American Research Journal of Computer Science and Information Technology

Volume 5, Issue 1, 1-5 Pages
Research Article | Open Access
ISSN (Online)- 2572-2921
DOI : 10.21694/2572-2921.22002

AMERICAN RESEARCH JOURNALS
An Academic Publishing House

Sentiment Analysis of Marathi Tweets: A Comparative Study of Transformer Based Models

Vedant Jadhav*[1], Shantanu Tripathi[1], Yashdeep Shetty[1], Aakash Tiwari[1] and Arun Chauhan[2]

[1]Department of Computer Science Engineering, Indian Institute of InformationTechnology, Dharwad.
[2]Department of Computer Science and Engineering, Graphic Era University.

ABSTRACT

The Internet is widely utilized as a platform for exchanging information and ideas. In these encounters, a large volume of textual data is produced. NLP requires a large amount of data for training, which can be found on social networking platforms such as Facebook, Twitter etc.(Shetty) Sarcasm identification, fake news detection, sentiment analysis, and other similar tasks have become possible and valuable. We have trained different transformer-based models on Marathi, a low-resource regional language, and compare their performance in this study (Magueresse et al., 2020). We present that MuRIL gives better performance than other Transformer based models.

INTRODUCTION

Social media sites such as Facebook, Twitter, etc. gained popularity in this digital age. Communication through these platforms is possible with friends, family, and other users around the globe. Users can freely express themselves on the internet using these social media platforms. These social platforms provide an enormous amount of textual data that is rich, diverse and covers a wide variety of topics. Researchers refer to this type of data as ”big data.” Big Data’s ”3V” refers to the data’s variety, volume, and velocity.(Big). The availability of such large data sets could be a tremendous opportunity, as many data points are required to train NLP models such as sentiment analysis, sarcasm recognition, and fake news detection. Our focus will be on sentiment analysis, which involves analyzing different types of speech in large data sets.

Sentiment Analysis is the method of determining a piece of content’s connotation. Sentiment analysis for tweets employs a combination of natural language processing (NLP) and machine learning approaches to provide weighted sentiment scores to the features addressed in the conversation. Sentiment analysis helps us gauge public reception of newly released products, public opinion(El Barachi et al., 2021), and understand customer experience(Miranda and Sassi, 2014).

Common NLP models require large amounts of data(Lauer, 1995) for training and testing and/or sophisticated language-specific engineering. Unfortunately, such amount of data is unavailable for most languages and in most cases trained speaker is not available to build language models(Mhaske and Patil, 2016). Therefore, traditional learning approach will not work due to lack of sufficient amount of data. One alternate approach is Cross-Lingual transfer Learning (Kim et al., 2017), where we exploit the commonalities between two languages. We train the model on a resource-rich language and transfer the model over to the lower resource language.

Marathi is considered as a low resource language despite it being the 4th most spoken language in India. 83 million people natively speak Marathi but it lacks a large, publicly available monolingual corpora or evaluation benchmarks. NLP tasks on Indic Languages become more important as the population of users consuming online content grows.

We trained multilingual models on Marathi tweet+ using MuRIL model. We chose MuRIL as the base model as it is very compact and easier to use in downstream tasks. On comparing the results we got from every model, MuRIL surpassed every other model it was up against and can be declared as the go-to model for Marathi NLP tasks.

DATASET

For our research, we used the data set provided in the paper (Kulkarni et al., 2021). The dataset was pre-divided into training, testing, and validation sets with 12114 tweets in the training set, 1500 tweets in the validation set, and 2250 tweets in the test set. The data is preannotated and balanced

between classes. All of the tweets are in Marathi, and they express a diverse spectrum of emotions.

Dataset Annotation

Positive, negative, and neutral labels are used to categorize the data. Table 1 shows the annotations for the sentiments.

Table 1. Sentiment annotation

Sentiment	Label
Positive	1
Neutral	0
Negative	-1

Tweets displaying positive emotions such as gratitude, happiness, support, inspiration, respect are tagged as positive. Tweets showing negative emotions such as disrespect hate, grief, disagreement, insult, opposition are tagged as negative. Tweets containing sarcasm and irony are tagged as also negative. Tweets that are critical in manner, or having very strong or harsh language are tagged as negative. Although, if the criticism is constructive, and stating possible improvements, then it is tagged as positive. Some tweets that consisted of mixed sentiments, were tagged by the more dominant sentiment expressed by them.

Additionally, tweets that did not express strong emotions but stated facts or made straightforward statements were assigned neutral sentiments.

