A FRAMEWORK FOR SOCIAL MEDIA OPINION MINING FOR LOW RESOURCE MARATHI TEXT

A.Y. 2021-22

A FRAMEWORK FOR SOCIAL MEDIA OPINION MINING FOR LOW RESOURCE MARATHI TEXT

Submitted in partial fulfilment of the requirements
of the degree of

B. E. Computer Engineering

By

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

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

2021-2022

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), "Manan Parikh" (60004180049), "Naitik Rathod" (60004180054) and "Nishit Mistry" (60004180066)** submitted to the University of Mumbai in partial fulfilment 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 for B.E.

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

Examiners

1. ____________________

2. ____________________

Date:

Place:

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.

Dhruv Talati
60004180022

Manan Parikh
60004180049

Naitik Rathod
60004180054

Nishit Mistry
60004180066

Date:

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 Tables

List of Abbreviations

Sr. No.	Abbreviations	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	ULMFiT	Universal Language Model Fine Tuning
vii	BERT	Bidirectional Encoder Representations From Transformers
viii	SVM	Support Vector Machine
ix	OWA	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	AI	Artificial Intelligence
xviii	TB	Tera Byte
xix	GUI	Graphical User Interface
xx	BPE	Byte-Pair Encoding

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 and 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

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.

[3] proposes 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.

The research done in [4] shows 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.

[6] has 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 sentiment analysis of low resource language Hindi is still lacking in richly populated linguistic resources, owing to the challenges involved in dealing with the Hindi language is displayed in the research done by authors in [7]. Hindi, is the fourth-most popular 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.

[9] presents 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

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 and also perform better than the models where Language Models are trained over news headings and articles.

Chapter 3

System Analysis

3.1 Functional Requirements

3.1.1 Get the Marathi Tweets

Download, filter, and store the required data in the local database. Structure the data with the labels as required for the sentiments.

3.1.2 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.

3.1.3 Requesting Sentiment

The users will enter their sentence to get the respective opinions generated by the model trained.

3.1.4 Feature Extraction and Learning

The system should be able to tokenize and understand the features of the sentence put to test to get the most appropriate result of the task.

3.1.5 Displaying Sentiment Probabilities

For each sentence tested, the probabilities of all the classes will be displayed and classified into the best fit sentiment over the GUI.

3.2 Non-Functional Requirements

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

3.2.1 Performance requirements

1. ***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.
2. ***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.
3. ***Reliability*** - Dataset used is taken from twitter and is processed according to our requirements.

3.2.2 Design constraints

1. ***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.
2. ***Software System Attributes*** - The system should be extensible to guarantee that it is useful for a reasonable period. Latest available dataset is being used for this task.
   (a) ***Usability*** - The model will be embedded in the backend of an application. It should be scalable designed to be easily adopted by a system.
   (b) ***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. The use-case diagram for our proposed system is displayed below in 3.1

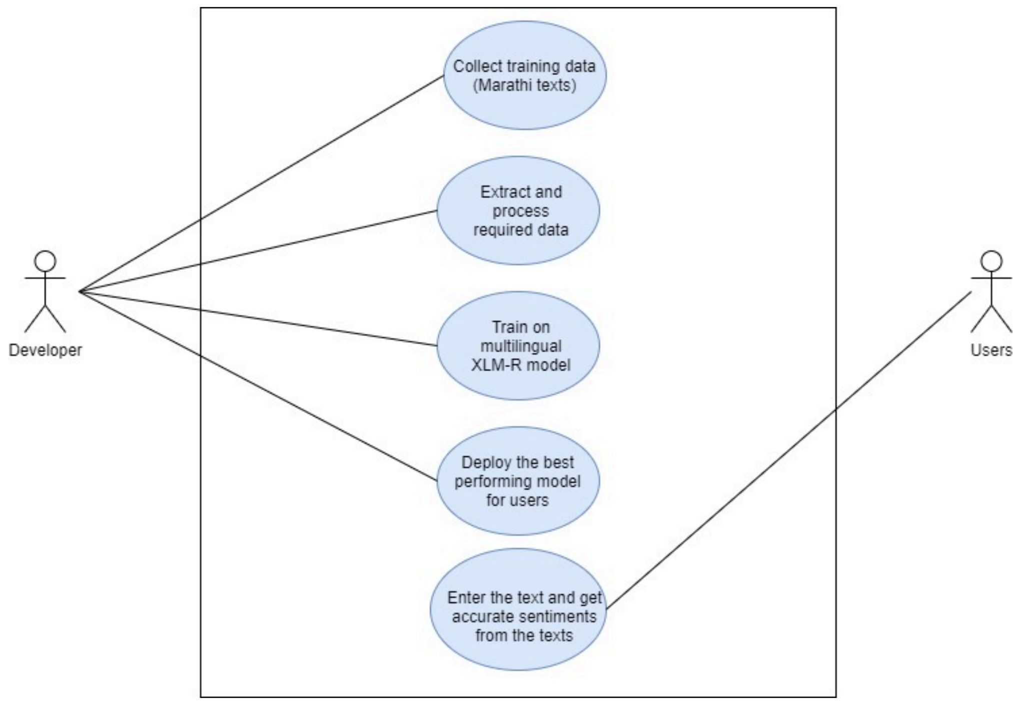

Figure 3.1: Use-case diagram of the proposed system.

