Sentiment Analysis of Marathi Texts using Deep Learning Models

Deepak Mane[1], Sarthak Pithe[2], Hrishikesh Potnis[3], Soham Nimale[4], Madhur Vaidya[5]

[1234567] Vishwakarma Institute of Technology, Pune, India
[1] deepak.mane@vit.edu
[2] sarthak.pithe22@vit.edu
[3] hrishikesh.potnis221@vit.edu
[4] soham.nimale22@vit.edu
[5] madhur.vaidya221@vit.edu

Abstract. The process of understanding the underlying subjective information and opinion in a given sentence is sentiment analysis. Accurate Sentiment analysis is possible with the use of Deep-Learning techniques. We have proposed three Deep-Learning based fine-tuned models for the three-class sentiment analysis task. It is complemented with an end-to-end pipeline that includes classical approaches like stemming, lemmatization in the data preprocessing, parameter tuning, and so forth for Marathi texts analysis. The proposed models are based on ANN, IndicBERT, and BiLSTM. The "L3CubeMahaSent" dataset was used for both training and testing of the proposed models. The IndicBERT based model achieved a superior F1-score of 85%, Precision of 85% and a recall score of 86% on the dataset. The overall weighted accuracy of this model was 81%. Additionally, all the proposed models achieved higher F1-scores and accuracy in predicting the negative sentiments compared to neutral and positive ones even in the class-balanced dataset.

Keywords: Marathi NLP · Sentiment Analysis · Deep Learning · IndicBERT · BiLSTM.

1 Introduction

Sentiment analysis is a technique used to predict the emotion in a given sentence into various classes. It is used to analyze the emotion behind the sentences and to classify them either in two-class (positive or negative sentences), three-class (positive, neutral or negative sentences) or multi-class [1, 2]. Thus, it is commonly used for applications like customer feedback analysis, and social media monitoring [3]. The two main methods for performing sentiment analysis are - dictionary-based and machine learning-based. Sentiment analysis has been extensively researched in the context of the English language. However, there is less research done on this topic, compared to other languages. Marathi is a language spoken by many people. It is the third-most spoken language in India and 13th in the world in terms of total native speakers, and has received little attention in this regard. This paper presents a comparative study of

various methods for three-class sentiment-analysis of sentences in Marathi language. Discussion about the implementations, advantages and disadvantages and how well they suit the environment of Marathi-language structure is also done. We used various deep learning methods to classify Marathi texts into positive, neutral or negative categories. We also propose an end-to-end pipeline for the Marathi Sentiment analysis that includes methods in Pre-processing, Model selection, Fine-tuning, getting Classification reports and revising the models.

The objectives of this paper are as follows:

- To perform accurate sentiment analysis using various classification models and preprocessing techniques on Marathi texts.
- To compare and evaluate the results obtained from each of them and with the existing models.
- To develop a preprocessing pipeline for the Marathi texts to obtain trainable features from raw dataset without much loss of sentiment.
- To understand the sentiments behind Marathi tweets using the models developed and its practical importance in various fields and in decision making.

The overview of this paper is as follows. The Section 1 and 2 includes the introduction and the related work respectively. It discusses the limitations of existing systems, and thus proves the novelty of our proposed solution. Section 3 discusses the general pipeline for sentiment analysis which is not language specific, along with the models proposed. Section 4 talks about the comparative analysis of results given by our proposed pipeline and previous results from papers cited in Section 2. Section 5 concludes the paper, followed by our future scope in Section 6.

2 Literature Review

In their paper 'L3CubeMahaSent', Kulkarni et al. [4] presented a new dataset of over 16,000 labelled Marathi tweets and applied various models, such as Convolutional Neural Network (CNN), Bidirectional long short-term memory (BiLSTM), and Bidirectional Encoder Representations from Transformers (BERT), for three-class and two-class sentiment analysis.

Divate [5] modified the long short-term memory (LSTM) method for Marathi sentences and used a BiLSTM model on a Marathi news dataset, achieving an F1-score of 0.72. Ansari and Govilkar [6] proposed a novel approach for Marathi and Hindi sentences using the Support Vector Machine (SVM) method, but they relied on Romanized translations instead of Devanagari scripts.

