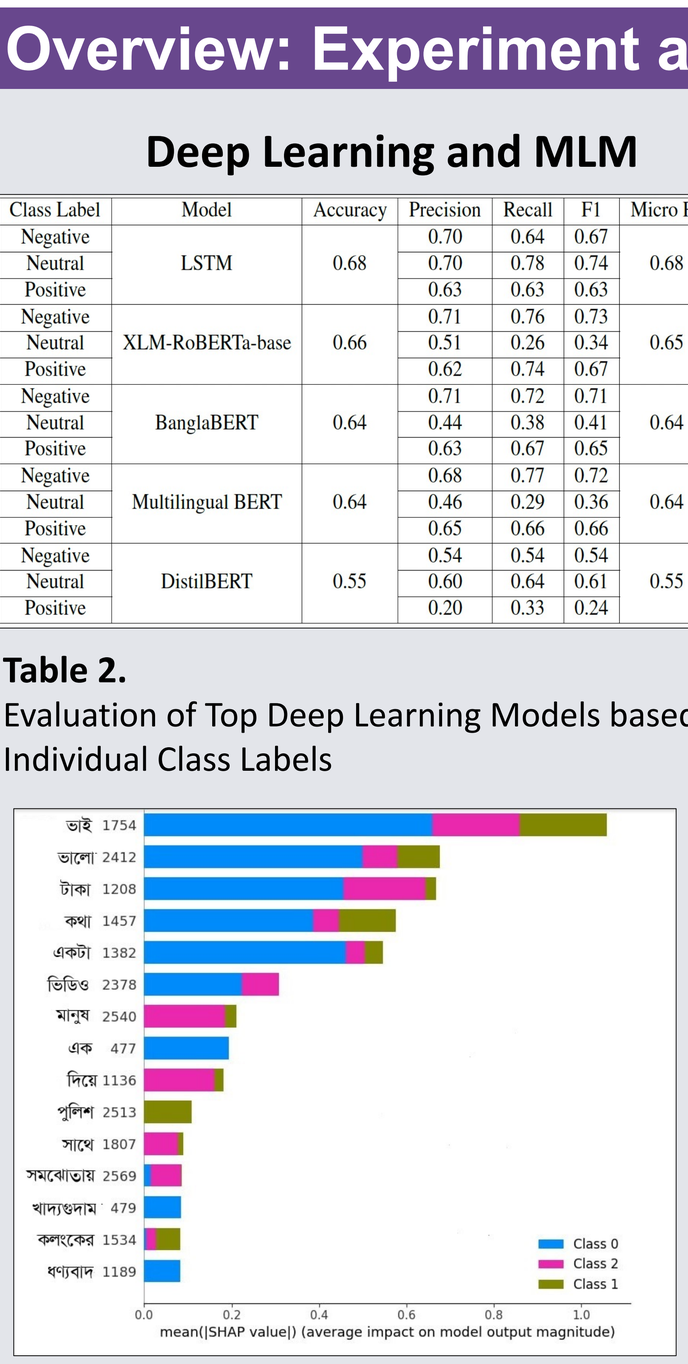

Figure 2.
SHAP on XLM-RoBERTa-base output- blue (Positive), green (Negative) and pink (Neutral)

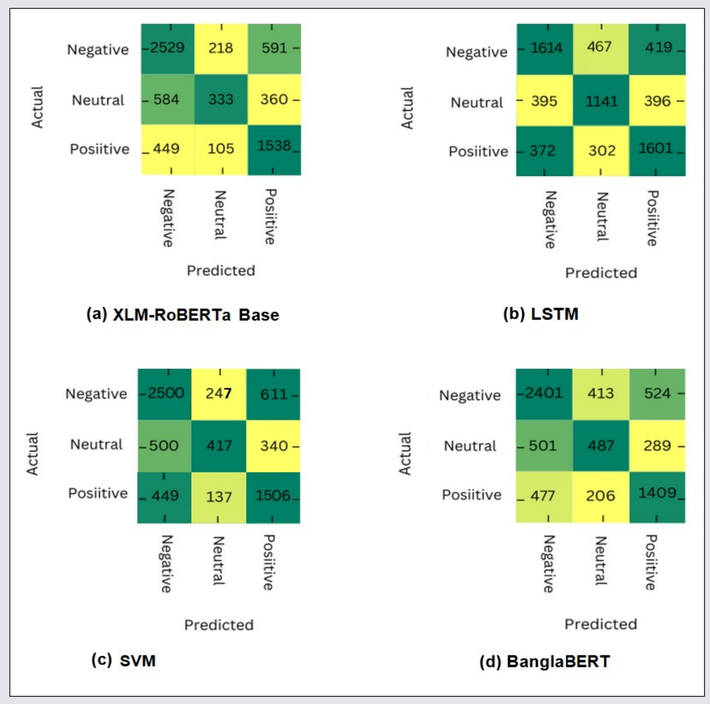

Figure 3.
Confusion Matrix

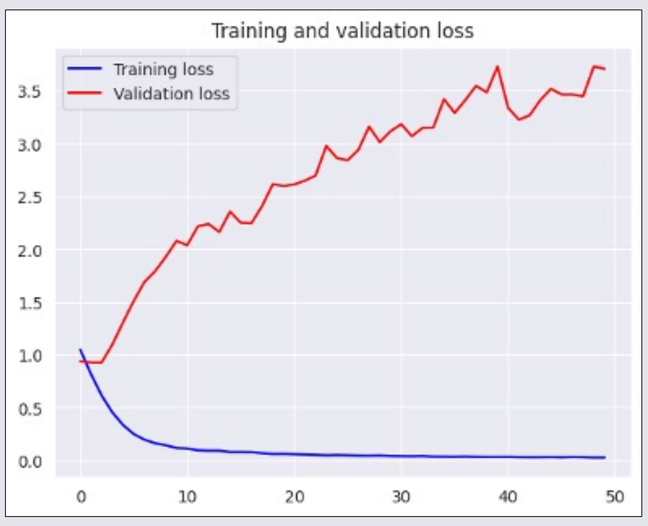

Figure 4.
Learning Curve for Underfit LSTM

Discussion: SHAP - XAI

SHAP plot combines the significance of features with their impacts. Y-axis lists features from most important to least important order. The labels on the Y-axis represent the most influential word features and associated indexing in the word vector. X-axis shows Shapely values from 0 to 1. Blue, green, and pink spectrum are Shapley values for "positive," "negative," and "neutral" classes. Not only the length of the spectrum but also the color has significance. E.g., the **"পুলিশ"**feature correlates less than 20% with the model output accuracy. The Shapely for blue is 70%, whereas for pink and green is 20% - having a **"ভাই"** word in a post mostly co-related to a positive post, which is also intuitively correct since it is a respectful salutation.

Limitations and Conclusion

- Vector assembler on huge data made the dimensions of the feature very large and computationally expensive, difficult to address with low computing resources.
- The highly imbalanced dataset has only 20% "Neutral" labels, which skewed the prediction against this class and caused some models to underfit. Developing MLM-based masked models with oversampled datasets improved the quality of the classification tasks for XLM-RoBERTa.

Our future work will focus on mitigating the challenge of Bangla sentiment analysis for lacking high-quality datasets, generalizable tools, comprehensive sentiment lexicons, and standardized evaluation metrics.