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 in 4.1 for the proposed system.

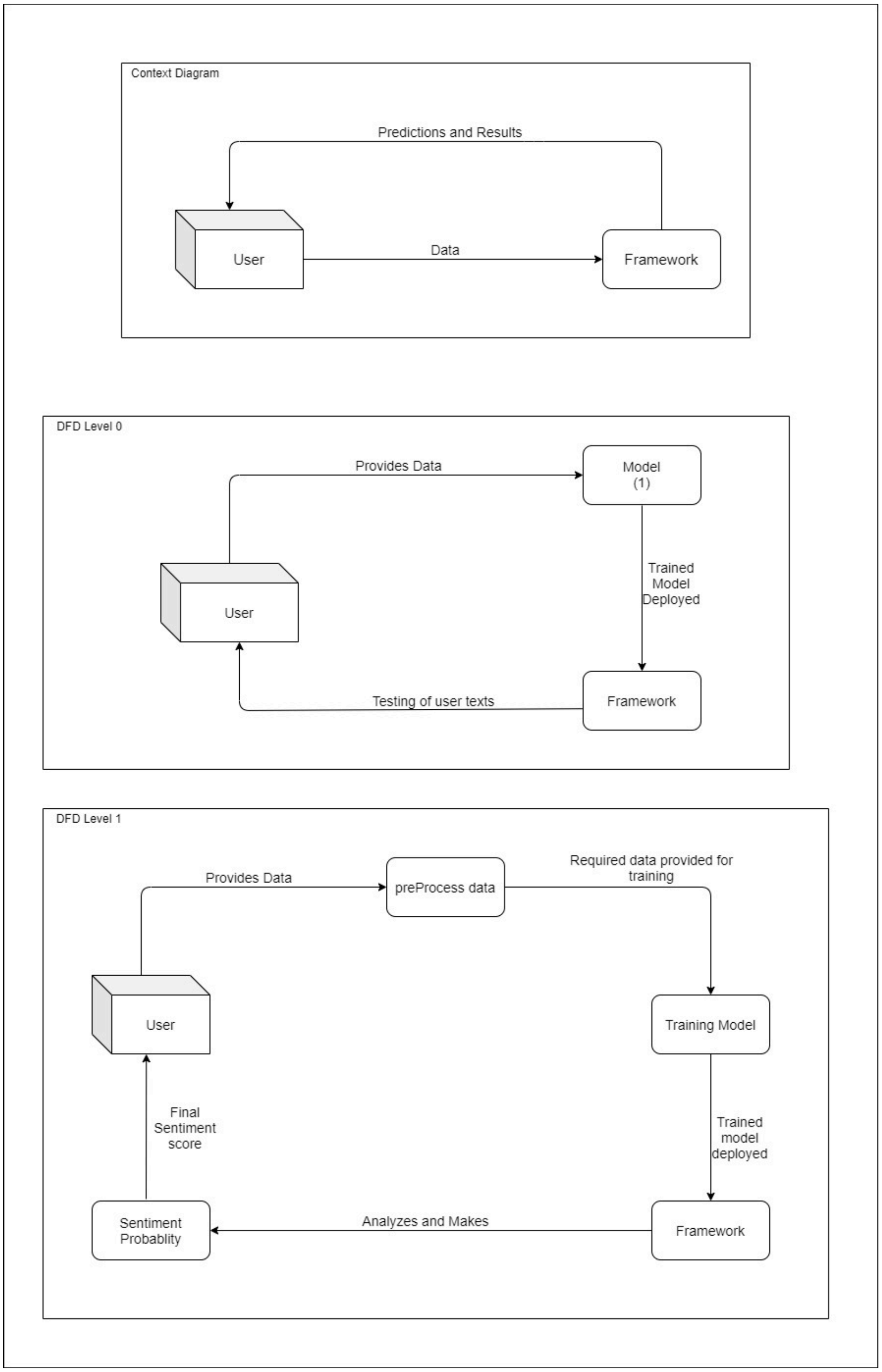

Figure 4.1: Data-Flow Diagram of the proposed system.

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. The class diagram for the proposed system is shown below in 4.2

