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

Opinion Mining is an essential subject in Natural Language Processing and one of the most significant jobs for any language. Popular and commonly used languages like as English, Russian, and Spanish have a large number of language models and datasets accessible for these purposes. But the research in Low Resource Languages like Hindi and Marathi is far behind. The Marathi language is one of India's most widely spoken languages, being the third most spoken language. It is mostly spoken by Maharashtra residents. The use of language on internet platforms has expanded dramatically during the last decade. Natural Language Processing (NLP) techniques for Marathi text, on the other hand, have gotten little 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 analysis of sentiments of the tweets using ML models. However there has been minuscule 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. According to the 2011 census, Marathi is India's third most spoken native language, with 83 million native speakers.

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 detailed review of the tools and approaches for Marathi text categorization that are currently available. On two datasets, namely the Marathi News Headline Dataset and the Marathi News Articles Dataset, the authors assessed several CNN, LSTM, BiLSTM, ULMFiT and BERT models. The presented model proposed by the authors in [1] wworks with a large amount of data from news sources to pre-train them. Because the target datasets are from the news domain, they have a greater level of accuracy, while the accuracy drops for non-news items.

The authors of [2] present an effective neural network-based technique for detecting aggression in hindi text. The authors tested a variety of models, including SVM, RF, BiLSTM, and pre-trained language models that are BERT versions. They selected a dataset of hostile and non-hostile hindi language from social media platforms such as Twitter, Facebook, and Whatsapp, among others, and annotated it with fine-grained categories such as false, hatred, defamation, and offensiveness. The datasets were split into four fine-grained labels using two alternative methods: OnevsRest(OvR) and Direct Multi-Label Multi-Classification(DMLMC), which were both utilised for training and testing, and the results were compared. On individual mBERT and XLM-R with their specified parameters, authors in [2] received 91.63 percent and 89.76 percent accuracy, respectively, while the hybrid received accuracy of 92.6 percent, which is the greatest performance among all the models used for 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 study in [4] shows how translation affects the sentiment classification job while moving from a resource-rich to a low-resource language. It classifies and categorises words that cause polarity shifts into five groups. It also establishes a link between languages with similar roots. Our research reveals that polarity shift as a result of translation from resource-rich to low-resource languages degrades sentiment categorization performance by 2-3 percentage points. To investigate how to generate a sentiment analysis dataset for low-resource languages using a translation ap-

proach. To compare the classification results of all languages after translating the English reviews into German, Urdu, and Hindi. The importance of managing Negation-affected terms has been studied by authors. The authors demonstrated that ”Faultless Production” was translated into Urdu as ”Bad Production” using Google Translate. This translation is erroneous, demonstrating yet again how Negation influences translation.

The authors of this study [5] show that pre-training multilingual language models at scale results in considerable performance enhancements for a variety of cross-lingual transfer tasks. The authors used more than two terabytes of filtered CommonCrawl data to train a Transformer-based masked language model on one hundred languages. On a range of cross-lingual benchmarks, the authors' model, termed XLM-R, surpasses multilingual BERT (mBERT). The authors found that XLM-R outperforms earlier XLM models in low-resource languages, improving XNLI accuracy by 15.7 percent for Swahili and 11.4 percent for Urdu. The authors present XLM-R, a new state-of-the-art multilingual masked language model, and demonstrate that it outperforms earlier multilingual models such as mBERT and XLM in terms of categorization and sequence labelling.

For the classification task of SemEval 2020, [6] has presented an effective neural network based solution for two code mixed languages: Hindi-English and Spanish-English. They used the SemEval dataset for these two mixed languages. The datasets were processed before being used in models such as BiLSTM, mBERT, and XLM-R. The Hindi-English dataset includes 17000 labelled social media texts, whereas the Spanish-English dataset has 15000 identified texts. All text is categorised as either good, negative, or neutral. The authors have demonstrated that good word embeddings can significantly improve performance, given that they already provide the model with information about the language. Because the authors are dealing with two languages instead of one, the problem becomes more challenging.

Low-resource language sentiment analysis The research done by writers in [7] shows that Hindi still lacks in highly filled linguistic resources due to the challenges associated in dealing with the Hindi language. Hindi is the fourth most widely spoken language in the world. To analyse the sentiment included in Hindi language text collected from Twitter, the authors first look at machine learning-based algorithms such as Nave Bayes, Support Vector Machine, Decision Tree, and Logistic Regression. The sentiment analysis data set used by the authors was obtained from Twitter. The authors used Twitter to extract tweets for movie and product reviews, using the language ”Hindi” in the search filters. They manually categorized 23,767 tweets into positive and negative categories. Ironic content, slang language, non-Hindi language, and English terms written in Hindi were all eliminated by the authors. The authors also removed the tweets with no subjectivity from the data set. There were 16,901 subjective tweets left after these were removed. Because of the quick translation available on the Internet, netizens find it appealing to write in their original languages. This

necessitates the use of sentiment analysis in other languages as well. On the Internet, there is a vast volume of content in other languages that must be examined in order to determine the opinions of non-English speakers. The authors' proposed CNN technique has an accuracy of 85 percent..

