A Framework for Social Media Opinion Mining for Low Resource Marathi Text

B.E PROJECT

By

Dhruv Talati	**60004180022**
Naitik Rathod	**60004180054**
Nishit Mistry	**60004180066**
Manan Parikh	**60004180049**

Guide:

Dr. Pratik Kanani
Assistant Professor

Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

University of Mumbai
2020-2021

CERTIFICATE

This is to certify that the mini project entitled **"*A Framework for Social Media Opinion Mining for Low Resource Marathi Text*"** is a bonafide work of **"Dhruv Talati(60004180022), Naitik Rathod(60004180054), Nishit Mistry(60004180066) and Manan Parikh(60004180049)"** submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of B.E. in Computer Engineering.

Dr. Pratik Kanani
Guide

Dr. Meera Narvekar
Head of Department

Dr. Hari Vasudevan
Principal

Project Report Approval

This mini project report entitled ***A Framework for Social Media Opinion Mining for Low Resource Marathi Text*** by ***Dhruv Talati, Naitik Rathod, Nishit Mistry and Manan Parikh*** is approved for the partial fulfillment of the degree of ***B.E. in Computer Engineering.***

Examiners

1.-------------------------------------

2. -------------------------------------

Date: 12th November 2021

Place: Mumbai

Declaration

I/We declare that this written submission represents my/our ideas in my/our own words and where others' ideas or words have been included, I/We have adequately cited and referenced the original sources. I/We also declare that I/We have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my/our submission. I/We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Talati

(Dhruv Talati – 60004180022)

Rathod

(Naitik Rathod - 60004180054)

Nishit

(Nishit Mistry - 60004180066)

M. Parikh

(Manan Parikh - 60004180049)

Date: 12th November 2021

Abstract

Sentiment Analysis is one of the most important tasks for any language and a very important domain in Natural Language Processing. Popular and widely used languages like English, Russian and Spanish have a great availability of language models for these tasks and widely available datasets too. But the research in Low Resource Languages like Hindi and Marathi is far behind. The Marathi language is one of the prominent languages used in India, being the third most spoken language. It is predominantly spoken by the people of Maharashtra. Over the past decade, the usage of language on online platforms has tremendously increased. However, research on Natural Language Processing (NLP) approaches for Marathi text has not received much attention. Therefore in this project we will be creating a framework that can be used for the opinion mining of the social media marathi texts without using any translations. Not using translations will not only get better results but also an error free model trained over the target language only. The multilingual model XLM-RoBERTa will be put under training over the Marathi tweets dataset for the task of opinion mining and classification. We aim at deploying the best performing model in our own GUI where users can test individual sentences where the whole analysis will be shown about the opinions generated.

Contents

List of Figures

List of Abbreviations

Sr. No.	Abbreviation	Expanded form
i	USA	United States of America
ii	XLM-R	XLM-RoBERTa
iii	CNN	Convolutional Neural Network
iv	LSTM	Long Short Term Memory
v	BiLSTM	BiDirectional Long Short Term Memory
vi	ULM FiT	Universal Language Model Fine Tuning
vii	BERT	Bidirectional Encoder Representations From Transformers
viii	SVM	Support Vector Machine
ix	RF	Ordered Weighted Averaging
x	OvR	One vs Rest
xi	DMLMC	Direct Multi-Label Multi-Classification
xii	mBERT	multilingual-BERT
xiii	TSA	Twitter Sentiment Analysis
xiv	NLP	Natural Language Processing
xv	UML	Unified Modelling Language

xvi	DFD	Data Flow Diagram
xvii	IEEE	Career Bowling Average of The Bowler
xviii	AI	Artificial Intelligence
xix	TB	Tera Byte
xx	GUI	Graphical User Interface

Chapter 1
Introduction

1.1 Description

The USA elections of 2020 and the fake news that was spread during and after the presidential campaigns shows us the importance of social media companies in the fight against fake news. The ability of Twitter to flag tweets considered as hateful or inciting violence was based on sentiment analysis of the tweets using Machine Learning models. However there has been miniscule research on opinion mining in low resource languages like Marathi, Gujarati and other Indian languages. Social media users in India are currently mainly from the urban areas mainly using the English language.However with the National Optical Fibre Mission and other initiatives to bring internet connectivity to rural areas, there is going to be a boom in the number of users using non-english native languages to text on social media. Hence the need for opinion mining in these languages is urgent.

1.2 Problem Formulation

The current models and systems available are designed to analyze the data of tweets in the English language. The accuracy of data converted from regional languages to English and then performing opinion mining was found to be too low. Hence,we here propose to create a system that is capable of social media opinion mining in the Marathi language.

1.3 Motivation

Huge surge of social media users is expected in India and 90% of these users will use Indian languages to communicate. This will lead to tremendous data generation in the regional languages. Marathi is the 3rd most spoken native language in India, with 83 million native speakers according to the 2011 census.

1.4 Proposed Solution

Current best available models for marathi text classification have been trained over news articles and news headlines data. This cannot give a very accurate analysis of the social media texts.

We will be using the dataset that is created from twitter tweets that were in marathi language.
We propose to train the XLM-RoBERTa model over the marathi tweets to achieve better accuracies for the opinion mining of the social media texts. Based upon the probabilities achieved from the final model, we will classify the text in not 3, but 5 categories, i.e., very positive, positive, neutral, negative and very negative. We will create a framework where the model will be deployed and can be used by multiple people for generating opinions out of their marathi text.

1.5 Scope of the project

The earlier models were mainly trained on the English language. However research on low resource languages like Marathi was minute. The proposed methodology will be beneficial for the governments of mainly Maharashtra and Goa where speakers of Marathi language are in abundance. The governments will be able to classify the text into positive, neutral and negative and can take action accordingly. It will also be beneficial to companies helping them address any grievances of their users mainly in rural areas who use Marathi language on social media.

Chapter 2
Review of literature

2.1 Previous work

Authors in [1] suggest a comprehensive overview of available resources and models for Marathi text classification.Authors have evaluated different CNN, LSTM, BiLSTM based models along with language models such as ULMFiT and BERT on two datasets viz. Marathi News Headline Dataset and Marathi News Articles Dataset. The presented model proposed by the authors in [1] works with a major chunk of data used to pre-train them coming from news sources.The target datasets also come from the news domain and hence achieve higher accuracy but the accuracy diminishes for non news articles.

In [2], proposes to provide an effective neural network based technique for the hostility detection in the hindi text. Authors have evaluated different models like SVM, RF, BiLSTM, and also pre-trained language models that are variants of BERT. They have curated the dataset on hostile and non-hostile hindi text from social media platforms like twitter, facebook and whatsapp, etc., which is further annotated into fine grained labels like fake, hate, defamation and offensive. The datasets have been processed into the four fine grained labels using two different methods, OnevsRest(OvR) and Direct Multi-Label Multi-Classification(DMLMC) and both of them are used for the training and testing and the results obtained are compared. Authors in [2] received 91.63% and 89.76% accuracies on individual mBERT and XLM-R with their set parameters and the hybrid received accuracy of 92.6% that is the best performance among all the models employed for the coarse-grained evaluations.

The authors in [3] propose a system for identification of low-toxic statements used by users on Educational and specialized web resources,which are characterized by a different type of user. The people using these sites are characterized by good manners, restraint in statements and expressions of emotion. Despite this fact, heated discussions also arise on these web resources, characterized not by highly toxic, but by low-toxic statements, ridicule, sharp jokes, provocative statements and hidden injections.The authors of this paper propose to annotate these low-toxic texts. Datasets are trained on XLM-RoBERTa by the authors in [3] because of its better performance for detection of low-toxic texts as compared to other models. Government agencies can detect low-toxic texts on educational and other related platforms helping them take any corrective actions if necessary.

