INTERNATIONAL CONFERENCE ON “EMERGING TRENDS AND TECHNOLOGIES ON INTELLIGENT SYSTEMS” (ETTIS-2024)

Marathi Text Sentiment Analysis Using Machine Learning

Research Paper Authors:

Deepak Mane

Sarthak Pithe

Hrishikesh Potnis

Soham Nimale (Presenting and Corresponding Author)

Madhur Vaidya

Department of Computer Engineering, Vishwakarma Institute of Technology, Pune, India

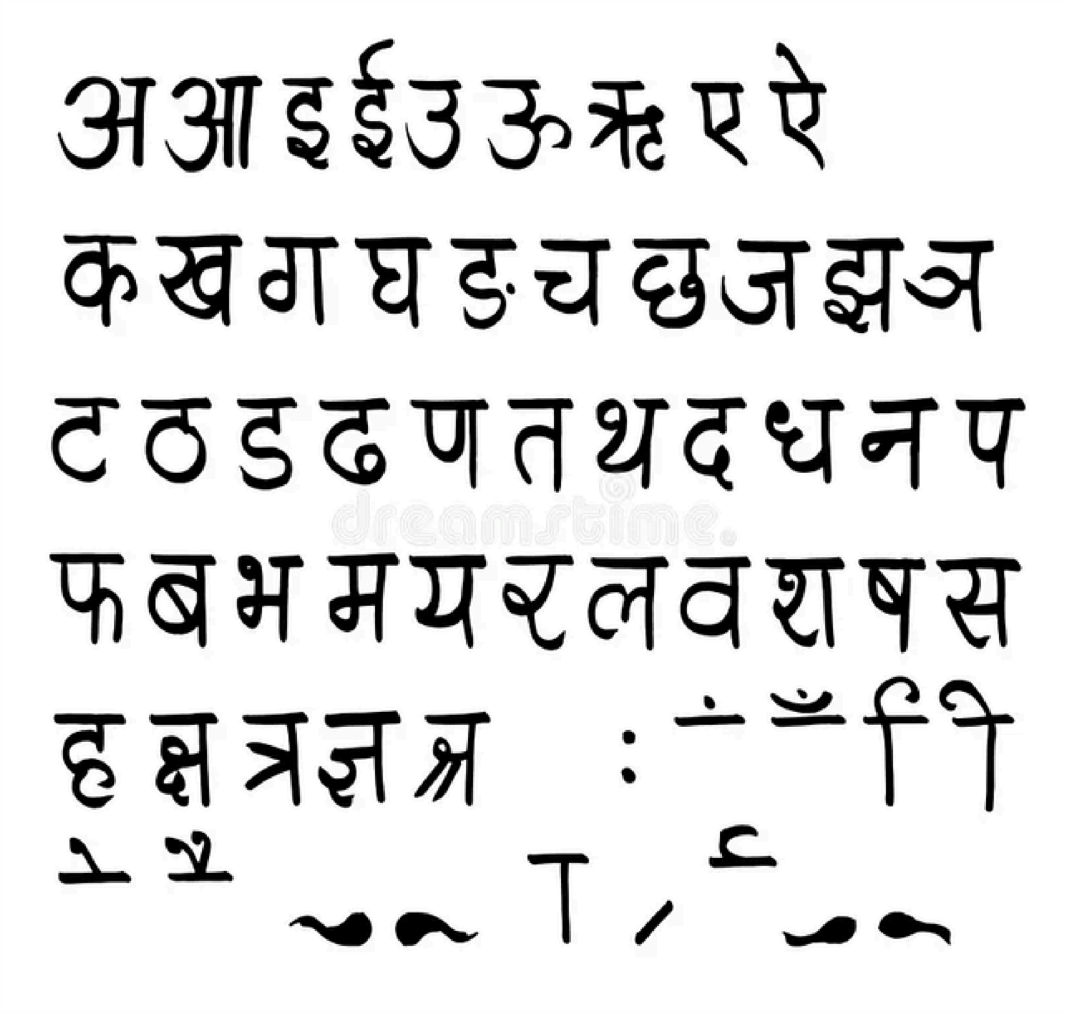
Contents

- Introduction
- Motivation
- Literature Review and Observations
- Problem Statement
- Objectives
- H/W and S/W requirements
- System Flow and Algorithm
- Preprocessing
- Training pipeline
- System Designs
- Model Evaluations
- Novelty of approach
- Conclusion
- Future Scope

Introduction

- Importance of Sentimental Analysis
- Sentimental Analysis - interesting
- Improvement in the quality of blended learning
- Upcoming topic
- Emotions

Motivation



- Marathi - 13th largest native speakers in world, 3rd in India
- Official Language of Maharashtra and Goa
- Little Machine Learning research on Marathi
- Combining - Sentimental Analysis of Marathi

Literature Review

Name	Year	Details	Accuracy	Dataset
Lexicon-Based vs. Bert-Based Sentiment Analysis	**2022**	• Compares sentiment analysis capabilities in Italian using BERT-based and lexicon-based methods. • Suggests lexicon methods for small datasets.	Accuracy 0.73 F1 score 0.67	600 reviews of about six different products
An Effective BERT-Based Pipeline for Twitter Sentiment Analysis: A Case Study in Italian	**2021**	• An approach for sentiment analysis of tweets using a two-step process: transforming tweet jargon into plain text and employing a BERT model pre-trained on plain text for classification.	F1 score 0.7500	Italian tweets

SENTIMENT ANALYSIS OF MIXED CODE FOR THE TRANSLITERATED HINDI AND MARATHI TEXTS	**2018**	Model - Naïve Bayes and Support Vector Machine	F1-score = 0.63	300 Marathi Documents
Cross-Lingual Sentiment Analysis for Indian Languages using Linked WordNets	**2012**	Combined Marathi + Hindi sentiment analysis Model - Naïve Translation Using Lexeme Replacement Accuracy -	97.87 (Words+Senses)	Sentences from blogs+editorals

