# Anti-social Behavior Detection using Multi-lingual Model


By

**Hafiz Zeeshan Ali**

**Fall-2018-MS-CS 275499 SEECS**

Supervisor

**Dr. Adnan Rashid**

**Department of Computing**

A thesis submitted in partial fulfillment of the requirements for the degree of Masters of Science in Computer Science (MS CS)

In

School of Electrical Engineering & Computer Science (SEECS)

National University of Sciences and Technology (NUST)

Islamabad, Pakistan.

(August 2022)

# Thesis Acceptance Certificate

Certified that final copy of MS/MPhil thesis entitled "**Anti-social Behavior Detection using Multi-lingual Model**" written by **Hafiz Zeeshan Ali**, (Registration No **Fall-2018-MS-CS 275499 SEECS**), of School of Electrical Engineering & Computer Science (SEECS)has been vetted by the undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial fulfillment for award of MS/M Phil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature: _____

Name of Advisor: **Dr. Adnan Rashid**

Date: _____

Signature (HoD): _____

Date: _____

Signature (Dean/Principal): _____

Date: _____

# Approval

It is certified that the contents and form of the thesis entitled "**Anti-social Behavior Detection using Multi-lingual Model**" submitted by **Hafiz Zeeshan Ali** have been found satisfactory for the requirement of the degree.

Advisor: **Dr. Adnan Rashid**

Signature: _____

Date: _____

Committee Member 1: **Dr. Rabia Irfan**

Signature: _____

Date: _____

Committee Member 2: **Dr. Arham Muslim**

Signature: _____

Date: _____

Committee Member 3:

Signature: _____

Date: _____

# Dedication

Dedicated to my elder brother whose enormous support and assistance led me to this wonderful achievement.

# Certificate of Originality

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at Department of Computing at School of Electrical Engineering & Computer Science (SEECS)or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at School of Electrical Engineering & Computer Science (SEECS)or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics which has been acknowledged.

Author Name: **Hafiz Zeeshan Ali**

Signature: _____

# Acknowledgments

I am thankful to **Almighty Allah** Who directed me to achieve this work. Who guided me at every single step taken to complete my thesis. I could have done nothing without His guidance and priceless support.

I would like to pay special thanks to my advisor **Dr. Adnan Rashid** who supported me, motivated me on every single step to complete my study. I would also like to pay my special thanks to GEC members **Dr. Rabia Irfan** and **Dr. Arham Muslim** for their valuable comments, suggestions, enormous support and assistance.

I am really thankful to my elder brother and family who made me worthy of this achievement and they continued to support me throughout every stage of my life.

Finally, I would like to pay special thanks to all individual who played a valueable role into my studies.

**Hafiz Zeeshan Ali**

# Contents

# List of Figures

# List of Tables

# Abstract

In the current era, social media has emerged as a very useful and reliable means of communication between different people and communities. However, with the leverage of communication platforms and billions of social media users, it became more challenging to stop hateful, abusive, or offensive content spread by extremists that are various aspects of Anti-social Behavior (ASB). Multiple users from several regions use different languages (a mix of native, local and other languages) to express their emotions. In the South Asia region, the frequently used languages on these platforms are Roman Urdu-English and Roman Hindi-English. Therefore, the ASB detection with multilingual model settings represents a wide area of interest for all kinds of social media platforms. Failing to properly address this issue over time on a global scale has already led to morally questionable real-life events, human deaths, and the perpetuation of hate itself. In this thesis, we perform a sentimental analysis of the Roman Urdu-English and Roman Hindi-English languages using Transformer based mBERT and XLM-R models. Moreover, we process the negatively classified sequences for detecting/analyzing the ASB.

Chapter 1

# Introduction and Motivation

This chapter presents an introduction to the research topic of the thesis. In particular, it includes a motivation of the topic, defining a problem statement and a description of the solution to the problem, proposed in this thesis alongside the main contributions of the thesis.

## 1.1 Motivation

Languages are considered as instruments for expressing ideas, believes and emotions of people all around the world and they are having a direct impact on their social and psychological lives. They have been used for communication between people in either verbal or written form. In this contemporary period, micro-media blogging platforms have attracted considerable interest of people to connect with each other by the usage of different languages in written form [1]. Over the past decade, the social media networks have been widely used to connect a variety of people from all over the world having different culture and speaking different languages.

Urdu and Hindi languages are the two frequently spoken languages in South Asia. Some of the statistics about the number and percentage of the people of the total population of the respective countries of South Asia speaking Urdu and Hindi languages are given in Table 1.1 [2, 3, 4].

| Country | Urdu Language Speaker (Numbers) | Urdu Language Speaker (Percentage) | Hindi Language Speaker (Numbers) | Hindi Language Speaker (Percentage) |
|---|---|---|---|---|
| Pakistan | 17,115,000 | 7.6 | - | - |
| India | 69,670,000 | 5.0 | 571,298,000 | 41.0 |
| Nepal | 772,000 | 2.6 | - | - |
| Singapore | - | - | 65,00 | 1.2 |
| Bangladesh | 250,000 | 0.1 | - | - |

**Table 1.1:** Urdu and Hindi Speakers in some South Asia Countries

According to Table 1.1, more than 17 million people speak Urdu language in their daily life for communication, which is 7.6% of the total population of Pakistan. Moreover, around 70 million and 8 thousand people of the total population of India and Nepal speaks Urdu which is 5% and 2.6%, respectively. Similarly, 571 million people speak Hindi in India in their daily lives, which is 41% of their total population. Similarly, 1.2% population of Singapore speak Hindi language. Furthermore, Bangladesh and Nepal are also located in the South Asia region and collectively around one million people of the total population of these two countries speak Urdu.

The frequent presence of different emotions in peoples' writing and sharing on social media trigger them for any psychosocial phenomenon including altruism, aggressiveness, and Anti-social Behavior (ASB). Therefore, understanding the behavior and the emotional state is required to categorize and classify the conduct (such as happiness, sadness, and anger or fear) of people [5]. The behavior of a person that harasses, alarms or distresses other who are not in their family is referred as ASB (by the Anti-Social Behaviour, Crime and Policing Act, 2014) [6]. For example, Bullying, strong reactions by a person, and the destructive and the impulsive behaviors are all different aspects of ASB [7]. Therefore, the following actions are deemed as ASB: (i) using the Internet to engage in criminal activities like selling fake items or objectionable sexually explicit content; (ii) using the Internet to bully others (i.e., cyber-bullying); (iii) using the Internet to defraud others (iv) engaging in an illegal gambling.

## 1.2 Problem Statement

People use informal and code-mixed languages on social media to post their ideas (Roman Urdu-English and Roman Hindi-English in our case). Young generation in the South Asia region commonly use Roman Urdu-English and Roman Hindi-English instead of Urdu or Hindi on social media as they feel difficultly in typing Urdu or Hindi languages. Thus, most of the short text produced on social media contains Roman Urdu-English and Roman Hindi-English dialect. Roman Urdu mixed with English or Roman Hindi mixed with English are considered as one of the mixed languages with poor resources [8]; As a result, it becomes more challenging to perform normal Natural Language Processing (NLP) activities, such as emotion or sentiment analysis based on the data about these languages collected from the social media platforms [9]. A regular use of social networking platforms has produced a huge amounts of data. The process of transforming an unstructured text to a structured format with the purpose of identifying significant patterns and fresh insights is known as text mining or also referred as text data mining. Mining of these social media platforms has a potential to cite illegal activities by detecting the ASB using NLP that may be useful to individuals, customers and businesses.