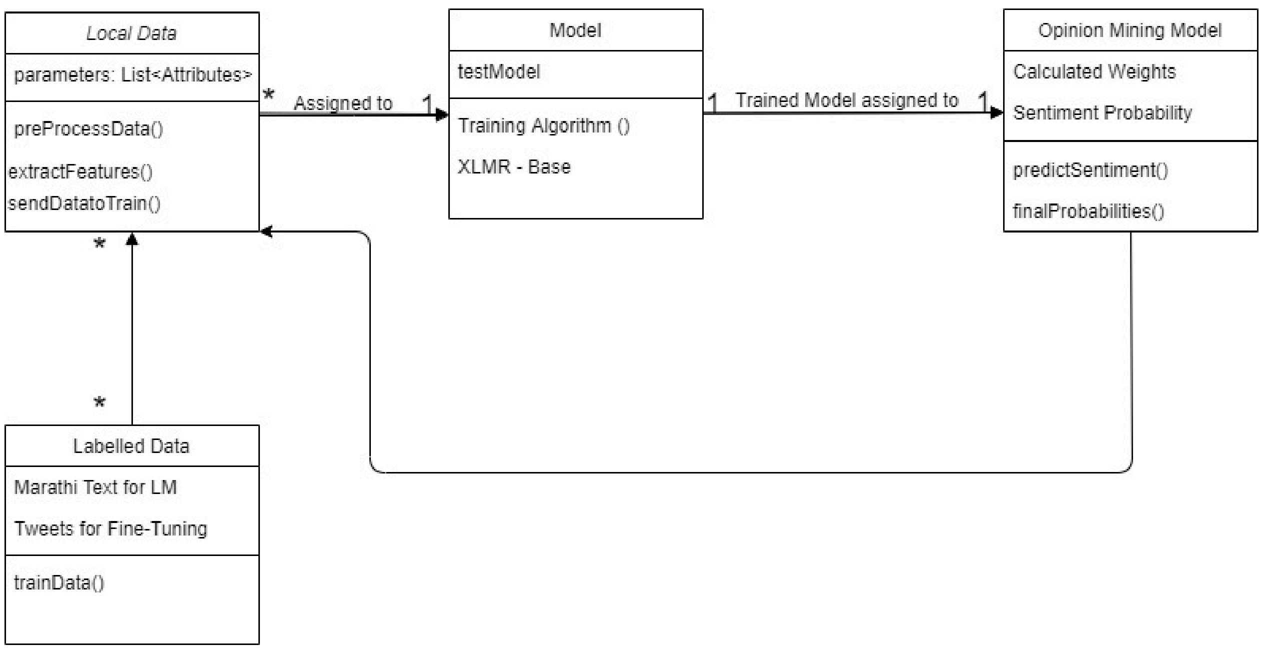

Figure 4.2: Class diagram of the proposed system.

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.

The State Diagram for our proposed system is displayed below in 4.3.

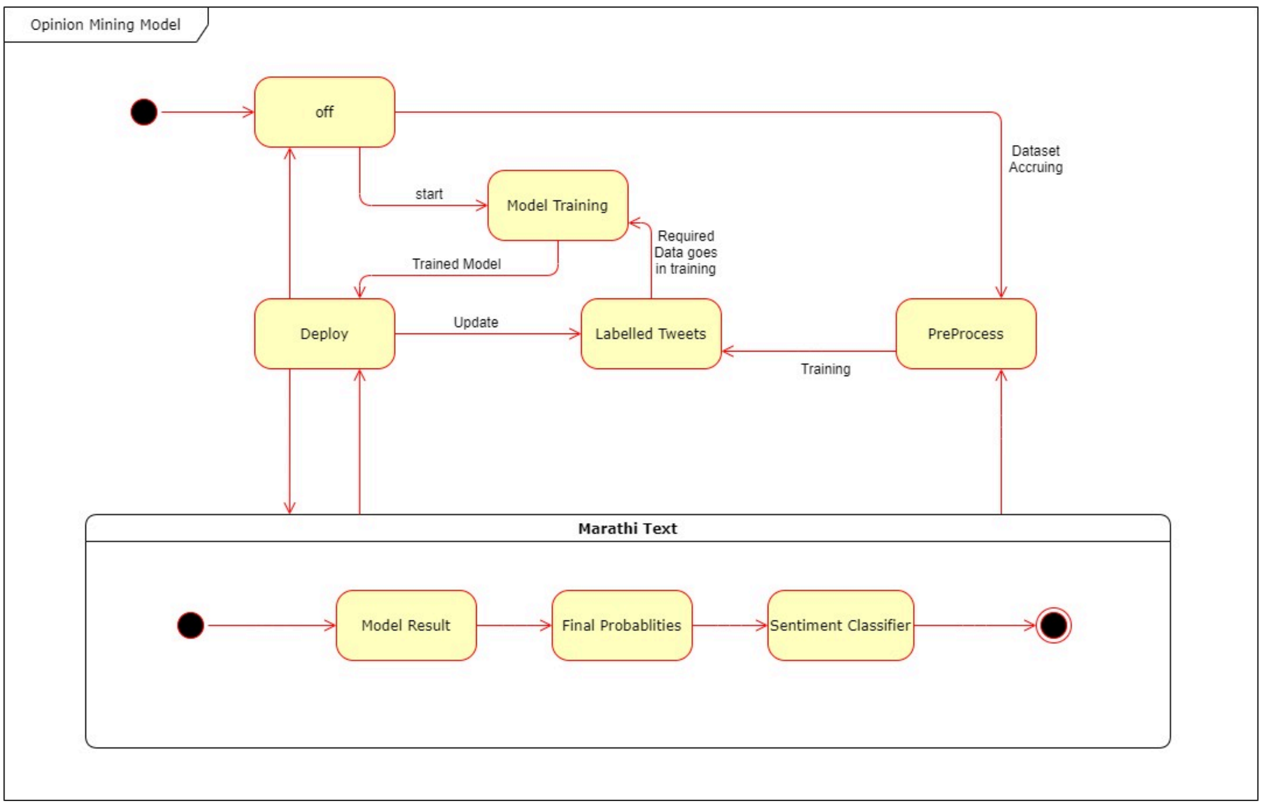

Figure 4.3: State diagram of the proposed system.

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.

The Activity Diagram for our system is shown below in 4.4.