Base Paper

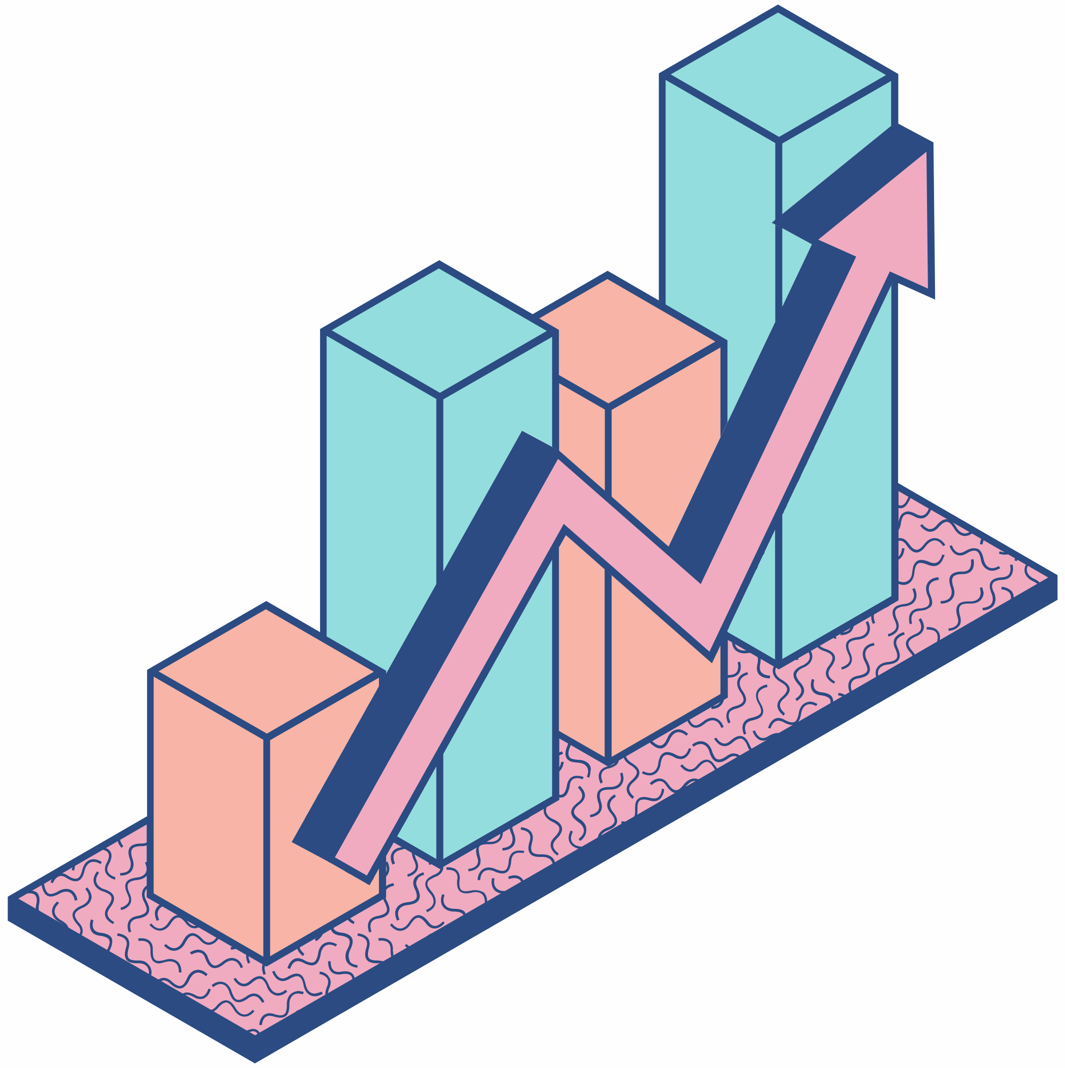
Name	Year	Details	Accuracy	Dataset
L3CubeMahaSent: A Marathi Tweet-based Sentiment Analysis Dataset Atharva Kulkarni, Meet Mandhane, Manali Likhitkar, Gayatri Kshirsagar and Raviraj Joshi	**2021**	Various models. Most successful - BERT (IndicBERT)	84.13 (3-class)	Twitter Dataset

Observations on Literature Survey

- Machine learning unresearched - Marathi
- No standardized dataset
- BERT, LSTM and variants - main approach

Problem Statement

To compare various machine learning models which perform sentimental analysis of Marathi sentences



Objectives

- To compare various machine learning models which perform sentimental analysis of Marathi sentences

- To analyze strengths and weaknesses of various sentimental-analysis models

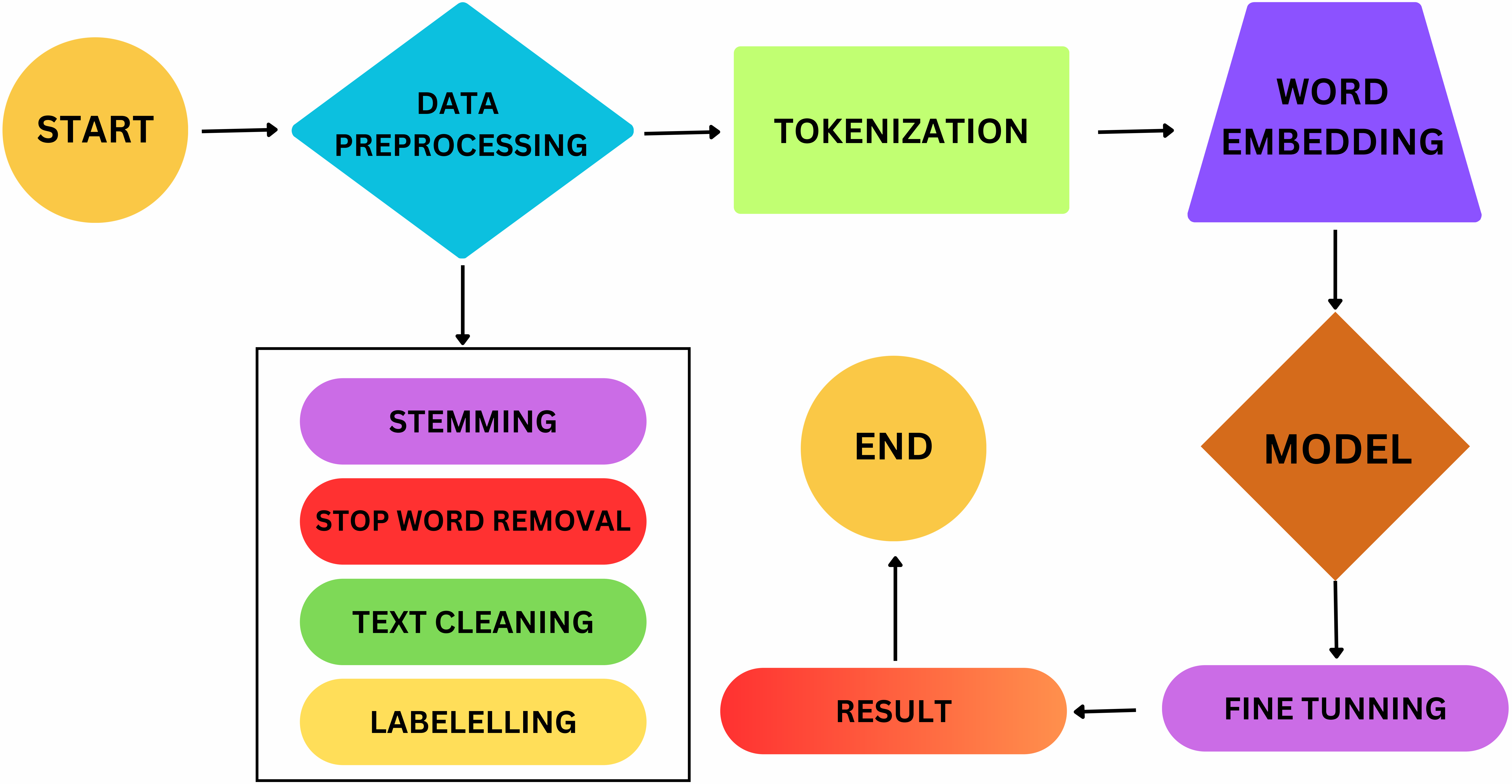
- To modify non-Marathi based models for increasing accuracy of Marathi analysis