The authors of [8] do sentiment analysis for the Manipuri language, categorising the text's direction as negative, positive, or neutral. Manipuri is the official language of Manipur, a northeastern Indian state. It is not only Manipur's official language, but it is also listed in the Indian Constitution's 8th Schedule. White space removal, stemming, removal of stop words, removal of numerals, removal of URL links, negation handling, substituting negative mentions, and reverting words that include repeated letters to their original form are some of the pre-processing procedures employed by authors. A goal standard data set for Manipuri sentiment analysis was collected and created by the authors from a local daily newspaper. Bengali script text is transliterated into Roman script text, while Meetei Mayek script text is transliterated into Roman script text using transliteration techniques. A major barrier for the writers was the lack of appropriate language-specific tool kits for the Manipuri language. The authors' transliterated gold standard data set could be useful in extending the study on the data set gathered from social media with adequate normalization.

The feature-based TSA system (incorporating an improved corpus-based negation modelling approach) classifies tweets based on syntactic and semantic elements derived from them, according to [9]. This paper adds to the development of a feature extraction system that aids in the construction of a variety of feature sets that can be used as input to classifiers. The authors present an algorithm for defining a set of rules for handling tweets in which the presence of negative does not always imply negation. The authors' study contributes to the presentation of a thorough research in the subject of TSA by examining the important parts of NLP, including tweet normalization and negation.

All previous work in this area has been done for languages with a lot of resources. In the subject of opinion mining, Marathi is one language where research is still lagging behind other languages like Hindi. As a result, we propose to develop a system in which we will deploy an XLM-RoBERTa-based model that has been fine-tuned using the Marathi tweets data set. This will outperform other models that require translation before performing Opinion Mining, as well as models that train Language Models on news headlines 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 for our proposed system are described below.

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 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*** - Latest available dataset is being used for this task for our proposed project.
   (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 not only have accurate results but also 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

he most frequent depiction of system/software specifications for an undeveloped software application is a UML use case diagram. The intended behavior (what) is specified in use cases, not the mechanism for achieving it (how). Once defined, use cases may be shown both textually and visually . The ability to construct a system from the viewpoint of the end user is one of the most essential notions in use case modeling. It's a good way to communicate behaviour of the system to users through their own words by detailing every outwardly apparent system operation. The figure 3.1 shows a use-case diagram for our suggested solution.

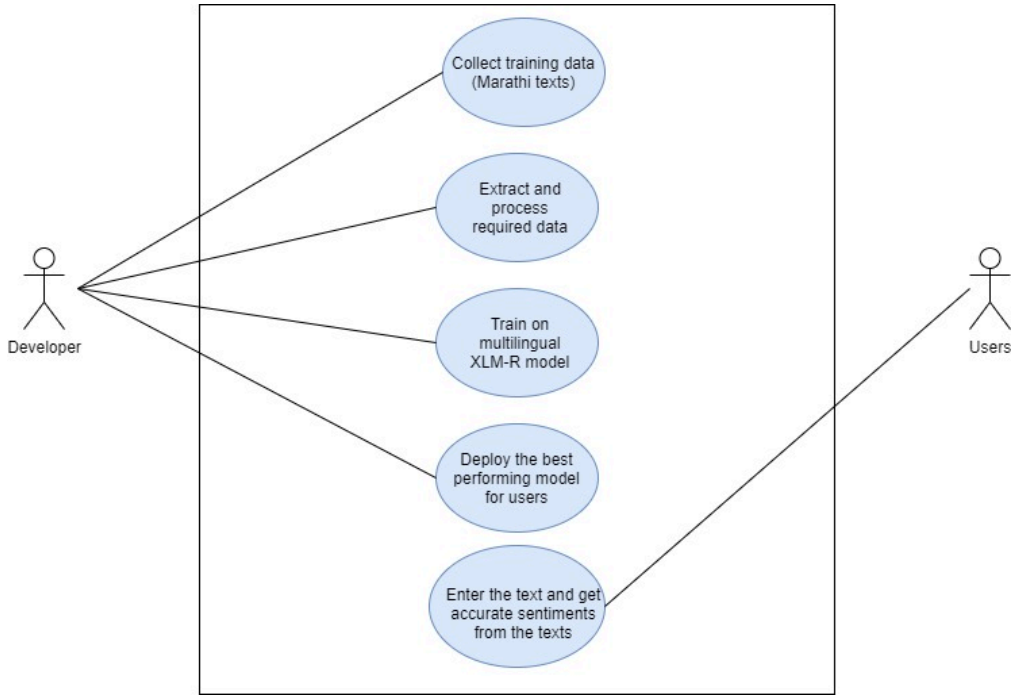

Figure 3.1: Use-case diagram of the proposed system.

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 modelling is the process of creating a visual representation of a full information system or portions of one in order to communicate links between data items and structures. The goal is to demonstrate the many types of data used and stored in the system, as well as the connections between them, how the data may be categorized and organized, and its forms and characteristics.At many degrees of abstraction, data may be modelled. Figure 4.1 shows the various degrees of abstraction at the context, level-0, and level-2 df-diagrams.