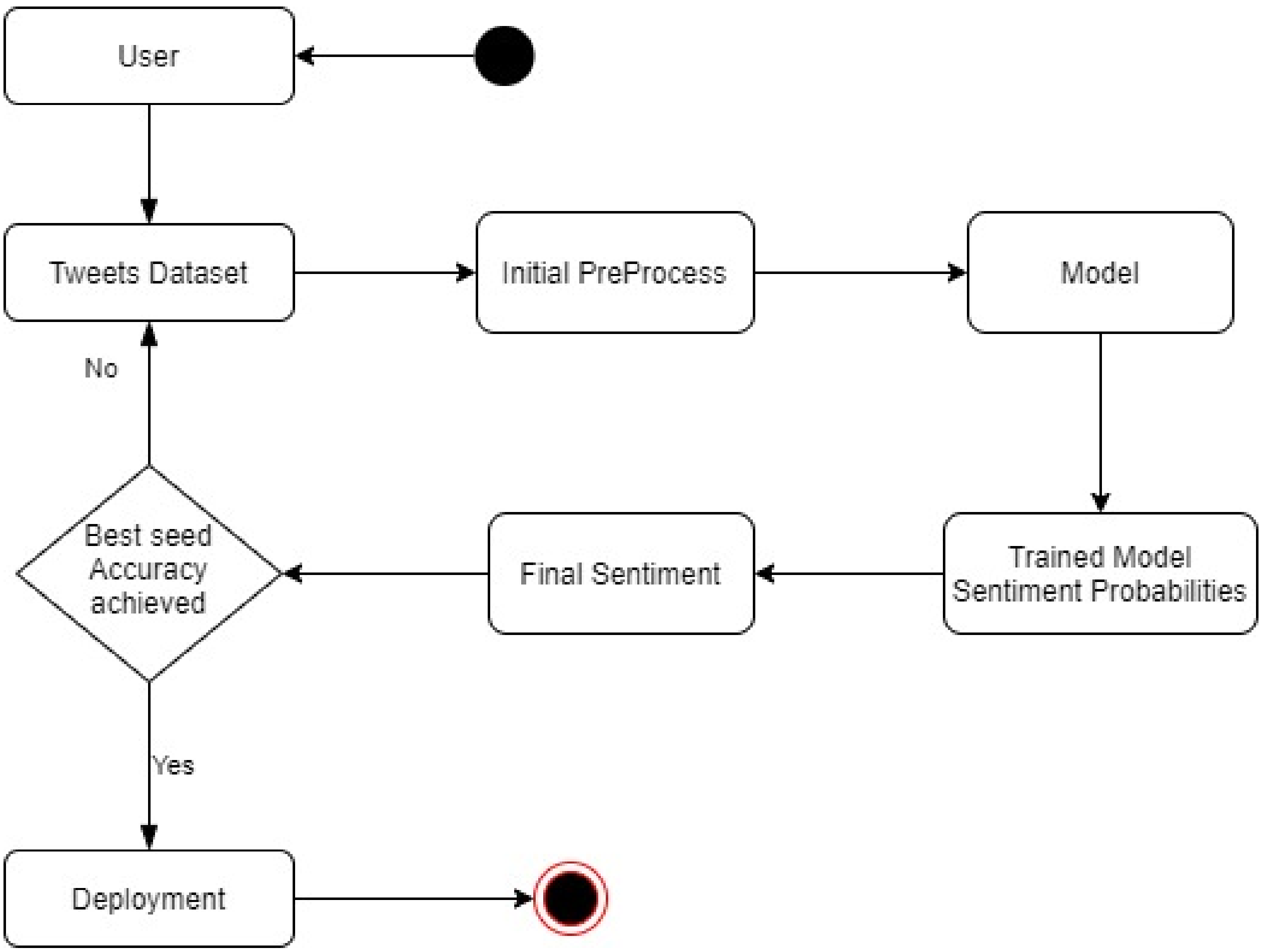

Figure 4.4: Activity diagram of the proposed system.

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.

The Sequence Diagram for our system is shown below in 4.5.