## 1.3 Proposed Solution

Approximately, 2.13 billion users are currently using the social media platforms that are bringing together people from different parts of the world having different thoughts, culture and living styles and speaking different languages [10]. Some popular social media platforms include Facebook, LinkedIn, Google+, and Twitter. With the immense increase of social media users, there is a dire need of analyzing/detecting the ASB. Manual Monitoring or traditional approaches such as Dictionary-based Approach and Corpus-based Approach [11] cannot produce desired results due to the involvement of the code-mixed text. To address this problem, we propose to use Multilingual Bidirectional Encoder Representations from Transformers (mBERT) and XLM-RoBERTa (XLM-R), which are bi-directional and multi-lingual machine learning models widely used for NLP.

## 1.4 Contributions

The main contributions of the thesis are:

### 1.4.1 Data Collection and Preprocessing

- Data Acquisition
- Lower casing, Removal of stop words, Word averaging
- Removal of nulls and one word based sequences

### 1.4.2 Model Training

- XLM-R hyperparameter tuning for Roman Urdu-English and Roman Hindi-English
- mBERT hyperparameter tuning for Roman-Urdu and Roman-Hindi

### 1.4.3 Classification

- Sentiment classification as positive or negative
- ASB classification

## 1.5 Thesis Outline

This section will give you an outline of the thesis by presenting a brief introduction of each chapter.

### 1.5.1 Literature Review

This chapters discusses the challenges involved in analyzing/detecting ASB and the state-of-the-art methods that have been used for detecting ASB. It also presents the deep learning techniques that have been used for detecting ASB and highlights the problems identities in these approaches.

### 1.5.2 Research Methodology

We provide the proposed methodology catering for the issues identified in the Chapter 2 of the thesis. In particular, we present an overview of the main blocks of the proposed framework, that include the data collection and processing, model development, training and its evaluation for the data.

### 1.5.3 Implementation and Results

This chapter provides the experimental design, its implementation and the corresponding results achieved in our proposed framework alongside a discussion providing a comparison of the results corresponding to different data sets.

### 1.5.4 Conclusion and Future Work

We provide the conclusion of the thesis along with highlighting some future directions of the work done in chapter of the this thesis.

Chapter 2

# Literature Review

In this chapter, we present the importance of analyzing the ASB and discusses about the challenges that are faced while detecting ASB. We also provide the traditional and deep learning based approaches used for analyzing ASB.

## 2.1 ASB Analysis

ASB refers to actions of a person that harasses or hurt others emotionally. The ASB analysis includes a detection of such harsh behaviours on social media platforms. Billions of people are connected on social media these days. A frequent sharing of feeling on any social or personal agenda on any social media platform and responses of millions of other users in the form of comments or retweets on these feelings generate a lot of textual data. This data has been further analyzed using different techniques, such as traditions techniques (dictionary and corpus-based approaches), machine learning based approaches and deep learning based approaches for detection of the ASB. ASB analysis/detection assists in stopping any harsh behavior of a social media user to reduce its negative impact on society. It is also helpful for businesses to improve the marketing of their products and services as Giatsoglou et al., [12]. Nowadays, comments on social media are posted in mixed languages. Therefore, the sentiment analysis on mixed languages is getting more attention for extracting information about the expressions Dashtipour at al., [13].

## 2.2 Challenges in ASB Analysis

A huge amount of data is created on social media platforms on daily basis. It is quite challenging to scrap this textual data to prepare it for ASB analysis. Next, this data is manually labeled or labeled with the help of lexical resources, which contains words dictionary with negative or positive meaning words collected from desired languages. Informal and code-mixed Roman Urdu-English and Roman Hindi-English languages may be viewed as a lot of the text is created on social media platforms in the South Asia region. These are the languages with little resources and have the leverage of writing style that one can chose different words and their combination for writing a sentence having same meaning, therefore performing NLP tasks like sentiment analysis on these languages becomes very difficult. Many researchers used informal code-mixed data for text classification using deep learning models for the text classification of the informal code-mixed data. However, another challenge for researchers is that these deep learning approaches are unable to use small word level embedding for classification of sentiments.

## 2.3 Traditional Approaches for Analyzing/Detecting ASB

The state-of-the-art techniques were proposed for performing sentimental analysis and analyzing/detecting ASB are Dictionary-based and Corpus-based Approachs [11]. These approaches are derived from the Lexicon-based approaches. In the Dictionary-based approach, a dictionary of a fixed set of sentimental words is created and a fixed set of operations is applied to analyze the sequence for sentiment analysis [14]. Dictionary is created by choosing negative, positive or neutral words from desired language. This dictionary can be used for the targeted language only. A method, called Part of Speech (POS) tagging is used to assign grammatical sense to each word, such as Adjective, Verb, Noun etc. The POS tagging makes it easier to detect an emotion or sentiment in a sequence after removing the stopwords. As languages evolved with the passage of time and these kind of approaches are unable to handle a mixed languages due to the limitation of performing the monolingual analysis only. Corpus based approach is further divided into two techniques, such as Statistical or Semantic approach. In Semantic approach, the whole document is classified as an opinion. However, in Statistical

approach, every single sequence is classified with a sentiment value based on some fixed rules of classification. Due to dictionary dependency, These approaches cannot predict sentiment value and can even ignore a sentiment completely if a word or a combination of words is not maintained in dictionary. Furthermore, they require to maintain the dictionary regularly to avoid loss of sentiment value.

## 2.4 Machine Learning-based Approaches

Multiple Machine Learning-based approaches have been used for ASB detection on code-mixed content of the social media to extract useful information. Mukund and Srihari et al., [15] performed sentiment analysis on code-mixed content of Urdu-Latin script and English language using Structural Corresponding Learning (SCL) that deals with domain adaptation (data distributions vary on training and testing sets). The authors applied the Part of Speech (POS) tagging on the given dataset and they used two oracles, one for the Urdu language to English language translation and other for incorporating the spelling variations. Similarly, Chen et al., [16] proposed Adversarial Deep Averaging Network (ADAN) to transfer knowledge learned from highly resourceful labeled data of English language to produce results for poorly resourceful unlabeled data of Arabic and Chinese language for sentimental analysis. Experimental study reveals that ADAN performs better on many baselines including domain adaptation models, cutting-edge cross-lingual text categorization techniques and competitive Machine Translation (MT) baselines.

## 2.5 Deep Learning-based Approaches

Nowadays, the deep learning approaches have been widely used for performing the sentimental analysis of code-mixed languages and achieved a high accuracy in the results. For example, Kumar et al., [17] used a deep learning based approach to perform the code-mixed sentiment analysis on Hindi-English languages by crafting a feature network based on the sentence vector. Authors used two different Bidirectional Long Short Term Memory (BiLSTM) Networks. One BiLSTM Network examines the sentiment of a sentence as a whole, whereas, the second BiLSTM Network uses a learning algorithm to concentrate on the specific sub-words that carry the sentence's sentiment. the proposed

method achieved F1-score of 0.827 as well as an accuracy of 83.54%. Similarly, Sabri et al., [18] applied BiLSTM networks on a dataset of 3640 tweets with Persian and English coding that was gathered using the Twitter API. Every tweet was subsequently assigned its appropriate polarity score, as well as, the polarity scores of all these data were learned using neural classification models. The code-mixed words in the intended texts were translated using the proposed models' Yandex and dictionary-based translation approaches. To further represent the data, the authors used pretrained BERT embeddings. The proposed models, accuracy on the data was 66.17%, and its F1 value was 63.66.

