

Multilingual Sentiment Analysis - A Deep Learning Approach using LSTM

Jo Cheriyan

Department of Computer Science and Engineering.
SAINTGITS College of Engineering, Kerala -686532.

Abstract

Social media significantly involves the daily life of people nowadays. Facebook, Whatsapp, and Twitter are mainly used in India, and posts and tweets greatly influence the social community. India ranked second in social media usage. Sentiment analysis is a social media analysis and acts as a powerful way to express and label the opinion toward news, events, products, and policies by crowd or community. These emotions vary from person to person, such as like, dislike, and neutral. Expressing feelings against any such event on a social platform would generate considerable information. There are multiple regional languages in India. Sentiment analysis(SA) from various languages is a requirement in our country. In this article, we suggest sentiment analysis using long-short-term memory (LSTM), an artificial neural network, which provides a diverse set of applications for the analysis. In addition, we focus on analyzing sentiment in Hindi and English by applying artificial intelligence (AI) and deep learning.

Keywords: social media, sentiment analysis, indigenous language, deep learning, LSTM

1 Introduction

The Sentiment analysis is a powerful way of expressing and labeling the sentiments that the crowd or community shows from the source of the text. Usually, people take to social networks like Facebook or Twitter to express their feelings about a topic. These emotions range from solid likes or dislikes towards

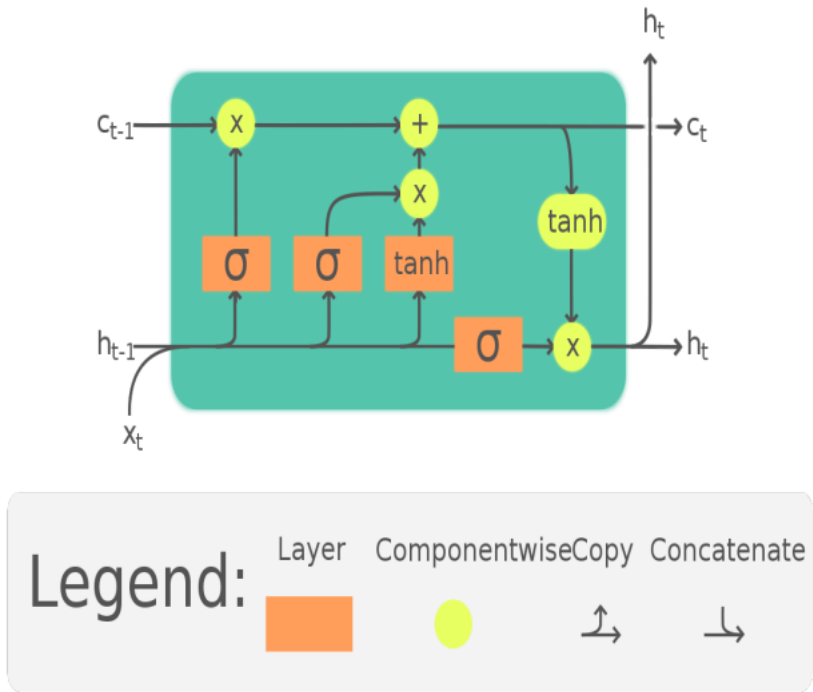


Fig. 1: LSTM Architecture

a product or a policy taken by a community [1]. Expressing emotions against any such event on a social platform would generate considerable information and is impractical to process manually [2]. Machine learning algorithms are adapted to process and analyze this information. For analyzing such extensive data, opinion mining helps in this scenario. The authors use supervised KNN to analyze tweets [3]. Sentiment analysis (or opinion mining) evaluates whether the data is positive, negative, or neutral. Sentiment analyzes helped improve domains such as business strategies, customer feedback, customer needs, and financial time-series forecasting.

Capturing tweets for sentiment analysis directly from social media platforms helps access the insights of social mentions at a level with time. Opinion mining combines computational linguistics and NLP to extract sentiments(positive, negative, or neutral). It helps to understand user or customer likes and dislikes and redesign the product or services. Opinion mining can be performed on structured or unstructured texts using appropriate natural language processing (NLP).