Authors in [4] have studied the effect of translation on the sentiment classification task from resource-rich language to a low-resource language. It identifies and enlists words causing polarity shifts into five distinct categories. It further finds the correlation between the languages with similar roots. Our study shows 2-3 percentage points performance degradation in sentiment classification due to polarity shift as a result of translation from resource-rich languages to low-resource languages. To explore the translation approach to develop a sentiment analysis dataset for low-resource languages. To study the effect of translating the English reviews into German, Urdu, and Hindi and compare the classification results of all languages. Authors have studied the importance of handling words affected by Negation. Authors have shown that google translator translated “Faultless Production” into Urdu which means “Bad Production”. This translation is incorrect, and this is another proof that Negation affects translation.

In this paper [5] authors have shown that pre training multilingual language models at scale leads to significant performance gains for a wide range of cross lingual transfer tasks.Authors have trained a Transformer based masked language model on one hundred languages, using more than two terabytes of filtered CommonCrawl data.Authors model, dubbed XLM-R, significantly outperforms multilingual BERT (mBERT) on a variety of cross-lingual benchmarks.Authors have shown that XLM-R performs particularly well on low-resource languages, improving 15.7% in XNLI accuracy for Swahili and 11.4% for Urdu over previous XLM models. Authors have introduced XLM-R, a new state of the art multilingual masked language model,they show that it provides strong gains over previous multilingual models like mBERT and XLM on classification and sequence labeling.

The authors in [6] have provided an effective neural network based technique for the classification task of SemEval 2020 for two code mixed languages: Hindi-English and Spanish-English. They have used the dataset of SemEval of these two mixed languages. The datasets have been processed and then employed under various models like BiLSTM, mBERT and XLM-R. The Hindi-English dataset consists of 17000 labelled texts from social media while the Spanish-English dataset consists of 15000 labelled texts. All text either labelled positive, negative or neutral.Authors have shown that proper word embeddings can boost performances by a large margin, considering the fact that it already offers the model an insight into that language. The problem becomes more complicated here as the authors deal with two languages instead of one.

The authors in [7] propose sentiment analysis of low resource language Hindi. Hindi, is the fourth-most popular language, still lacking in richly populated linguistic resources, owing to the challenges involved in dealing with the Hindi language. In this article authors first explore the machine learning-based approaches—Naïve Bayes, Support Vector Machine, Decision Tree, and Logistic Regression—to analyze the sentiment contained in Hindi language text derived from Twitter. The dataset employed by authors for sentiment analysis has been fetched from Twitter. Authors have downloaded tweets for movie and product reviews from Twitter, selecting the language "Hindi" in the search filters.They have manually labelled 23,767 tweets into positive or negative. Authors removed the tweets with ironic content, slang language, non-Hindi language, and English words written in Hindi. The tweets without subjectivity were also dropped from the dataset by the authors. After removing these tweets, 16,901 subjective tweets were left.The availability of easy translation provided on the Web, netizens find it interesting to write in their native languages. This pushes for the requirement to perform sentiment analysis in other languages also. Large amounts of content in different languages are available on the Web, which needs to be analyzed to determine the opinion of non-English speaking masses. The proposed CNN approach by the authors gives an accuracy of 85%.

Authors in [8] perform sentiment analysis for Manipuri language where orientation of the text is classified into either negative, positive or neutral sentiment. Manipuri is the lingua franca of Manipur, a northeastern state of India. It is not only the official language of Manipur but also included in the 8th Schedule of Indian Constitution. Pre-processing methods used by authors include white space removal, stemming, removal of stop words, removal of numbers, removal of URL links, negation handling, replacing negative mentions, reverting words that contain repeated letters in their original form. Authors have collected and prepared a goal standard dataset for Manipuri sentiment analysis from a local daily newspaper. Transliteration systems are implemented to transliterate Bengali script text to Roman script text and Meetei Mayek script text to Roman script text.Limited availability of good language-specific toolkits for Manipuri language acted as a major constraint for the authors. The transliterated gold standard dataset prepared by the authors could be of use in extending the work on the dataset collected from social media with proper normalization.

The authors in [9] present a detailed description of the feature-based TSA system (incorporated with an improved corpus-based negation modeling approach), which classifies tweets based on syntactic and semantic features extracted from them.This work contributes in presenting a feature extraction system that would help in generation of varieties of feature sets, which can be used as an input to classifiers. Authors provide an algorithm for implementing a set of rules for handling those tweets where negation occurrence does not necessarily mean negation.This article by the authors contributes in presenting a comprehensive research in the field of TSA by looking into the critical aspects of NLP that are tweet normalization and negation.

.

2.2 Research Gap

All the previous work done in this domain has been done for the languages with plenty resources. Marathi is one language where the research done is still way behind other languages like Hindi in the field of opinion mining. Hence we propose to create a framework where we will be deploying XLM-RoBERTa based model fine-tuned over marathi tweets dataset. This will achieve better accuracies than other models where there is a need of translation before the task of Opinion Mining.

Chapter 3

System Analysis

3.1 Functional Requirements

Get the Marathi Tweets Dataset - download, filter, and store the required data in the local database. Structure the data with the labels as required for the sentiments.

Analysis Strategy - The tweets positive, neutral and negative will be mentioned as 1, 0, -1. This will be used for training the language model over the tweets dataset.

Displaying Sentiment Probabilities: For each sentence tested, the probabilities of all the classes will be displayed and classified into the best fit sentiment.

3.2 Non-Functional Requirements

The non-functional requirements of the system are explained below as performance requirements and design constraints.

1. Performance requirements:
   a) Accuracy - Since we will give priority to the accuracy of the model, the performance of the framework will be better and accurate results will be obtained.
   b) Openness - The system should be extensible to guarantee that it is useful for a reasonable period. Latest available dataset is being used for this task.
   c) Reliability - Dataset used is taken from twitter and is processed according to our requirements.

2. Design constraints:

a) Hardware Constraints - The model will be integrated with a web application. To use the opinion mining model, the user should enter from a personal computer or access website from mobile where the sentences can be tested and the probabilities and sentiment will be displayed.

b) Software System Attributes

- Usability - The model will be embedded in the backend of an application. It should be scalable designed to be easily adopted by a system.
- Reliability - The system should have accurate results and fast responses when user checks for social media texts.

3.3 Specific Requirements

Dataset used for training the models should be secured and no manipulation should be done to the data after the model is trained as it will lead to faulty results.

Data obtained from various sources should be stored in similar manner for faster training and prediction purposes.

3.4 Use-Case Diagrams and description

A UML use case diagram is the primary form of system/software requirements for a new software program underdeveloped. Use cases specify the expected behavior (what), and not the exact method of making it happen (how). Use cases once specified can be denoted both textual and visual representation (i.e., use case diagram). A key concept of use case modeling is that it helps us design a system from the end user's perspective. It is an effective technique for communicating system behavior in the user's terms by specifying all externally visible system behavior.