Karim et al., [19] used three feature learners (cascades) in the proposed deep learning model, named as Multi-cascaded model (McM), for the sentiment classification of informal short text. This model is compared with three other multilingual models, such as ConvNets, AttentionLSTM, and SimpleConv. The proposed model outperformed the other three models in terms of accuracy, achieving 0.69% with fine tuning and 0.68% without it on the multilingual short text. Furthermore, Tho et al, [20] have selected low-resource code-mixed language, Bahasa Indonesian and Javanese languages for the study. Using Google Machine Translation, the input dataset is first translated to English. A lexicon-based sentiment analysis technique applied on Tto English lexicon label datasets, Sentiwordnet and VADER. Additionally, a Sentence-BERT model trained and used to classify the English-translated input text. The dataset for this investigation is divided into positive and negative categories. The experimentation showed that the collective Google machine translator and Sentence-BERT model achieved average accuracy 83%, average precision 90%, average recall 76%, and average F1 Score 83%. However, none of the above-mentioned contributions provide the sentimental and ASB analysis of the Roman Urdu-English and Roman Hindi-English languages, which is the scope of this thesis.

Chapter 3

# Research Methodology

This chapter provides a research methodology proposed in this thesis for detecting ASB for the multi-lingual models.

## 3.1 General Architecture

We provide a general architecture, capturing every step of our proposed research framework in Figure 3.1. Our proposed framework starts with the acquisition of the data and ends with the detection of the ASB.

We start with acquisition of the code mixed Roman Urdu-English and Roman Hindi-English datasets. These datasets are then preprocessed using different preprocessing techniques, such as lowercasing, removing nulls and outliers, stop word removal, average word count and are prepared for the further experimentation. Then, the XLM-R and mBERT models are used for the sentiment classification. Initially these models are trained with the pre-processed data. However, for improving the accuracy of the results, they are trained again after finetuning their hyperparameters. Next, we use the accuracy metrices for evaluating models. Finally, the ASB classification is done using Negative Words Polarity score, which is based on classification of the *negative/positive* content. The negative content is processed to find out the polarity of negativity. If Negative Words Polarity score reaches a threshold value, the sequence is classified as ASB.

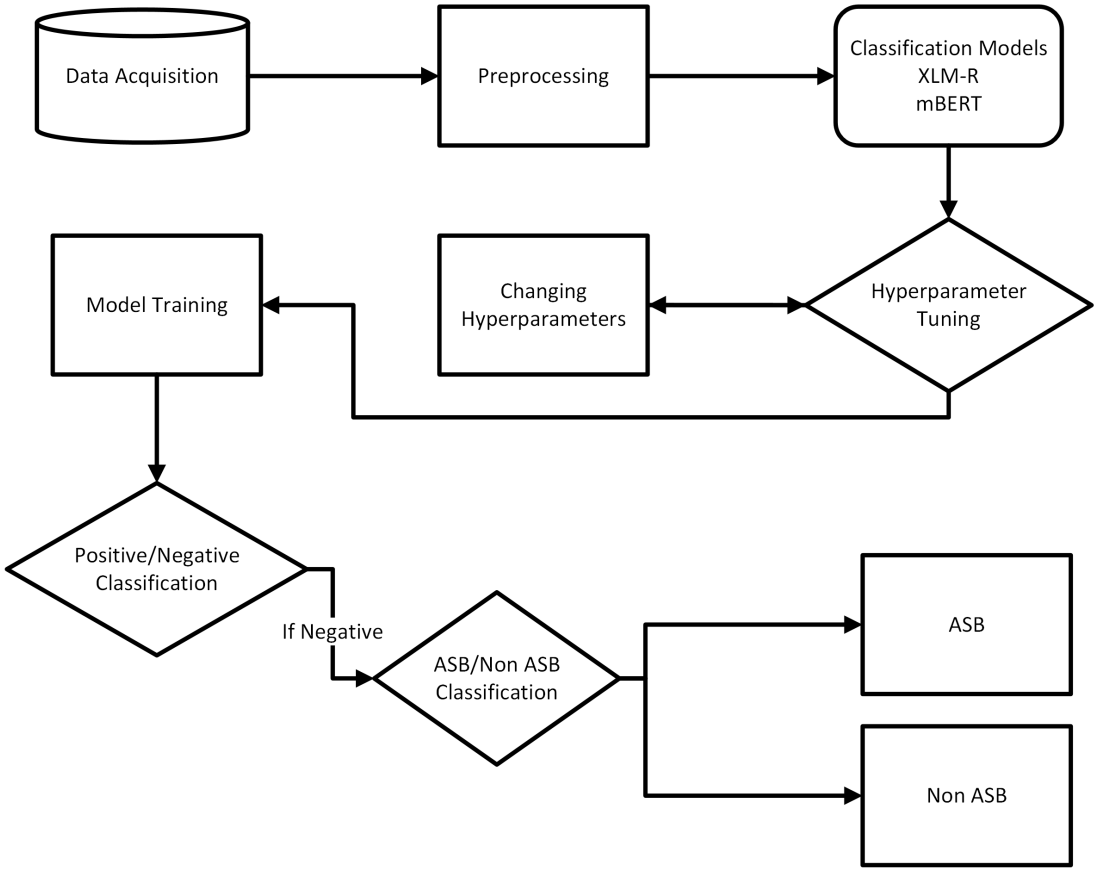


**Figure 3.1:** General Architecture

## 3.2 Dataset Acquisition

The experimentation is carried out on two different datasets, i.e., Roman Urdu-English and Roman Hindi-English datasets. Roman Urdu-English dataset is publicly available at [21, 22] and Roman Hindi-English dataset is shared by Bohra et al., [23] on request for educational purpose only. Roman Urdu-English dataset contains code-mixed Roman Urdu and English tweets data and Roman Hindi-English consists of code-mixed Roman Hindi and English YouTube comments data. Both datasets contains the monolingual and multilingual sequence of text, i.e., Roman Urdu and English and Roman Hindi and English. Moreover, the Roman Urdu-English dataset contains three sentiment classes and Roman Hindi-English dataset contains two sentiment classes. The total number of sequences with respect to a sentiment class is shown in Table 3.1.

| Class Label | Roman Urdu-English | Roman Hindi-English |
|---|---|---|
| *Positive* | 6,013 | 2,914 |
| *Negative* | 5,286 | 1,661 |
| *Neutral* | 8,929 | - |

**Table 3.1:** Class Label Distribution

## 3.3 Data Preprocessing

Both datasets are preprocessed separately, and various data operations are applied to preprocess the data for experimentation. Initially all the text sequences are lower cased and null or one word-based sequences are removed from the dataset. Next, outliers are removed and sentiment classes are finalized.

| processed | length | num_words | words_not_stopwords | avg_word_length |
|---|---|---|---|---|
| sai kha ya her kisi kay bus ki bat nhi hai lak... | 97 | 25 | 19 | 3.000000 |
| sahi bt h | 9 | 3 | 3 | 2.333333 |
| kya bt hai | 10 | 3 | 1 | 2.000000 |
| wah je wah | 10 | 3 | 3 | 2.666667 |
| are wha kaya bat hai | 20 | 5 | 4 | 3.250000 |

**Figure 3.2:** Roman Urdu-English after Preprocessing

| processed | length | num_words | words_not_stopwords | avg_word_length |
|---|---|---|---|---|
| knowing ki vikas kitna samjhata hai priyanka a... | 124 | 25 | 17 | 4.705882 |
| i am muhajir aur mere lye sab se pehly pakist... | 188 | 41 | 35 | 3.828571 |
| doctor sab sahi me ke phd in hate politics wa... | 157 | 31 | 28 | 4.285714 |
| poore desh me patel obc me aate hain sirf gujr... | 256 | 49 | 38 | 4.631579 |
| sarkar banne ke bad hindu hit me ek bhi faisla... | 136 | 26 | 20 | 4.700000 |

**Figure 3.3:** Roman Hindi-English After Preprocessing

Since in ASB, the negative sentiment is more important, therefore, to simplify the analysis, the *Neutral* class in Roman Urdu-English is skipped and stop words are removed from the sequences for both datasets to achieve high accuracy. In addition, we computed

some other useful parameters, such as the total number of words, the number of stop words, the number of not stop words and the average word length in a sequence. Finally, the class labels are encoded from *Positive* and *Negative* to 1 and 0. A few samples of sequences from the Roman Urdu-English and Roman Hindi-English datasets are shown in Figures 3.2 and 3.3, respectively.