- To perform Text Sentiment Analysis in Marathi to understand the current market trends, customer reviews and social media monitoring

Hardware/Software Requirements

H/W	S/W
• 12GB RAM • GPU • 8GB VRAM	• Google colab • Libraries: Tensorflow, sklearn

System Flow


START
DATA PREPROCESSING
TOKENIZATION
WORD EMBEDDING
STEMMING
STOP WORD REMOVAL
TEXT CLEANING
LABELELLING
END
MODEL
RESULT
FINE TUNNING

Algorithm/Pseudo code

- Pre-processing
  1. Input: Text Data (Corpus of sentences)
  2. Output: Cleaned preprocessed data
- Feature Extraction
  1. Input: Preprocessed data
  2. Output: Feature vectors for machine learning
- Model Training
  1. Input: Feature vectors, labeled sentiment data
  2. Output: Trained sentiment analysis model
- Sentiment Prediction
  1. Input: New text input for trained model
  2. Output: Predicted sentiment label (Positive, Negative, Neutral)

Preprocessing

Handling: Punctuations, Conjunctions, English Letters, English Numbers, Marathi Numbers, Stop Words, Pronouns, Special Characters, Stemming, Tokenization.

#BETIBACHAO चा फक्त नारा देऊन उपयोग नाही. महिला अत्याचाराच्या आरोपींना वेळीच कठोर शासनही झालं पाहिजे. पण 'गहुंजे' खटल्यात अक्षम्य दिरंगाई झाली आहे. महिला सुरक्षेबाबत तत्परतेचे दावे फोल ठरले आहेत. 'गहुंजे'च्या आरोपींना फाशी होणेबाबत सरकारने तातडीने कायदेशीर पावले उचलली पाहिजे.
PIC.TWITTER.COM/X6GOZJM6TK

फक्त नारा देऊन उपयोग नाही महिला अत्याराच्या आरोपींना वेळीच कठोर शासनही झालं पाहिजे गहुंजे खटल्यात अक्षम्य दिरंगाई महिला सुरक्षेबाबत तत्परतेचे दावे फोल ठरले त गहुंजे च्या आरोपींना फाशी होणेबाबत सरकारने तातडीने कायदेशीर पावले उचलली पाहिजे

Training Pipeline

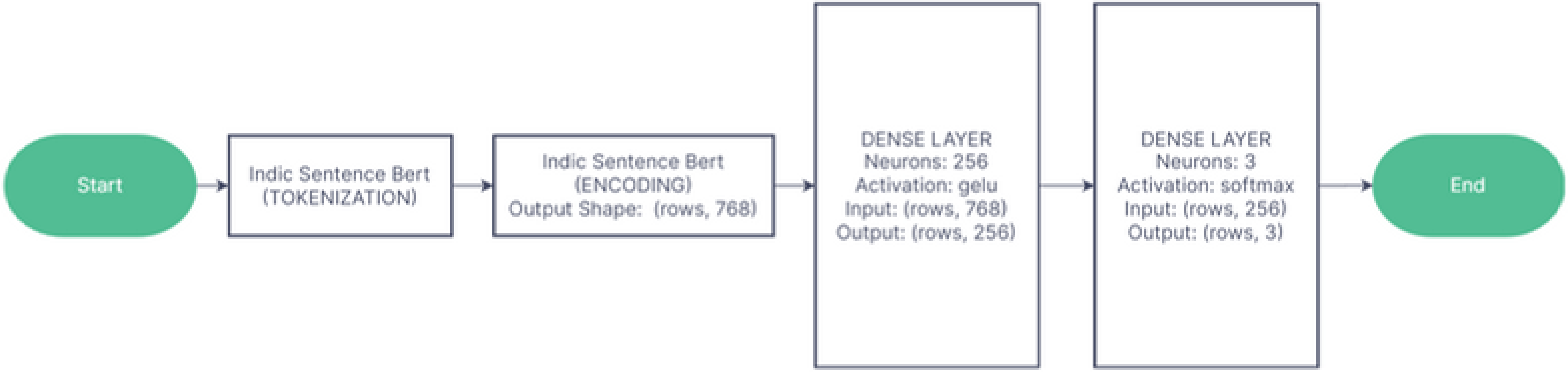
Text converted to Tokens, then further encoded using One-Hot-Encoding/Index Encoding /Pre-Trained Model Encoding.

Then we produced embeddings through Default Keras embeddings/Pre-Trained Model Embeddings, Other available ones.

Finally, we created multiple Models for final Classification purposes, most of them consisting of Dense layers + Hyper-parameters Tuning

Model Architecture

INDIC (SENTENCE) BERT BASED MODEL



Start

Indic Sentence Bert
(TOKENIZATION)

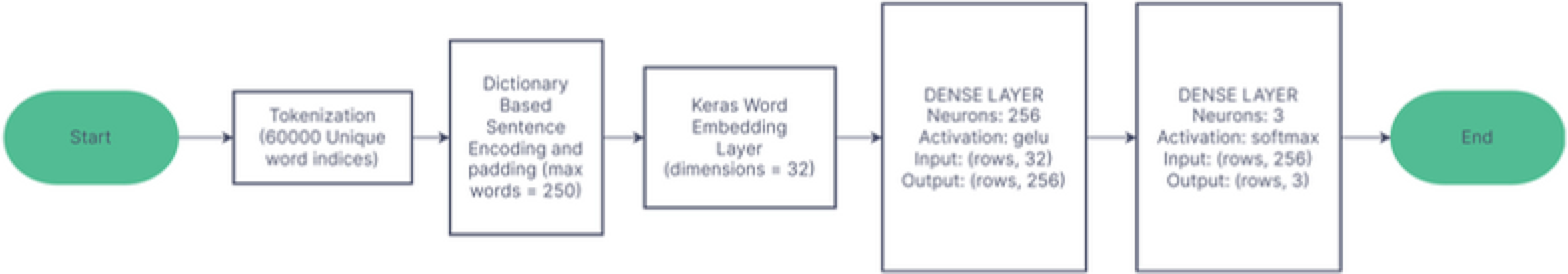
Indic Sentence Bert
(ENCODING)
Output Shape: (rows, 768)

DENSE LAYER
Neurons: 256
Activation: gelu
Input: (rows, 768)
Output: (rows, 256)

DENSE LAYER
Neurons: 3
Activation: softmax
Input: (rows, 256)
Output: (rows, 3)

End

SIMPLE ANN MODEL WITH KERAS EMBEDDINGS



Start

Tokenization (60000 Unique word indices)

Dictionary Based Sentence Encoding and padding (max words = 250)

