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

Naitik Rathod
Department of Computer Engineering
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India
naitikrathod18@gmail.com

Dhruv Talati
Department of Computer Engineering
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India

Nishit Mistry
Department of Computer Engineering
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India

Manan Parikh
Department of Computer Engineering
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India

Aniket Kore
Department of Computer Engineering
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India

Pratik Kanani
Department of Computer Engineering
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India

***Abstract* — Sentiment Analysis is one of the most important tasks for any language and a very important domain in Natural Language Processing which has shown remarkable progress in recent years. 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 presenting the results of multiple XLM-R models over the Marathi tweets dataset for the task of opinion mining.**

***Keywords*- Sentiment Analysis, Low Resource Language, Marathi, XLM-R,**

I. INTRODUCTION

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. The current models and systems available are designed to analyse 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.

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.

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 display the class of sentiment. 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.

In Marathi language the location of the words can be changed without changing the meaning of the sentence. For example, "मला मिठाई आवडते" can be changed to "मिठाई मला आवडते". Marathi follows the subject-object-verb

format most of the times. Sometimes, it merges verb and object into a single word. For Example, "मी उद्यानात आहे".

II. Literature Review

Social Media plays a significant role in determining the opinion of people and hence is an important NLP task for detecting and analysing the text. This will help in knowing the polarity of texts and understanding people's opinions on various issues. The majority of existing works for sentiment analysis in the Marathi language have used a limited dataset based on news articles as in [1]. They achieve higher accuracy for sentiment analysis of news headlines and news articles, but the accuracy diminishes for non-news article. Hence, there is a need for using a wider dataset and training it on models which yield better results on low resource languages, like the XLM-R model.

There has been research work on generating sentiment analysis of low resource languages by translating them into the English language. However authors in [4] have studied the effect of translating the English reviews into German, Urdu, and Hindi and compared the classification results of all languages, their research shows low accuracy is yielded when sentiment analysis is performed through translation There are various Machine Learning models that can be used for NLP, however authors in [5] have shown that XLM-R, significantly outperforms multilingual BERT (mBERT) on a variety of cross-lingual benchmarks, the model performs particularly well on low-resource languages, improving 15.7% in XNLI accuracy for Swahili and 11.4% for Urdu over previous XLM models.

There has been a tremendous rise in events of hate speech on social media, authors in [2] provide an effective neural network-based technique for the hostility detection in the low resource language Hindi text. It is very important to classify the social media texts in categories like very negative, negative, neutral, positive and very positive to locate the hostile texts. They have used word embeddings for deciding the polarity of texts as authors in [6] have shown that proper word embeddings can boost performances by a large margin. The dataset used for training was cleaned of any emojis, English language text along with white space removal, stemming, removal of stop words, removal of numbers, removal of URL links to ensure higher accuracy as shown by authors in [7] and [8] where they had performed sentiment analysis for the Hindi and the Manipuri language respectively. In sentiment analysis, there has been the problem of dealing with tweets and social media texts where negation occurrence does not necessarily mean negation, authors in [9] have presented comprehensive research in the field of sentiment analysis by looking into tweet normalization and negation which are the critical aspects of NLP. In recent years educational and specialized web resources have seen heated discussions. People using these sites are characterized by restraint in statements and expressions of emotion characterised by ridicule, sharp jokes, provocative statements and hidden injections. Hence there is also a need for annotating these low toxic statements as done by authors in [3]. Hence we propose XLM-R based models for the task of opinion mining in Marathi language.

III. Methodology

A. Dataset

In this work, we used the publicly available L3CubeMahaSent [1] Twitter corpus, which is 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 goal is tweet polarity classification, that is, classifying a tweet into positive, negative, or neutral classes. Table 1 provides a statistic on training and testing data sets, which clearly shows the perfect balance of the classified tweets among them. So, there is lesser chance of a bias in the training and testing of the models.