## 3.4 Classification Models

We use most recent multilingual deep learning-based classification models i.e., mBERT [24] and XLM-R [21]. The mBERT model implementation is based on BERT base and XLM-R is built on BERT Large. BERT base has a total number of 12 layers of transformers with 768 hidden layers. BERT large uses dynamic masking and avoids the same masking for training examples in each epoch. Both models are pre-trained, with BERT Large for a longer duration than BERT base. In comparison to BERT base, BERT Large consists of 24 layers of transformers. In NLP, the transformer is an architecture to perform sequence-wise task and is consist of encoders and decoders. Encoders and Decoders are stacked together in a transformer and encoder consists of Multi-Head Attention and a Feed Forward Neural Network. Also, decoder contains an extra Masked Multi-Head Attention along with Multi-Head Attention Feed Forward Neural Network. Attention gives the context score for a word based on other words in a sequence and is computed using the following equation.

$$Attention(Q, K, V) = softmax\left(\frac{Qk^T}{\sqrt{d_k}}\right)v \tag{3.4.1}$$

The attention mechanism receives three primary inputs namely queries (Q), keys (K) and values (V). The word embeddings or the output of the previous encoder in the stack is multiplied by three separate matrices that are trained throughout the training phase to form these vectors. The attention score, which instructs the model where to focus while encoding a word, must first be computed after the vector has been calculated. By adding up the dot products of the query and each additional key vector in the sequence, the score is calculated. As once score has been calculated, employ the softmax function to get the weights on the value vectors by dividing each score by the square root of the dimension of the key vectors. Finally, add the weights associated with each value vector

to form the output of the self-attention layer at this place.

Figure 3.4 presents a BERT classification for a single sequence. BERT masks the sequence partially and predicts the value of masked token with the context of non-masked tokens. Special character <SEP> is used to denote the end of sequence and <CLS> is used to encode the whole sequence. Consider the sequence from Roman-Urdu data set, i.e., "kya bt hai", it is embedded word by word by encoders. From the sequence, E1 is the encoded representation of the first word i.e., "kya". Moreover, the final representation of the first word from the sequence is T1. Next, classification layer using the Gaussian Error Linear Unit (GELU) as an activation funtion. GELU is a non-convex, non-monotonic function in comparison with other activation functions. Finally, the classification layers assigns the label *Negative* or *Positive* to each sequence as shown in Figure 3.4.

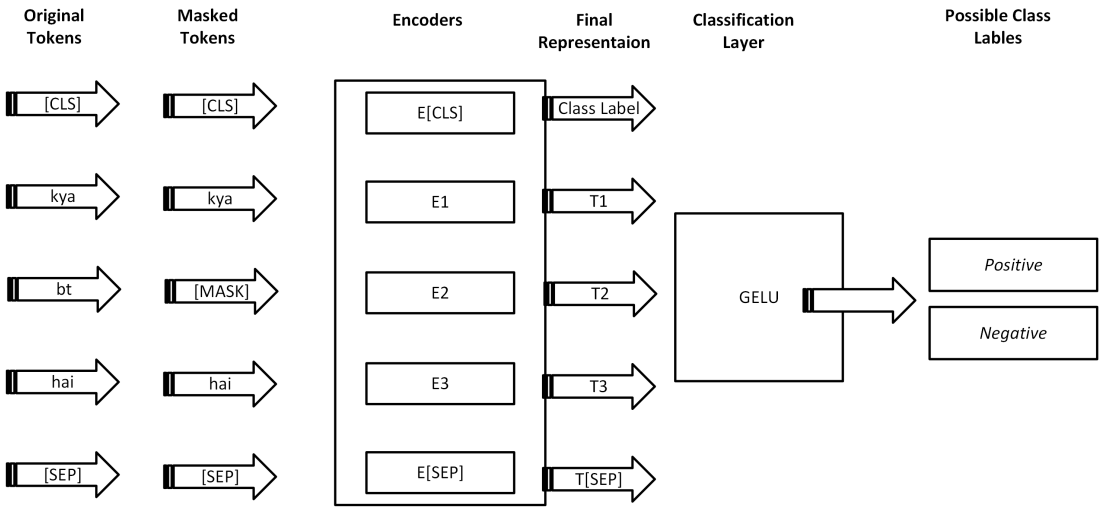


**Figure 3.4:** BERT Classification for a Single Sequence

## 3.5 Hyperparameter Tuning

Hyperparameters are tuned to get better results for both models on both datasets (i.e., Roman Urdu-English and Roman Hindi-English). To find the best parametric values, multiple combinations of parametric settings are tried. The best found combination of hyperparameters is listed for all four experiments (which are XLM-R with Roman Urdu-English, XLM-R with Roman Hindi-English, mBERT with Roman Urdu-English

and mBERT with Roman Hindi-English dataset) in Table 3.2. Models are trained on code-mixed datasets after tuning the hyperparameters.

| **Model** | **Dataset** | **Number of Epochs** | **Batch Size** | **Learning Rate** |
|---|---|---|---|---|
| XLM-R | Roman Urdu-English | 8 | 8 | 1e-05 |
| mBERT | Roman Urdu-English | 8 | 8 | 1e-05 |
| XLM-R | Roman Hindi-English | 7 | 8 | 1e-05 |
| mBERT | Roman Hindi-English | 8 | 8 | 1e-05 |

**Table 3.2:** Hyperparameters

Chapter 4

# Results and Discussion

This chapter provides details of the environmental setup for analyzing/detecting ASB and results of the experiments, such as, time consumed for the model training and the accuracy matrices of results.

## 4.1 Environmental Setup

Python is used as a Coding language for experimentation on Google Colaboratory Environment (Colab) with Graphics Processing Unit (GPU) enabled. Many third-party libraries, such as 'pandas', 'numpy', 'torch', 'seaborn', 'tqdm', 'sklearn', 'transformers' and 'logging' are included for data profiling, preprocessing, visualization, error logging, model training and for computing the result metrics.

### 4.1.1 Google Colaboratory

Google Colab environment is a web-based Integrated Development Environment (IDE) that consists of Jupyter notebooks running on the cloud and is highly integrated with Google Drive, making it easier to implement, access, share and collaborate with other researchers/developers on the same project. Therefore, it completely removes the system dependency with the drive integration. A researcher/developer can place his/her data on Drive and can process using Colab anywhere. The IDE requires zero configuration, supports thousands of third-party libraries, can markdown cells for explaining the experiment steps and provides Central Processing Units (CPU's), GPU's and Tensor Processing Units (TPU's) with multi sessions at a time.

## 4.2 Model Training

In this step, the pretrained XLM-R and mBERT sequence classification models are trained on both Roman Urdu-English and Roman Hindi-English datasets acquired from the sources described in Chapter 2.2. After training on two different datasets, complete experimentation produces four trained model (i.e., XLM-R on Roman Urdu-English and on Roman Hindi-English, and mBERT on Roman Urdu-English and Roman Hindi-English). Moreover, training these large models from scratch on small datasets cause overfitting, which cause modeling error that typically occurs when any function relates too close to a specific set of data, therefore the pretrained models are used purposely.