Kale, Sunil D., et al. [7] discussed various Indian languages and the corresponding Sentiment Analysis work done in that area. About the Marathi language, they compared a few previously done research works with their methods and obtained their results.

One of the pioneers in this field was Balamurali et al. [8]. They suggested a cross-lingual approach and a 2-class classification method (positive and negative). Linked-WordNets were used by them, with both Hindi and Marathi texts in their dataset.

Deshmukh, Sujata, et al. [9] used WordNet in their approach and calculated the sentiment using a dictionary-based method. Keywords from the given sentence are detected and mapped to a previously assigned polarity. The average polarity for each sentence was then computed using the individual polarity of each keyword in a sentence. The sentiment of a sentence is decided based on this generated polarity.

In [10], Snehal and Swati used the Naïve Bayes (NB) probabilistic model and SVM for lexicon-based sentiment analysis. The classifier uses a positive and negative word set for reference. Their approach is optimized for the analysis of customer reviews and feedback.

M. K. Patil et al. [11] present various problems faced by researchers in Marathi sentiment analysis and discuss a semantic-corpus based model for the Marathi sentiment analysis and opinion-mining.

Catelli et al. [12] and Pota, Marco, et al. [13] suggested various methods for BERT-based analysis, which are useful for training the BERT machine learning model in the case of non-English languages. Pota, Marco, et al. [14] proposed a BERT-based pipeline for sentiment analysis for both English and Italian sentences.

3 Methodology

In the first section, the input dataset is pre-processed using the proposed pipeline as depicted in the fig. 1. A detailed description of the same is provided in section 3.2. The pre-processed texts were then used to generate word and sentence embeddings. Embeddings are needed to extract numerical features from the input Marathi texts. The following techniques were used to generate the embeddings for the proposed models: fastText for BiLSTM based model, IndicBERT for IndicBERT based model and Keras word embeddings for Artificial Neural Network (ANN) based model.

The proposed models are then trained on the extracted word or sentence embeddings. Each model is associated with hidden fully-connected layers before the final output layer, to further enhance feature extraction. The IndicBERT based model was run for 100 epochs, while the ANN and BiLSTM based models were run for less than 10 epochs to obtain the optimal accuracy and avoid overfitting. The optimal learning rate for the IndicBERT based model was 0.0001%. Dropout rates of 10% were used for the fully connected layers. Validation dataset was used to validate the trained models and to avoid overfitting. After the validation, models were individually tested on the test dataset. Classification report was then generated from the results obtained.

3.1 Preprocessing

We have used the dataset from [4] for the sentiment analysis. It is a tweet-based, class-balanced dataset with over 16,000 tweets in the Marathi language, each corresponding to one of these three sentiment output classes: positive, negative, and neutral. The dataset was available in training, testing, and validating splits. Each split consists of a tweet column with the texts in the Marathi language and a sentiment

column with the appropriate class identity for the corresponding sentiment. Some of the features of the dataset include:

- The training, testing, and validation datasets consist of 12114, 2250, and 1500 rows. Each split has unique, class-balanced Marathi tweets in it.
- The maximum length of a tweet ignoring Uniform Resource Locators (URLs) and stop words is rounded off to 250 words. We used this length to pad the strings to ensure a consistent input shape while training.
- The number of unique words after removing the URLs and stop words is approximately equal to 60,000. This data is helpful during tokenization.
- The tweets also consisted of English words in the Marathi texts. English words were observed the most in the hashtags part of tweets.

The available dataset had not undergone any prior preprocessing; hence, we performed a series of steps on the training and testing dataset independently before the model training for the sentiment classification task [15, 16]. These steps include:

- We conducted a comprehensive analysis of the dataset before starting with the preprocessing to gain insights, some of which are presented at the start of this section.
- After analyzing the data, the first step in the preprocessing involved using regular expressions to eliminate punctuation marks and special characters. Examples include: "@", "#", and "?". This step also involved the removal of excessive spaces, tabs, or newline characters with the use of the same.
- We also removed the alpha-numeric English letters from the Marathi texts. Examples include: "100", "#politics", and "https". This step, together with the previous one ensures the removal of the URLs from the text.
- The lists of conjunctions, stop-words, pronouns, and numbers in the Marathi language were manually cumulated from various sources available on the internet by native speakers. We used these lists to remove parts of the sentences with less sentiment significance. Some examples from each list are provided in Table I along with their English translation.
- Because the Marathi language does not consist of lower-case and upper-case letters as in the English Language and all the English letters were removed from the texts in step three, the case conversion step in the English text preprocessing was skipped [17].
- Finally, we stem the words in the Marathi language into their root form. We used a dictionary-based suffix removal algorithm for the same.