Long short-term memory (LSTM) networks are involved in deep learning. Many recurring neural networks (RNNs) can learn long-term dependencies, particularly in sequence prediction tasks. This behavior is essential to solve complicated problems such as speech recognition and machine translation [4], [5]. Sequence to Sequence modeling is one of the many fascinating applications

of natural language processing. The application for natural language processing and language translation systems extensively uses it. Methods based on sequence-to-sequence (Seq2Seq) conversion, such as text in Hindi converted to English, involve converting sequences from one domain to sequences of another [6]. The LSTM encoder and decoder carry out this Seq2Seq modeling. The Fig. 1 shows the LSTM architecture.

Furthermore, most sentiment analysis research focuses on text written in English, and there is a significant lack of information sources for other languages [7]. As a result, most resources, such as sentiment lexicons and corpora, have been created for the English language [8], [9]. A practical sentiment analysis approach should handle a wide range of languages, allowing it to detect content or specific words in different languages and improve overall sentiment classification in the data [10].

The major contributions of this paper include:

- Categorizes the different sentiments and analyses according to the acceptance using LSTM.
- Recommend the most effective sentiment analysis methodology.

The remaining part of the article is organized as follows, Section 2 discusses the literature on various sentiment analysis processes and uses by machine learning and the deep learning approach. The proposed sentiment analysis using LSTM is explained in Section 3, and its validation procedure and results are depicted in Section 4. Finally, we conclude the article in Section 5.

2 Literature Study

The first study on review categorization, based on the sentiment orientation of the text, was found in the literature. Using the Point-wise Mutual Information (PMI) -Information Retrieval (IR) algorithm, the similarity between the two words is calculated [11]. The algorithm initially extracts the adjective and adverb phrases present in the sentence. The semantic orientation of each phase is calculated and the classification of the review is then made based on the average semantic value of the phrase [12], [13].

There is no specific method for evaluating sentiment analysis in multilingual contexts [13], [21]. The selection of the evaluation model without academic justification and the evaluation criteria were determined solely by the researcher. Although these evaluation models are widely used in practice, it is critical to identify a set of generic evaluation criteria that can accommodate various languages without producing bias toward a specific dataset.

According to the findings of the systematic review of the literature, most models supported two languages, with English being the most widely used language in sentiment analysis studies [14], [5]. None of the reviewed literature has addressed the combination of languages for multilingual sentiment analysis in

English, Chinese, Malay, and Hindi. Tokenization, normalization, capitalization, N-grams, and machine translation are standard multilingual preprocessing techniques. Meanwhile, hybrid sentiment analysis, which includes localized language analysis, unsupervised topic clustering, and multilingual sentiment analysis, is the sentiment analysis classification technique for multilingual sentiment [16], [17], [18]. In terms of evaluation, most studies used precision, recall, and accuracy as a benchmark of the results. More hybrid systems and sophisticated deep learning models are required to address all Hindi SA-related problems [19], [20]. Additionally, a larger dataset and resources create for the construction of practical SA systems.

When previous work is examined, it is found that LSTM performed well with multilingual sentiment analysis [22]. We used LSTM to train the data set and used that model to create a web application that will determine whether the given sentence is positive [23]. Furthermore, the imbalance in the number of instances in different classes significantly affects accuracy. As a result, combining other evaluation models, such as accuracy and precision, or accuracy and f-measure, could lead to correct conclusions about the performance of sentiment analysis models. Our proposed model performs sentimental analysis on multilingual data. So, sentiment analysis of non-English speakers can also be analyzed.

3 The Proposed System

Our work aims to translate a given text from any Indian language to English and then analyze its sentiments.

3.1 Methodology

Sentiment analysis is textual contextual analysis that identifies and extracts personal information from source material, helping businesses understand the social sentiment of their brand, product, or service while monitoring online conversations. LSTM is an abbreviation for extended short-term memory networks used in Deep Learning [24]. It is a class of recurrent neural networks (RNNs) that can learn long-term dependencies, particularly in sequence prediction problems. In addition to single data points, such as images, LSTM has feedback connections, which means it can process the entire data sequence [25]. The LSTM is used in speech recognition, machine translation, and other areas. The different stages of the sentiment analysis process are depicted in Fig. 2.

The utilized dataset was retrieved from Twitter. Pre-processing methods are used to sanitize the data. We use the Google Translate function to translate our text into English. It is a useful tool for comparative researchers when using bag-of-words text models. TextBlob library is used to extract sentiments of the text. The performance evaluation of several machine learning classifiers utilizing their suggested feature set is the contribution of this study. Positive and negative tweets are categorized. Classifiers' accuracy, precision, recall, and F-1 score are used to evaluate their performance.