### 4.2.1 Training XLM-R

Two XLM-R models are trained one for each datasets with and without finetuning the hyperparameters. Models are trained multiple times to find the best parametric combination values for hyperparameters. The final versions of XLM-R with Roman Urdu-English and XLM-R with Roman Hindi-English consumed 1.75 and 1.15 hours of training time, as given in Table 4.1, on GPU.

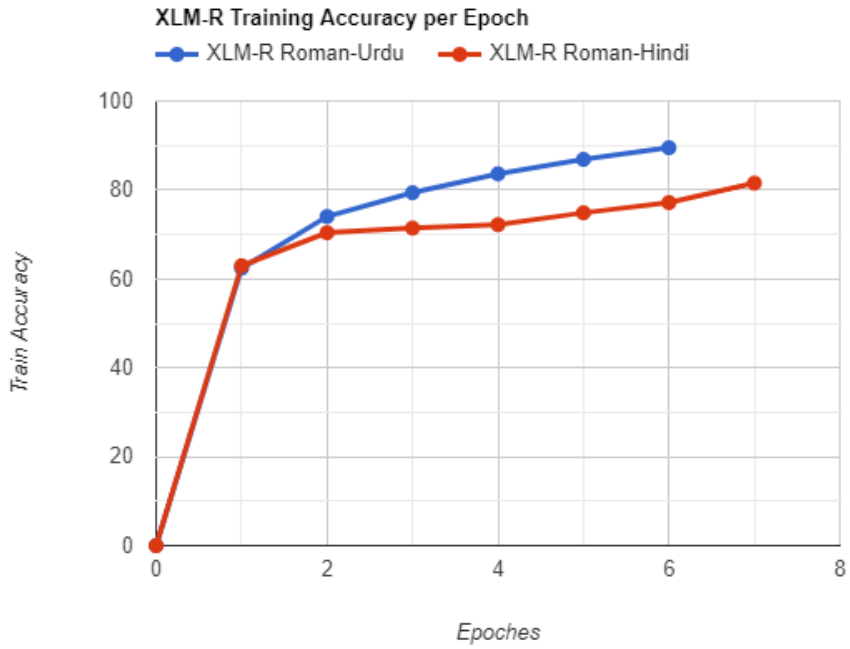


**Figure 4.1:** XLM-R Train Accuracy

To minimize the loss on each epoch *crossentropy loss function* is used for both datasets. Accuracy and loss optimization per epoch for both datasets are given in Figures 4.1 and 4.2, respectively, where x-axis in both figures represents the number of epochs. The y-axis in Figure 4.1 represents the train accuracy from 1 to 100 in percentage, whereas the y-axis in Figure 4.2 provides the train loss from 0 to 1. Initially, at Epoch = 1, the trained model achieved almost same accuracy on both datasets as shown in Figure 4.1. However, for Epoch = 2, accuracy of XLM-R with Roman Urdu-Rnglish got a slight rise and maintained the rise till the end of training. Therefore the XLM-R performed better for Roman Urdu-English than Roman Hindi-English.

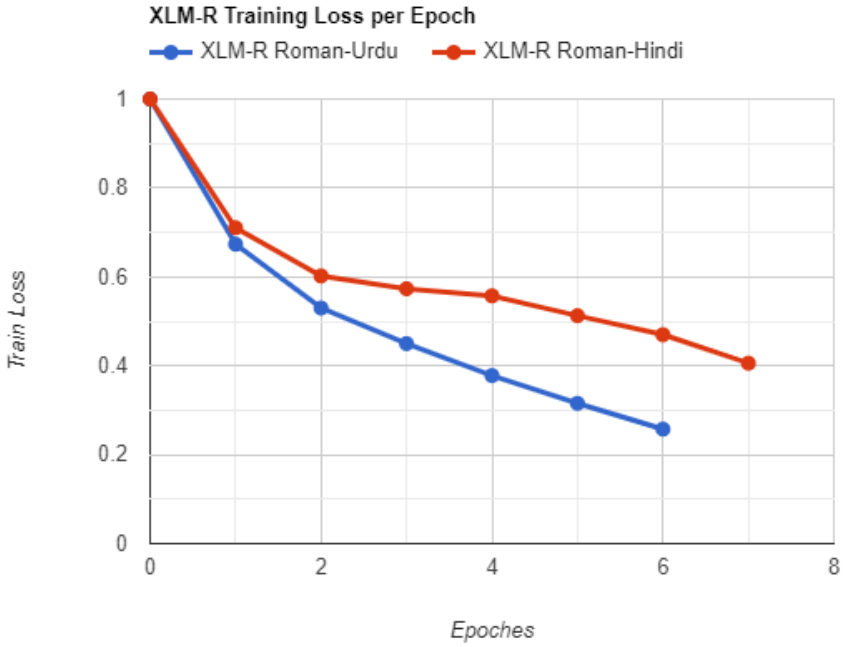


**Figure 4.2:** XLM-R Train Loss

Similarly, at Epoch=1, there was a slight decrease in the training loss for Roman Urdu-English as compared to Roman Hindi-English for the XLM-R model. For Epochs > 1, the XLM-R training loss kept on decreasing for Roman Urdu-English, which concludes that XLM-R performed better for Roman Urdu-English than Roman Hindi-English. From Figures 4.1 and 4.2, it can be clearly seen that the maximum training accuracy and loss optimization (minimum loss) is achieved by XLM-R for the Roman Urdu-English dataset, which is 89.5 percent and 0.25, respectively.

### 4.2.2 Training mBERT

Similarly, two mBERT models are trained, one for each data set, with and without fine-tuning the hyperparameters. The final versions of mBERT with Roman Urdu-English and Roman Hindi-English datasets consumed a little more time in comparison with XLM-R, which is 2.33 and 1.5 hours, respectively, as shown in Table 4.1.

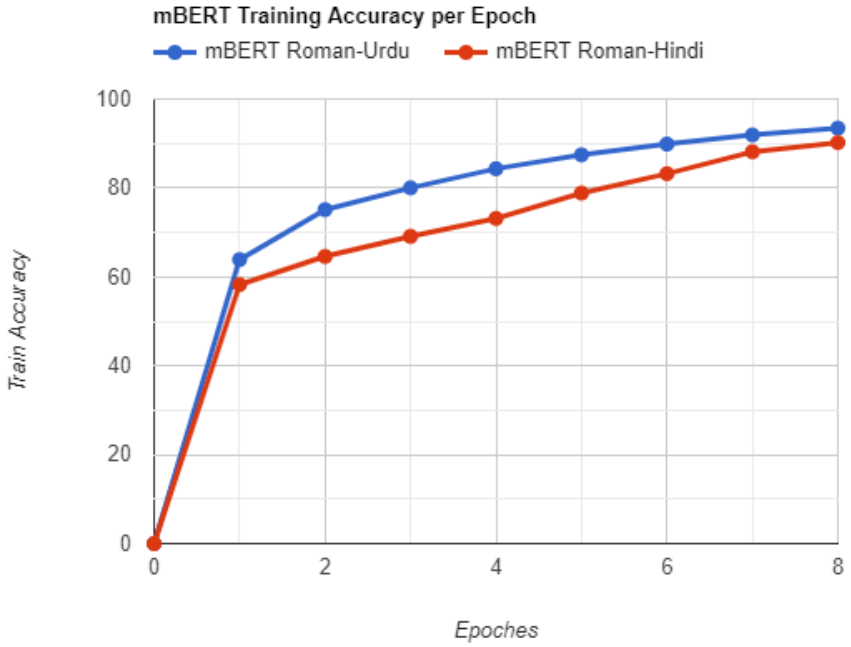


**Figure 4.3:** mBERT Train Accuracy

Accuracy and loss optimization per epoch for both datasets are given in Figures 4.3 and 4.4, respectively, where x-axis in both figures represents the number of epochs. The y-axis in Figure 4.3 represents the train accuracy in percentage, whereas the y-axis in Figure 4.4 provides the train loss from 0 to 1. Figure 4.3, for Epoch=1, the training accuracy mBERT for Roman Urdu-English is better as compared to Roman Hindi-English. For Epochs $>$ 1, the mBERT for Roman-Urdu performed better and accuracy increased gradually in comparison to mBERT for Roman Hindi-English.