Table 1. Examples of words with less sentiment significance

S. No	Example of Numbers	Examples of Pronouns	Examples of Stop-words
1	११ - Eleven	आम्ही - We	त्यामुळे – Therefore
2	१२ - Twelve	तुझे - Yours	झाली – Done
3	१३ - Thirteen	तो - He	होता – Was

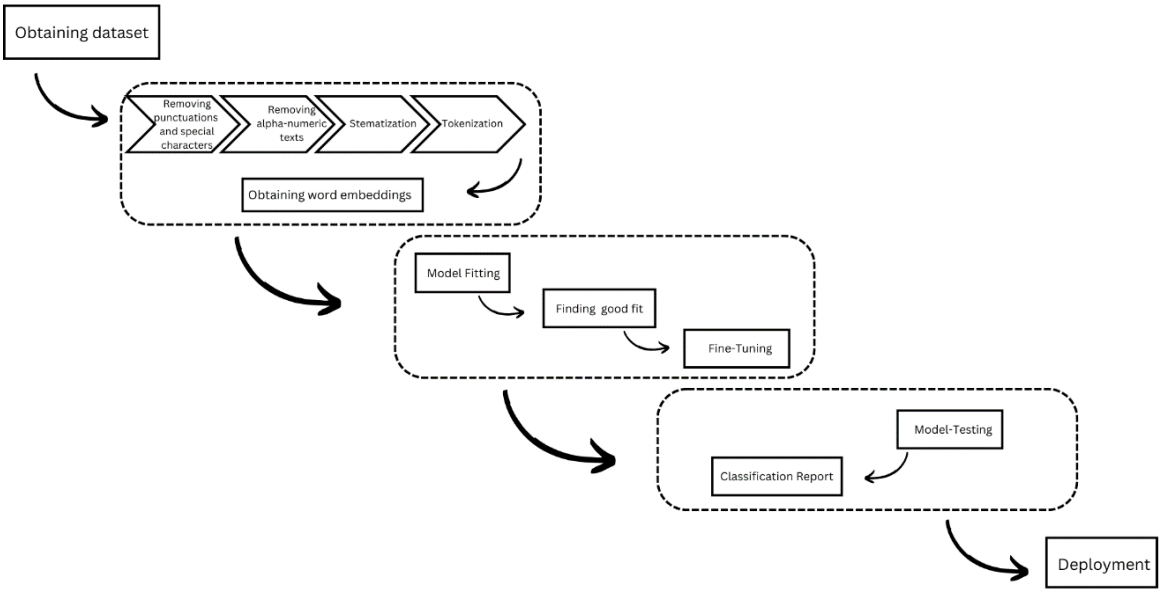


Fig. 1. Proposed Marathi sentiment analysis pipeline.

3.2 Classification Models

All the described models make use of dense hidden layers for the classification purpose including the output layer integrated with the Softmax activation function. The mathematical equation representing the dense layer is as follows.

$$\alpha_j = \beta_j + \sum_{i=0}^{n} (\alpha_i \times w_{ij})$$

"j" is the index of node under consideration that is present in the current dense layer. "i" is the index of node from the previous dense layer. "i" varies from the 0^{th} index to the n^{th} index, n being the number of nodes present in the previous layer. "ß" is the bias associated with the j^{th} node of current layer, and w_{ij} is the weight of the edge from i^{th} node of previous layer to the j^{th} node of the current layer.

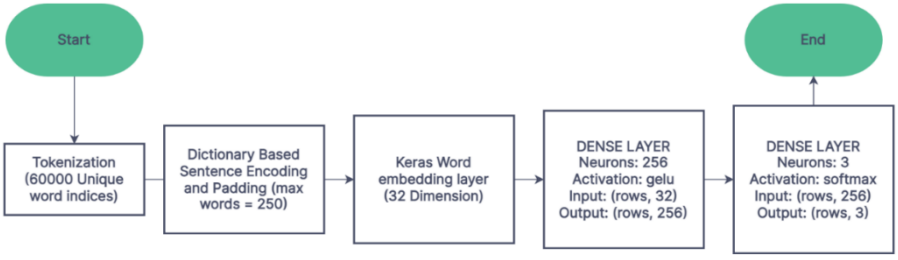


Fig. 2. Block diagram of the model using an ANN.