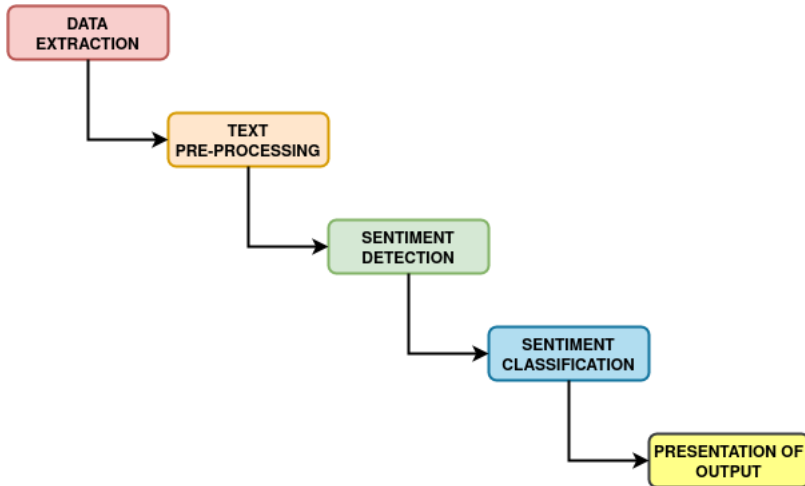


Fig. 2: Stages of Sentiment Analysis Process

4 Experimentation and Result

4.1 Experiment

The competence of our proposed method is validated through experiments. We have developed a python based simulator using Python 3.10.0, and the following python packages, Pandas, Numpy, and Sklearn, were used to build the system. The result of each analysis is found to be promising. Different global metrics compare the accuracy of the proposed method.

4.1.1 Translation Process

The steps involved in the process are translating the text into English from various Indian languages. In the experiment, we translated the data available in the indigenous language, mainly Hindi, into English using each machine translation system. Each sentence is considered a vector whose features marked the presence/absence (1 or 0) of the unigrams and bigrams in the corresponding training set. We obtained the unigrams and bigrams in all the sentences in the training set obtained by pre-processing the translated data in English. Fig. 3 shows the translation process, translating from Malayalam to English.

The translated sentence further processed the cleaning, legalization, and was finally classified into positive or negative. The given process is shown in Fig. 5. The proposed method of sentiment analysis of non-English and Indian languages produced an accuracy of 90 4%. Hence LSTM is a good Deep learning model for sentiment analysis.

```
[ ] translator = Translator()
data="ദൃഢ നിശ്ചയം ചെയ്തവർക്ക് എന്തിനെയും മറികടക്കാൻ കഴിയും"
result = translator.translate(data)

[ ] print(result.src)
print(result.dest)
print(result.origin)
print(result.text)
print(result.pronunciation)

ml
en
ദൃഢ നിശ്ചയം ചെയ്തവർക്ക് എന്തിനെയും മറികടക്കാൻ കഴിയും
Surely those who are determined can overcome
None
```

Fig. 3: Translation Process

Translated Text (En):

This car is so powerful, it's so easy to drive.

Your sentiment:)

Positive

Fig. 4: Experimentation of classifying positive or negative sentiments

```
[ ] model.evaluate(X_test,y_test)

30/30 [=====] - 3s 65ms/step - loss: 0.2793 - accuracy: 0.9844
[0.27932828664779663, 0.9843570756912231]
```

Fig. 5: Accuracy of LSTM performing sentiment analysis

4.2 Result

The result for multilingual analysis in an indigenous language is a challenging problem, and the result is promising. Most of the population in India uses their regional languages. The sentiment analysis on regional languages is more significant in social network analysis.

The loss function is calculated in the training data throughout an epoch and provides the quantitative loss measure at the specified epoch. In Fig. 6, the loss is plotted on the y-axis, and the number of epochs is plotted on the x-axis. The loss decreases after each epoch. There is a slight difference between the performances of the sentiment analysis system using the translated data in English. In the worst case, there is a maximum drop of 8 percent.