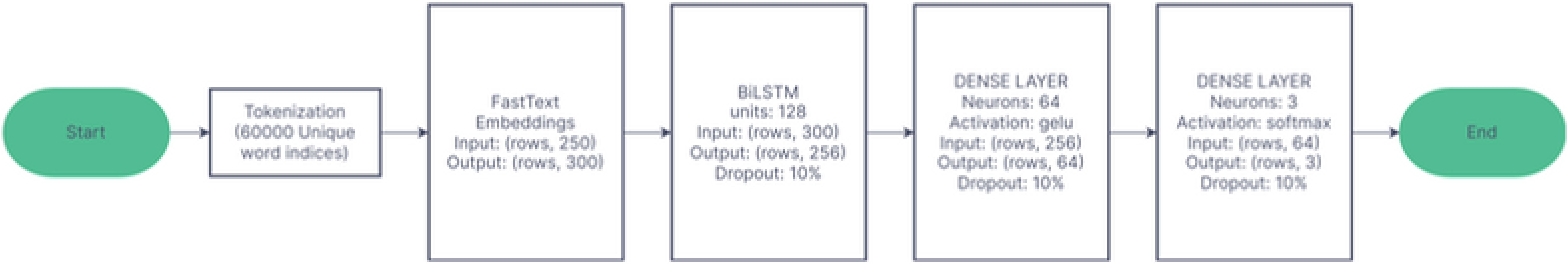
Keras Word Embedding Layer (dimensions = 32)

DENSE LAYER
Neurons: 256
Activation: gelu
Input: (rows, 32)
Output: (rows, 256)

DENSE LAYER
Neurons: 3
Activation: softmax
Input: (rows, 256)
Output: (rows, 3)

End

FAST TEXT EMBEDDINGS + BILSTM



Start

Tokenization (60000 Unique word indices)

FastText Embeddings
Input: (rows, 250)
Output: (rows, 300)

BiLSTM
units: 128
Input: (rows, 300)
Output: (rows, 256)
Dropout: 10%

DENSE LAYER
Neurons: 64
Activation: gelu
Input: (rows, 256)
Output: (rows, 64)
Dropout: 10%

DENSE LAYER
Neurons: 3
Activation: softmax
Input: (rows, 64)
Output: (rows, 3)
Dropout: 10%

