# Sentiment Analysis of code-mixed Telugu-English text using Sequence Models

P.Ashok kumar

E4-CSE @RGUKT-Nuzvid

# ABSTRACT

Feedback of the customers on a product is very important to the developer or manufacturer or company. Nowadays all the social media websites enables code-mixed feedback i.e., giving feedback in their local language. To read and analyze all the comments or the reviews of a product, and generate insights is a hectic task. Especially, India is a diverse country where people speak multiple languages. As people prefer to share opinions in local languages, there are no.of challenges while extracting information from that data. To reduce the complexity and help companies to enhance the customer experience and improve customer service, this model is trained. It gives the overall feedback by analyzing the thousands of customers' feedback.

# Table of Contents

## List of figures

# CHAPTER 1

# INTRODUCTION

Mostly, at present, sentiment analysis on Unicode data of languages like English, Hindi etc is available. As India is a multilingual country, a code-mixed data analyzer has more significance. But the research on analyzing the code-mixed data of Telugu-English languages is comparatively less. So developing a sentiment analyzer of code-mixed data would be more useful. Nowadays, all the social media platforms enables communication in code-mixed text. Code-mixed text means representing a text of one language in another language i.e, writing a sentence in one language but the utterance would be in another language. In this project, code-mixed comments of some videos and posts are collected from various social-media platforms like YouTube and Twitter. Each video or post consists of hundreds of comments. Collecting each and every comment in copy paste method is a hectic task. So to collect those data, Selenium is used. Selenium is an automated tool which makes the task of collecting comments easier. The collected data have to be labeled i.e, categorizing all the comments into different fields. The categorization is performed by giving a tag to every comment as positive, negative or neutral as per the emotion of the comment. With the help of Label studio, all the comments were labeled. Label studio provides a structured platform suitable to tabulate the comment and its respective tag in a understandable manner. Python is an easy and quick programming language to understand and learn. It makes various complex tasks easier with its libraries and in-built methods. In this, python is the language used. The collected data, the comments, comes under sequential data. Bi-LSTM model is one of the the best models to get trained with sequential data in order to give the best performance.

## 1.1 Natural Language Processing Tool Kit (NLTK):

The Natural Language Toolkit (NLTK) is a platform used for building Python programs that work with human language data for applying in statistical Natural Language Processing (NLP). It contains text processing libraries for tokenization, parsing, classification, stemming, tagging and semantic reasoning. It also includes graphical demonstrations and sample data sets as well as accompanied by a cook book and a book which explains the principles behind the underlying language processing tasks that NLTK supports.

The Natural Language Toolkit is an open source library for the Python programming language originally written by Steven Bird, Edward Loper and Ewan Klein for use in development and

education.

NLTK comes with a hands-on guide that introduces topics in computational linguistics as well as programming fundamentals for Python which makes it suitable for linguists who have no deep knowledge in programming, engineers and researchers that need to delve into computational linguistics, students and educators.

## 1.2 Regular Expressions (RE)

The RE(Regular Expressions) module provides a set of powerful regular expression facilities, which allows you to quickly check whether a given string matches a given pattern (using the match function), or contains such a pattern (using the search function). A regular expression is a string pattern written in a compact (and quite cryptic) syntax.The functions in this module let you check if a particular string matches a given regular expression.

## 1.3 Recurrent Neural Network (RNN)

RNN works on the principle of saving the output of a particular layer and feeding this back to the input in order to predict the output of the layer.Below is how you can convert a Feed-Forward Neural Network into a Recurrent Neural Network:

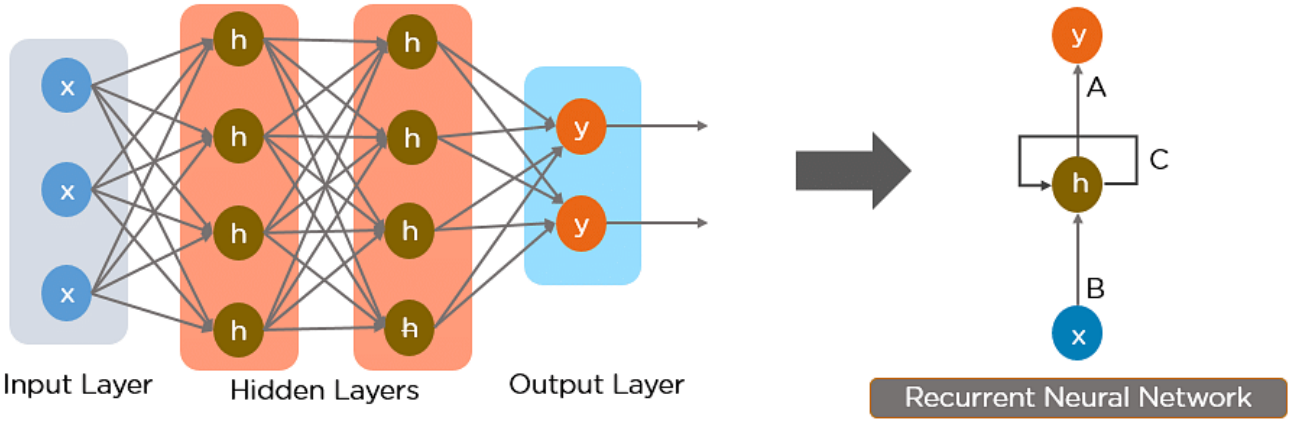



Figure 1: Simple Recurrent Neural Network

## 1.4 Convolution Neural Network (CNN)

A Convolution Neural Network is a type of artificial neural network, which is widely used for image/object recognition and classification. It follows the concept of convolution, which is a matrix that will extract features from the input matrix. It can be used for sentiment analysis also.

## 1.5 Long Short Term Memory (LSTM)

Long Short Term Memory Network(LSTM) is an advanced RNN, a sequential network, that allows information to persist. It is capable of handling the vanishing gradient problem faced by RNN.

## 1.6 Bi-Directional LSTM (Bi LSTM)

Bidirectional recurrent neural networks(RNN) are really just putting two independent RNNs together. This structure allows the networks to have both backward and forward information about the sequence at every time step.Using bidirectional will run your inputs in two ways, one from past to future and one from future to past and what differs this approach from unidirectional is that in the LSTM that runs backward you preserve information from the future and using the two hidden states combined you are able in any point in time to preserve information from both past and future.

# CHAPTER 2

# REQUIREMENTS AND ANALYSIS

## 2.1 Hardware components:

- Processor: 64-bit, quad-core, 2.5 GHz minimum per core
- RAM: 4 GB or more.
- HDD: 20 GB of available space or more.
- Display: Dual XGA (1024 x 768) or higher resolution monitors.
- Keyboard: A standard keyboard

## 2.2 Software components:

- **Python:**Python offers concise and readable code. While complex algorithms and versatile workflows stand behind machine learning and AI, Python's simplicity allows developers to write reliable systems.
- **Numpy:** Numpy is a Python library used for working with arrays.
- **Tensor flow:**TensorFlow is an open-source framework developed by Google researchers to run machine learning, deep learning and other statistical and predictive analytic workloads. Like similar platforms, it's designed to streamline the process of developing and executing advanced analytic applications for users such as data scientists, statisticians and predictive modeler
- **Selenium**:Selenium is an open-source, automated testing tool used to test web applications across multiple browsers. It’s primarily built in Java and supports several browsers and programming languages.The Selenium test suite comprises of four tools. One of the tool is Selenium Web driver.It was the first cross-platform testing framework that could configure and control the browsers on the OS level.It served as a programming interface to create and run test cases.Web driver performs actions on web elements. It supports various programming languages like Java,C#, PHP, Python , among others.

- **Pandas:**Pandas is an open-source library that is made mainly for working with relational or labeled data both easily and intuitively. It provides various data structures and operations for manipulating numerical data and time series. This library is built on top of the NumPy library. Pandas is fast and it has high performance & productivity for users.

- **Seaborn:**Seaborn is a library that uses Matplotlib underneath to plot graphs. It will be used to visualize random distributions.

- **Label Studio:** Label Studio is an open source tool for labeling and exploring multiple types of data. Different types of labeling can be performed with many data formats. Label Studio can also be integrated with machine learning models to supply predictions for labels (pre-labels), or perform continuous active learning.To perform labeling the data file has to be imported and after labeling it can be exported in desired file format.