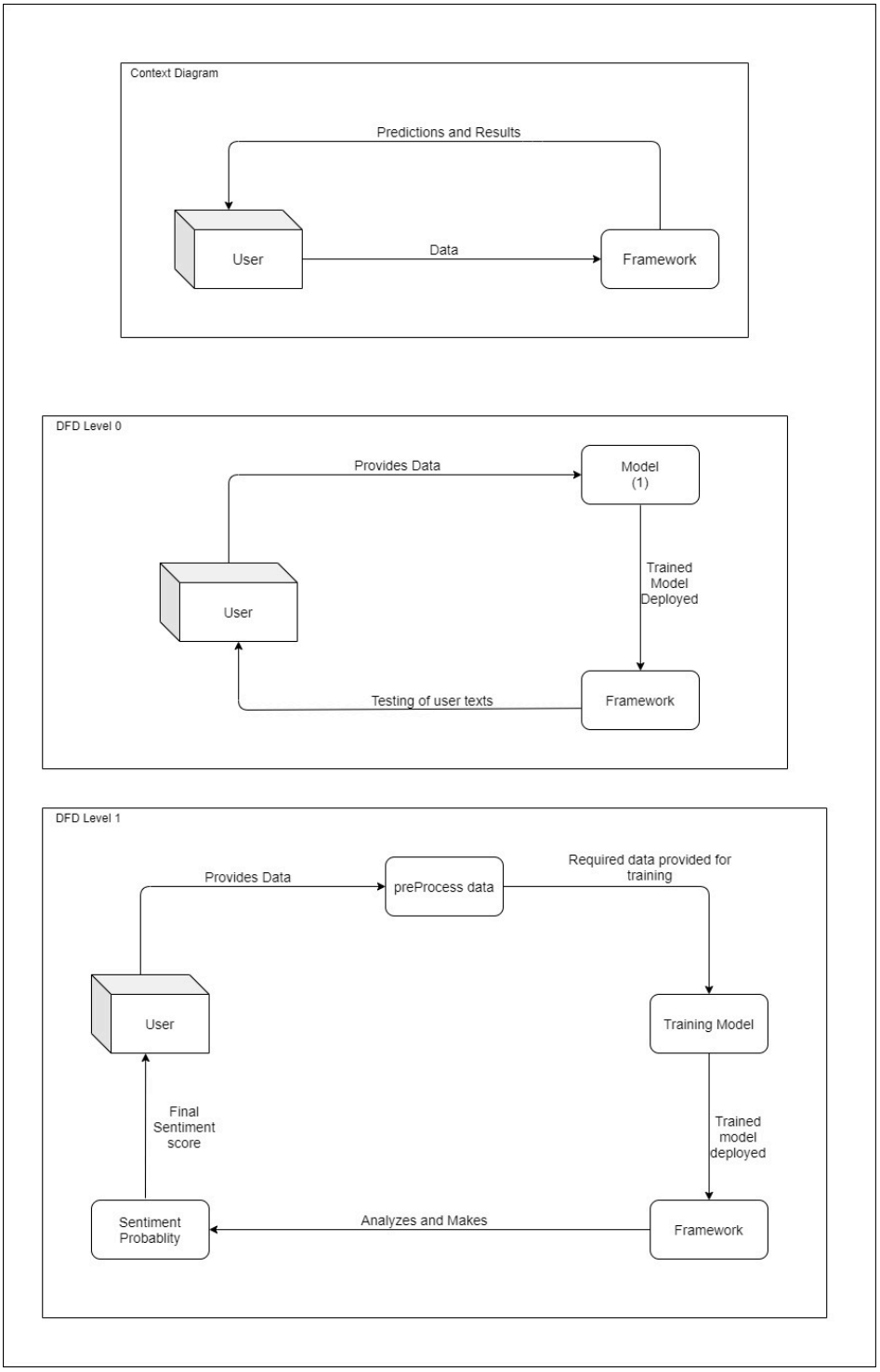

Figure 4.1: Data-Flow Diagram of the proposed system.

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

4.2.1 Class Diagram

A static diagram is a class diagram.It displays an application's static view. A class diagram can be used to visually represent, explain, and record numerous parts of a system, and to generate executables for a software program.

A class diagram depicts the traits and characteristics of a class, and also the system's constraints. Since they're the only UML diagrams which can be linked directly with object-oriented languages, class diagrams are often employed in the design of object-oriented systems.

A class diagram is a visual representation of a group of classes, interfaces, connections, collaborations, and constraints. Our suggested system's class diagram is depicted in 4.2.