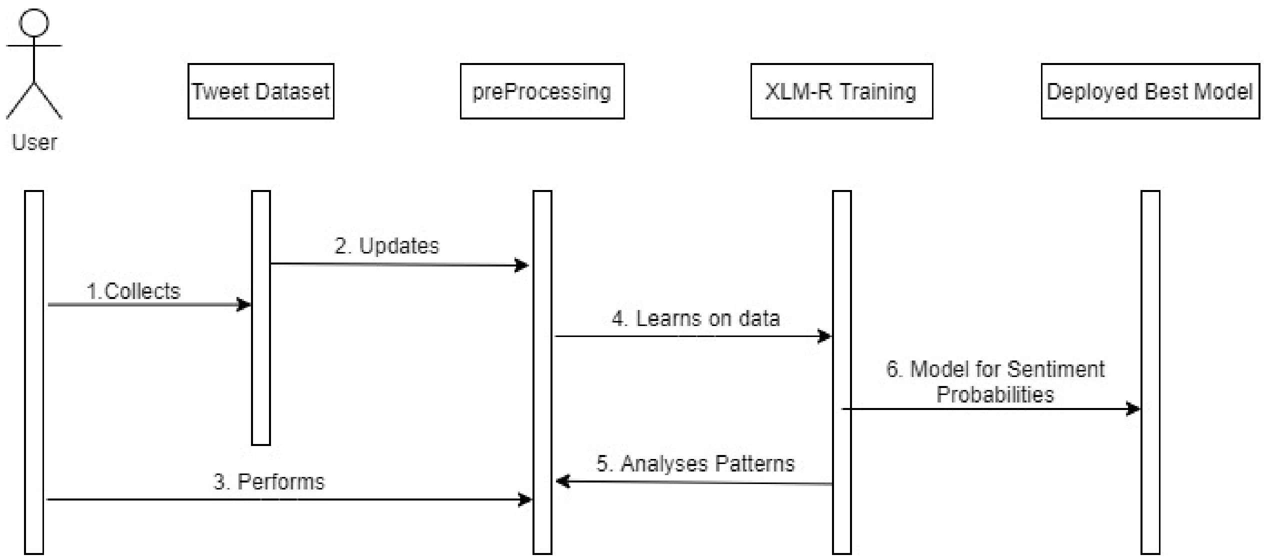

Figure 4.5: Sequence diagram of the proposed system.

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.

The Collaboration Diagram for our system is shown below in 4.6.

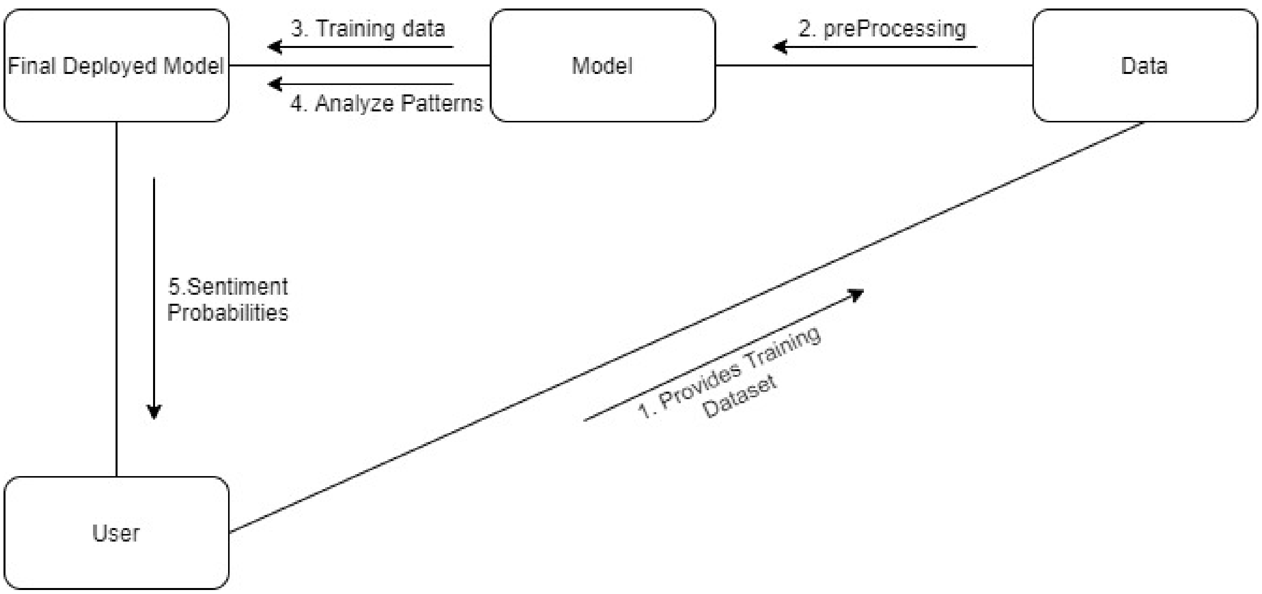

Figure 4.6: Collaboration diagram of the proposed system.

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.