- **Colaboratory Notebook:** Colaboratory or "Colab" is a product from Google Research. Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education.

- **Windows OS 64-bit.**

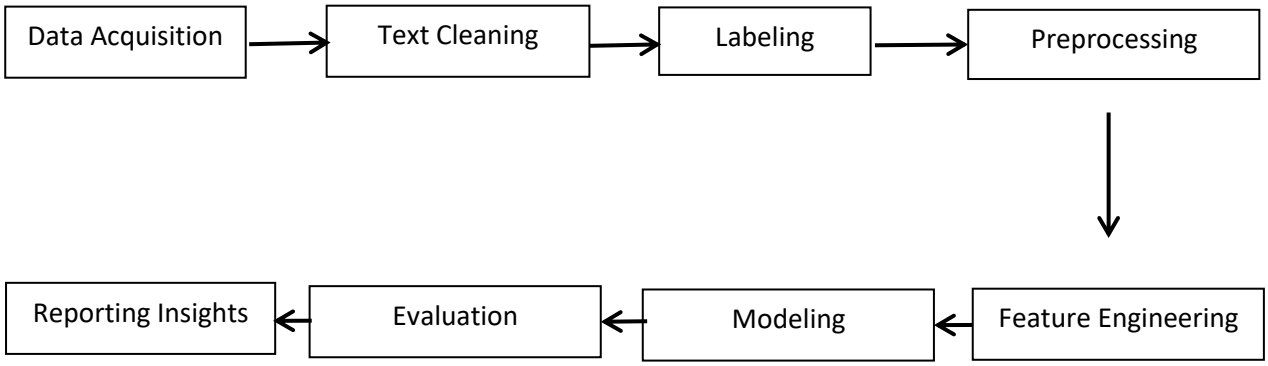
# CHAPTER 3

# PROPOSED MODEL AND FLOW OF THE PROJECT

## 3.1 Proposed model

Deep learning models are suitable to analyze sequential data. The collected code-mixed data would be labeled and preprocessed. The preprocessed data would be divided into training and testing data sets. The model would be trained with the training data set. Based on the knowledge gained during the training phase, the sentiment of a given statement will be analyzed.

## 3.2 Flow of the project




## Data Acquisition:

The data is the comments collected through web scraping with the help of Selenium. Selenium is an open source which provides libraries to support browser automation. The comments were collected from various categories of videos like videos related to politics, news channels, movie reviews, product reviews etc. And publicly available data set was also taken into consideration.

## Text Cleaning:

Cleaning of the text includes the removal of some comments, considering the following constraints:

- Non-code mixed comments.
- comments less than 2-3 words.
- Lower casing the data
- Removing extra spaces
- Removing redundant data and incorrect data(missing values)

## Labeling the data:

Labeling means deciding whether the comment is positive, negative or neutral. For that, Label Studio was used. Label Studio makes the work easy. It provides a user-friendly way to label each comment so that the switching from one comment to another comment is easier while labeling.

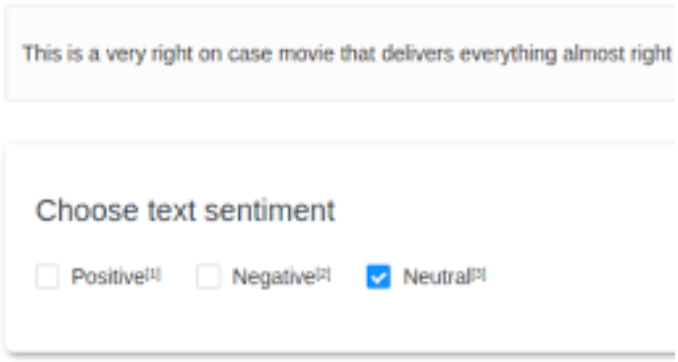



Figure 2:Labeling in Label studio

In this project,double-labeling was performed i.e.,each comment was labeled by two individual members of the team separately in order to take multiple opinions into consideration. After completion of double-labeling, the files were exported and compared. Later, the comments with matched labels were kept and those with unmatched labels were discarded.

## Preprocessing the data:

Preprocessing the labeled data means making it ready in such a format that the model can understand the data, without any information lose.