An example of tweets:

1. Tweet: ज्येष्ठ पत्रकार अनंत दीक्षित यांच्या नि- धनाचे वृत्त दु:खद आहे. चार दशकं त्यांनी आप- ल्या परखड लेखणीने पत्रकारितेत अमूल्य योगदान दि- ले. दीक्षित यांच्या मागर्दशर्नाखाली पत्रकारांची पि- ढी घडली. अनंत दीक्षित यांना भावपूणर् श्रद्धांजली! आम्ही त्यांच्या परिवाराच्या दु:खात सहभागी आहोत. pic.twitter.com/s3gnQQLtpk

Translation: The demise of senior journalist Mr. Anant Dixit is a sad occasion. His career spanning over four decades, his forthright and honest articles have immensely contributed to the journalism world. Under his guidance and mentorship, a new generation of journalists has flourished. May he rest in peace. We grieve with his family. pic.twitter.com/s3gnQQLtpk

Label: -1

2. ***Tweet***: आगामी निवडणुकांच्या पार्श्वर्भूमीवर आज म्हसळा तालुक्यातील कार्यकर्त्यार्ंशी संवाद साधला व त्यांना मागर्दशर्न केले. #Mhasala #Raigad #maharashtra #assemblyelection 2019 pic.twitter.com/ptNKKbwHPG

Translation: Given forthcoming elections, met with, interacted, and guided the party workers of Mhasala taluka. #Mhasala #Raigad #maharashtra #assemblyelection 2019 pic.twitter.com/ptNKKbwHPG

Label: 0

3. ***Tweet***: वाढदिवसाच्या हार्दिक शुभेच्छा उद्धव ठाकरे जी. तुम्हाला निरोगी व दीघार्युष्य लाभो या सदिच्छा. @OfficeofUT

Translation: Wish you a very happy birthday Mr. Uddhav Thackeray. Wishing you a long and healthy life. @OfficeofUT

Label: 1

Dataset statistics

As mentioned above, we have a total of 15,864 tweets, which are divided into training, validation, and testing sets. All the datasets are class balanced, with each sentiment having an equal number of tweets. Table 2 shows the class-wise distribution and the train, test, validation split.

Table 2. Dataset statistics

Label	Train	Valid	Test
Positive(1)	4038	500	750
Neutral(0)	4038	500	750
Negative(-1)	4038	500	750
Total	**12114**	1500	2250

Preprocessing

The tweets are processed and cleaned before feeding them to the next module. All white spaces and mentions (@) are removed from the tweets. However, removing punctuations poses a challenge. In English, for showing the combinations of consonants, we just write them one after the other, for example, in the word "plethora", the consonants 'p' and 'l' are pronounced together.

But in Marathi, writing symbols one after the other is not the right manner. To pronounce a word, we first write half symbol(s) followed by the full symbol. This system is called जोडाक्षर (Jōḍakṣara)(Lele, 2012)

Such words in Marathi contain symbols, which can be considered as punctuations in English. So, to preserve such words, we have to manually remove such punctuations from the dataset. Apart from these, all the URLs are removed from the tweets. The hashtags are preserved by removing the # and keeping the rest of the text. After preprocessing we made an observation, that some tweets are empty, as they consisted of tokens that were removed during preprocessing. But, it comprised only 0.03% of the dataset, and hence can be omitted from dataset. The new dataset statistics are:

Table 3. Dataset statistics after preprocessing

Label	Train	Valid	Test
Positive(1)	4037	500	750
Neutral(0)	4034	499	750
Negative(-1)	4038	500	750
Total	**12109**	1499	2250

METHODOLOGY

Traditional machine learning models (Naive Bayes, Linear SVM, RBF SVM, Random Forest) were trained on top of tfidf

vectors created from textual data for our research. We worked on certain NLP architectures such as Dense Neural Network, LSTM, and Bi-LSTM to utilise the time series aspects of the data on top of the embeddings created from transformer models like mBERT, IndicBERT, and MuRIL to get better results.[1] Traditional machine learning models (Naive Bayes, Linear SVM, RBF SVM, Random Forest) are trained on top of TF-IDF vectors created from textual data for our study. We have worked on certain deep learning architectures such as Dense Neural Network, LSTM, and Bi-LSTM to utilize the time series aspects of the data on top of the embeddings created from transformer models like mBERT, IndicBERT, and MuRIL to get better results.