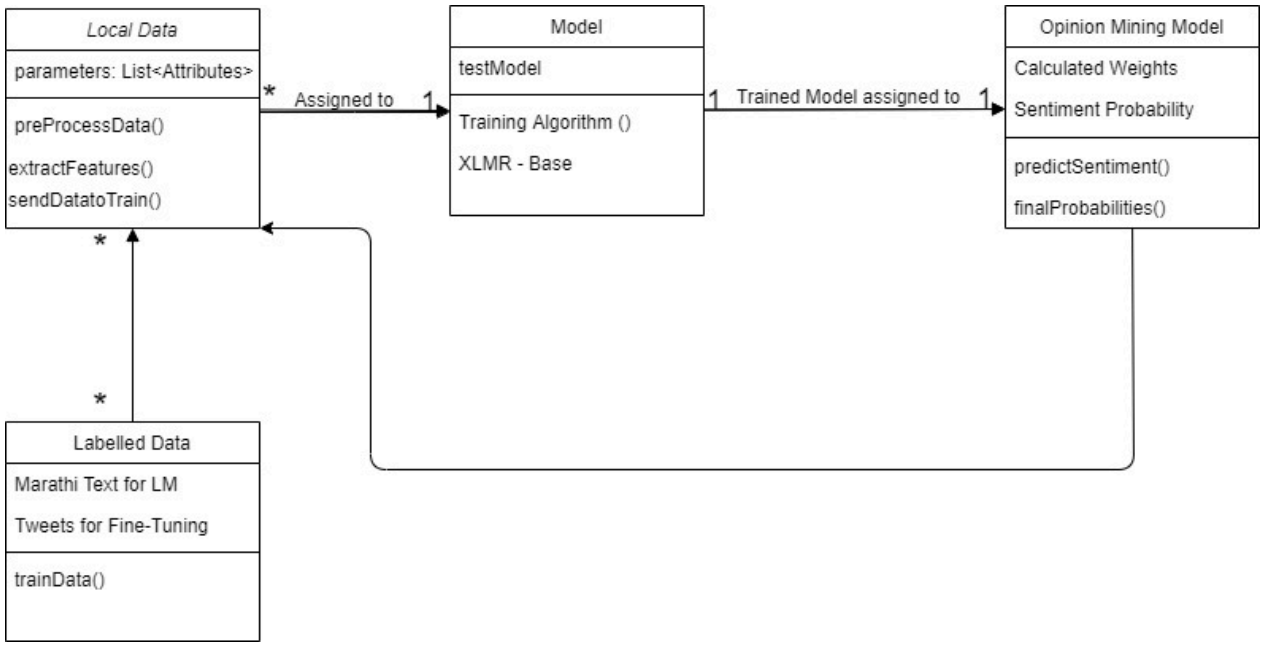

Figure 4.2: Class diagram of the proposed system.

4.2.2 State Diagram

A state diagram is a visual representation of the state of a system or a part of a system at a certain point in time. It's a behavioral diagram that depicts the action using only a few state transitions. State diagrams, often known as state machines or state chart diagrams, are a type of state diagram. These terms are commonly misunderstood. Simply expressed, a state diagram depicts the dynamic behavior of a class in response to time and changing external inputs. We may say that every class has a state, but we don't utilize State diagrams to represent all of them. Modeling states in groups of three or more is something we enjoy doing. Figure 4.3 depicts the state diagram for our proposed system.

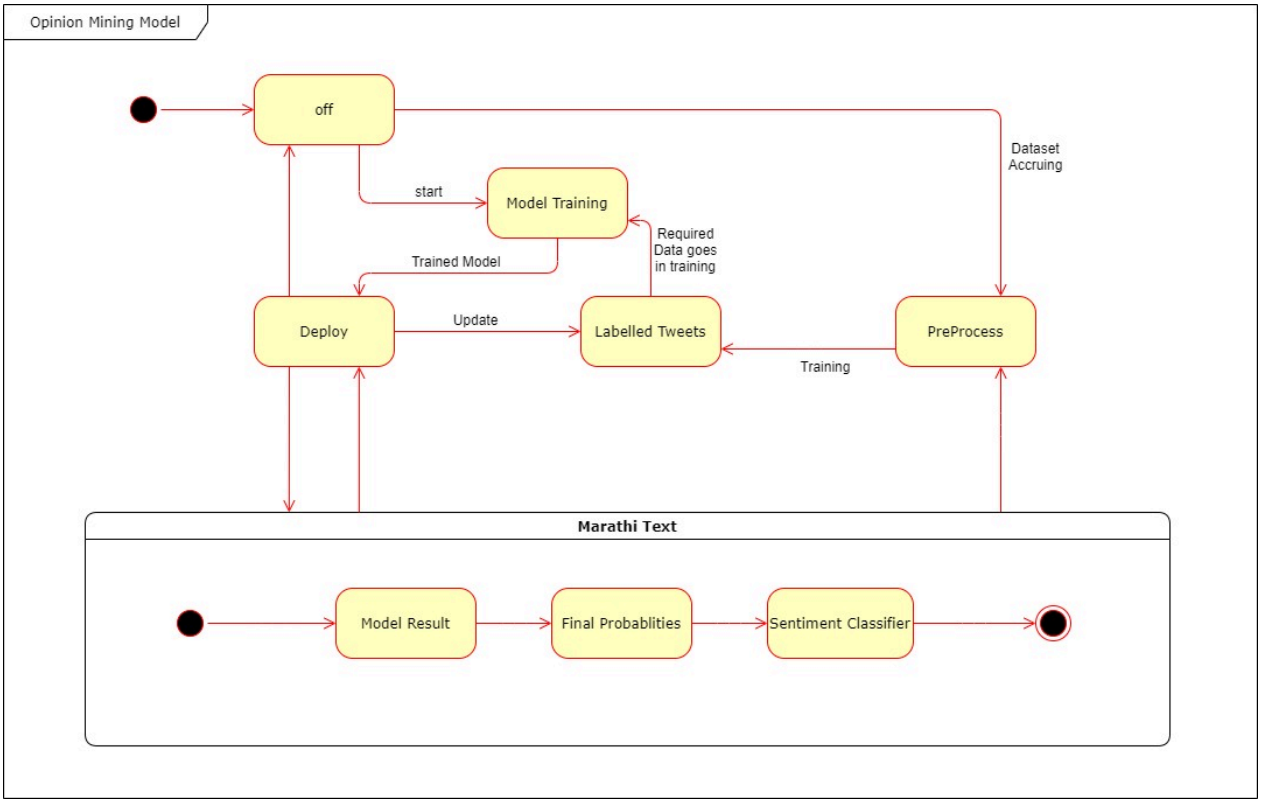

Figure 4.3: State diagram of the proposed system.