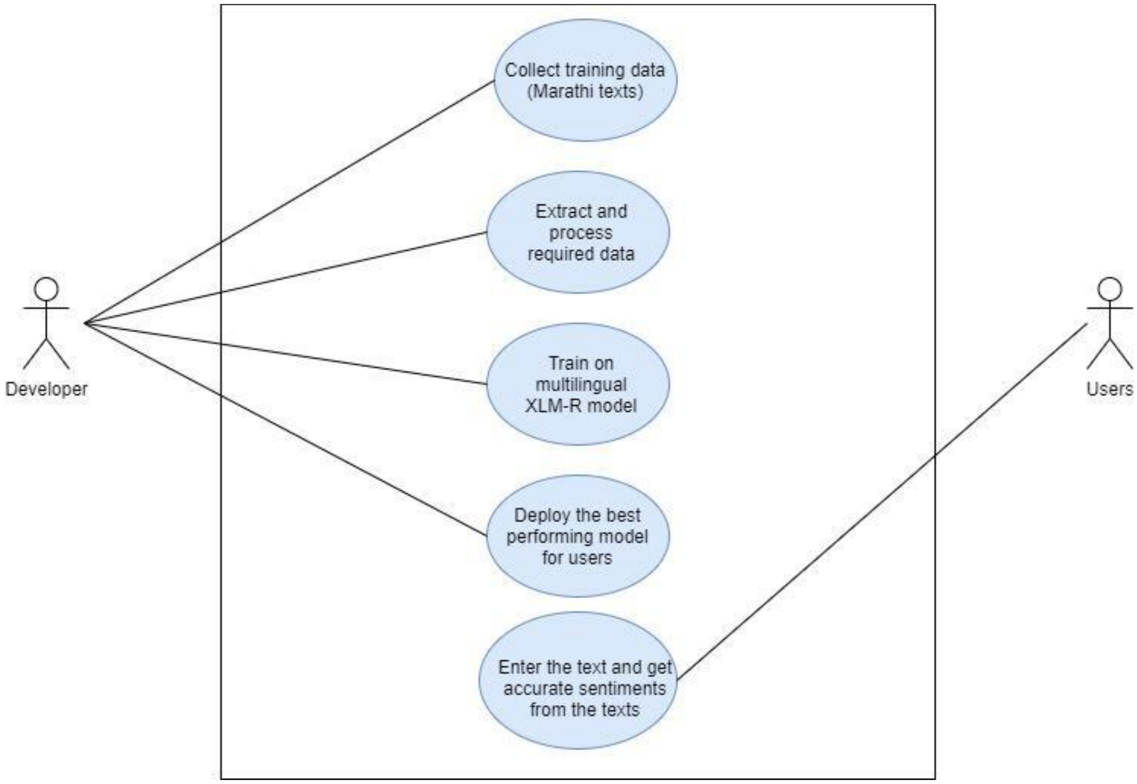

Collect training data (Marathi texts)
Extract and process required data
Train on multilingual XLM-R model
Deploy the best performing model for users
Enter the text and get accurate sentiments from the texts
Developer
Users

Fig 3.1 Use-case diagram of the proposed system

1. The initial training dataset (tweets dataset) is acquired.
2. The preprocessing and cleaning of data is done.
3. Training is done over the XLM-R model for the opinion mining.
4. Model makes its probabilities and final sentiment is generated.
5. The user gets to test their own social media texts for sentiments.

Chapter 4

Analysis Modeling

4.1 Data Modeling

Data modeling is the process of creating a visual representation of either a whole information system or parts of it to communicate connections between data points and structures. The goal is to illustrate the types of data used and stored within the system, the relationships among these data types, the ways the data can be grouped and organized and its formats and attributes. Data can be modeled at various levels of abstraction.

The DFD diagram is displayed below for the proposed system.

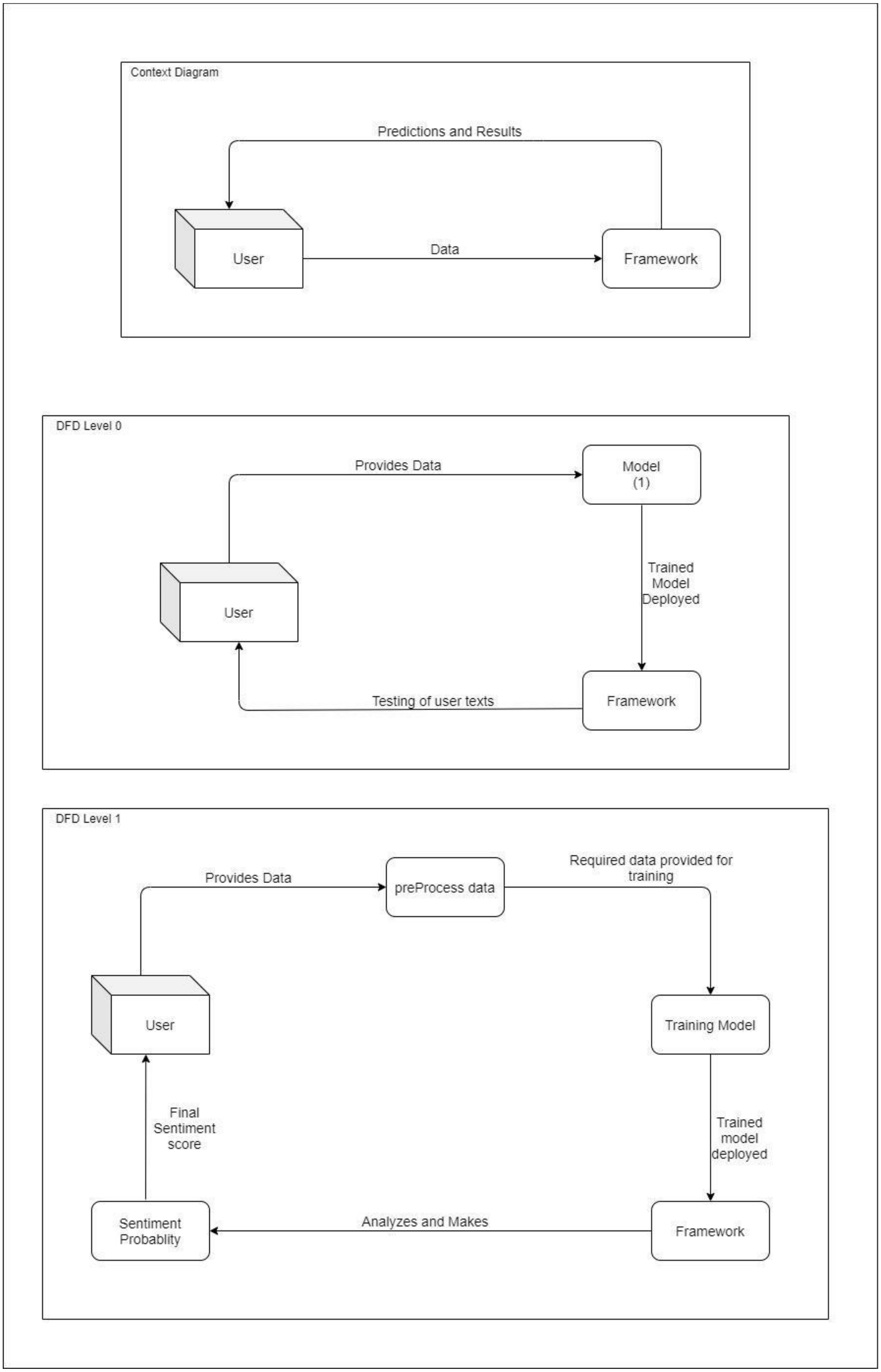


Fig 4.1 Data Flow Diagram of the proposed system

4.2 Activity Diagrams / Class Diagram / Sequence / Collaboration / State

4.2.1 Class Diagram

Class diagram is a static diagram. It represents the static view of an application. Class diagram is not only used for visualizing, describing, and documenting different aspects of a system but also for constructing executable code of the software application.

Class diagram describes the attributes and operations of a class and also the constraints imposed on the system. The class diagrams are widely used in the modelling of object-oriented systems because they are the only UML diagrams, which can be mapped directly with object-oriented languages.

Class diagram shows a collection of classes, interfaces, associations, collaborations, and constraints. It is also known as a structural diagram.