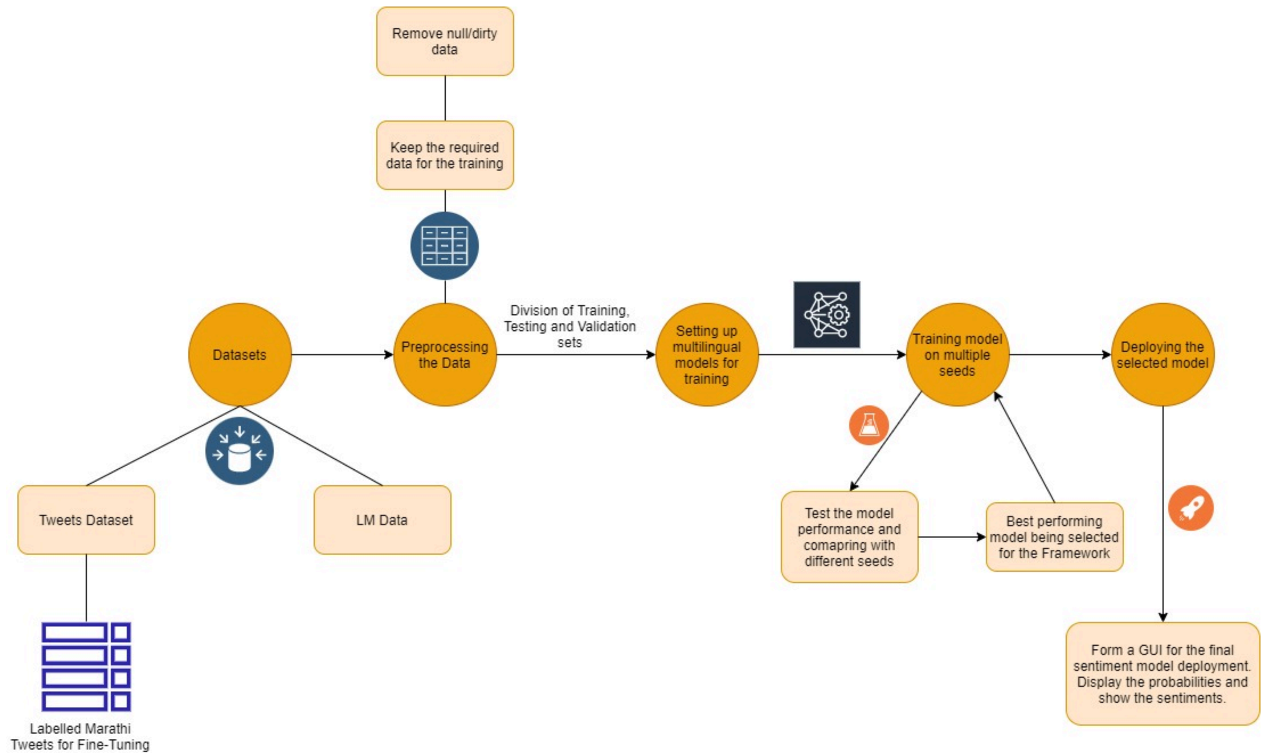

Figure 5.1: System Architecture Diagram of the proposed system.

The Figure 5.1 is the final architecture for the proposed framework for sentiment analysis of Marathi Social Media texts. We will be processing the dataset we have used, that is created from scraping tweets from twitter. XLM-RoBERTa Models will be setup for the training over the same dataset and tested for accuracies so that it can be deployed for the users to use for their analysis of Marathi texts.

Chapter 6
Implementation

6.1 Algorithms / Methods Used

6.1.1 Dataset used

We have used the publicly available L3CubeMahaSent [1] Twitter dataset, which happens to be the first publicly available dataset in Marathi language for the task of Twitter Sentiment Analysis. This corpus was released in 2021 alongside their experiments on the baseline models available for sentiment analysis. This includes approximately 15900 Marathi tweets manually classified into the 3 classes. OUr end goal is tweet polarity classification, by classifying a tweet into three categories according to their polarity, the three categories being positive, neutral and negative.

6.1.2 Model Description

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 will use the base and even the large variant of XLM-R for our task. 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. The Figure 6.1 shows the architecture of the XLM-RoBERTa model.

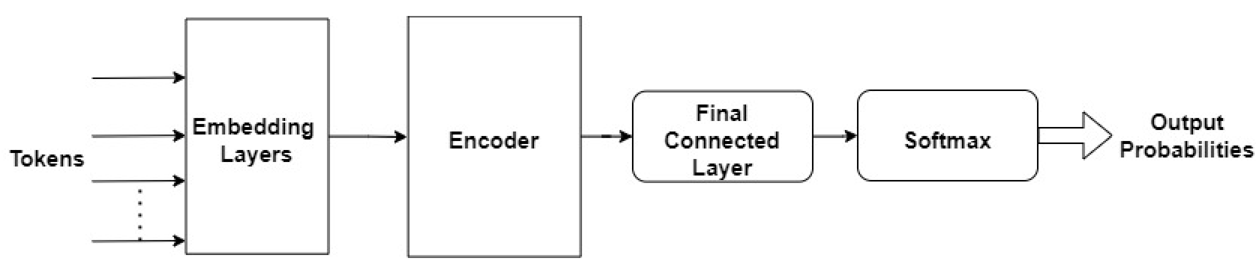

Figure 6.1: Architecture of XLM-RoBERTa

6.2 Working of the Model

The tweets were tokenized using Roberta Tokenizer. Roberta Tokenizer is used for tokenizing the tweets before training of the models. It uses byte level BPE as a tokenizer. It treats spaces as parts of the tokens so it is treated differently at the front of a word and the back of a word. This tokenizer is derived from the GPT-2 tokenizer and is commonly used for tokenizing for the language models. 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 Training Architecture for our system is shown below in 6.2.