4.2.3 Activity Diagram

Another essential UML diagram for capturing the system's dynamic features is the activity diagram. An activity diagram is a flowchart that depicts the movement of data from one activity to the next. The action can be described using a system operation. The control flow is depicted as it moves from one activity to the next. This flow might well be sequential, branching, or parallel. Various elements, including such fork, join, and others, are used in activity diagrams to deal with various sorts of flow control. Figure 4.4 depicts the activity diagram for our suggested system.

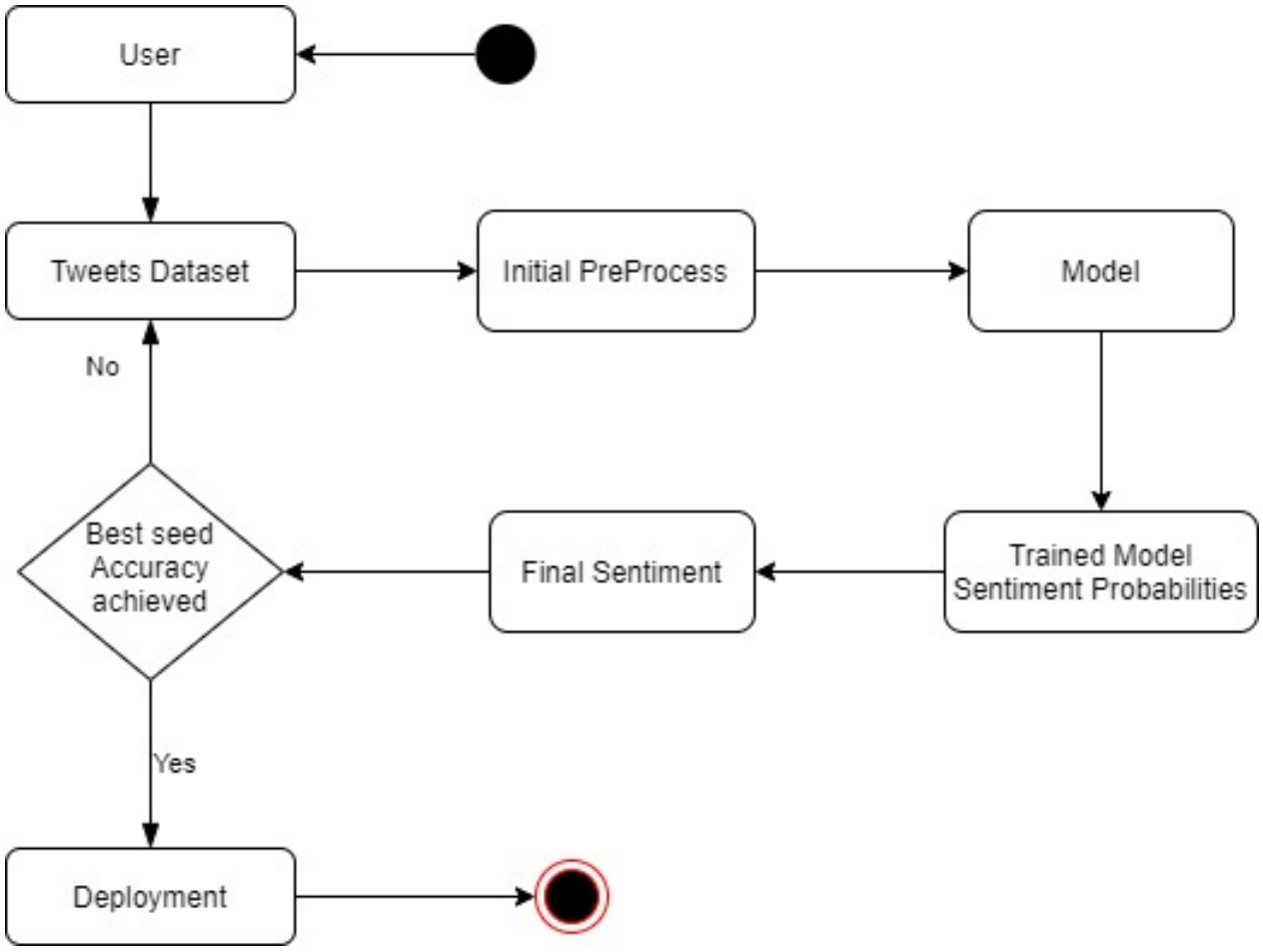

Figure 4.4: Activity diagram of the proposed system.

4.2.4 Sequence Diagram

A sequence diagram depicts the order in which things communicate, or the sequence in which these components interact. Business people and software developers frequently use these diagrams to illustrate and grasp needs for new and current systems. A sequence diagram, often called an event diagram or an event scenario, illustrates a succession of occurrences. Sequence diagrams show how the different elements of a system interact and in what sequence they do so.Figure 4.5 depicts the sequence diagram for our proposed system.