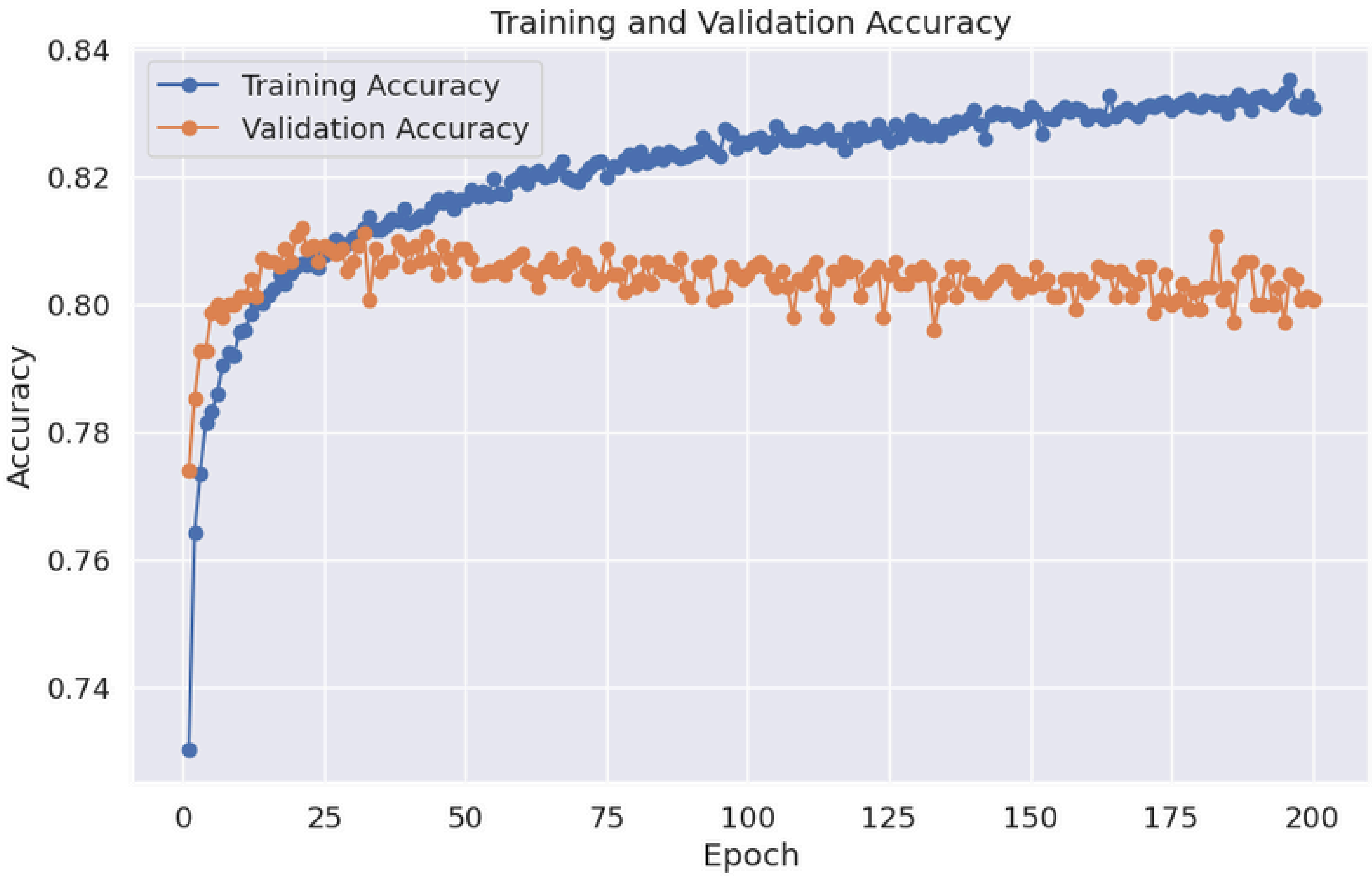
End

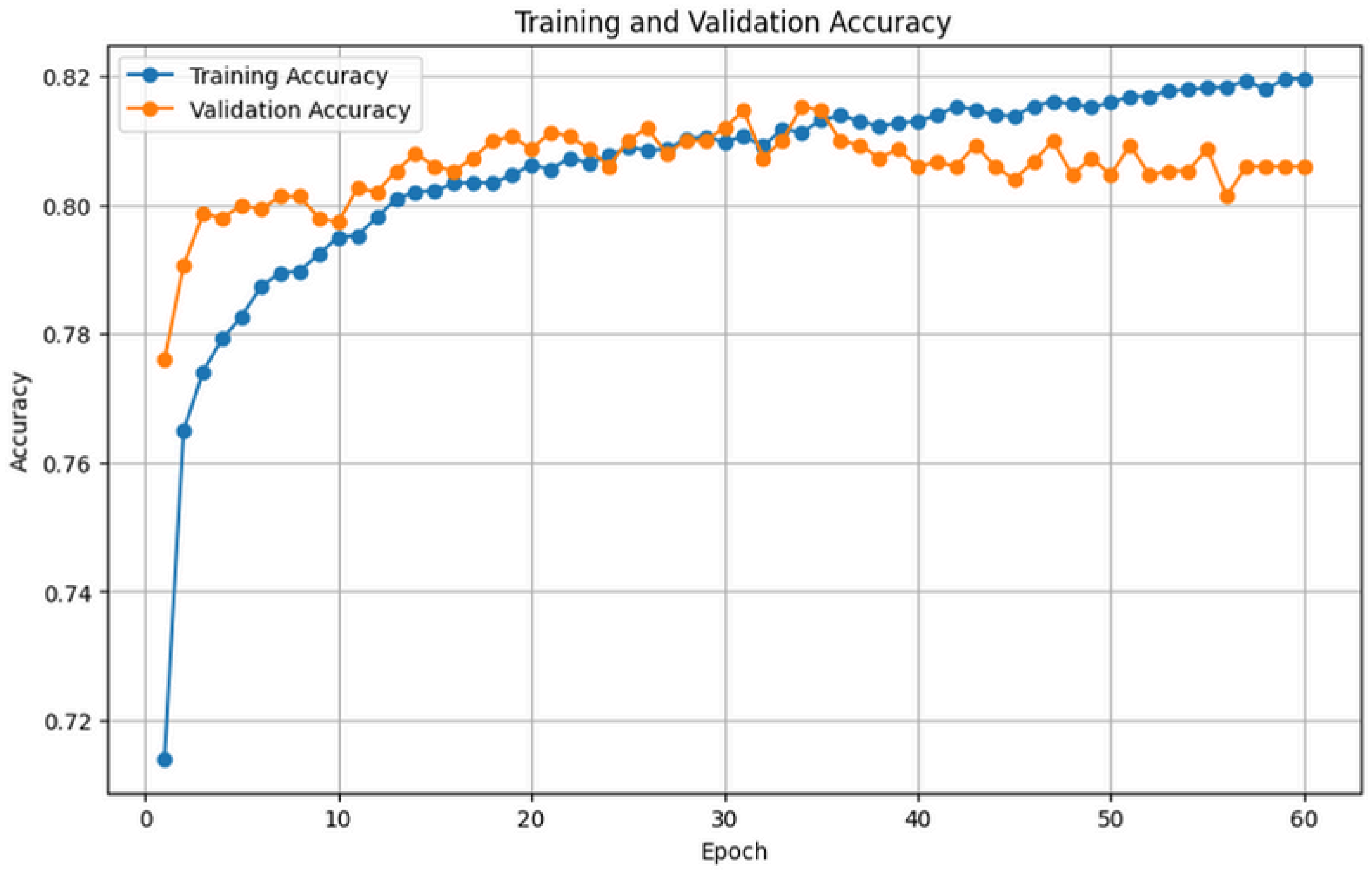
Finding the right balance between

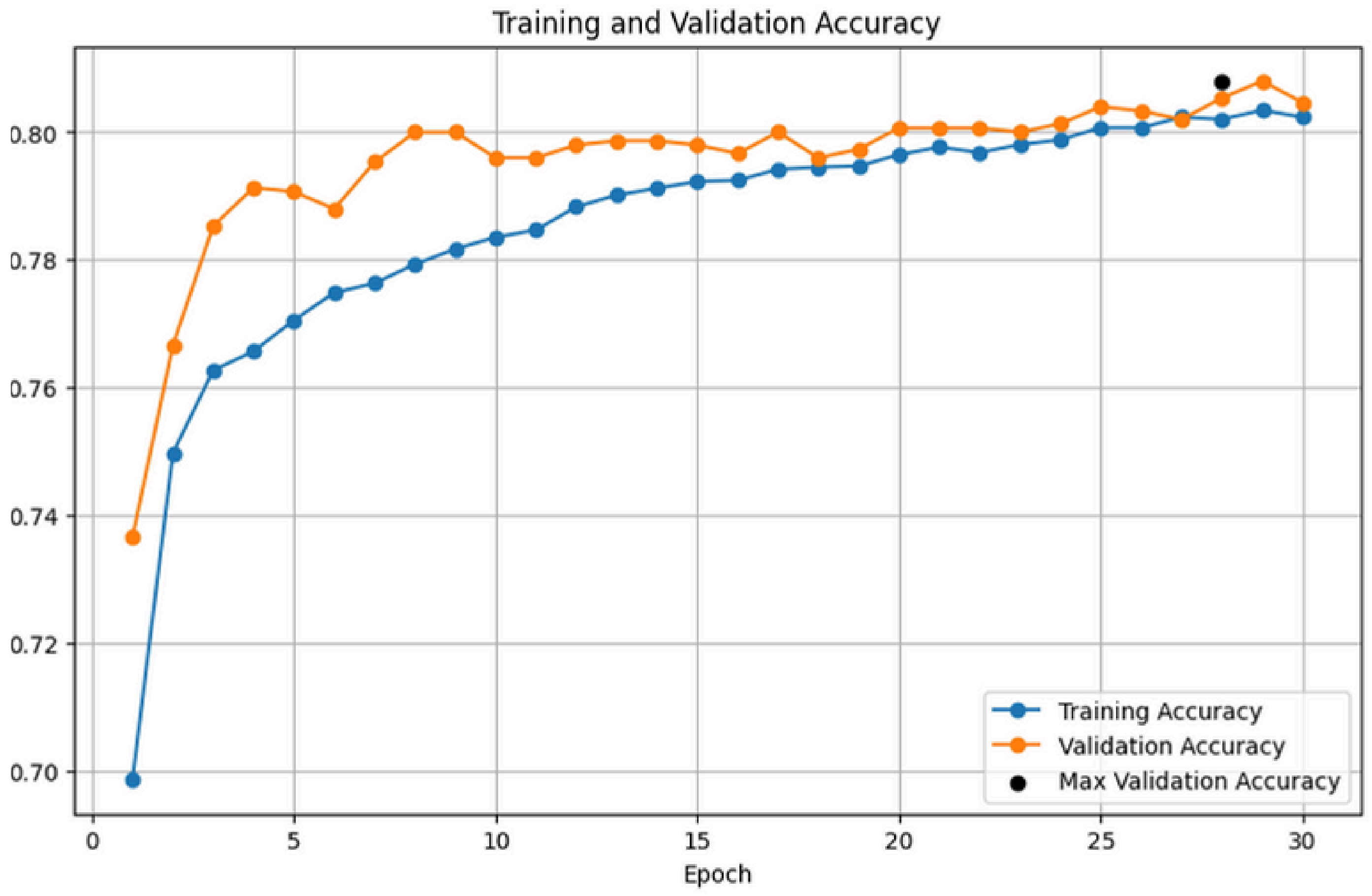
OVER-FITTING

and

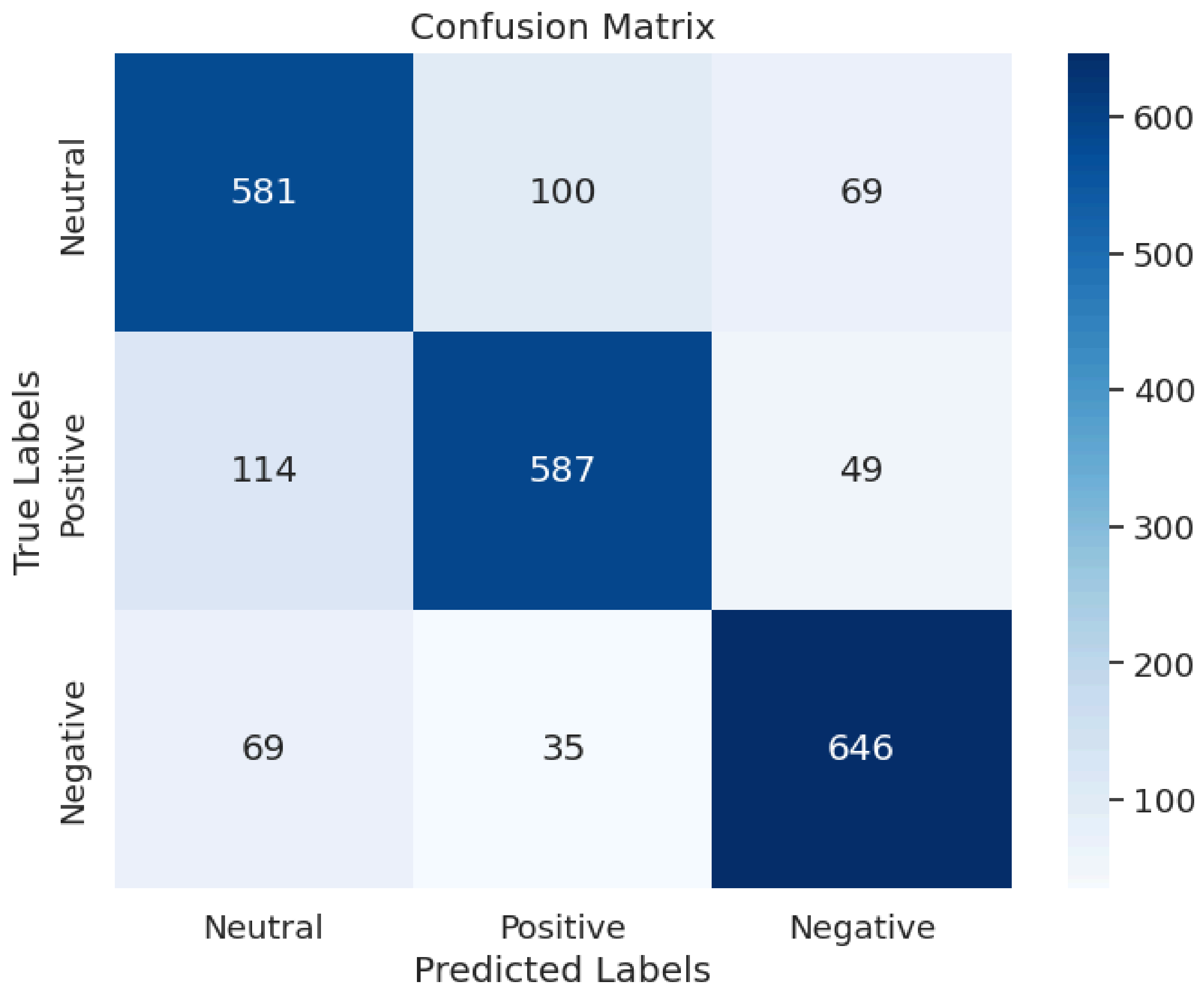
UNDERFITTING


Training and Validation Accuracy


Training and Validation Accuracy


Training and Validation Accuracy

MODEL EVALUATION - INDIC BERT 30-35 EPOCH

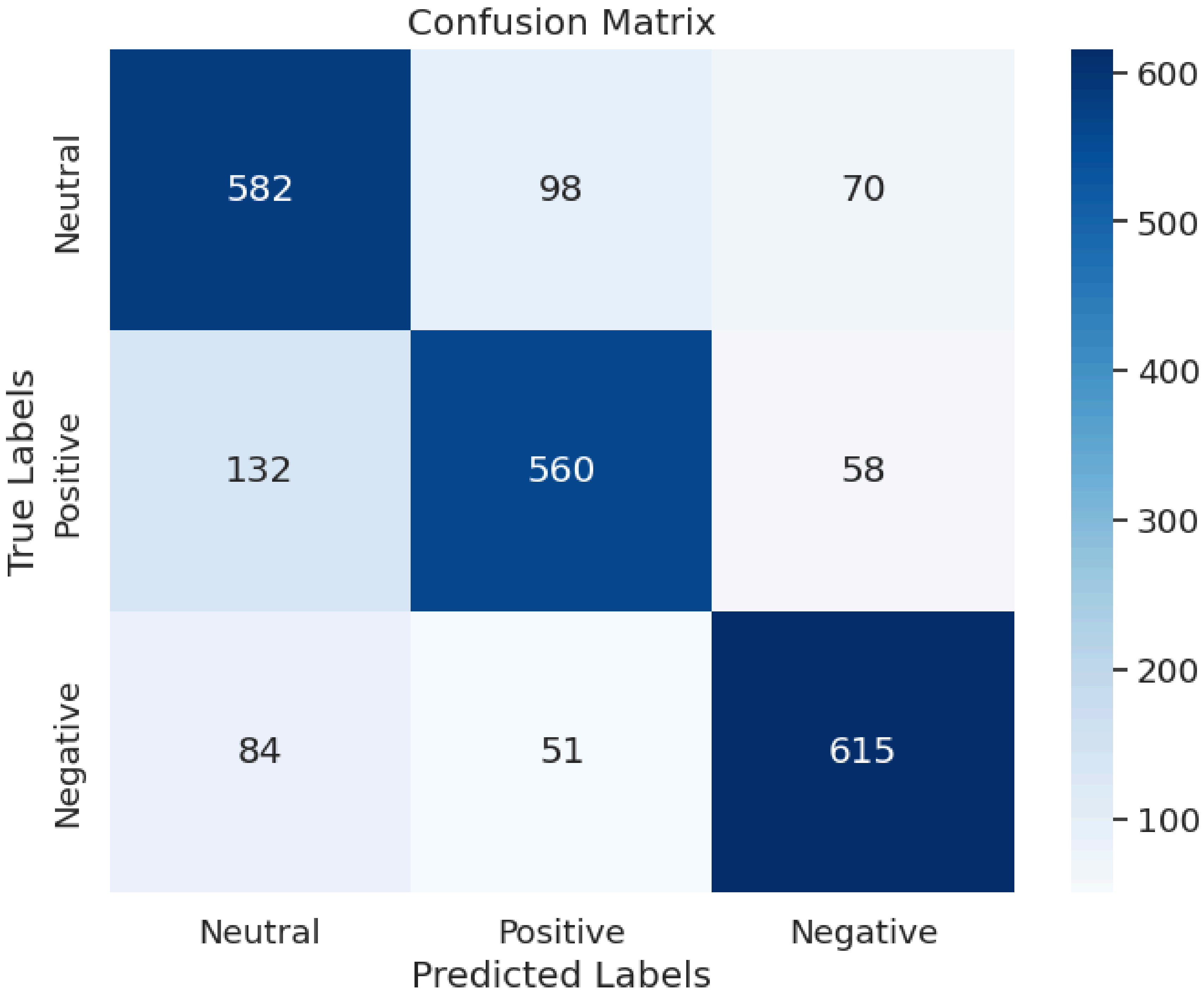


```
              precision    recall  f1-score   support

           0       0.76      0.77      0.77       750
           1       0.81      0.78      0.80       750
           2       0.85      0.86      0.85       750

    accuracy                           0.81      2250
   macro avg       0.81      0.81      0.81      2250
weighted avg       0.81      0.81      0.81      2250
```