Recurrant Neural Networks (RNNs)

RNN is a type of deep learning model and it is a supervised learning algorithm. The neurons here are time-connected to one another. RNN's goal is to remember what information was in prior neurons so that these neurons can send information to themselves in the future for further analysis. It indicates that data from one time instance (T1) is used as input for the following time instance (T2).(T, 2021)

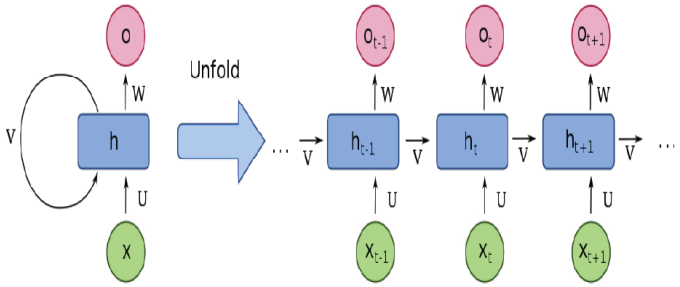

Figure 1. RNN unfolded

LSTM

LSTM(Long Short Term Memory) is a specific type of Recurrent Neural Network, capable of learning long-term dependencies. It is designed to overcome the vanishing gradient problem faced by RNNs. The architecture of LSTM is depicted in the figure below. It has a memory cell at the top that aids in transferring information from one-time instance to the next in an efficient manner. When compared to RNN, it can recall a lot of information from prior states and avoids the vanishing gradient problem. (T, 2021)

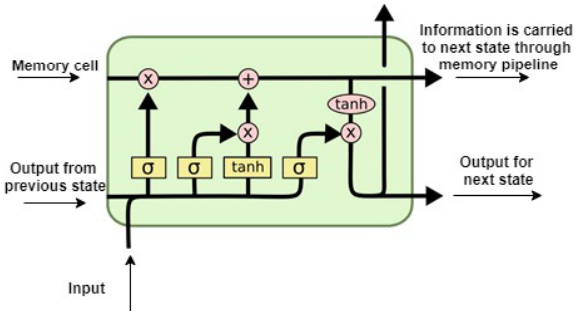

Figure 2. LSTM Architecture

[1]https://github.com/Daishinkan002/Sentiment-Analysis-of-Marathi-Tweets

BiLSTM

Recurrent neural networks, such as Bidirectional Long Short-Term Memory (BiLSTM), works with two hidden layers, it processes data in two directions. This is where LSTM and BiLSTM diverge the most. In natural language processing, BiLSTM has shown to be effective. (Rhanoui et al., 2019)

Transformer

Transformer is a deep learning framework that adopts a mechanism of self-attention. Transformers allow for better parallelism and can reach a new state of the art with comparatively less training. (Vaswani et al., 2017), Various models make use of Transformers in NLP tasks. BERT or Bidirectional Encoder Representations from Transformers is a language representation model. The innovative approach in BERT is its bidirectional training of the language model, which forms a better understanding of the deeper context and language flow.(Devlin et al., 2018a)

mBERT

Multilingual-BERT(mBERT) is a flavor of BERT, which is pre-trained on the Wikipedia pages of the top 104 languages, with a shared word piece vocabulary.(Pires et al., 2019)

Indic-BERT

IndicBERT(Kakwani et al., 2020) is an ALBERT model with a corpus of over 9 billion tokens that have been pre-trained on 12 main Indian languages. For several NLP tasks, it performs as well as other multilingual models with far fewer parameters.

MuRIL

MuRIL(Multilingual Representations for Indian Languages) is a BERT model that has been pre-trained on 17 Indian languages, which includes Marathi, as well as their transliterated equivalents. This model uses a BERT based architecture (Devlin et al., 2018b) pre-trained from scratch using the Wikipedia (?), Common Crawl (com) , PMINDIA (PMINDIA) and Dakshina (Roark et al., 2020) corpora for 17 Indian languages.

When an input sentence is provided to the MuRIL encoder, 3 sets of embeddings are generated.

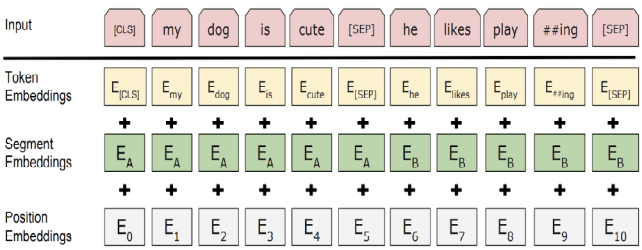

Figure 3. MuRIL encoder : embeddings generated

1. Token embeddings: A CLS token appears at the start of the first sentence, and a SEP token (the separator token) appears at the conclusion of each sentence.