**Data Cleaning:**

It includes removing punctuation marks, commas, special characters excluding (A-Z,a-z,0-9), making the sentence lower-case and removing stopwords. Stopwords are the words of a sentence which are needed for a human to understand the meaning of the sentence clearly. For example, is, are, at, of a, an, the,etc., are the words which provides proper structure to the sentence but don’t impact the sentiment of the sentence.

**Data Transformation:**

The major part is that the text-data should be transformed into something that the Machine Learning model can understand. For that, Word-Embedding is used. It is a beautiful way of representing the relationship among the words of the sentence. It converts the text-sentence into

an array of vector embedding. Before that each unique word is given with a unique number. Based on the unique numbers vectors' dimension is decided.

## Feature Engineering:

Feature Engineering is a process of extracting meaningful information (insight features and patterns) from the pre-processed data to make it usable for Deep Learning models. Features are the essential inputs for any model.

The features extracted will help the model understand it well and provide efficient results.

## Model Selection:

Different deep learning sequential models are experimented. The outperformed model is selected for the implementation.

## Evaluating the model:

The model is evaluated using some metrics called performance metrics.

- Confusion Matrix
- Accuracy
- Precision
- Recall
- F1Score

## Confusion Matrix:

A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model.

For a binary classification problem, we would have a 2 x 2 matrix as shown below with 4 values:

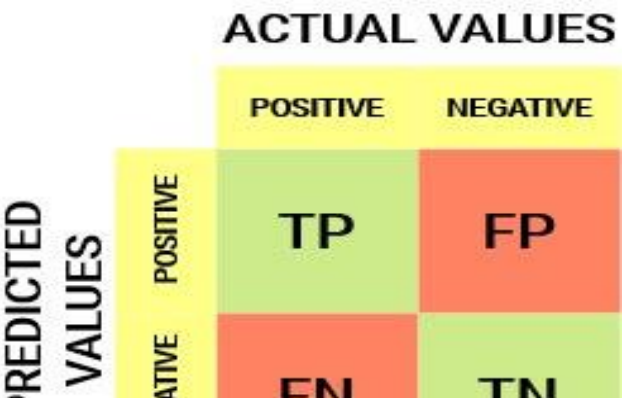



Figure 3: 2x2 Confusion Matrix

**True Positive (TP)**: The predicted value matches the actual value. The actual value was positive and the model predicted a positive value

**True Negative (TN)**: The predicted value matches the actual value. The actual value was negative and the model predicted a negative value

**False Positive (FP)**: Type 1 error. The predicted value was falsely predicted. The actual value was negative but the model predicted a positive value Also known as the Type 1 error

**False Negative (FN)** : Type 2 error. The predicted value was falsely predicted.The actual value was positive but the model predicted a negative value. Also known as the Type 2 error

**Accuracy:** This is how we'll calculate the validation accuracy:

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

**Precision:** Precision tells us how many of the correctly predicted cases actually turned out to be positive.Here's how to calculate Precision:

$$Precision = \frac{TP}{TP + FP}$$

**Recall:**Recall tells us how many of the actual positive cases we were able to predict correctly with our model. And here's how we can calculate Recall:

$$Recall = \frac{TP}{TP + FN}$$

**F1 score:** F1 score is the weighted average of precision and recall

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

Here, we have 3 output labels. So 3x3 confusion matrix has to be used.

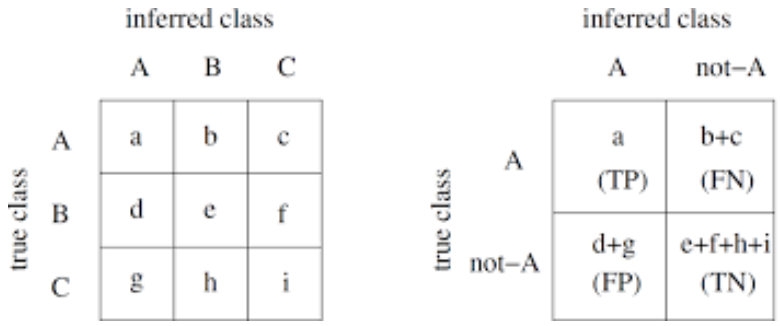



Figure 4: 3x3 Confusion Matrix

For a multi class classification problem, we need to calculate TP, TN, FP, FN for each individual class.