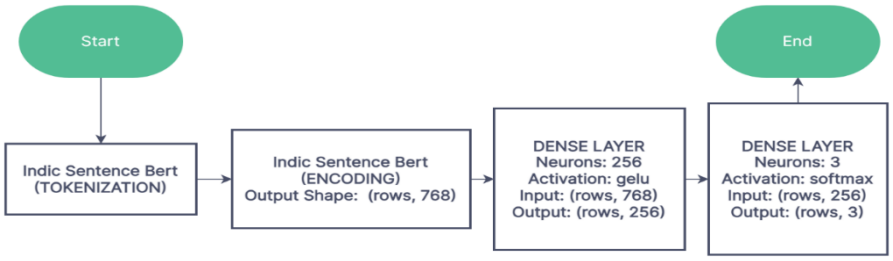


Fig. 3. Block diagram of the model using IndicBERT.

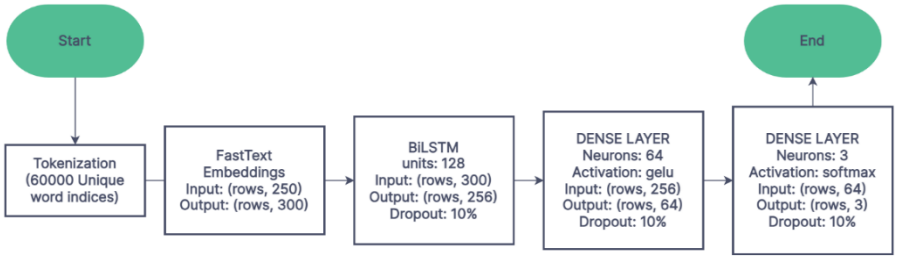


Fig. 4. Block diagram of the model using FastText and BiLSTM.

ANN. We used tokenization to create a vocabulary set of 60000 words. Each word in the sentence was replaced with the corresponding token number in the vocabulary set; Padding with zeros was done for consistent sentence length. Next, a Keras Word Embedding Layer was used to transform these numerical values into 32-dimensional vectors. These vectors were then passed through a Dense Layer with 256 units, using the Gaussian Error Linear Units (GELU) activation function for pattern learning. Finally, another Dense Layer with 3 units and a Softmax activation function is used to classify the text into one of 3 categories. The architecture of the ANN-based model is presented in the fig. 2.

IndicBERT. For BERT based models like IndicBERT, the input should consist of tokens with special labels. This input format was processed through IndicBERT that assigned a set of 768 numbers to each sentence. These vectors were passed through a dense layer with the GELU activation and output 256 numbers. Lastly, another dense layer with Softmax activation classifies the sentence among 3 categories. The architecture of the IndicBERT-based model is presented in the fig. 3.

FastText + BiLSTM. Similar to the ANN model we have used a tokenization that creates a vocabulary of 60000 distinct words. We used different word embedding – fastText. It assigns 300 unique numerical values to each word and encodes it in a 300-dimensional space to capture the semantic relationships between the words. Then, we have passed these vectors through the BiLSTM neural network of 128 units that generates an output of 256-dimensional space (each BiLSTM cell has two sets of LSTM

layers). Then through a dense layer of 64 units with the GELU activation function. The last 3-Node dense layer is the same as the IndicBERT based model. The architecture of the BiLSTM-based model is presented in the fig. 4.

4 Experimental Results

We discuss the outcomes of each of the three models used. Given in Table II is the classification report for each of the models using our proposed pipeline. A graphical comparison of performance of each model is given in Fig. 5. It is evident that our second model—which combines IndicBERT with dense layers—performs better than the other two models in terms of accuracy, with an accuracy rate of 81%.