Model Evaluation - ANN Default Embeddings



Confusion Matrix

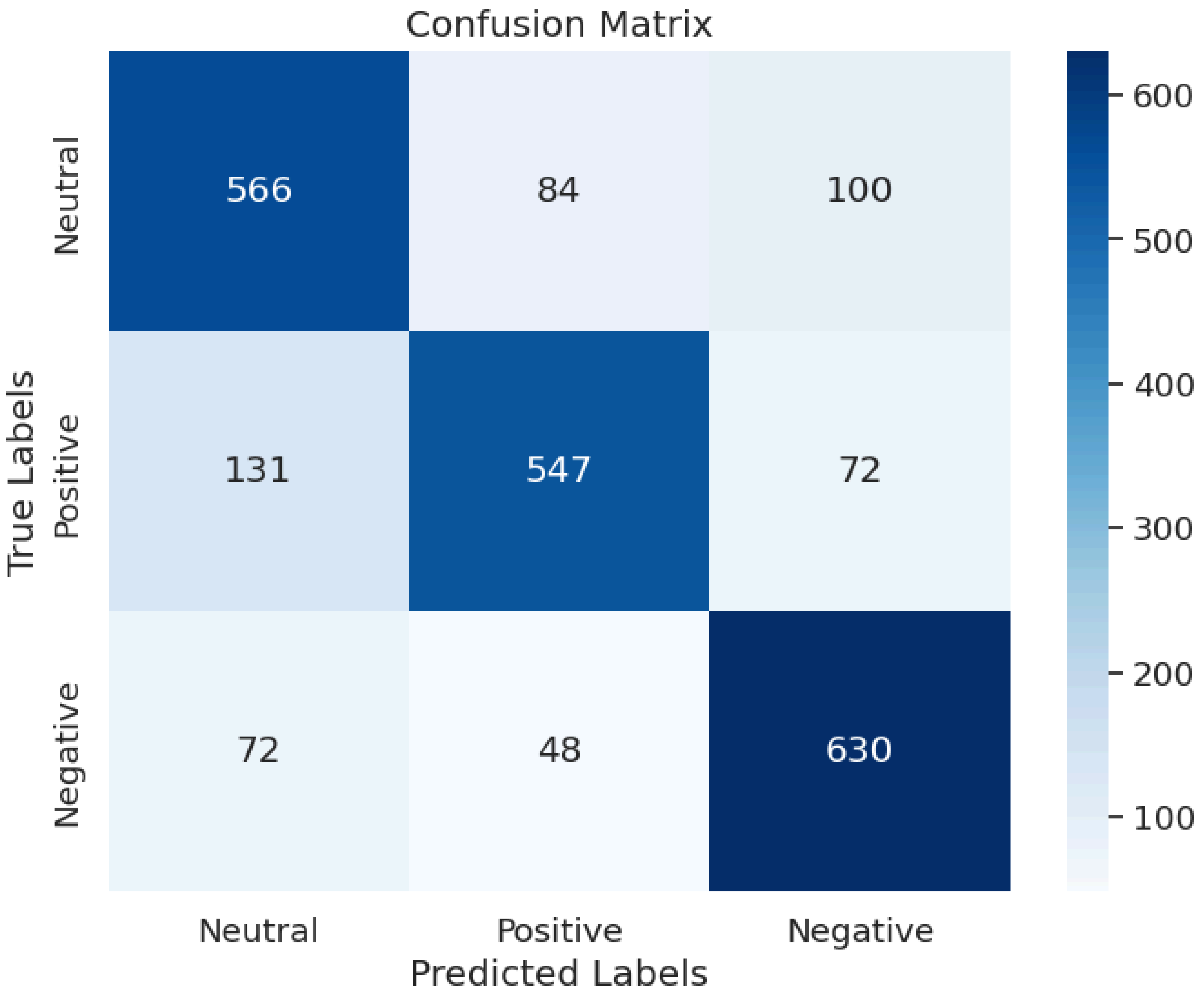
True Labels \ Predicted Labels	Neutral	Positive	Negative
Neutral	582	98	70
Positive	132	560	58
Negative	84	51	615

```
              precision    recall  f1-score   support

           0       0.74      0.81      0.77       750
           1       0.84      0.75      0.79       750
           2       0.84      0.86      0.85       750

    accuracy                           0.80      2250
   macro avg       0.81      0.80      0.80      2250
weighted avg       0.81      0.80      0.80      2250
```

Model Evaluation - BiLSTM Fast-Text Embeddings



	precision	recall	f1-score	support
0	0.74	0.75	0.75	750
1	0.81	0.73	0.77	750
2	0.79	0.84	0.81	750
accuracy			0.77	2250
macro avg	0.78	0.77	0.77	2250
weighted avg	0.78	0.77	0.77	2250

Comparison with Base Paper

MODELS:	INDIC BERT	FASTTEXT BILSTM
L3CubeMahaSent: A Marathi Tweet-based Sentiment Analysis Dataset	Accuracy: 84% F1-Score: 78.9%	Accuracy: 79%
L3Cube-MahaCorpus and MahaBERT: Marathi Monolingual Corpus, Marathi BERT Language Models, and Resources	Accuracy: 79%	-
Our Results	Accuracy: 81% F1-Score: 81%	Accuracy: 77% F1-Score: 77%

Novelty of the proposed model

- To perform Text Sentiment Analysis in Marathi to understand the current market trends, customer reviews and social media monitoring
- To compare various machine learning models which perform sentimental analysis of Marathi sentences
- To analyze strengths and weaknesses of various sentimental-analysis models
- To modify non-Marathi based models for increasing accuracy of Marathi analysis