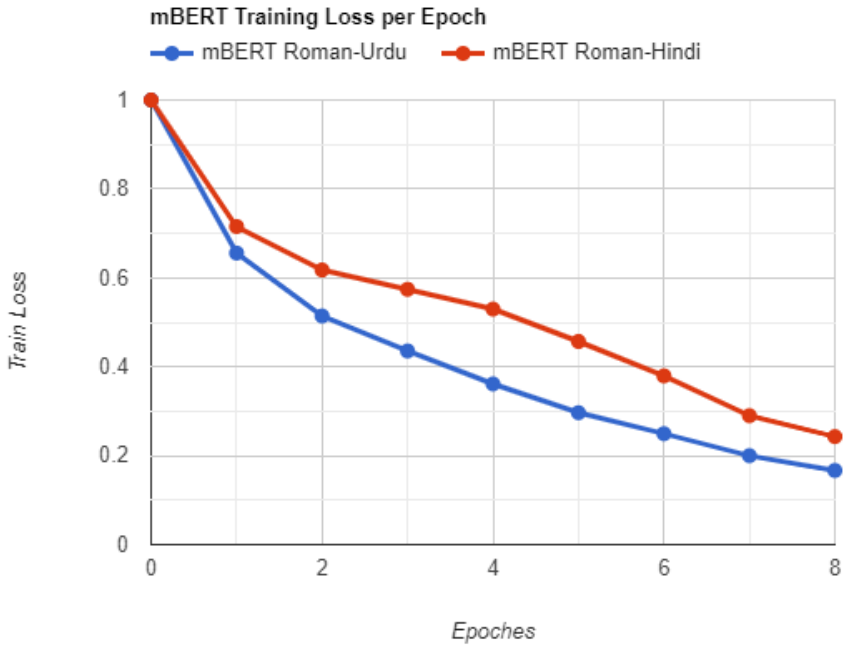


**Figure 4.4:** mBERT Train Loss

Similarly, in the Figure 4.4, mBERT train loss gradually decreased with the increment of epochs for Roman Urdu-English which concludes that XLM-R performed better for Roman Urdu-English than Roman Hindi-English.

### 4.2.3 Training Comparison

Table 4.1 provides a comparison of the trained XML-R and mBERT models with respect to their various parameters, such as training time, accuracy of the results and training loss. Overall, the maximum training accuracy and loss optimization (minimum loss) is achieved by the mBERT model with Roman Urdu-English dataset. Moreover, it consumed a total of 2.33 hours of time to train.

| **Classification Model** | **Dataset** | **Train Time in Hours** | **Ephochs** | **Accuracy in Percentage** | **Loss 0 to 1** |
|---|---|---|---|---|---|
| XLM-R | Roman Urdu-English | 1.75 | 8 | 89.50 | 0.25 |
| XLM-R | Roman Hindi-English | 1.15 | 8 | 81.53 | 0.40 |
| mBERT | Roman Urdu-English | **2.33** | 7 | **93.45** | **0.16** |
| mBERT | Roman Hindi-English | 1.50 | 8 | 90.19 | 0.24 |

**Table 4.1:** Training Stats

Figure 4.5 compares the training accuracy for all trained models, where x-axis represents the number of epochs from 0 to 8 and y-axis represents the accuracy from 1 to 100 in percentage. Here, we can see that XLM-R with Roman-Urdu (Displayed in orange) and mBERT with Roman Urdu-English (Displayed in blue) almost performed similar for every epoch but mBERT with Roman Urdu-English achieved maximum accuracy for epoch 7.

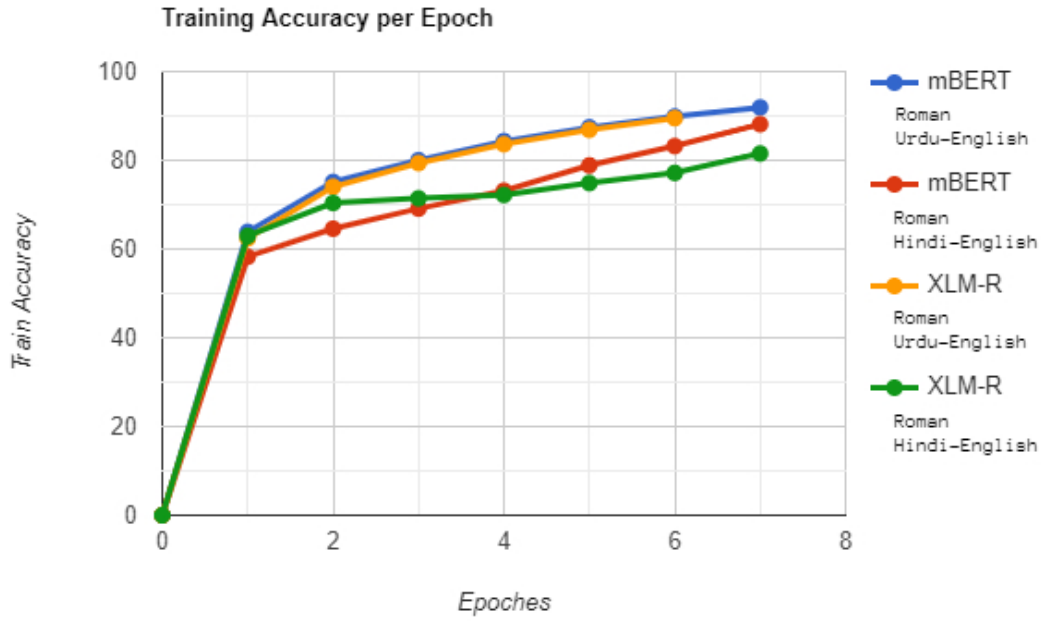


**Figure 4.5:** All Models Train Accuracy

Similarly in Figure 4.6, x-axis represents the number of epochs from 0 to 8 and y-axis

represents the loss from 0 to 1. Here, mBERT with Roman Urdu-English performed well and optimized to 0.16 which is maximum optimization in comparison with all other models.