## 3.3 Advantages:

- A faster way of getting insights from customer data.
- The ability to act on customer suggestions.
- Identifies an organization's Strengths,Weakness, Opportunities & Threats (SWOT Analysis).
- As 80% of all data in a business consists of words , the Sentiment Engine is an essential tool for making sense of it all.
- More accurate and insightful customer perceptions and feedback.

## 3.5 Applications:

Business and Organizations:

- Brand analysis
- New product perception
- Product and service bench marking
- Business spends a huge amount of money to find consumer sentiments and opinions.

Social Media:

- finding general opinion about recent hot topics in society.

Ads placements:

- Placing ads based on the user-generated content.

# CHAPTER 4

# IMPLEMENTATION

## 4.1 Implementation of the model

Various models are implemented to get the efficient results.

Experimented models:

Simple_RNN : Accuracy(70%)

LSTM : Accuracy(70%)

CON1D : Accuracy(73%)

Bi LSTM : Accuracy(74%)

Bi-directional LSTM outperformed among all the experimented models.

## 4.2 Bi-Directional Long Short Term Memory (Bi LSTM)

Bi LSTM is the process of making any neural network to have the sequence information in both directions backwards (future to past) and forward (past to future).

In bidirectional, our input flows in two directions, making a Bi-LSTM different from the regular LSTM. With the regular LSTM, we can make input flow in one direction , either backwards or forward. However, in bi-directional, we can make the input flow in both directions to preserve the future and the past information. For a better explanation,

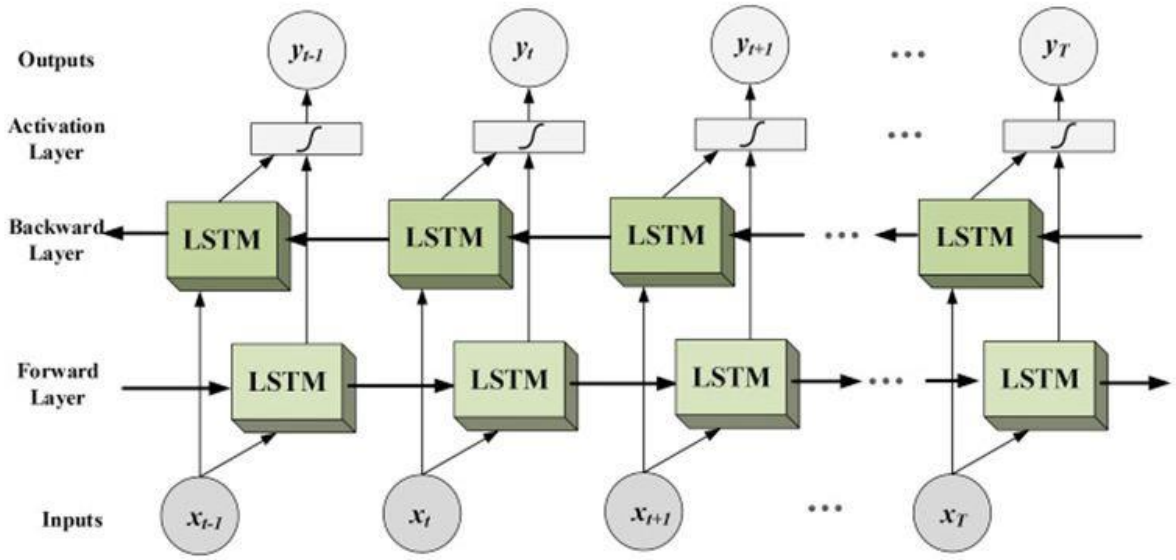



Figure 5 Image for Bi-LSTM

In the diagram, we can see the flow of information from backward and forward layers. Bi-LSTM is usually employed where the sequence to sequence tasks are needed. This kind of network can be used in text classification, speech recognition and forecasting models.