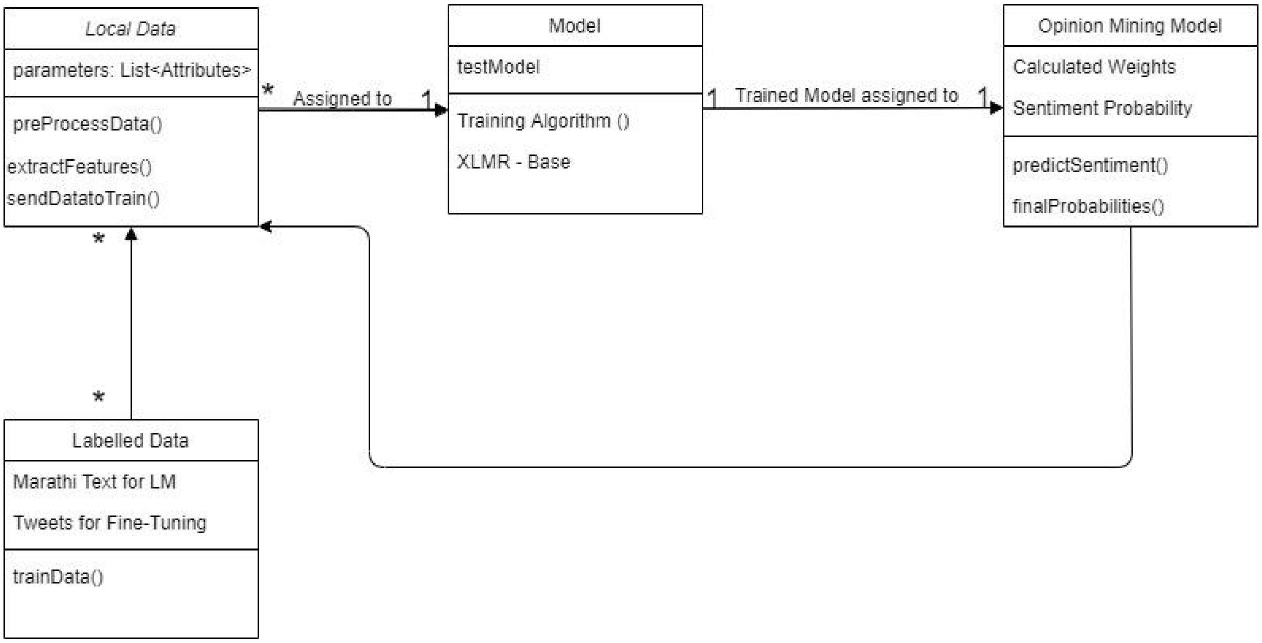


Fig 4.2 Class diagram of the proposed system

4.2.2 State Diagram

A state diagram is used to represent the condition of the system or part of the system at finite instances of time. It's a behavioral diagram and it represents the behavior using finite state transitions. State diagrams are also referred to as State machines and State-chart Diagrams. These terms are often used interchangeably. So simply, a state diagram is used to model the dynamic behavior of a class in response to time and changing external stimuli. We can say that each and every class has a state but we don't model every class using State diagrams. We prefer to model the states with three or more states.

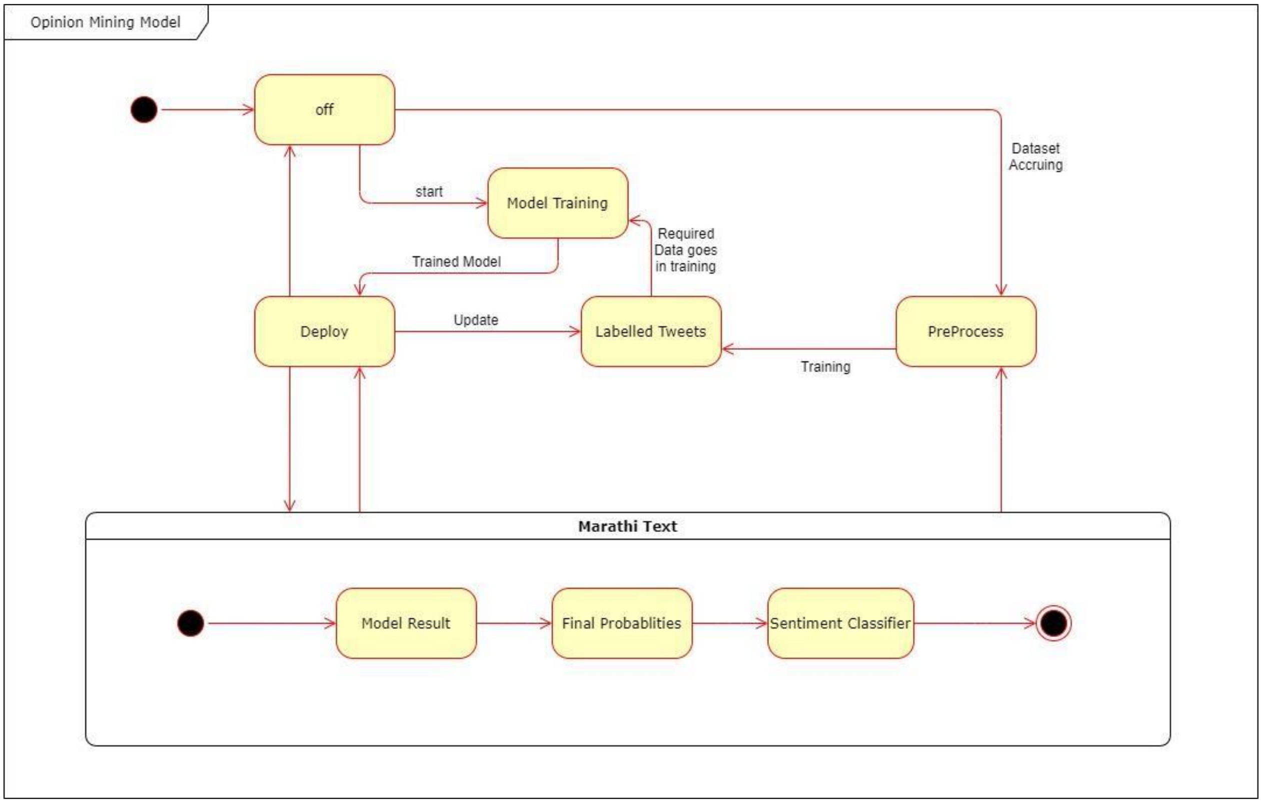


Fig 4.2 State diagram of the proposed system

4.2.3 Activity Diagram

Activity diagram is another important diagram in UML to describe the dynamic aspects of the system. Activity diagram is basically a flowchart to represent the flow from one activity to another activity. The activity can be described as an operation of the system. The control flow is drawn from one operation to another. This flow can be sequential, branched, or concurrent. Activity diagrams deal with all type of flow control by using different elements such as fork, join, etc.