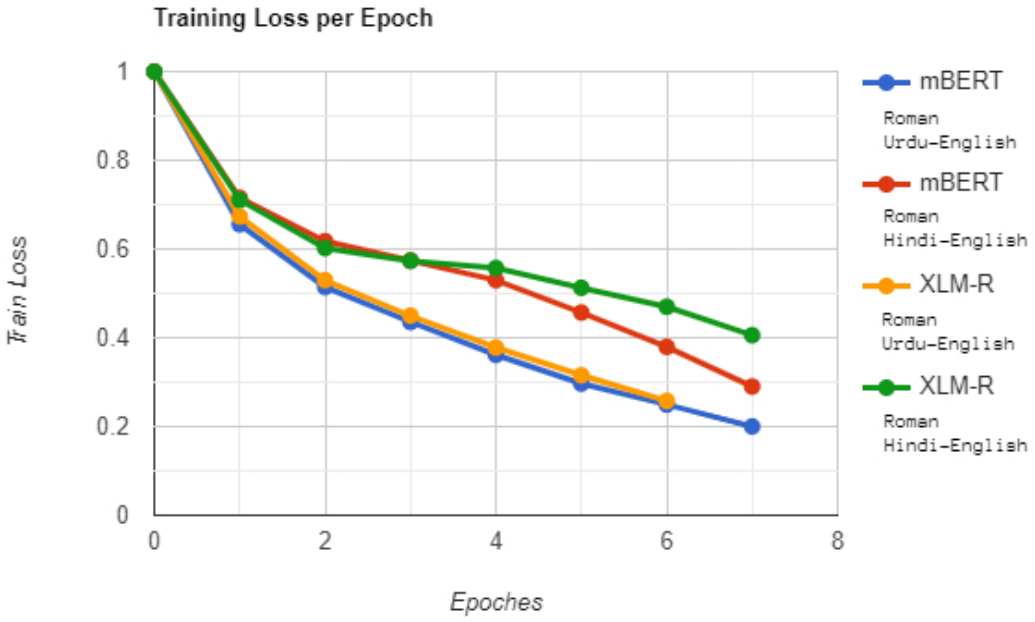


**Figure 4.6:** All Models Train Loss

## 4.3 Accuracy

After training all models, accuracy is calculated with the test data (20% of the total data) on each model for both datasets separately. Models are applied as a binary classifier, where we have only two prediction classes, i.e., *Positive* and *Negative*. All predictions (TP, TN, FP, FN) are recorded to compute the accuracy using Equation (4.3.1).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.3.1}$$

- TP = True Positive
- TN = True Negative
- FP = False Positive

- FN = False Negative

Model prediction first creates a range of the true positive and true negative values for the accuracy metric. Next, on the second segment, it gives false positive and false negative values on the validation data. Accuracy on validation data is given in Table 4.2 for all the models. Maximum validation accuracy is achieved by mBERT model on the Roman Hindi-English dataset.

| **Classification Model** | **Dataset** | **Accuracy in Percentage** |
|---|---|---|
| XLM-R | Roman Urdu-English | 71.48 |
| XLM-R | Roman Hindi-English | 76.59 |
| mBERT | Roman Urdu-English | 74.25 |
| mBERT | Roman Hindi-English | **76.84** |

**Table 4.2:** Validation Accuracy

## 4.4 ASB Classification

mBERT classification model with Roman Urdu-English performed well on validation data and achieved an accuracy of 76.84. As ASB analysis is to carry out on negatively classified sequences only. So, we filtered out negatively classified sequences from the predicted data. mBERT with Roman Urdu-English predicted *negative* class for 430 sequences out of 4000 test data sequences. Further, we performed Negative Words Polarity analysis on negatively classified sequences by our classification models. A dictionary of highly negative and abusive words is created to find the highly abusive words in the classified sequences. Negative Words Polarity is calculated using Equation (4.4.1).

$$NegativeWordsPolarity = \frac{\sum M}{\sum W} \tag{4.4.1}$$

- M = Word matched with Dictionary
- W = Word not stop-word

Equation (4.4.1) always returns a value between 0 and 1. If the value is created to find the highly abusive words in classified sequences then the Negative Words polarity is

calculated using Equation (4.4.1). A threshold value of 0.3 is set to declare a sequence as ASB. For example, a sequence having a polarity of greater than 0.3 is declared as ASB as shown in Figure 4.7.

| | processed | Negative Words Polarity | ASBNotASB |
|---|---|---|---|
| **43** | hum qaum ke liye qurbani de rahe hai | 0.250000 | Not ASB |
| **157** | ye pyar nai ye gangbang hai | 0.250000 | Not ASB |
| **324** | lavay mai jal rhi thi | 0.250000 | Not ASB |
| **322** | mar mar k na is ko words bh ne mil rhy mujy | 0.285714 | Not ASB |
| **288** | bilkul band kro is sho ko | 0.250000 | Not ASB |
| **426** | wo tou sbse pehle kaategi | 0.333333 | ASB |
| **428** | me tu sochti ho ye drama khatam hojayega tu hu... | 0.375000 | ASB |
| **90** | lanat hay is byzameer anker per | 0.250000 | Not ASB |
| **58** | kuch be nahi hoga hum saray america k ghulam hain | 0.285714 | Not ASB |
| **94** | ya rabbe kaynaat musharraf ko tabah o barbad k... | 0.285714 | Not ASB |

**Figure 4.7:** Sequences Classified as ASB or Not-ASB

Figure 4.7 represents the 10 negatively classified sequences by mBERT on Roman Urdu-English with negative words polarity score calculated using Equation (4.4.1), and ASB detection over threshold value of 0.3 as explained in Figure 4.8. ASB classification is based on negative words polarity score which is further dependent on negative words dictionary.

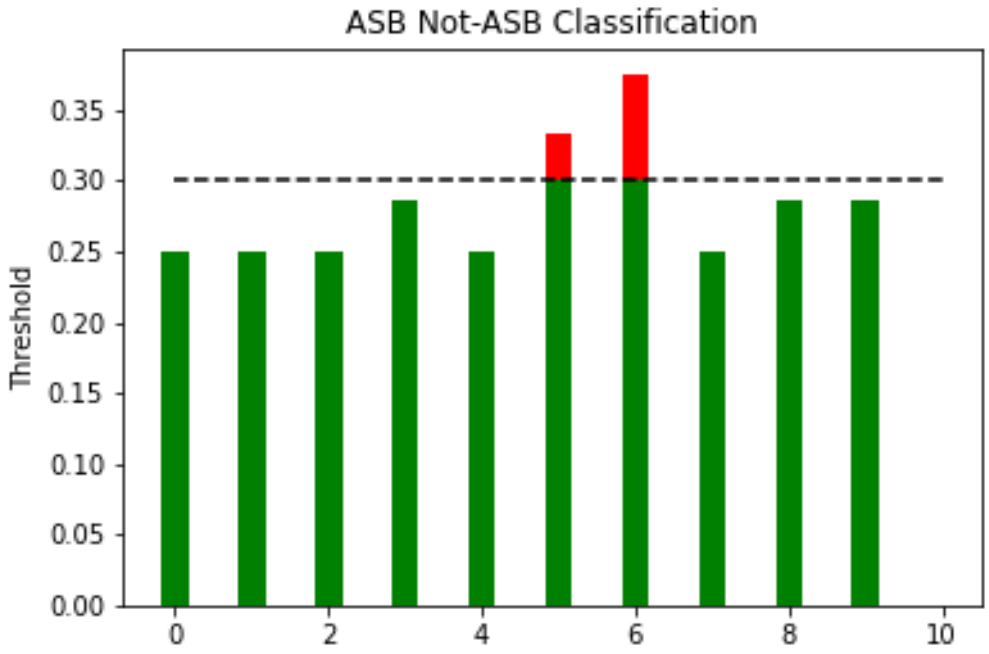


**Figure 4.8:** Threshold Reached

Same 10 sequences (shown in Figure 4.7) are presented in Figure 4.8 using bar chart. In figure 4.8, x-axis represents the number of sequence and y-axis represents the negative polarity score of sequences. Dotted bar-intersecting line represents the threshold value i.e., 0.3. Sequences classified as ASB are colored red above the threshold intersecting dotted line. 2 out 10 sequences are classified as ASB in Figure 4.8.