Table 2. Results obtained using all proposed pipelines

Models	Positive	Negative	Neutral	Weighted Average	Macro Average	Overall Accuracy
ANN	Precision:84%	Precision: 84%	Precision: 74%	81%	81%	
	Recall: 75%	Recall: 86%	Recall: 81%	80%	80%	80%
	F1-score: 79%	F1-score: 85%	F1-score: 77%	80%	80%	
IndicBert + Dense Layers	Precision: 81%	Precision: 85%	Precision: 76%	81%	81%	
	Recall: 78%	Recall: 86%	Recall: 77%	81%	81%	81%
	F1-score:80%	F1-score:85%	F1-score:77%	81%	81%	
FastText + BiLSTM	Precision: 81%	Precision: 79%	Precision: 74%	78%	78%	
	Recall: 73%	Recall: 84%	Recall: 75%	77%	77%	77%
	F1-score: 77%	F1-score:81%	F1-score:75%	77%	77%	

Fig. 6. a), b), c) depict the confusion matrix for the Artificial Neural Network (ANN), IndicBERT. BiLSTM neural network with fastText word embeddings respectively. The IndicBERT model, which uses dense layers and IndicBert phrase encodings, achieved the maximum accuracy of 81%. Both ANN and IndicBERT based models showed higher accuracies in predicting negative and positive sentiments as compared to neutral sentiment. Finally, the BiLSTM based model that uses the aforementioned characteristics to predict sentiments, achieved an accuracy of 77%.

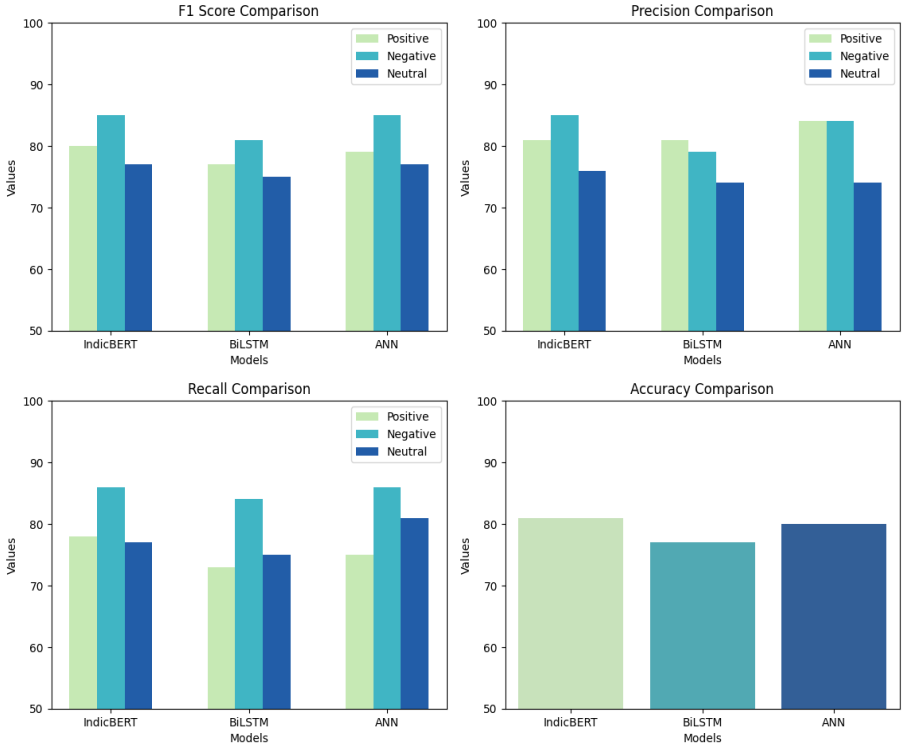
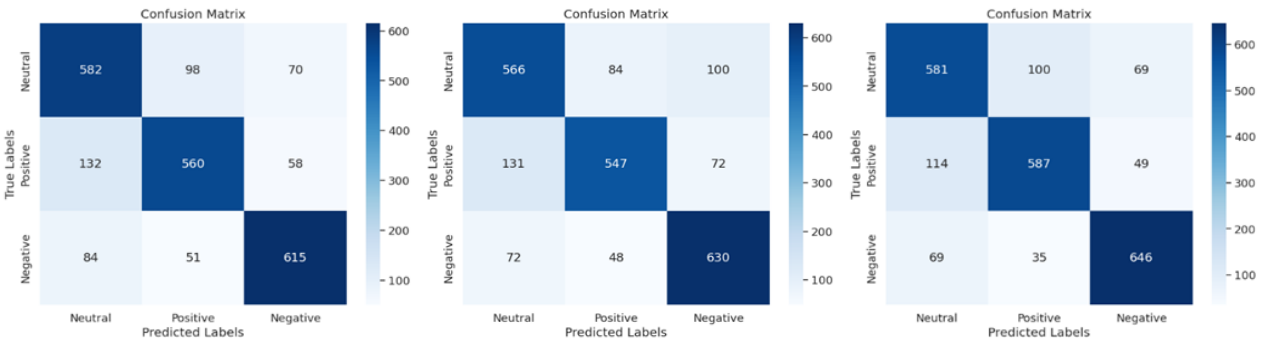