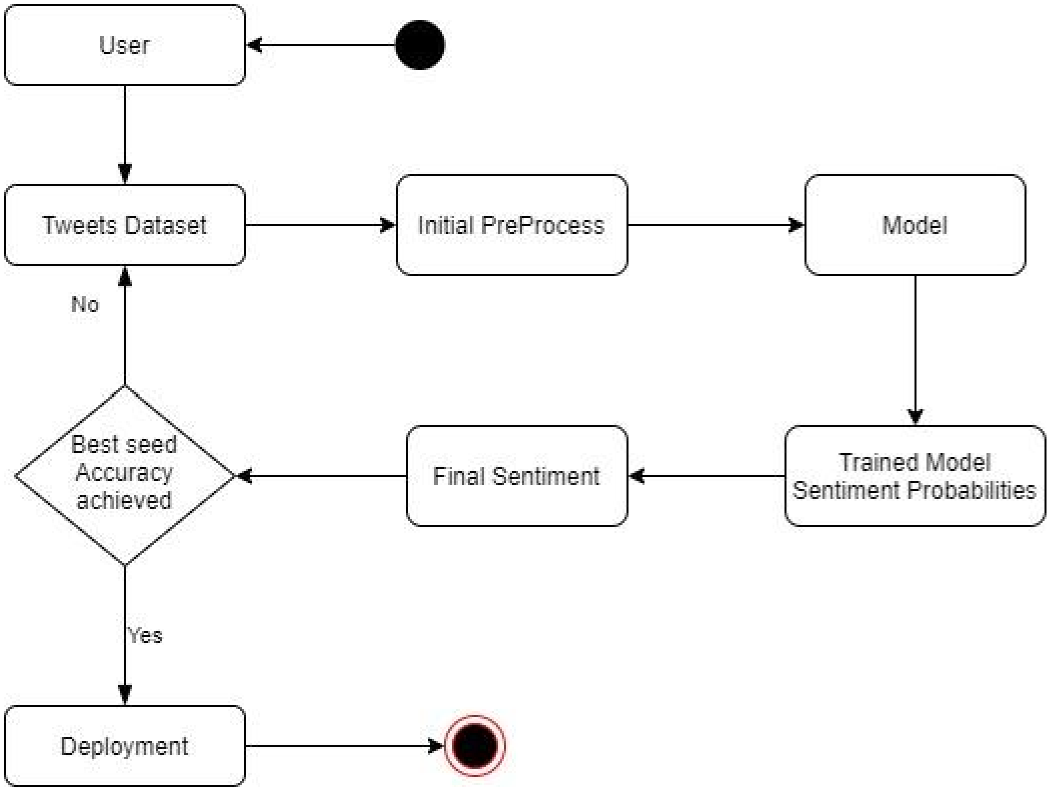


Fig 4.3 Activity diagram of the proposed system

4.2.4 Sequence Diagram

A sequence diagram simply depicts interaction between objects in a sequential order i.e. the order in which these interactions take place. We can also use the terms event diagrams or event scenarios to refer to a sequence diagram. Sequence diagrams describe how and in what order the objects in a system function. These diagrams are widely used by businessmen and software developers to document and understand requirements for new and existing systems.

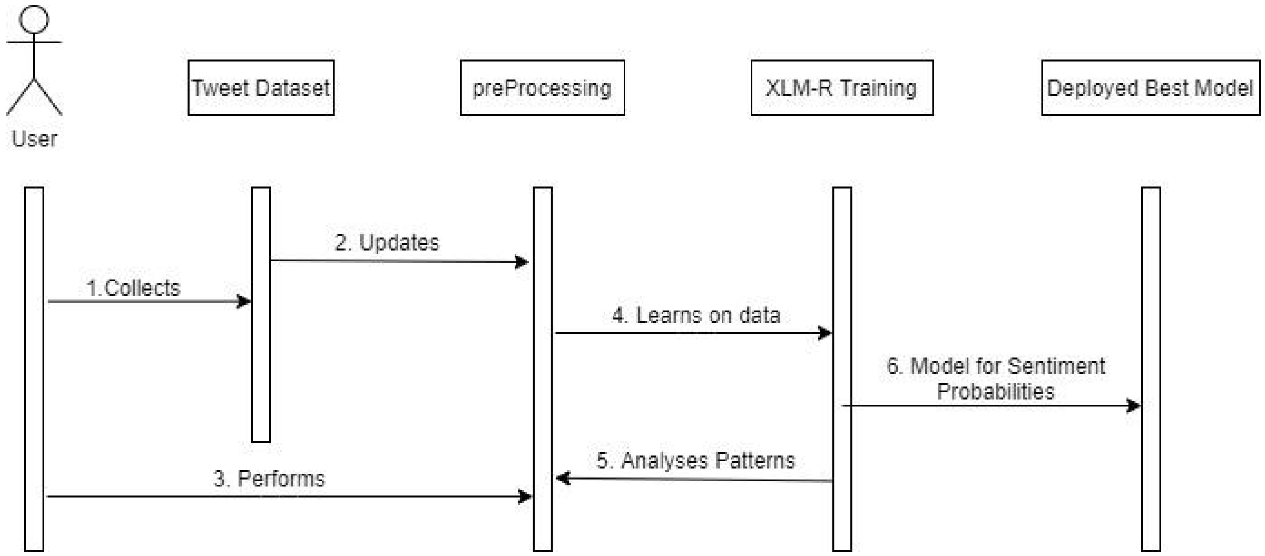


Fig 4.4 Sequence diagram of the proposed system

4.2.5 Collaboration Diagram

The collaboration diagram is used to show the relationship between the objects in a system. Both the sequence and the collaboration diagrams represent the same information but differently. Instead of showing the flow of messages, it depicts the architecture of the object residing in the system as it is based on object-oriented programming. An object consists of several features. Multiple objects present in the system are connected to each other. The collaboration diagram, which is also known as a communication diagram, is used to portray the object's architecture in the system.

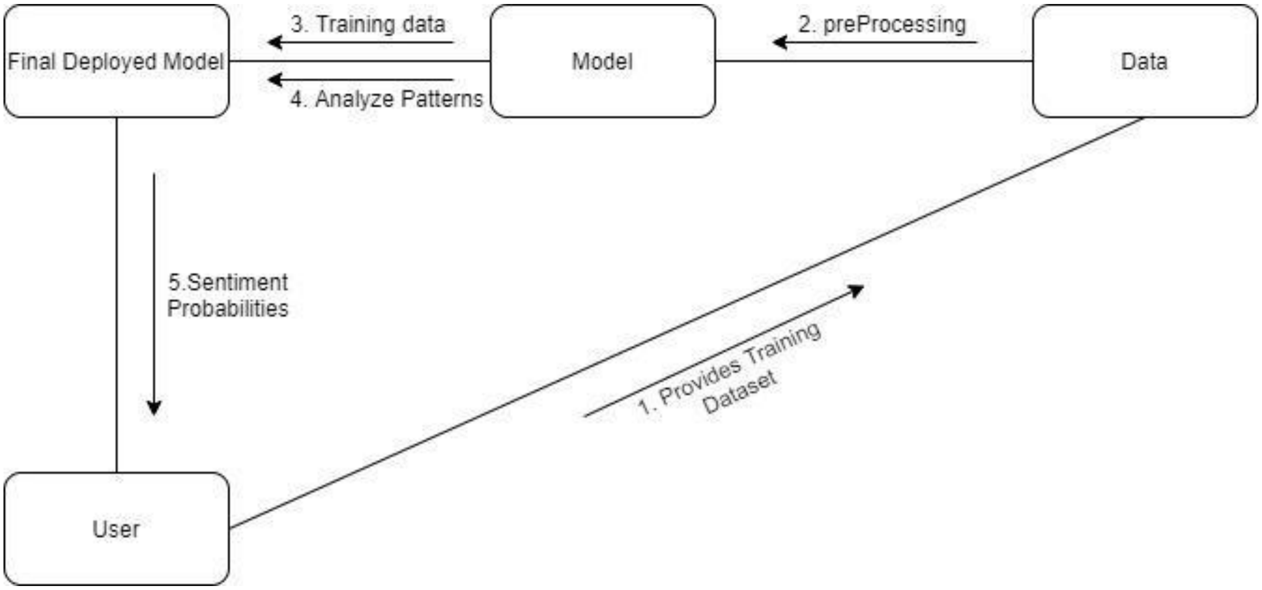


Fig 4.5 Collaboration diagram of the proposed system

Chapter 5
Design

5.1 Architectural Design for proposed system

The architectural design of our proposed system would represent the software needs and design of the system. IEEE defines architectural design as “the process of defining a collection of hardware and software components and their interfaces to establish the framework for the development of a computer system.” The software that is built for computer-based systems can exhibit one of these many architectural styles.