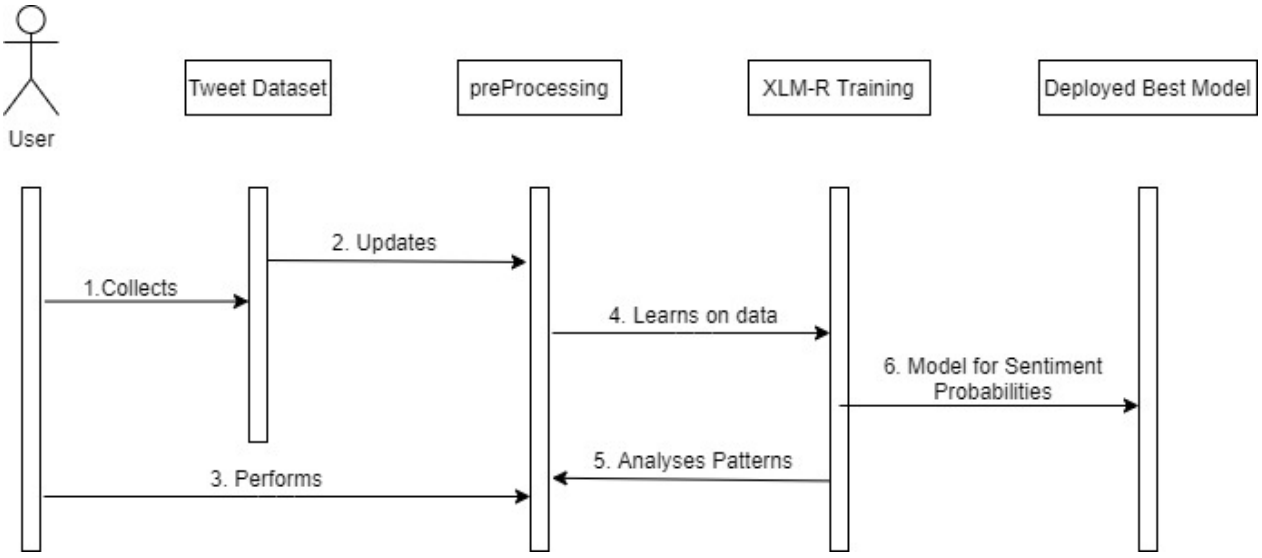

Figure 4.5: Sequence diagram of the proposed system.

4.2.5 Collaboration Diagram

The collaboration diagram is a diagram that shows how objects in a system are connected to one another. Both the sequence and collaboration diagrams show the same information, but in different ways. Because it is based on object-oriented programming, it depicts the architecture of the object existing in the system rather than the flow of messages. A feature is one of several components that make up a product. Several components of the system are interconnected. A collaboration diagram, also known as a communication diagram, is a diagram that shows the architecture of a system component. Figure 4.6 shows the cooperation diagram for our proposed system.

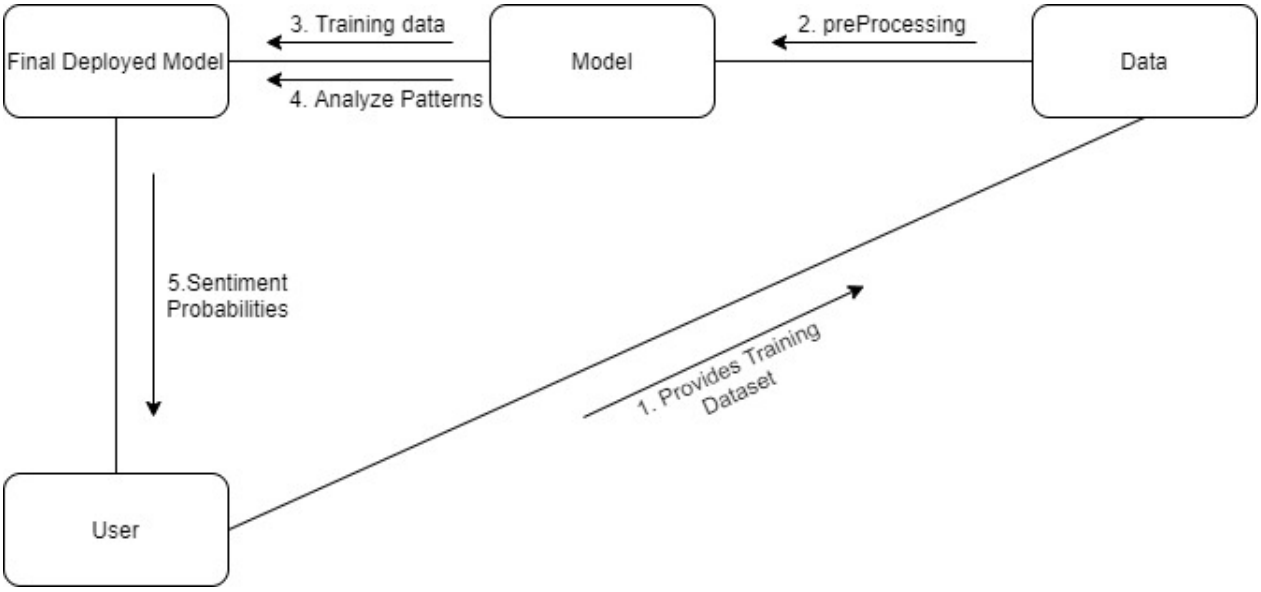

Figure 4.6: Collaboration diagram of the proposed system.

