XLM-R for Social Media Opinion Mining of Marathi Texts

Naitik Rathod[1, a)], Dhruv Talati[1, b)], Nishit Mistry[1, c)], Manan Parikh[1, d)], Aniket Kore[1, e)] and Pratik Kanani[1, f)]

Author Affiliations
[1]*Department of Computer Engineering, D.J Sanghvi College of Engineering, Mumbai, India*
Author Emails
[a)] naitikrathod18@gmail.com
[b)] dhruvtalati51000@gmail.com
[c)] nishitmistry07223@gmail.com
[d)] dhruvil153@gmail.com
[e)] aniketkore24@gmail.com
[f)] pratik.kanani@djsce.ac.in

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 recently. 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 languages mentioned in the 8th schedule of the constitution of India and is the third most spoken language in India which is mainly used in the Deccan region which includes Maharashtra and Goa. There has been low research for sentiment analysis approaches based on the Marathi text. Therefore, in this project proposes use of XLM-RoBERTa (XLM-R) models 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-R and its versions will be put under training over the Marathi tweets dataset after tokenizing using RoBERTa tokenizer for the purpose of opinion mining and classification. Authors aim at presenting the results of multiple XLM-R models over the Marathi tweets dataset for the task of opinion mining.

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 possible due to huge amount of research which has already taken place in the English language 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 as shown in Gupta et al [1]. 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, authors 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 has 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 as in Kulkarni et al., [2]. This cannot give a very accurate analysis of the social media texts. Authors in this research will be using the dataset that is created from twitter tweets that were in Marathi language. Authors in this work propose to train the XLM-R models over the Marathi tweets as shown in Conneau et al., [3] where XLM-R gives higher accuracies for the opinion mining of low resource languages. Based upon the probabilities achieved from the final model, authors will classify the text and display the class of sentiment as in Zaharia et al., [4]. Authors 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".

LITERATURE REVIEW

Social Media has now become an integral part of our life with users on an average spending more than 2 hours a day on it. One can gauge the opinion of users on the basis of their social media texts. Using Natural Language Processing (NLP) on social media texts 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 Kulkarni et al., [2] where 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. 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 Ghafoor et al., [5] have studied the results of translating text from English language to a low resource language like Hindi. Their research shows that low accuracy is yielded when sentiment analysis is performed through translation. Hence authors in this work have avoided use of translation into English from the Marathi text for sentiment analysis.

There are various Machine Learning models that can be used for NLP, however authors in Conneau et al [3] have trained a Transformer based masked language model on one hundred languages, using more than two terabytes of filtered CommonCrawl data, their research has shown that XLM-R model performs better than multilingual BERT (mBERT) on a various benchmarks. XLM-R model achieves better performance particularly on low resource languages like Urdu and Swahili improving accuracy by over 10 percentage points compared to other models.

Anti-social elements have taken to social media and our spreading hate speech against religious minorities and other weaker sections like Scheduled Tribes. Authors in De et al., [6] have evaluated different models and have curated the dataset on hostile and non-hostile hindi text from social media platforms, which is further annotated into fine grained labels like fake, hate, defamation and offensive. The research provides an effective neural network-based technique for the hostility detection in the low resource language Hindi text which can be further used in sentiment analysis of other low resource languages like Marathi.

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 Zaharia et al., [4] have shown that using word embeddings shoots up accuracy by a significant 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 Gupta et al., [1]. The dataset employed by authors in Gupta et al., [1] for sentiment analysis has been fetched from Twitter.

Similar method has been used by authors in Meetei et al., [7] where they had performed sentiment analysis for the Manipuri language. Authors in Meetei et al [7] have prepared a goal standard dataset for Manipuri sentiment analysis from a local daily newspaper.

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 Gupta et al., [8] have presented extensive research on sentiment analysis mainly by dwelling 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.

Hence there is also a need for annotating these low toxic statements as done by authors in Seliverstov et al., [9]. Hence authors propose XLM-R based models for the task of opinion mining in Marathi language.

METHODOLOGY

Dataset

Authors in their research have used the publicly available L3CubeMahaSent [2] 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 16000 Marathi tweets manually classified into the 3 classes. Author's goal in this research is tweet polarity classification, by classifying a tweet into three categories according to their polarity, the three categories being positive, neutral and negative. Table 1 provides a statistic on training, testing and validation data sets, which 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. Detailed information of dataset

Category	Number of Tweets
Training	12000
Testing	2250
Validation	1500

Data-Analysis

Number of tweets for each sentiment: There are a total of 5250 tweets available for positive, negative, and neutral classes. Each of which is divided into 75%, 10%, and 15% for training, validation and testing sets respectively. The three sentiments count is kept equal in all subsets to avoid any bias in training of the models.