Each style will describe a system category that consists of:

1. A set of components (for example: a database, computational modules) that will perform a function required by the system.
2. The set of connectors will help in coordination, communication, and cooperation between the components.
3. Conditions that how components can be integrated to form the system.
4. Semantic models that help the designer to understand the overall properties of the system.

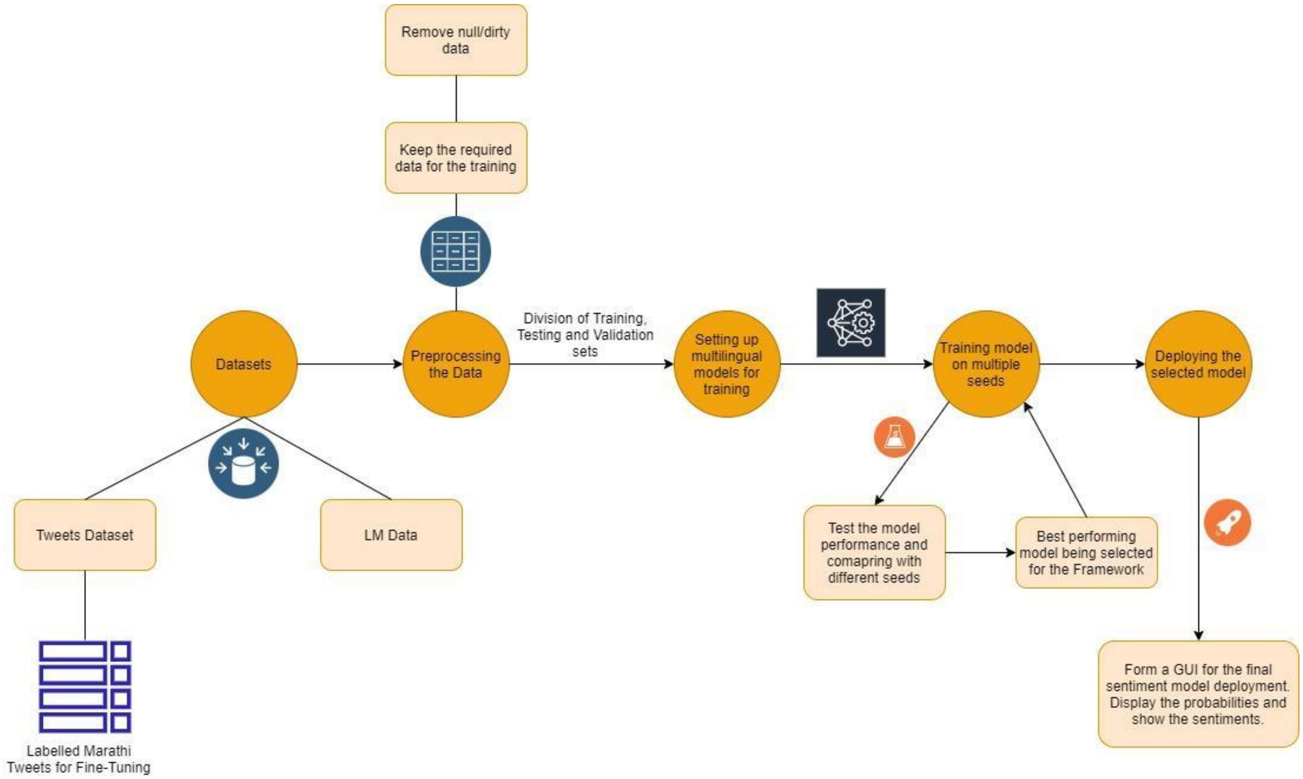

Remove null/dirty data
Keep the required data for the training
Datasets
Preprocessing the Data
Division of Training, Testing and Validation sets
Setting up multilingual models for training
Training model on multiple seeds
Deploying the selected model
Tweets Dataset
LM Data
Labelled Marathi Tweets for Fine-Tuning
Test the model performance and comapring with different seeds
Best performing model being selected for the Framework
Form a GUI for the final sentiment model deployment. Display the probabilities and show the sentiments.

Fig 5.1 Architectural design for the proposed system

This is the final architecture for the proposed framework. We will be processing the dataset we have used, that is created from scraping tweets from twitter. Models will be setup for the training and tested to achieve higher accuracies so that it can be deployed for the users to test.

Chapter 6
Implementation

6.1 Algorithms / Methods Used

XLM-RoBERTa - a multilingual language model, trained on 100 different languages, is used for our proposed task of opinion mining. The XLM-R model is created by FacebookAI using 2.5TB of CommonCrawl data over these 100 languages.

We fine-tune the XLM-R for marathi tweets multiple times over different seeds to achieve different and better results.

Best working model will be deployed over the GUI created for the users to use.

6.2 Working of the project

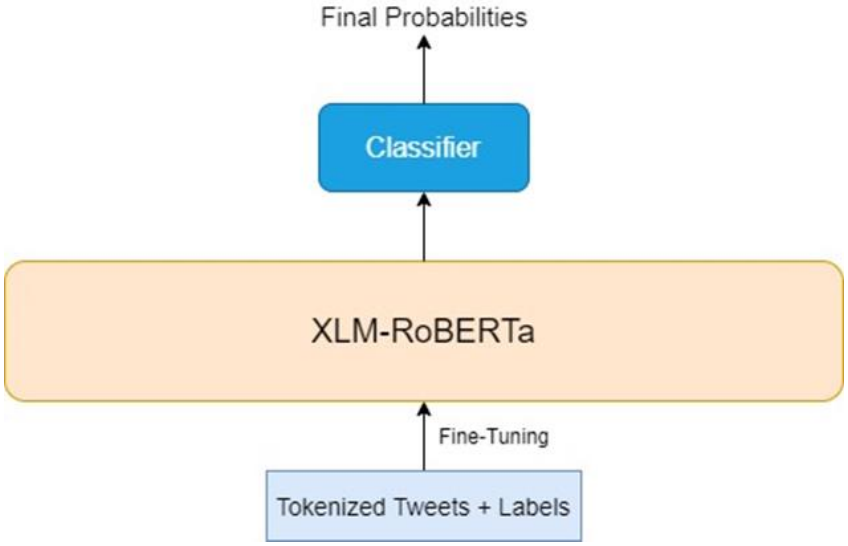


Fig 6.1: Training of XLM-R

The tweets were tokenized using Roberta Tokenizer. The dataset is divided in 3 parts as training, testing and validation datasets as 75%, 15% and 10% of the total data of around 16000 tweets.

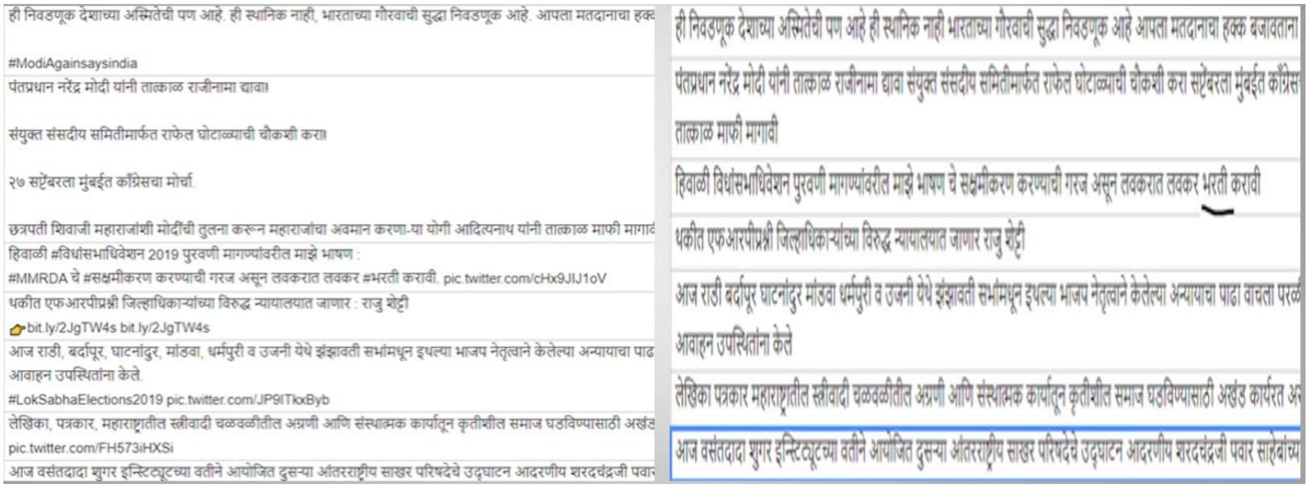
The models were run for 25 epochs to achieve best performance and avoiding overfitting of the models.