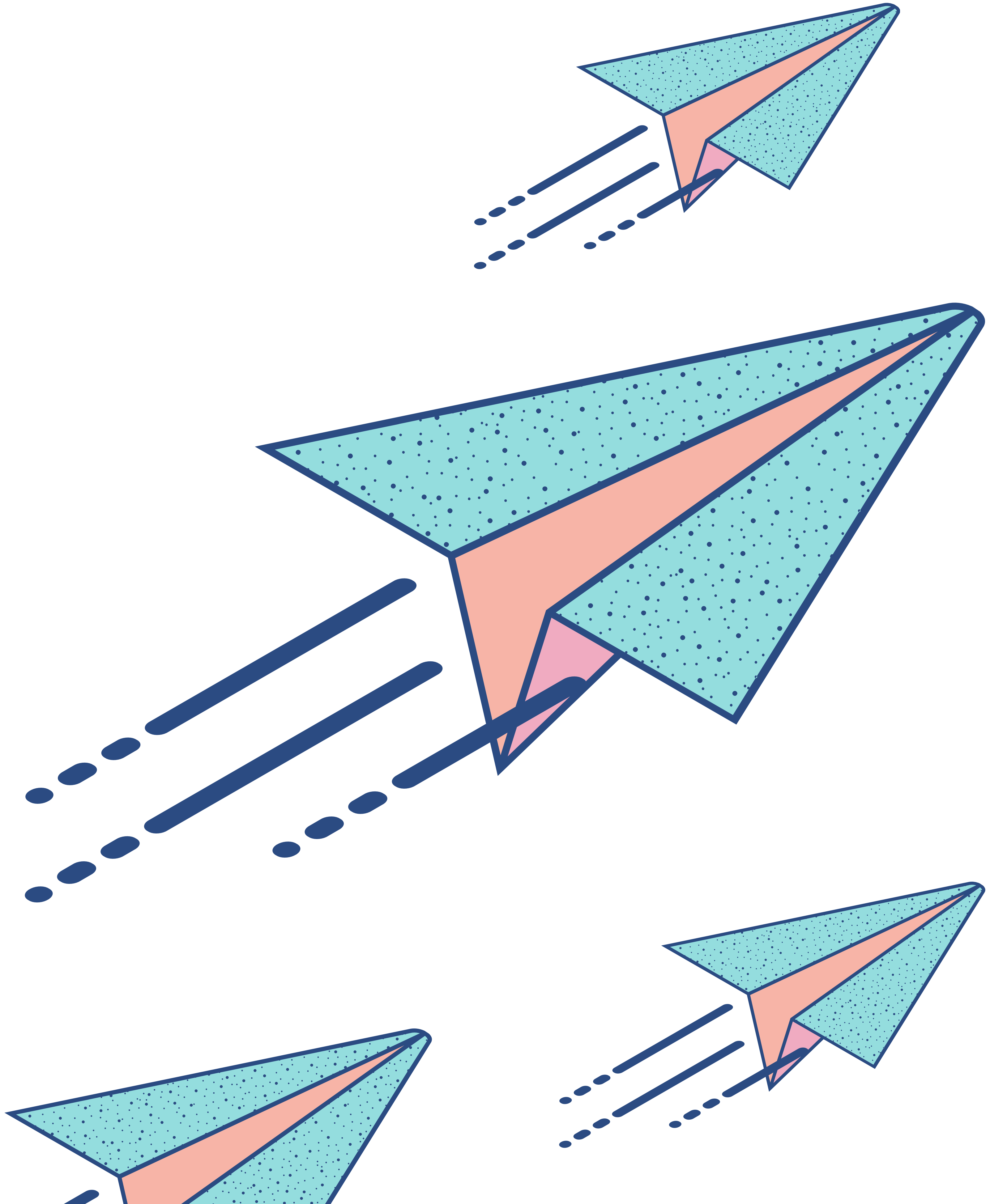
Future Scope

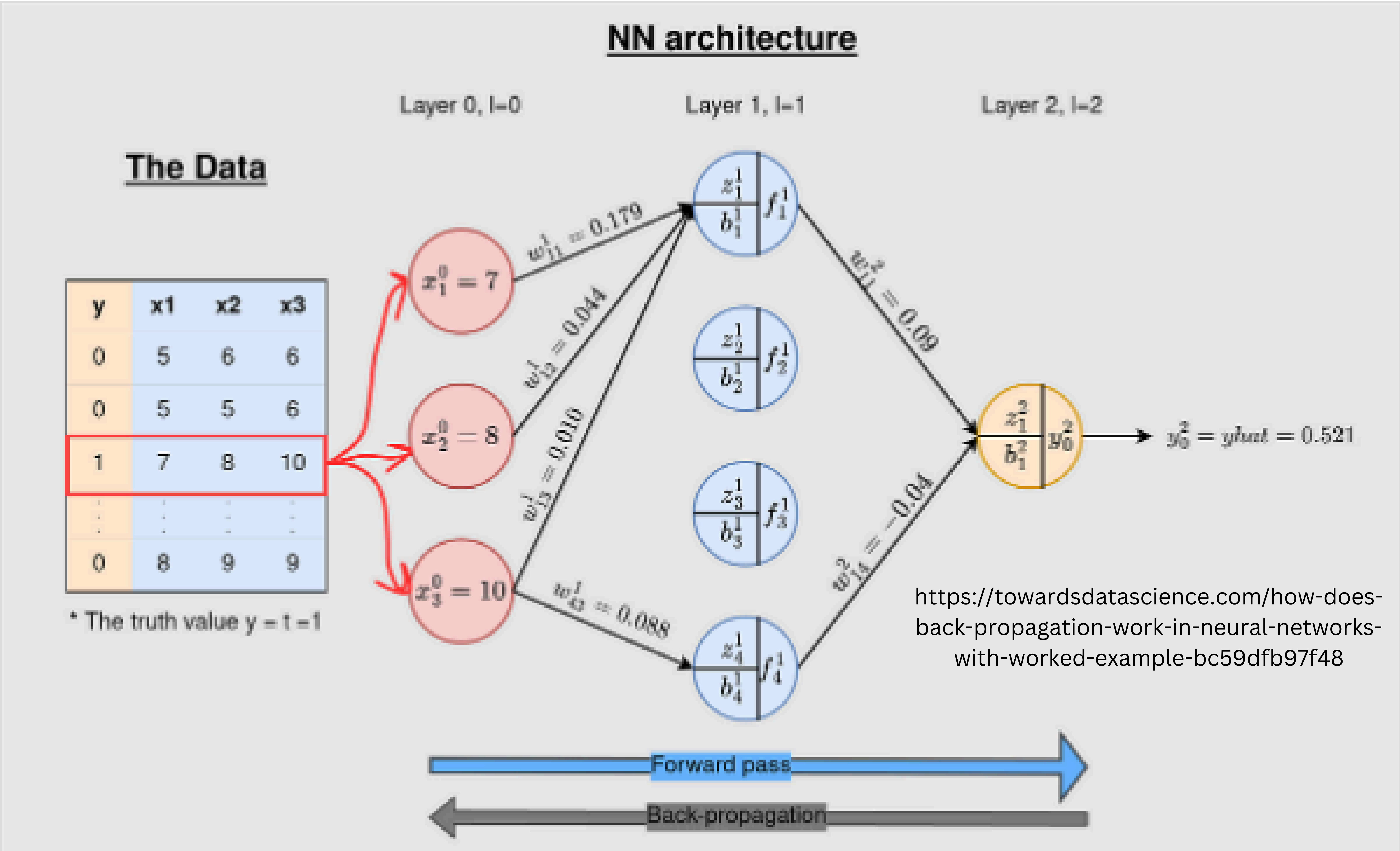
- Larger datasets could be integrated to enhance the accuracy of the model.
- New pre-trained models can be tested out for the same purpose.
- Additional parameter tuning can be done to get a precise vector of parameter values for optimal accuracy.

Conclusion

- One stop solution for effective sentiment analysis of Marathi texts
- Beneficial to solve the real world issues and leverage the social media monitoring

Thank You





The Data

y	x1	x2	x3
0	5	6	6
0	5	5	6
1	7	8	10
⋮	⋮	⋮	⋮
0	8	9	9

* The truth value y = t = 1

https://towardsdatascience.com/how-does-back-propagation-work-in-neural-networks-with-worked-example-bc59dfb97f48