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. This explains the whole architecture of how the proposed system for opinion mining will go about.

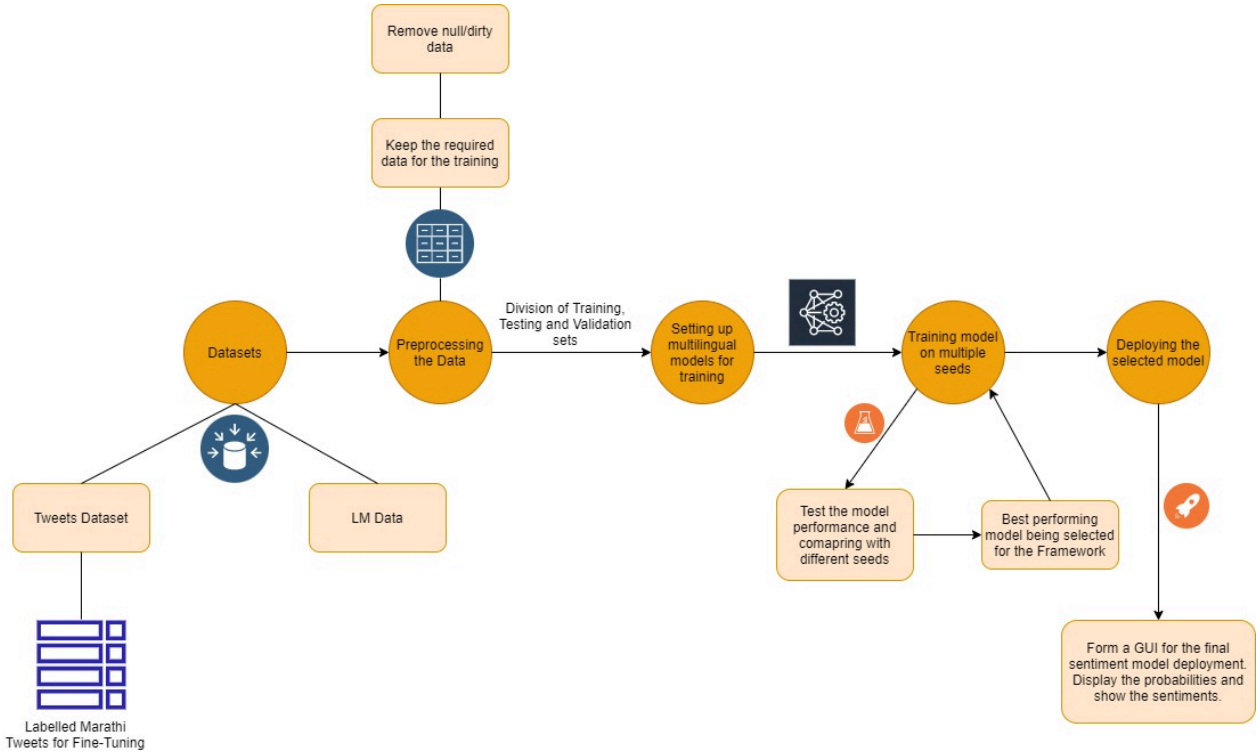

Figure 5.1: System Architecture Diagram of the proposed system.

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.

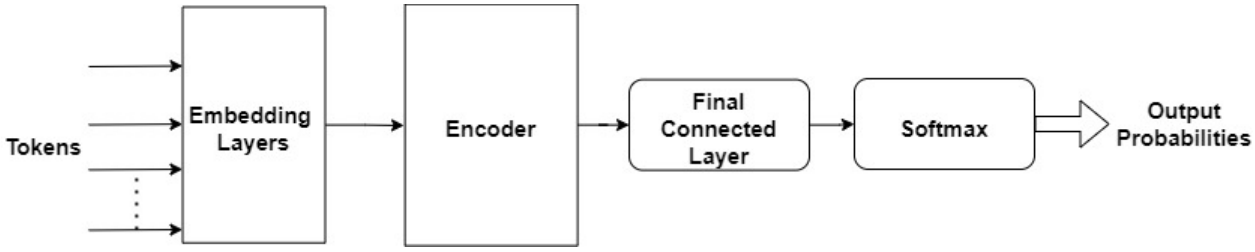

Figure 6.1: Architecture of XLM-RoBERTa

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.