2. Segment embeddings: It assigns a token to each term in the statements to distinguish them as sentence A, B, or C, respectively.

3. Position embeddings: To represent the location of words in the input sequence, each input token is assigned a unique positional token starting at zero.

RESULTS

The dataset obtained from (Maha paper) proved to be efficacious for our deep learning model architectures. All the aforementioned transformer-based models are trained on Marathi tweet data, obtained from Twitter,

Table 4. Traditional methods

Model	Accuracy
Naive Bayes	72.08
Linear SVM	71.82
RBF SVM	72.80
Random Forest	65.38

Table 5. Comparison of metrics

Embedding	Architecture	Accuracy
mBERT	Dense Layer	80.04
	LSTM	80.92
	BiLSTM	81.58
IndicBERT	Dense Layer	74.26
	LSTM	74.31
	BiLSTM	74.71
MuRIL	Dense Layer	82.71
	LSTM	83.60
	BiLSTM	84.04

for sentiment classification analysis, and they have outperformed the traditional Machine Learning models.Owing to LSTM, BiLTSM's ability to output based on previous, bidirectional sequential information respectively, the performance of the models increases as the complexity of the models increase. It is evident, as seen in Table 5 below, MuRIL surpasses mBERT but marginally exceeds IndicBERT, outperforming all models taken into consideration in this task and hence, as claimed, is the go-to choice.

CONCLUSION

Detecting positive, pleasant information on social media is critical during these trying times to assist persons suffering from despair, anxiety, depression, and other mental illnesses. This study discusses several ways for detecting sentiment in social media comments. Multiple state-of-the-art transformer-based models (mBERT, IndicBERT, MuRIL) are integrated with deep learning architectures such as Dense Neural Net, LSTM, and BiLSTM to detect sentiments in the low-resource Marathi language. According to the data, MuRIL and BiLSTM looked to outperform each combination.

REFERENCES

1. What is big data?
2. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018a. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805.
3. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018b. Bert: Pretraining of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
4. May El Barachi, Manar AlKhatib, Sujith Mathew, and Farhad Oroumchian. 2021. A novel sentiment analysis framework for monitoring the evolving public opinion in real-time: Case study on climate change. Journal of Cleaner Production, 312:127820.
5. Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4948–4961, Online. Association for Computational Linguistics.
6. Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. 2017. Cross-lingual transfer learning for POS tagging without crosslingual resources. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2832–2838, Copenhagen, Denmark. Association for Computational Linguistics.
7. Atharva Kulkarni, Meet Mandhane, Manali Likhitkar, Gayatri Kshirsagar, and Raviraj Joshi. 2021. L3cubemahasent: A marathi tweetbased sentiment analysis dataset.
8. Mark Lauer. 1995. How much is enough?: Data requirements for statistical nlp. arXiv preprint cmp-lg/9509001.
9. Kaushik Lele. 2012. Combining consonants in marathi.
10. Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. CoRR, abs/2006.07264.
11. Neelima Mhaske and Ajay Patil. 2016. Issues and challenges in analyzing opinions in marathi text. International Journal of Computer Science Issues (IJCSI), 13(2):19.
12. Marcelo Drudi Miranda and Renato José Sassi. 2014. Using sentiment analysis to assess customer satisfaction in an online job search company. In International

Conference on Business Information Systems, pages 17–27. Springer.

13. Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.
14. PMINDIA. Pmindia.
15. Maryem Rhanoui, Mounia Mikram, Siham Yousfi, and Soukaina Barzali. 2019. A cnn-bilstm model for document-level sentiment analysis. Machine Learning and Knowledge Extraction, 1(3):832– 847.
16. Brian Roark, Lawrence Wolf-Sonkin, Christo Kirov, Sabrina J. Mielke, Cibu Johny, Işin Demirşahin, and Keith Hall. 2020. Processing South Asian languages written in the Latin script: the Dakshina dataset. In Proceedings of The 12th Language Resources and Evaluation Conference (LREC), pages 2413–2423.
17. Badreesh Shetty. Natural language processing(nlp) for machine learning.
18. Santhosh Kumar T. 2021. Natural Language Processing – Sentiment Analysis using LSTM.
19. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. CoRR, abs/1706.03762.

Citation: Vedant Jadhav, Shantanu Tripathi, et al., "Sentiment Analysis of Marathi Tweets: A Comparative Study of Transformer Based Models", American Research Journal of Computer Science and Information Technology, Vol 5, no. 1, 2022, pp. 1-5.