Word Cloud: A word cloud shows how often the words have occurred in the data. Larger the words, more is its occurrence. Authors created the word clouds for the total tweets available in the dataset and divided it among the three sentiments of the dataset. The most used words in the tweets after removing the stop words are displayed the biggest while the words used a smaller number of times have a relatively smaller size in the word cloud. We have used the word cloud library by Andreas Mueller [10] where we can set multiple parameters for the final image. This library takes the text and parameters as input, the frequency of the words is calculated, and we get our required word clouds as the output as shown below in the Figure 1.

(a) (b) (c)

FIGURE 1: (a) Positive Word Cloud (b) Neutral Word Cloud (c) Negative Word Cloud

Pre-Processing

The data that is fetched directly from twitter is not clean and contains many unwanted text and noise. This needs to be cleaned as in Gupta et al., [1] and Meetei et al., [7] before putting the data under training for the models to properly understand the language one has trained it on.

Links: Every tweet scraped from any API contains the link to that tweet followed by the tweet itself. Authors have 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. Authors have 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. Authors have 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, authors 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.

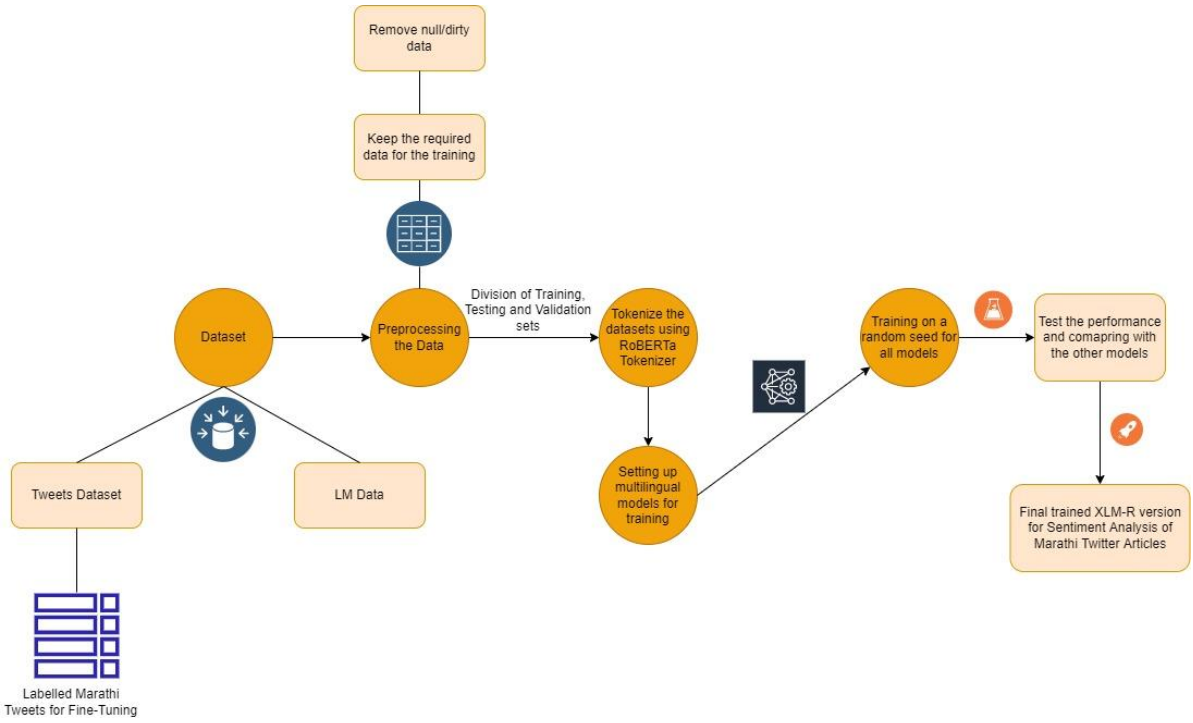

FIGURE 2: Model Architecture

Proposed Model

Tokenization (Roberta)

The task of breaking the text data into smaller chunks of words or small texts is called tokenizing. It is an integral part of processing of data before training any model so that the model being trained can understand the text.

Multiple ways of tokenizing exist for different applications in text processing. 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.

XLM – RoBERTa – Base

XLM-R is a transformer-based multilingual masked model pre-trained on CommonCrawl data in 100 languages with base having 250M parameters, which obtains state-of-the-art results in the tasks of question answering, sequence labeling, cross-lingual classification. The base variant 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, Conneau et al., [3]. Authors propose to fine-tune all the available XLM-R models over the Marathi tweets dataset and compare the accuracies achieved.

XLM – RoBERTa – 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 as cited in Goyal et al., [11]. 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 but being a model with such large numbers of parameters, it requires huge computing power to use and train that model hence it is kept out of consideration for the current comparison.