## 4.3 Import all necessary libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import tensorflow as tf
import nltk
from nltk.corpus import stopwords
nltk.download('stopwords')
print(tf.__version__)
import os

os.environ['TF_CPP_MIN_LOG_LEVEL'] = '2'
tf.random.set_seed(0)

from tensorflow.keras.preprocessing.text import Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, Embedding, Dropout, SimpleRNN
from tensorflow.keras.layers import MaxPooling1D
from tensorflow.keras.layers import GRU, LSTM, Bidirectional
from tensorflow.keras.layers import Conv1D, Flatten
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelBinarizer
from tensorflow.keras.optimizers import Adam
from tensorflow.keras.callbacks import EarlyStopping, ModelCheckpoint
import os
from sklearn import metrics
```

4.4 Bi-LSTM Definition

```
def bi_lstm():
    model = Sequential()
    model.add(Embedding(input_dim=vocab_size, output_dim=embedding_dim, input_length=max_len))
    model.add(Bidirectional(LSTM(512, return_sequences=True)))
    model.add(Bidirectional(LSTM(256)))
    model.add(Dropout(0.3))
    model.add(Dense(3, activation='softmax'))
    model.compile(loss='categorical_crossentropy', optimizer=Adam(0.0001), metrics='accuracy')
    return model
```

## 4.5 Experimental settings:

- Training data - 80%
- Testing data - 20%

- Batch size - 128
- Epochs - 20
- Early stopping - 5
- Loss function - categorical_crossentropy
- Activation Function - soft max
- Vocab size - 15000
- Embedding_dim - 50
- Max_len - 100
- Optimizer - Adam

# CHAPTER 5

# RESULT

## 5.1 Performance of the model

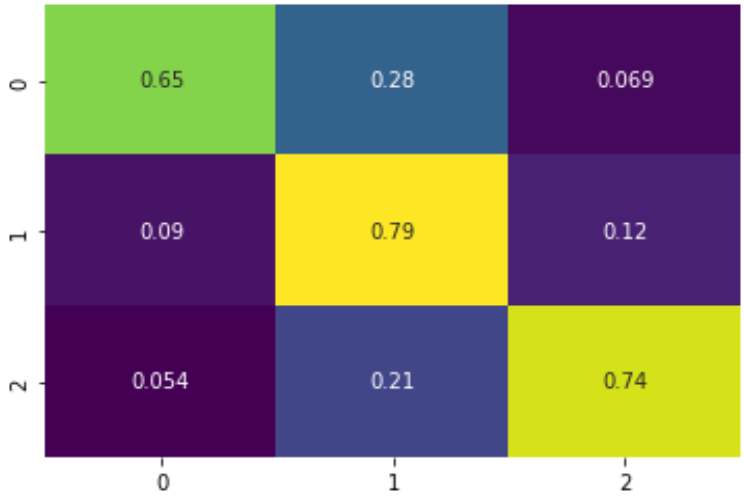



Figure 6:Confusion matrix

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| -1.0 | 0.74 | 0.65 | 0.69 | 3047 |
| 0.0 | 0.72 | 0.79 | 0.75 | 5327 |
| 1.0 | 0.77 | 0.74 | 0.75 | 3858 |
| accuracy | | | 0.74 | 12232 |
| macro avg | 0.74 | 0.73 | 0.73 | 12232 |
| weighted avg | 0.74 | 0.74 | 0.74 | 12232 |

Figure 7:Performance metrics

## 5.2 Graphs

### Accuracy Graph:

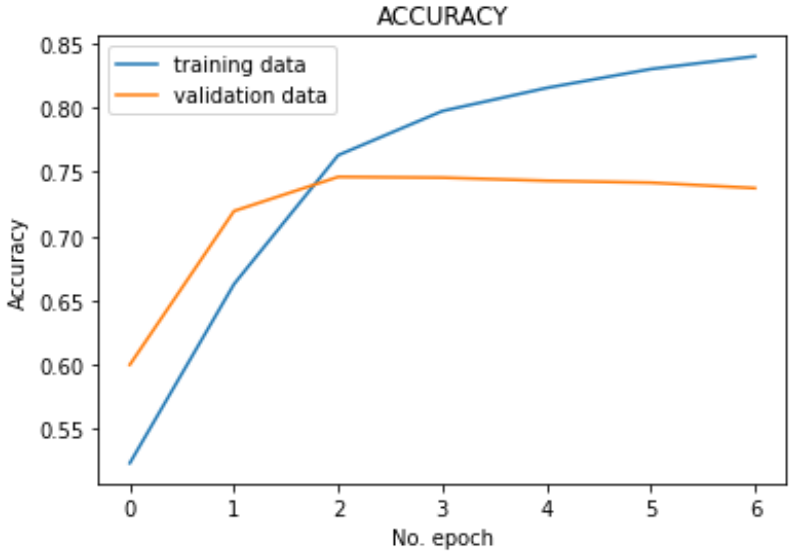



Figure 8: Accuracy vs No.of epoch graph

### Accuracy when model is tested using Training data set:

After one epoch the accuracy is raised to 0.60. The accuracy has raised according to the increase in no.of epochs. Finally it reached to 0.83 approximately after 6 epochs.