Fig. 5. Comparison of f1-score, precision, recall, and accuracy for IndicBERT, BILSTM, and ANN model for each of the three sentiment classes.



a) b) c)

Fig. 6. Confusion matrices after model testing. a) Confusion matrix for ANN model, b) Confusion matrix for IndicBERT model, c) Confusion matrix for BiLSTM model.

4.1 Overall performance

Fig. 7. depicts the percentage of tweets in each of the predicted classes, along with the actual distribution in the test dataset. The red slice represents the negative sentiment, the green slice represents the positive sentiment and the yellow slice represents the neutral sentiment. We can observe from the confusion matrices that the sentences associated with negative sentiments are predicted more accurately compared to the positive and the neutral ones. Examples of sentiments of sentences in each of the three classes along with the predictions made by our models is given in Table III. A sentiment analysis model must be resilient to unknown datasets in order to accurately forecast the sentiment of a given input text, and our model meets this requirement.

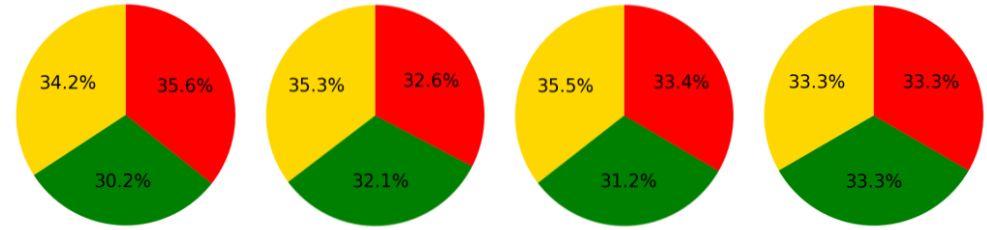


Fig. 7. Distribution of predicted and actual sentiments of the tweets in positive, negative and neutral classes. a) BiLSTM based distribution, b) IndicBERT based distribution, c) Simple ANN based distribution, and d) Actual distribution in the test dataset.

Table 3. Comparing few sentiment predictions with the actual sentiment

S. No	Original Tweet	English Translation (Translated using google translate)	Sentiment predicted by our model	Actual Sentiment
1	सत्तेत येण्यापुर्वी तरुणांना ठोस रोजगार देण्याचं स्वप्न दाखवलंत, अजूनही नोकर भरतीची गाजरं दाखवतच आहात आणि आता म्हणताय सरकारी नोकरी हवी कशाला?	Before coming to power, you showed the dream of providing concrete employment to the youth, you are still showing the carrots of recruitment and now you are saying why do you need a government job?	Negative	Negative
2	Met Hon Union HM @AmitShah ji in Parliament in New Delhi. केंद्रीय गृहमंत्री मा. अमित शाहजी यांची आज संसद भवन येथे सदिच्छा भेट घेतली. pic.twitter.com/92jX8Ox5vG	Met Hon Union HM @AmitShah ji in Parliament in New Delhi. Union Home Minister Hon. Satichha met Amit Shahji today at Sansad Bhavan. pic.twitter.com/92jX8Ox5vG	Neutral	Neutral
3	@INCMaharashtra च्या माजी अध्यक्ष स्व. प्रभाताई राव यांना जयंतीनिमित्त विनम्र अभिवादन! pic.twitter.com/Ztvgy1nvxt	Former president of @INCMaharashtra. Greetings to Prabhatai Rao on his birth anniversary! pic.twitter.com/Ztvgy1nvxt	Positive	Positive