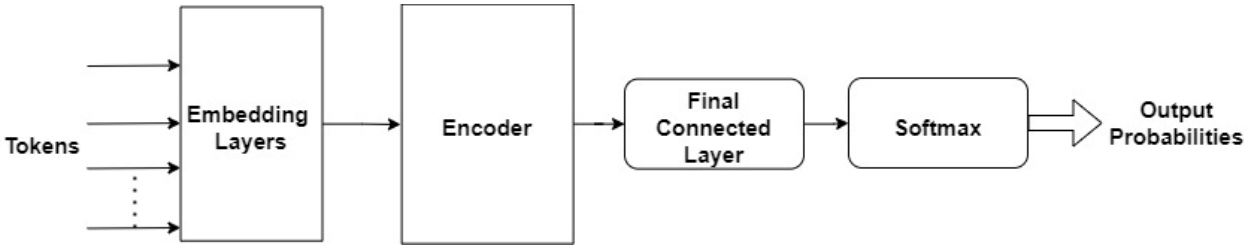

FIGURE 3: Model Architecture

RESULTS AND OBSERVATION

Experimental Setup

Figure 1 shows the architectural diagram designed for the training of the models while the Figure 3 shows the XLM-R architecture. Authors had divided the dataset into training, testing and validation sets and 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. Authors have tokenized 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.

.

Results

The results from Table 2 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 Kulkarni et al., [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 2. Results

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

Observations

Authors observed from the training loss graph, Figure 4(a), 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.

The precision-recall curve as shown below in Figure 4(b) shows values for the precision and the recall for different threshold. The high area under the curve represents that both recall, and precision 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.

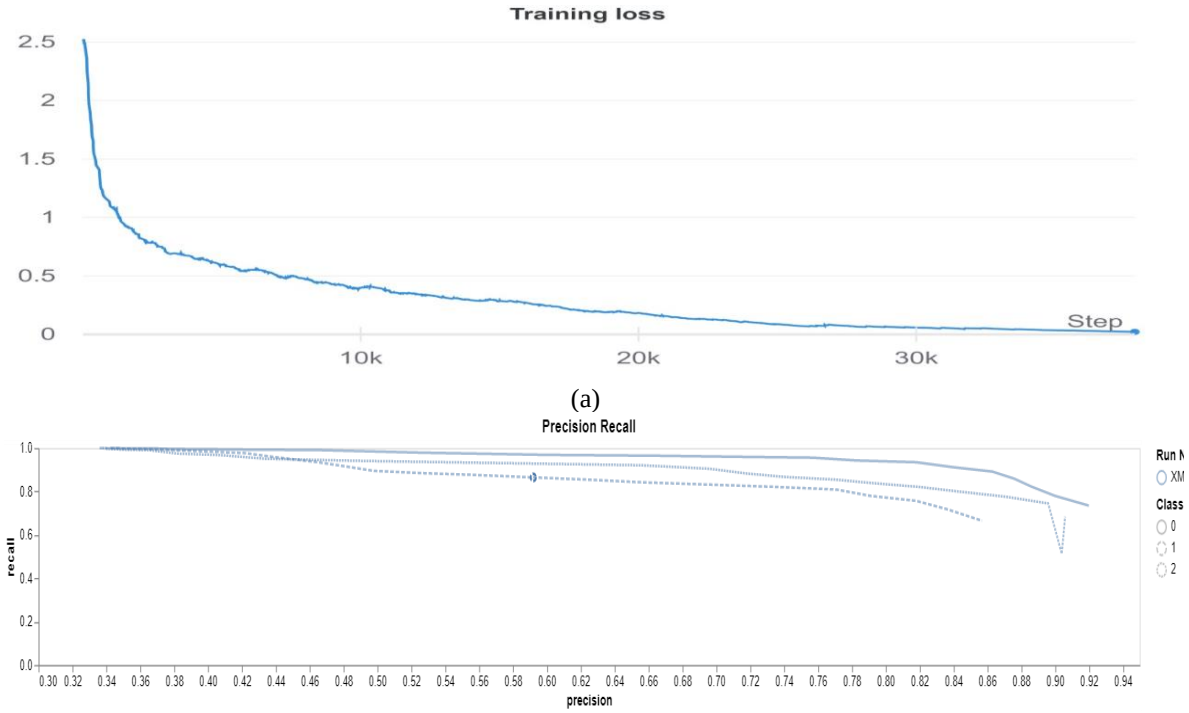

Figure 4: (a) Training Loss Graph (b) Precision-Recall Curve

CONCLUSION

This paper explored training of Marathi Opinion Mining system using transformer XLM-R 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.

REFERENCE

1. 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
2. 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/
3. 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
4. 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
5. 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
6. 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 Hostile 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
7. 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
8. 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
9. 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
10. https://amueller.github.io/word_cloud/
11. A Goyal, Naman, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. "Larger-scale transformers for multilingual masked language modeling." *arXiv preprint arXiv:2105.00572* (2021). https://aclanthology.org/2021.repl4nlp-1.4/