Three different seeds were tested for the XLM-R and we got the best results in the seed 40 training. The confusion matrix and training loss is displayed below for reference.

The model will display the probabilities of all three classes for the input text and the final sentiment will be displayed accordingly.

Chapter 7

Results and Discussions



ही निवडणूक देशाच्या अस्मितेची पण आहे. ही स्थानिक नाही, भारताच्या गौरवाची सुद्धा निवडणूक आहे. आपला मतदानाचा हक्

#ModiAgainsaysindia

पंतप्रधान नरेंद्र मोदी यांनी तात्काळ राजीनामा द्यावा।

संयुक्त संसदीय समितीमार्फत राफेल घोटाळ्याची चौकशी करा।

२७ सप्टेंबरला मुंबईत काँग्रेसचा मोर्चा.

छत्रपती शिवाजी महाराजांशी मोदींची तुलना करून महाराजांचा अवमान करणा-या योगी आदित्यनाथ यांनी तात्काळ माफी मागाव

हिवाळी #विधांसभाधिवेशन 2019 पुरवणी मागण्यांवरील माझे भाषण :
#MMRDA चे #सक्षमीकरण करण्याची गरज असून लवकरात लवकर #भरती करावी. pic.twitter.com/cHx9JIJ1oV

धकीत एफआरपीप्रश्नी जिल्हाधिकाऱ्यांच्या विरुद्ध न्यायालयात जाणार : राजु शेट्टी
👉bit.ly/2JgTW4s bit.ly/2JgTW4s

आज राडी, बर्दापूर, घाटनांदुर, मांडवा, धर्मपुरी व उजनी येथे झंझावती सभांमधून इथल्या भाजप नेतृत्वाने केलेल्या अन्यायाचा पाढ
आवाहन उपस्थितांना केले.
#LokSabhaElections2019 pic.twitter.com/JP9ITkxByb

लेखिका, पत्रकार, महाराष्ट्रातील स्त्रीवादी चळवळीतील अग्रणी आणि संस्थात्मक कार्यातून कृतीशील समाज घडविण्यासाठी अखंड
pic.twitter.com/FH573iHXSi

आज वसंतदादा शुगर इन्स्टिट्यूटच्या वतीने आयोजित दुसऱ्या आंतरराष्ट्रीय साखर परिषदेचे उद्घाटन आदरणीय शरदचंद्रजी पवा

ही निवडणूक देशाच्या अस्मितेची पण आहे ही स्थानिक नाही भारताच्या गौरवाची सुद्धा निवडणूक आहे आपला मतदानाचा हक्क बजावताना

पंतप्रधान नरेंद्र मोदी यांनी तात्काळ राजीनामा द्यावा संयुक्त संसदीय समितीमार्फत राफेल घोटाळ्याची चौकशी करा सप्टेंबरला मुंबईत काँग्रेस
तात्काळ माफी मागावी

हिवाळी विधांसभाधिवेशन पुरवणी मागण्यांवरील माझे भाषण चे सक्षमीकरण करण्याची गरज असून लवकरात लवकर भरती करावी

धकीत एफआरपीप्रश्नी जिल्हाधिकाऱ्यांच्या विरुद्ध न्यायालयात जाणार राजु शेट्टी

आज राडी बर्दापूर घाटनांदुर मांडवा धर्मपुरी व उजनी येथे झंझावती सभांमधून इथल्या भाजप नेतृत्वाने केलेल्या अन्यायाचा पाढा वाचला परळ
आवाहन उपस्थितांना केले

लेखिका पत्रकार महाराष्ट्रातील स्त्रीवादी चळवळीतील अग्रणी आणि संस्थात्मक कार्यातून कृतीशील समाज घडविण्यासाठी अखंड कार्यरत अ

आज वसंतदादा शुगर इन्स्टिट्यूटच्या वतीने आयोजित दुसऱ्या आंतरराष्ट्रीय साखर परिषदेचे उद्घाटन आदरणीय शरदचंद्रजी पवार साहेबांच्या

Fig 7.1: Final Dataset

The dataset before and finally obtained after the pre-processing is shown above in Fig 7.1. Links, mentions, emojis and text in different languages was cleaned and removed before the dataset is used. This final set is divided for training of the models.

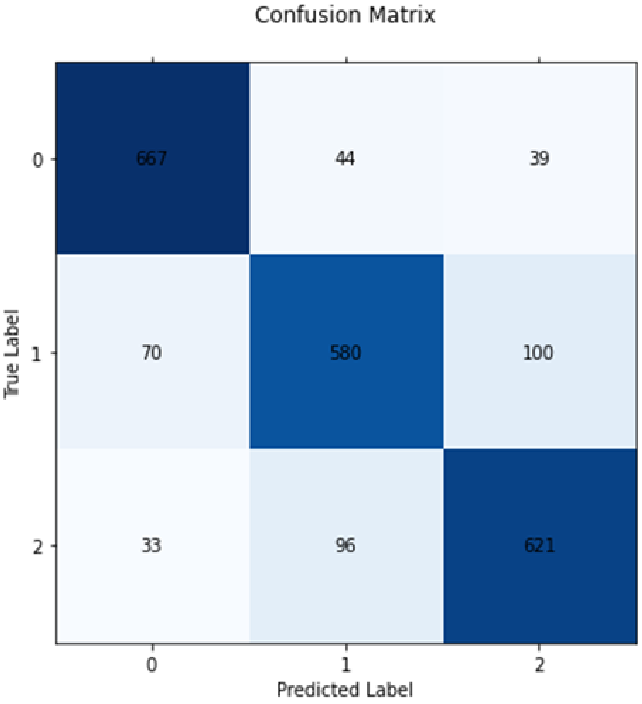


Fig 7.2: Confusion Matrix for Seed 40 - XLM-R

The Fig 7.2 shows the confusion matrix for the testing of seed 40 model. The model correctly classifies over 83% of the data and gives better accuracy than seeds 10 and 20.