Table 4. Comparison with existing models

Year & Ref.	Method	Comparison parameters with values
2023 [19]	MahaEmoSen - Indicbert	Accuracy: 75.67%
2023 [19]	MahaEmoSen - BiLSTM	Accuracy: 77%
2021 [4]	L3CubeMahaSent-Indicbert	F1-Score: 78.9%
2021 [22]	IndicFed - mBert	Accuracy: 69.98% F1-Score: 69.97%
-	**Our Proposed Pipeline**	**Accuracy: 81%** **F1-Score: 81%**

4.2 Comparison with existing models

We provide a thorough comparison of models using our pipeline and other existing ones in Table IV. It is important to note that depending on the type of comparison parameter (example: accuracy/F1 score) and the model used, the results of some papers outperform our/others models and vice versa [18, 19]. For instance, L3CubeMahaSen has superior accuracy for IndicBERT that is around 84%, while MahaEmoSen has better accuracy for its Multilingual Representations for Indian Languages (MuRIL) model that is around 82%.

In [20] authors proposed a methodology that translates Devnagari Marathi texts into English using google translate to perform sentiment analysis on the English texts and compare the accuracy with the models trained for Devnagari texts. Our methodology outperforms the proposed models in the paper with accuracy of 65.41% and 64.16% for Random Forest and SVM based models respectively.

In [21], authors proposed the use of data augmentation of the dataset using back translation along with the extraction of visual features using visual encoders from the tweets in order to improve the overall accuracy of models. Despite this, our models present a higher accuracy than the proposed models in [21].

The authors in [4] introduced the 'L3CubeMahaSent' dataset for 3-class sentiment analysis of Marathi sentences. The following approaches/models were used – CNN, BiLSTM+GlobalMaxPool, ULMFiT (Based on LSTM) and BERT (mBERT and IndicBERT). Their results concluded that IndicBERT was best-performing model for 3-class sentiment analysis, which is also true in our case.

5 Future work

The future scope for this research includes various directions. One of them being the development of an accurate stemmer for Marathi language. The primary objective of this Marathi stemmer would be to produce root words without losing the contextual meaning of the text. Unlike English, Marathi words undergo intricate morphological changes based on various factors such as tense, gender, number, and case. Therefore, a stemmer must be accurate enough in recognizing and handling such linguistic details effectively.

MuRIL is a state-of-the-art multilingual model developed by Google AI that is specifically designed to handle the linguistic complexities of Indian language. Therefore, fine-tuning MuRIL for sentiment analysis could help achieve better results.

Expansion of the dataset using techniques like scraping and annotation [22] could also lead to the enhancement of the results obtained using the models. Scraping involves extraction of Marathi text data from online sources like social media platforms. Whereas, annotation requires human annotators to label Marathi text samples. Application of Institution-level Federated Learning for Marathi sentiment analysis could help expand the dataset without the need to make the datasets available publicly [23].

Another potential direction is the development of a large-scale stop-words dictionary for the removal of appropriate words from the Marathi texts that do not carry any sentimental meaning. Machine learning based models could be developed for the same purpose [24], making the removal of such stop-words easy and help sustain the sentimental and contextual meaning of the sentence.

6 Conclusion

This paper focuses on sentiment analysis of Marathi-language sentences. From experimentation, we conclude that the optimal model for performing sentiment analysis was IndicBERT (combined with dense hidden layers) with the highest accuracy of 81%. All the models have performed very well in predicting negative sentiments which is a key factor in sentiment analysis. The accuracy of predicting the neutral sentiment was generally the lowest. The development of an end-to-end pipeline for Marathi text sentiment analysis is the novelty of our work.

References

1. Medhat W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: A survey. Ain Shams engineering journal. 2014 Dec 1;5(4):1093-113.
2. Prabowo R, Thelwall M. Sentiment analysis: a combined approach. Journal of Informetrics. 2009 Apr 1;3(2):143-57.
3. Alessia D, Ferri F, Grifoni P, Guzzo T. Approaches, tools and applications for sentiment analysis implementation. International Journal of Computer Applications. 2015 Jan 1;125(3).
4. Kulkarni A, Mandhane M, Likhitkar M, Kshirsagar G, Joshi R. L3cubemahasent: a marathi tweet-based sentiment analysis dataset. arXiv preprint arXiv:2103.11408. 2021 Mar 21.
5. Divate MS. Sentiment analysis of Marathi news using LSTM. International journal of Information technology. 2021 Oct;13(5):2069-74.
6. Ansari MA, Govilkar S. Sentiment analysis of mixed code for the transliterated hindi and marathi texts. International Journal on Natural Language Computing (IJNLC) Vol. 2018;7.
7. Kale SD, Prasad R, Potdar GP, Mahalle PN, Mane DT, Upadhye GD. A comprehensive review of sentiment analysis on Indian regional languages: techniques, challenges, and trends. International Journal on Recent and Innovation Trends in Computing and Communication. 2023;11(9s):93-110.