### Accuracy when model is tested using Validation data set:

Initially the accuracy was 0.60.After one epoch the accuracy is raised to 0.70. After 2 epochs the accuracy reached to 0.75. From there the accuracy is maintained same with slight difference. Finally after 6 epochs the accuracy was 0.74.

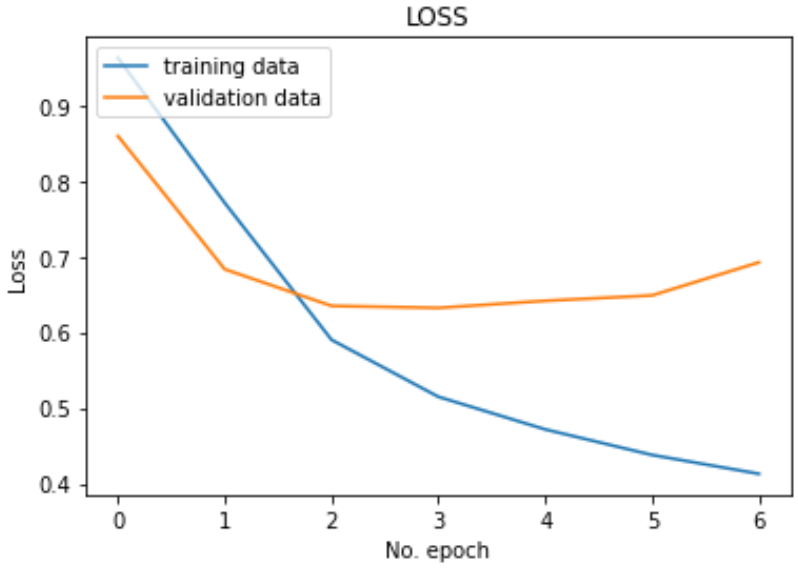



Figure 9: Loss vs No.of epoch graph

## Loss when model is tested using Training data set:

Initially the loss was very high. Gradually it was reduced as the no.of epochs increased. Finally it was reduced to approx 0.42 after 6 epochs.

## Loss when model is tested using Validation data set:

Initially the loss is approx 0.86. Gradually the loss was reduced as the no.of epochs increased. Finally it was reduced to 0.7 after 6 epochs.

# CHAPTER 6

# CONCLUSION

## 6.1 Summary

The code-mixed data was collected from various social media platforms. Text cleaning was performed and the data was labeled. The labeled data was preprocessed. After experimenting with different sequential models, Bi-LSTM model was implemented as it outperformed all the experimented models. After training and testing, the model is able to analyze the sentiment of the given code-mixed text successfully with 74 percent accuracy.

## 6.2 Future work (Model Deployment)

- Usage of multilingual transformer models.
- Model deployment is the process of placing finished ML model into a live environment where it can be used for intended purpose.
- Models can be deployed in a wide range of environments and they are often integrated with apps through an API so they can be accessed by end users from any where.
- This can take an input and return an output that can be used in making practical business decisions.

## 6.2 References


- .https:/aclanthology.org/2021.ranlp-1.86.pdf
- https://machinelearningmastery.com/develop-bidirectional-lstm-sequence-classification-python-keras/
- https://www.analyticsvidhya.com/blog/2022/06/an-end-to-end-guide-on-nlp
- pipeline/#:~:text=Text%20preprocessing%20step%20is%20the,model%20based%20on%20evaluation%20metrics
- https://analyticsindiamag.com/complete-guide-to-bidirectional-lstm-with-python-codes/#:~:text=Bidirectional%20long%2Dshort%20term%20memory(bi%2Dlstm)%20is,different%20from%20the%20regular%20LSTM