TABLE 1. Information of Dataset

Category	Total Tweets
Training	12114
Testing	2250
Validation	1500

B. Data Analysis

- Number of tweets for each sentiment
  There are a total of 5288 tweets available for positive, negative, and neutral classes. Each of which is divided into 75%, 15%, and 10% for training testing and validation sets respectively. The three sentiments count is kept equal in all subsets to avoid any bias in training of the models.
- Average word length per tweet for each sentiment
- Average sentence length per tweet each sentiment

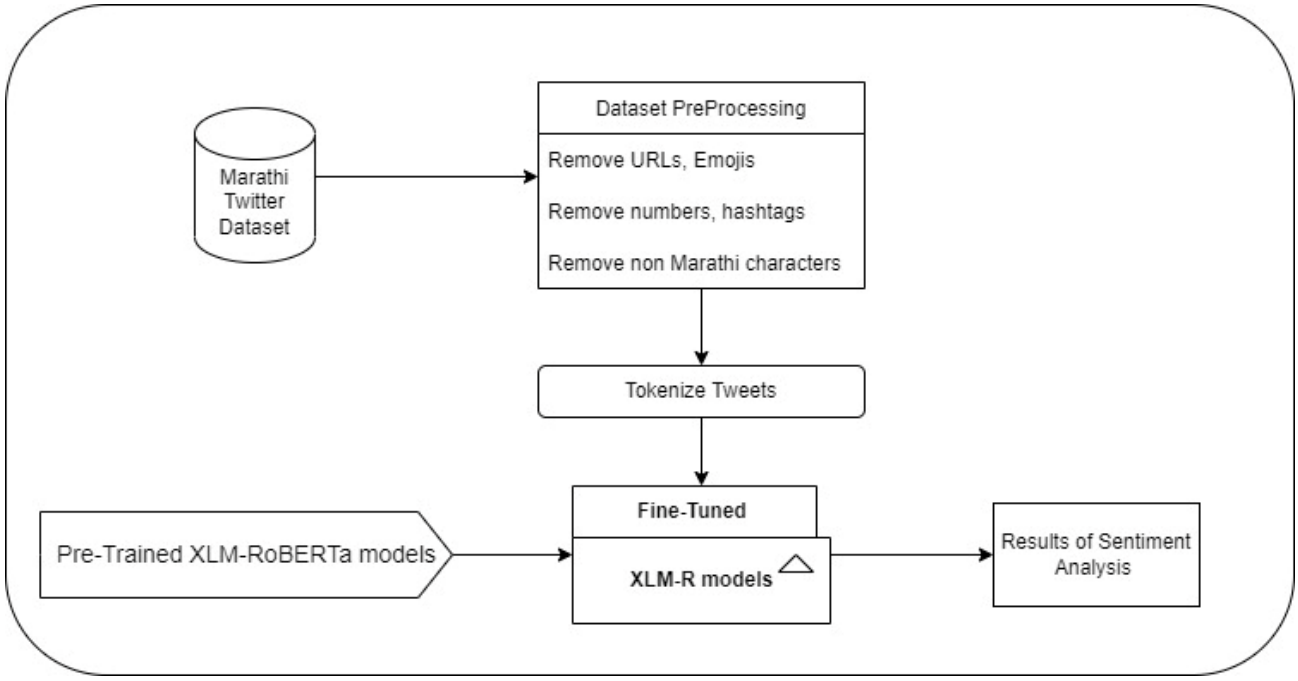


FIGURE: Model Architecture

- Word Cloud
  We created the word clouds for the total tweets available and divided it among the three sentiments of the dataset. The most used words in the tweets after removing the stopwords are displayed the biggest while the words used less number of times have a relatively smaller size in the word cloud.

Figure: Positive word cloud

Figure: Neutral word cloud

Figure: Negative word cloud

C. Preprocessing

The data that is fetched directly from twitter is not clean and contains many unwanted text and noise. This needs to be cleaned [7] [8] before putting the data under training for the models to properly understand the language we train it on.

Links: Every tweet scraped from any API contains the link to that tweet followed by the tweet itself. We removed all the links and URLs present in the tweets as they have no significance in our required task.

Hashtags and Mentions:

Hashtags are words that are preceded by #(symbol), these are used when referring to a known or popular topic or keyword. Hashtags serve as URL to a page displaying posts about that same topic. We removed all the hashtags except the one's in Marathi language as they might have a significant meaning in the tweet. So, in the Marathi hashtags, only the symbol # was removed.

Mentions are words that are preceded by @(symbol), containing another twitter user's username in the tweet body and are used when talking to or about someone. We removed all the mentions in the tweets as they serve no purpose in the sentiment analysis task.

Emojis: People these days use the social media creatively and this increases the usage of the emojis in the tweets, messages, and posts. Although these emojis can be replaced with its meaning in the English datasets, it is not possible to do so in Marathi yet. Hence, we completely removed all the emojis present in the tweets.

Spaces: The extra spaces from the tweets were removed and replaced with a single space.

Numbers and Punctuations: Numbers have no role in the sentiment analysis; hence numbers and the punctuations of the tweets were removed.