8. Balamurali AR, Joshi A, Bhattacharyya P. Cross-lingual sentiment analysis for indian languages using linked wordnets. InProceedings of COLING 2012: Posters 2012 Dec (pp. 73-82).
9. Deshmukh S, Patil N, Rotiwar S, Nunes J. Sentiment analysis of Marathi language. International Journal of Research Publications in Engineering and Technology IJRPET.. 2017 Jun;3:93-7.
10. Snehal V. P. Swati M. Sentiment analysis in Marathi language. International Journal on Recent and Innovation Trends in Computing and Communication Volume: 5 Issue: 8. Aug 2017.
11. M. K. Patil, N. Chaudhari, B. V. Pawar and R. Bhavsar, "Exploring various emotion-shades for Marathi Sentiment Analysis," 2021 Asian Conference on Innovation in Technology (ASIANCON), PUNE, India, 2021, pp. 1-5, doi: 10.1109/ASIANCON51346.2021.9544961.
12. Catelli R, Pelosi S, Esposito M. Lexicon-based vs. Bert-based sentiment analysis: A comparative study in Italian. Electronics. 2022 Jan 26;11(3):374.
13. Pota M, Ventura M, Fujita H, Esposito M. Multilingual evaluation of pre-processing for BERT-based sentiment analysis of tweets. Expert Systems with Applications. 2021 Nov 1;181:115119.
14. Pota M, Ventura M, Catelli R, Esposito M. An effective BERT-based pipeline for Twitter sentiment analysis: A case study in Italian. Sensors. 2020 Dec 28;21(1):133.
15. De Clercq O, Lefever E, Jacobs G, Carpels T, Hoste V. Towards an integrated pipeline for aspect-based sentiment analysis in various domains. InProceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis 2017 Sep (pp. 136-142).
16. Agarwal A, Sabharwal J. End-to-end sentiment analysis of witter data. InProceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data 2012 Dec (pp. 39-44).
17. Camacho-Collados J, Pilehvar MT. On the role of text preprocessing in neural network architectures: an evaluation study on text categorization and sentiment analysis. arXiv preprint arXiv:1707.01780. 2017 Jul 6.
18. Naukarkar RA, Thakare AN. A review on recognition of sentiment analysis of Marathi tweets using machine learning concept. Int J Scienti Resear Sci, Eng Tech (IJSRSET). 2021 Mar;8(2):190-3.
19. Bhagat C, Mane D. Survey on text categorization using sentiment analysis. Int. J. Sci. Technol. Res. 2019;8(8):1189-95.
20. Gavali, Harry. Text sentiment analysis of Marathi language in English And Devanagari script. Diss. Dublin Business School, 2020.
21. Chaudhari P, Nandeshwar P, Bansal S, Kumar N. MahaEmoSen: Towards emotion-aware multimodal Marathi sentiment analysis. ACM Transactions on Asian and Low-Resource Language Information Processing. 2023 Sep 22;22(9):1-24.
22. Pingle A, Vyawahare A, Joshi I, Tangsali R, Joshi R. L3Cube-MahaSent-MD: a multi-domain Marathi sentiment analysis dataset and transformer models. arXiv preprint arXiv:2306.13888. 2023 Jun 24.
23. Mehta J, Gandhi D, Rathod N, Bagul S. IndicFed: A Federated Approach for Sentiment Analysis in Indic Languages. InProceedings of the 18th International Conference on Natural Language Processing (ICON) 2021 Dec (pp. 487-492).
24. VIDYAVIHAR M. Sentiment analysis in Marathi language. International Journal on Recent and Innovation Trends in Computing and Communication. 2017;5(8):21-5.