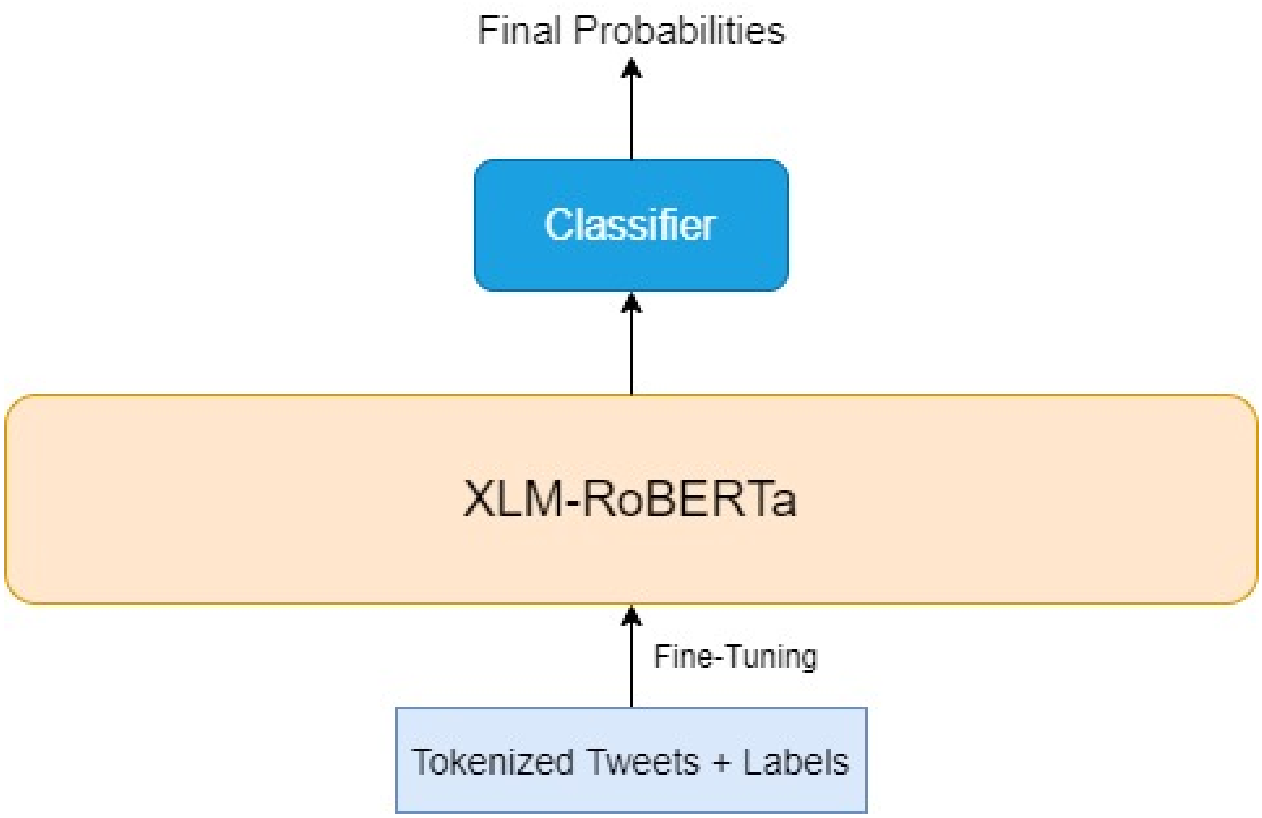

Figure 6.2: Training Architecture of the proposed system.

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 base and large. 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

The results from Table 7.1 show us that large variant of the XLM-R performs the best over this Marathi dataset and has comparable results with other models as in [1]. Additionally, accuracy achieved in this work using XLM-R large is better than the base model for the task of Marathi sentiment analysis using XLM-R models for three class classification.

Table 7.1: Model Results

Model	Accuracy(%)
XLM-R base	82.5
XLM-R large	83.82

Figure 7.1: Training Loss graph

Authors observed from the training loss graph in Figure 7.1 that there is no overfitting taking place during the training of the model. Limiting the number of epochs takes care of the case of overfitting in the language model training.

The precision-recall curve as shown in the Figure 7.2 shows values for the precision and the recall for different threshold. The high area under the curve represents that both recall and precision

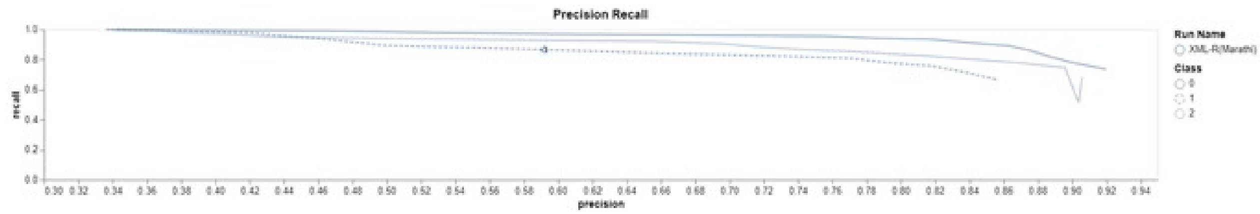

Figure 7.2: Precision Recall curve of XLM-R training

are high; where high precision proves there are less false positives, and high recall proves there are low false negatives. Summarizing the curve, high scores for both precision and recall can be used to conclude that the classifier performance is accurate to a very high extent due to higher precision and a maximum of all positive results due to high recall.

Chapter 8

Conclusion & Future Scope

This project explains the training of Marathi Opinion Mining framework using the transformer XLM-R and its variants without use of translations. The model can generate sentiments for the Marathi texts we test it for. The quality of Opinion Mining system can be increased by using larger models given larger computing power is available. The goal of this research was to create a system which can be trained using less data and low resources. With this research, multiple languages models can be prepared for similar tasks given we have the availability of the labelled datasets.

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 and higher trained language models.

Literature Cited

[1] A. Kulkarni, M. Mandhane, M. Likhitkar, G. Kshirsagar, J. Jagdale, and R. Joshi, "Experimental evaluation of deep learning models for marathi text classification," *CoRR*, 2021. arXiv: 2101.04899. [Online]. Available: arxiv.org/abs/2101.04899.