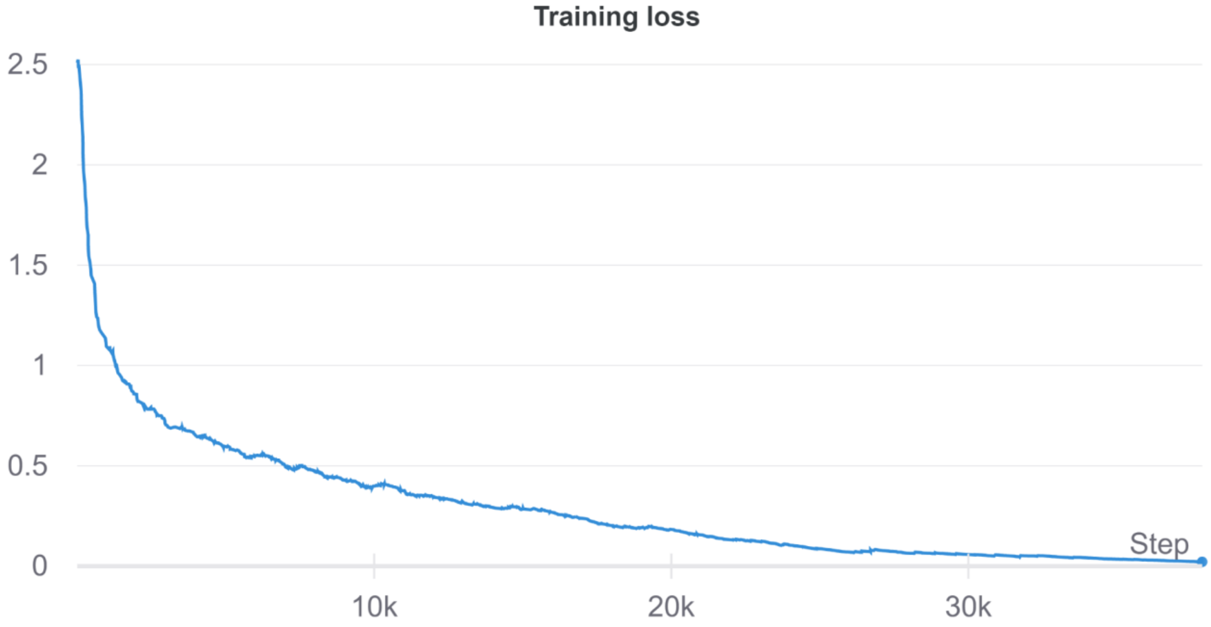


Fig 7.3: Training loss for seed 40 - XLM-R

The Fig 7.3 above shows the training loss decreasing gradually as we train for the 25 epochs. Overfitting is avoided by limiting epochs to 25 in all the models.

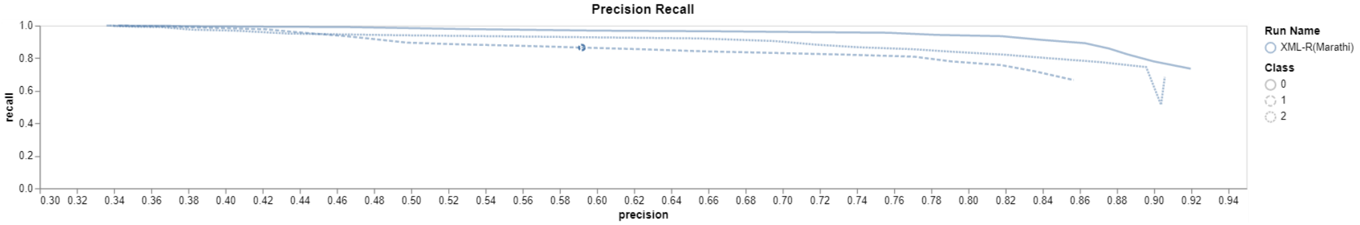
This final model will be deployed in the GUI that will be created where people will be able to test it for the marathi social media texts.

Chapter 8

Conclusions and Future Scope

In this paper, we have used the XLM-R model due to its higher accuracy and better performance for NLP on low resource languages like Marathi.We use three different seeds and received the highest accuracy of 83% on seed 40. We aim to deploy the model over a GUI for sentiment analysis of Marathi text where users will get the feedback on the input text with the sentiment probabilities and a final sentiment according to these values.This will help governments in states like Maharashtra and Goa where Marathi is the most widely spoken language to analyse responses on government schemes and make necessary changes if required.This can also be deployed by social media intermediaries to flag the hateful content helping in removing of these toxic texts help in maintaining social harmony along with saving the modesty of a person especially women who bear the unequal burden of social media bullying.Researchers further can strive to achieve higher accuracy by using expanded datasets,higher trained language models.

Appendix



Precision and Recall graph for the best performed model - XLM-R-seed 40

Literature Cited

[1] Kulkarni, Atharva, Meet Mandhane, Manali Likhitkar, Gayatri Kshirsagar, Jayashree Jagdale, and Raviraj Joshi. "Experimental evaluation of deep learning models for marathi text classification." *arXiv preprint arXiv:2101.04899* (2021).
https://arxiv.org/abs/2101.04899

[2] De, Arkadipta, Venkatesh Elangovan, Kaushal Kumar Maurya, and Maunendra Sankar Desarkar. "Coarse and fine-grained hostility detection in Hindi posts using fine tuned multilingual embeddings." In *International Workshop on Combating On line Ho st ile Posts in Regional Languages during Emergency situation*, pp. 201-212. Springer, Cham, 2021.
https://link.springer.com/chapter/10.1007/978-3-030-73696-5_19

[3] Seliverstov, Yaroslav A., Andrew A. Komissarov, Eleonora D. Poslovskaia, Alina A. Lesovodskaya, and Artur V. Podtikhov. "Detection of Low-toxic Texts in Similar Sets Using a Modified XLM-RoBERTa Neural Network and Toxicity Confidence Parameters." In *2021 XXIV International Conference on Soft Computing and Measurements (SCM)*, pp. 161-164. IEEE, 2021.
https://ieeexplore.ieee.org/abstract/document/9507117

[4] Ghafoor, Abdul, Ali Shariq Imran, Sher Muhammad Daudpota, Zenun Kastrati, Rakhi Batra, and Mudasir Ahmad Wani. "The Impact of Translating Resource-Rich Datasets to Low-Resource Languages Through Multi-Lingual Text Processing." *IEEE Access* 9 (2021): 124478-124490.
https://ieeexplore.ieee.org/abstract/document/9529190

[5]Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. "Unsupervised cross-lingual representation learning at scale." *arXiv preprint arXiv:1911.02116* (2019). https://arxiv.org/abs/1911.02116

[6] Zaharia, George-Eduard, George-Alexandru Vlad, Dumitru-Clementin Cercel, Traian Rebedea, and Costin-Gabriel Chiru. "Upb at semeval-2020 task 9: Identifying sentiment in code-mixed social media texts using transformers and multi-task learning." *arXiv preprint arXiv:2009.02780* (2020).
https://arxiv.org/abs/2009.02780

[7] Gupta, Vedika, Nikita Jain, Shubham Shubham, Agam Madan, Ankit Chaudhary, and Qin Xin. "Toward Integrated CNN-based Sentiment Analysis of Tweets for Scarce-resource Language—Hindi." *Transactions on Asian and Low-Resource Language Information Processing* 20, no. 5 (2021): 1-23.
https://dl.acm.org/doi/abs/10.1145/3450447

[8] Meetei, Loitongbam Sanayai, Thoudam Doren Singh, Samir Kumar Borgohain, and Sivaji Bandyopadhyay. "Low resource language specific pre-processing and features for sentiment analysis task." *Language Resources and Evaluation* (2021): 1-23.
https://link.springer.com/article/10.1007/s10579-021-09541-9

[9] Gupta, Itisha, and Nisheeth Joshi. "Feature-Based Twitter Sentiment Analysis With Improved Negation Handling." *IEEE Transactions on Computational Social Systems* (2021).
https://ieeexplore.ieee.org/abstract/document/9399630

Acknowledgements

We would like to express our special thanks of gratitude to our project guide Asst Prof Dr.Pratik Kanani for their able guidance, support and suggestions which helped us in completing this project. We would also like to extend our gratitude to our Principal, Dr. Hari Vasudevan and the Head of the Computer Engineering Department, Dr. Meera Narvekar for providing us with all the facility that were required for completion of this project.