D. Proposed Model

1) Tokenization (Roberta)

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.

2) XLM-RoBERTa – base

XLM-R is a transformer-based multilingual masked language model pre-trained on 2.5 TB of CommonCrawl data in 100 languages with base having 250M parameters, which obtains state-of-the-art performance on cross-lingual classification, sequence labeling and question answering. The base is trained over the BERT-base architecture along with XLM. The XLM-R outperforms other significant models by over 20% for the task of text classification due to the larger size of the training of the XLM-R [5]. We propose to fine-tune all the available XLM-R models over the Marathi tweets dataset and compare the accuracies achieved.

3) XLM-R – Large

XLM-R large is trained with 560M parameters and XL and XXL versions of XLM-R have been trained with 3.5B and 10.7B parameters in 100 languages. Large model has been trained with BERT-Large architecture with 250K being the vocabulary size. XXL is the largest XLM-RoBERTa model currently available that gives high accuracies for most of the tasks for most of the languages of the 100 available.

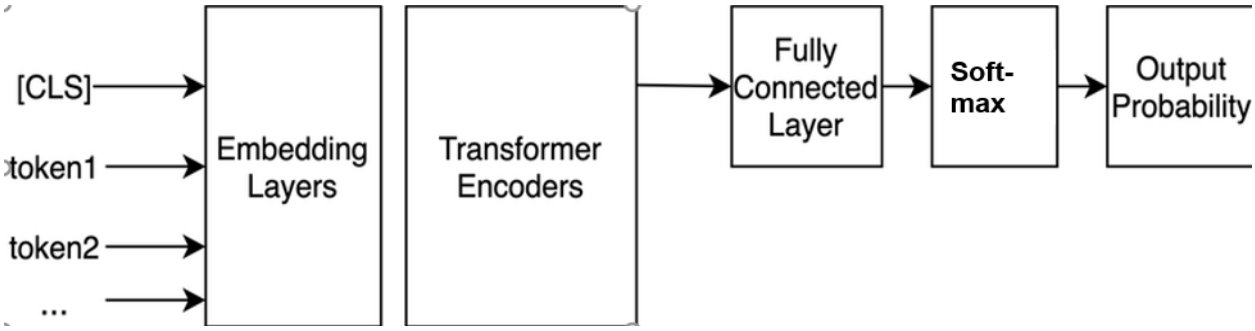


Figure: XLM-R architecture

IV. Results and Observation

A. Experimental Setup

We had divided the dataset into training, testing and validation sets and we used them for the respective parts. The testing part was used for the determination of the accuracies of the models displayed below in the Table 2. We tokenize the data using the Roberta tokenizer and use it for the training purpose. The accuracies are tested over the models that are trained with 25 epochs over the same training set.

B. Results

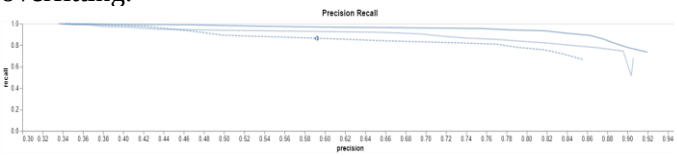
The results from Table 2 show us that _____ varant of the XLM-RoBERTa performs the best over this Marathi dataset and has comparable results with other models as in Kulkarni et al [1]. Additionally our accuracy of _____ is the best accuracy possible for the task of sentiment analysis using XLM-R model for three class classification.

TABLE 2. Results

Model	Accuracy
XLM-R base	
XLM-R Large	
XLM-R XL	
XLM-R XXL	

C. Observations

We observe from the training loss graph that there is no overfitting taking place during the training of the model. Limiting to 25 epochs takes care of the case of overfitting.



The precision-recall curve shows values for precision and recall for different threshold. The high area under the curve represents that both recall and precision are high; high precision means there are less false positives, and high recall means there are low false negatives. Summarizing the curve, high scores for both show us that the classifier is returning accurate results due to higher precision and a maximum of all positive results due to high recall.

Acknowledgment

This research did not receive any specific funding from donors in the public, commercial, or not-for-profit sectors.

Reference

[1] Kulkarni, Atharva, Meet Mandhane, Manali Likhitkar, Gayatri Kshirsagar, and Raviraj Joshi. "L3CubeMahaSent: A Marathi Tweet-based Sentiment Analysis Dataset." In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 213-220. 2021. https://aclanthology.org/2021.wassa-1.23/

[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