[2] A. De, V. Elangovan, K. K. Maurya, and M. S. Desarkar, "Coarse and fine-grained hostility detection in hindi posts using fine tuned multilingual embeddings," in *Combating Online Hostile Posts in Regional Languages during Emergency Situation*, T. Chakraborty, K. Shu, H. R. Bernard, H. Liu, and M. S. Akhtar, Eds., Cham: Springer International Publishing, 2021, pp. 201–212, ISBN: 978-3-030-73696-5.

[3] Y. A. Seliverstov, A. A. Komissarov, E. D. Poslovskaia, A. A. Lesovodskaya, and A. 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)*, 2021, pp. 161–164. DOI: 10.1109/SCM52931.2021.9507117.

[4] A. Ghafoor, A. S. Imran, S. M. Daudpota, *et al.*, "The impact of translating resource-rich datasets to low-resource languages through multi-lingual text processing," *IEEE Access*, vol. 9, pp. 124 478–124 490, 2021. DOI: 10.1109/ACCESS.2021.3110285.

[5] A. Conneau, K. Khandelwal, N. Goyal, *et al.*, "Unsupervised cross-lingual representation learning at scale," *CoRR*, vol. abs/1911.02116, 2019. arXiv: 1911.02116. [Online]. Available: http://arxiv.org/abs/1911.02116.

[6] G. Zaharia, G. Vlad, D. Cercel, T. Rebedea, and C. Chiru, "UPB at semeval-2020 task 9: Identifying sentiment in code-mixed social media texts using transformers and multi-task learning," *CoRR*, vol. abs/2009.02780, 2020. arXiv: 2009.02780. [Online]. Available: https://arxiv.org/abs/2009.02780.

[7] V. Gupta, N. Jain, S. Shubham, A. Madan, A. Chaudhary, and Q. Xin, "Toward integrated cnn-based sentiment analysis of tweets for scarce-resource language—hindi," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 20, no. 5, Jun. 2021, ISSN: 2375-4699. DOI: 10.1145/3450447. [Online]. Available: https://doi.org/10.1145/3450447.

[8] L. Meetei, T. D. Singh, S. Borgohain, and S. Bandyopadhyay, "Low resource language specific pre-processing and features for sentiment analysis task," *Language Resources and Evaluation*, vol. 55, Dec. 2021. DOI: 10.1007/s10579-021-09541-9.

[9] I. Gupta and N. Joshi, "Feature-based twitter sentiment analysis with improved negation handling," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 4, pp. 917–927, 2021. DOI: `10.1109/TCSS.2021.3069413`.

[10] N. Goyal, J. Du, M. Ott, G. Anantharaman, and A. Conneau, "Larger-scale transformers for multilingual masked language modeling," in *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, Online: Association for Computational Linguistics, Aug. 2021, pp. 29–33. DOI: `10.18653/v1/2021.repl4nlp-1.4`. [Online]. Available: `https://aclanthology.org/2021.repl4nlp-1.4`.

[11] P. B. Bafna and J. R. Saini, "Scaled document clustering and word cloud-based summarization on hindi corpus," in *Progress in Advanced Computing and Intelligent Engineering*, C. R. Panigrahi, B. Pati, P. Mohapatra, R. Buyya, and K.-C. Li, Eds., Singapore: Springer Singapore, 2021, pp. 398–408, ISBN: 978-981-15-6353-9.

[12] X. Li, L. Bing, P. Li, and W. Lam, "A unified model for opinion target extraction and target sentiment prediction," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 6714–6721, Jul. 2019. DOI: `10.1609/aaai.v33i01.33016714`. [Online]. Available: `https://ojs.aaai.org/index.php/AAAI/article/view/4643`.

[13] R. Joshi, P. Goel, and R. Joshi, "Deep learning for hindi text classification: A comparison," in *Intelligent Human Computer Interaction*, U. S. Tiwary and S. Chaudhury, Eds., Cham: Springer International Publishing, 2020, pp. 94–101, ISBN: 978-3-030-44689-5.

[14] M. S. Divate, "Sentiment analysis of marathi news using lstm," *International Journal of Information Technology*, 2021.

[15] R. Joshi, R. Karnavat, K. Jirapure, and R. Joshi, "Evaluation of deep learning models for hostility detection in hindi text," *CoRR*, vol. abs/2101.04144, 2021. arXiv: `2101.04144`. [Online]. Available: `https://arxiv.org/abs/2101.04144`.

[16] A. Wani, I. Joshi, S. Khandve, V. Wagh, and R. Joshi, "Evaluating deep learning approaches for covid19 fake news detection," in *Combating Online Hostile Posts in Regional Languages during Emergency Situation*, T. Chakraborty, K. Shu, H. R. Bernard, H. Liu, and M. S. Akhtar, Eds., Cham: Springer International Publishing, 2021, pp. 153–163, ISBN: 978-3-030-73696-5.

[17] V. Gupta, N. Jain, P. Katariya, *et al.*, "An emotion care model using multimodal textual analysis on covid-19," *Chaos, Solitons Fractals*, vol. 144, p. 110 708, 2021, ISSN: 0960-0779.

DOI: https://doi.org/10.1016/j.chaos.2021.110708. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960077921000618.

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 and the publication for 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.