## 4.5 Discussion

In comparison with other studies, our methodology outperformed on cross-lingual or multi-lingual models with Roman Urdu-English and Roman Hindi-English datasets and achieved the highest accuracy. In [14], accuracy on XLM-R with Roman Urdu-English is very close and have a difference of 0.48% only but in comparison with mBERT on Roman Urdu-English, our proposed approach got a rise of 5.35% in total. the accuracy of XLM-R and mBERT with French-English is very low [24]. This is due to the lack of resources in French-English languages. In comparison with XLM-R and mBERT on French-English, our methodology outperformed in all combinations of models and

language pairs. Accuracy comparison with closely related studies is given in Table 4.3.

| **Study** | **Model** | **Dataset** | **Accuracy** |
|---|---|---|---|
| Sohail et. al., [14] | XLM-R | Roman Urdu-English | 71.00% |
| Sohail et. al., [14] | mBERT | Roman Urdu-English | 69.00% |
| Tiţa et. al., [24] | XLM-R | French-English | 51.00% |
| Tiţa et. al., [24] | mBERT | French-English | 41.00% |
| Proposed Methodology | XLM-R | Roman Urdu-English | **71.48%** |
| Proposed Methodology | mBERT | Roman Urdu-English | **74.25%** |
| Proposed Methodology | XLM-R | Roman Hindi-English | **76.59%** |
| Proposed Methodology | mBERT | Roman Hindi-English | **76.84%** |

**Table 4.3:** Accuracy Comparison

Chapter 5

# Conclusion and Future Work

This chapter summarizes the complete thesis including contribution to NLP community and a brief debate on limitation and future work.

## 5.1 Summary

In this work, we performed sentiment analysis and ASB detection on code-mixed Roman Urdu-English and Roman Hindi-English datasets using pretrained multilingual models, i.e., XLM-R and mBERT. Dictionary of highly abusive words is formatted to detect negative words polarity to classify sequences as ASB. Our experimentation found that mBERT with Roman-Urdu performed well and achieved highest accuracy in comparison with other models. Further, our study revealed that it is more convenient to use pre-trained models for a small dataset to achieve better accuracy. Training large models like mBERT and XLM-R from scratch will cause overfitting on small datasets.

## 5.2 Contribution

Mining of social media platforms has the potential to cite illegal activities that may be useful to security departments, individuals, customers, and businesses. Study achieved a remarkable accuracy score on frequently used code-mixed Roman Urdu-English and Roman Hindi-English languages in South-Asia region.

## 5.3 Limitations and Future Work

Code-mixed languages are most spoken languages on social media. Commonly, a code-mixed language is mixed with English, special characters, emojis and regional language. Every regional language has its own grammar rules and user have freedom to write a word in his/her style. So, it becomes more difficult to create association rules based on other words. Roman Urdu-English and Roman Hindi-English languages are very similar in writing but have lack of resources. In future we will train a single model on both languages together and will increase the size of Dictionary containing abusive words. Combining Roman Urdu-English and Roman Hindi-English abusive word to one single Dictionary to perform classification using one model on both languages will produce exciting results.

# References


[1] S. Yu, L. Jiang, and N. Zhou, "The impact of l2 writing instructional approaches on student writing motivation and engagement," *Language Teaching Research*, p. 1362168820957024, 2020.

[2] "Hindi Speaking Countries kernel description," https://www.worlddata.info/languages/hindi.php, accessed: 2022-08-05.

[3] "Hindi Speaking Countries kernel description," https://www.worlddata.info/languages/hindi.php, accessed: 2022-08-05.

[4] "Urdu Speaking Countries kernel description," https://www.worlddata.info/languages/urdu.php, accessed: 2022-08-05.

[5] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval)," *arXiv preprint arXiv:1903.08983*, 2019.

[6] V. Heap, "Exploring the effects of long-term anti-social behaviour victimisation," *International review of victimology*, vol. 27, no. 2, pp. 227–242, 2021.

[7] A. Millie, J. Jacobson, and E. McDonald, *Anti-social behaviour strategies: Finding a balance*. Policy Press, 2005.

[8] M. Alam and S. U. Hussain, "Roman-urdu-parl: Roman-urdu and urdu parallel corpus for urdu language understanding," *Transactions on Asian and Low-Resource Language Information Processing*, vol. 21, no. 1, pp. 1–20, 2022.

[9] M. H. Shakeel, S. Faizullah, T. Alghamidi, and I. Khan, "Language independent sentiment analysis," in *2019 International Conference on Advances in the Emerging Computing Technologies (AECT)*. IEEE, 2020, pp. 1–5.

[10] C. Blackburn, "The policy-to-practice context to the delays and difficulties in the acquisition of speech language and communication in the first five years," Ph.D. dissertation, Birmingham City University, 2014.

[11] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams engineering journal*, vol. 5, no. 4, pp. 1093–1113, 2014.

[12] M. Giatsoglou, M. G. Vozalis, K. Diamantaras, A. Vakali, G. Sarigiannidis, and K. C. Chatzisavvas, "Sentiment analysis leveraging emotions and word embeddings," *Expert Systems with Applications*, vol. 69, pp. 214–224, 2017.

[13] K. Dashtipour, S. Poria, A. Hussain, E. Cambria, A. Y. Hawalah, A. Gelbukh, and Q. Zhou, "Multilingual sentiment analysis: state of the art and independent comparison of techniques," *Cognitive computation*, vol. 8, no. 4, pp. 757–771, 2016.

[14] M. Sohail, A. Imran, H. U. Rehman, and M. Salman, "Anti-social behavior detection in urdu language posts of social media," in *2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*. IEEE, 2020, pp. 1–7.

[15] S. Mukund and R. K. Srihari, "Analyzing urdu social media for sentiments using transfer learning with controlled translations," in *Proceedings of the second workshop on language in social media*, 2012, pp. 1–8.

[16] X. Chen, Y. Sun, B. Athiwaratkun, C. Cardie, and K. Weinberger, "Adversarial deep averaging networks for cross-lingual sentiment classification," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 557–570, 2018.

[17] V. Kumar and M. Dhar, "Looking beyond the obvious: Code-mixed sentiment analysis (cmsa)," 2018.

[18] N. Sabri, A. Edalat, and B. Bahrak, "Sentiment analysis of persian-english code-mixed texts," in *2021 26th International Computer Conference, Computer Society of Iran (CSICC)*. IEEE, 2021, pp. 1–4.

[19] M. H. Shakeel and A. Karim, "Adapting deep learning for sentiment classification of code-switched informal short text," in *Proceedings of the 35th annual ACM symposium on applied computing*, 2020, pp. 903–906.

[20] C. Tho, Y. Heryadi, I. H. Kartowisastro, and W. Budiharto, "A comparison of lexicon-based and transformer-based sentiment analysis on code-mixed of low-resource languages," in *2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI)*, vol. 1. IEEE, 2021, pp. 81–85.

[21] A. Younas, R. Nasim, S. Ali, G. Wang, and F. Qi, "Sentiment analysis of code-mixed roman urdu-english social media text using deep learning approaches," in *2020 IEEE 23rd International Conference on Computational Science and Engineering (CSE)*. IEEE, 2020, pp. 66–71.

[22] "Roman Urdu Dataset kernel description," https://github.com/Smat26/Roman-Urdu-Dataset, accessed: 2022-09-17.

[23] A. Bohra, D. Vijay, V. Singh, S. S. Akhtar, and M. Shrivastava, "A dataset of hindi-english code-mixed social media text for hate speech detection," in *Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media*, 2018, pp. 36–41.

[24] T. Tiţa and A. Zubiaga, "Cross-lingual hate speech detection using transformer models," *arXiv preprint arXiv:2111.00981*, 2021.