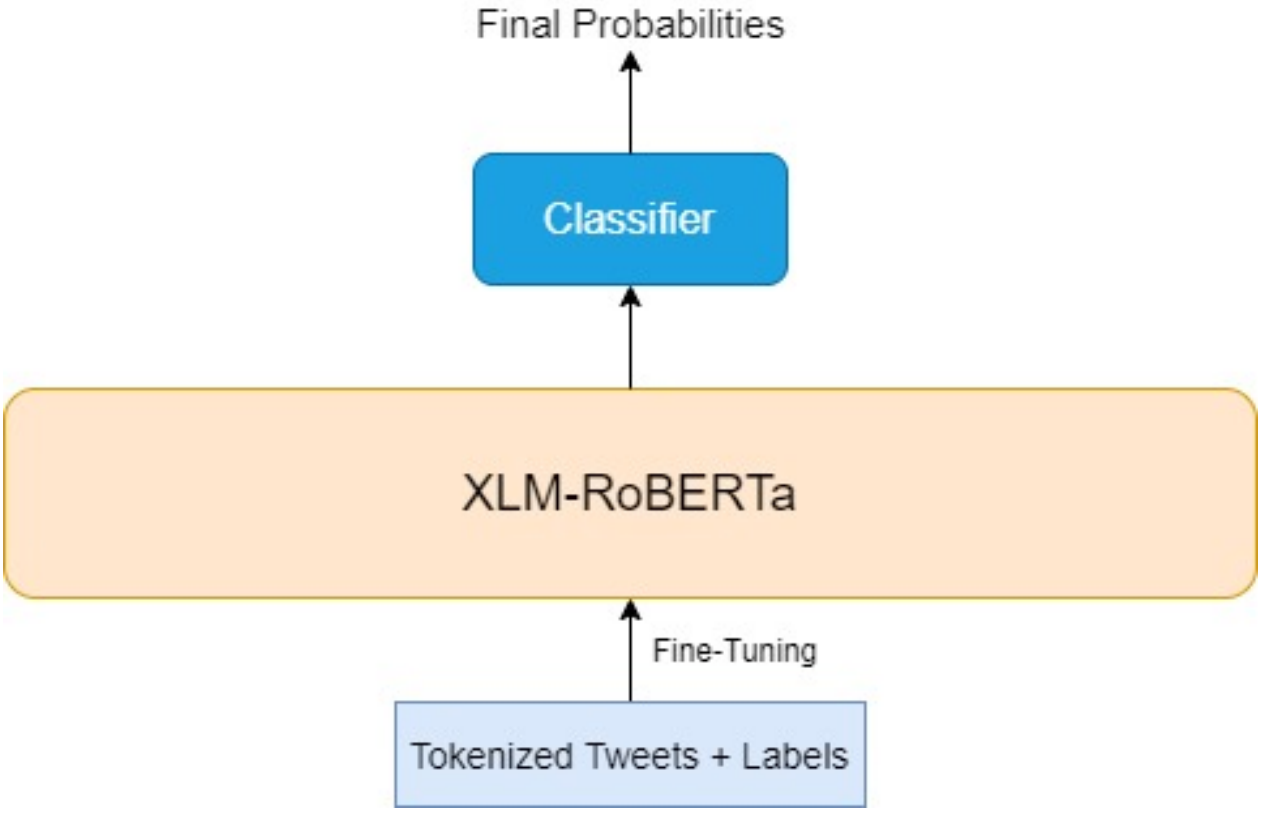

Figure 6.2: Training Architecture of the proposed system.

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 levels. The high area of the curve shows that both precision and also the recall

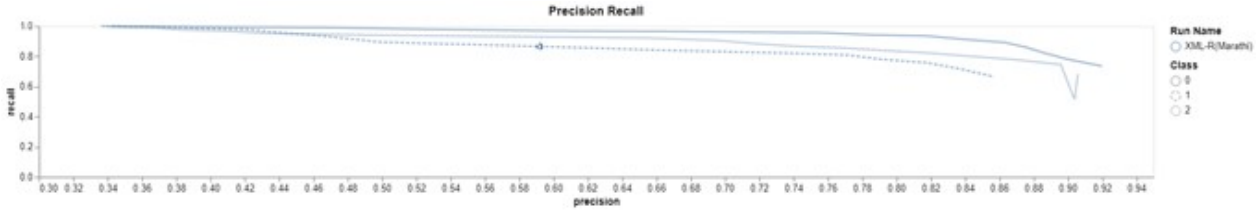

Figure 7.2: Precision Recall curve of XLM-R training

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.

Appendix

The publication by our group for this project is as follows:

- Naitik Rathod, Nishit Mistry, Dhruv Talati, Manan Parikh, Aniket Kore, Pratik Kananu, 'Marathi Social Media Opinion Mining using XLM-R', in *2022 International Conference on Applied Artificial Intelligence and Computing(ICAAIC 2022)*,2022.

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.

Naitik_Marathi_Report

ORIGINALITY REPORT

9%	2%	7%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	link.springer.com Internet Source	2%
2	Vedika Gupta, Nikita Jain, Shubham Shubham, Agam Madan, Ankit Chaudhary, Qin Xin. "Toward Integrated CNN-based Sentiment Analysis of Tweets for Scarce-resource Language—Hindi", ACM Transactions on Asian and Low-Resource Language Information Processing, 2021 Publication	2%
3	Submitted to University of Wales Institute, Cardiff Student Paper	2%
4	Itisha Gupta, Nisheeth Joshi. "Feature-Based Twitter Sentiment Analysis With Improved Negation Handling", IEEE Transactions on Computational Social Systems, 2021 Publication	1%
5	Submitted to Asia Pacific International College Student Paper	1%

6	Yaroslav A. Seliverstov, Andrew A. Komissarov, Eleonora D. Poslovskaia, Alina A. Lesovodskaya, Artur V. Podtikhov. "Detection of Low-toxic Texts in Similar Sets Using a Modified XLM-RoBERTa Neural Network and Toxicity Confidence Parameters", 2021 XXIV International Conference on Soft Computing and Measurements (SCM), 2021 Publication	1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		