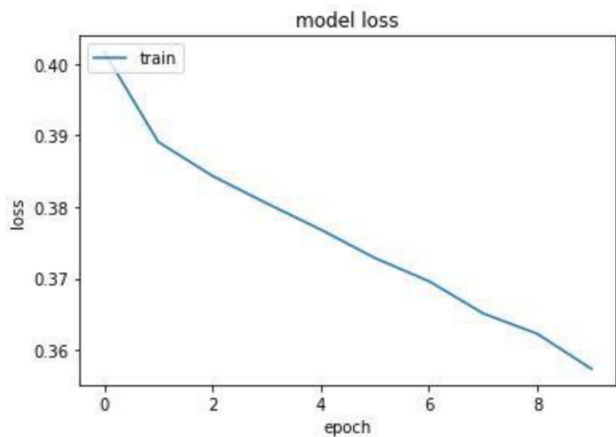


Fig. 6: Performance Loss on sentiment analysis on translated English dataset.

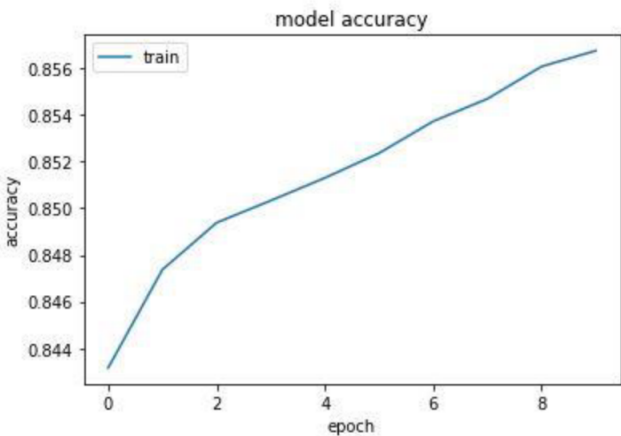


Fig. 7: Accuracy of Sentiment analysis with different epoch.

The accuracy is measured for the sentiment analysis and is plotted in Fig. 7. The graph plots the accuracy on the y-axis, and the number of epochs plotted on the x-axis. Accuracy improves with an increase in the epoch size. Adding all the translated training data together drastically increases the noise level in the training data, creating harmful effects on classification performance.

5 Conclusion

India is one of the most linguistically and culturally diverse countries on earth. Many people prefer to express themselves in their native language. Using our sentimental analyzer, brands can better understand how their customers respond to a specific product and identify their strengths(positive sentiment)

and weaknesses(negative sentiment). Our study also analyzed the accuracy of the LSTM model. The accuracy could be improved by adding more data to the training set. Our study can be extended to be used in live Twitter web crawlers to analyze people's emotions, which can be used to detect early warning signs of suicide and even help prevent it.

We can broaden the analysis by incorporating different classifications and clusters and other data analysis. In addition to emotional analysis, allowing examinations and comparisons from a new perspective may provide an opportunity to further support the current results and compare the conclusions.

References

- [1] Kim, S.M. and Hovy, E., 2004. *Determining the sentiment of opinions. In COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics (pp. 1367-1373).*
- [2] Liu, B., 2012. *Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), pp.1-167.*
- [3] Phillips, M.S., 2000. *Society and Sentiment. In Society and Sentiment. Princeton University Press.*
- [4] Wang, Jin, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. "Dimensional sentiment analysis using a regional CNN-LSTM model." In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers), pp. 225-230. 2016.*
- [5] Murthy, G.S.N., Allu, S.R., Andhavarapu, B., Bagadi, M. and Belusonti, M., 2020. *Text based sentiment analysis using LSTM. Int. J. Eng. Res. Tech. Res, 9(05).*
- [6] Sun, X., Ma, X., Ni, Z. and Bian, L., 2018, December. *A new lstm network model combining textcnn. In International Conference on Neural Information Processing (pp. 416-424). Springer.*
- [7] Yao, Lirong, and Yazhuo Guan. "An improved LSTM structure for natural language processing." In *2018 IEEE International Conference of Safety Produce Informatization (IICSPI), pp. 565-569. IEEE, 2018.*
- [8] Tripathi, M., 2021. *Sentiment analysis of nepali covid19 tweets using nb svm and lstm. Journal of Artificial Intelligence, 3(03), pp.151-168.*
- [9] Dashtipour, Kia, Soujanya Poria, Amir Hussain, Erik Cambria, Ahmad YA Hawalah, Alexander Gelbukh, and Qiang Zhou. "Multilingual sentiment analysis: state of the art and independent comparison of techniques." *Cognitive computation* 8, no. 4 (2016): 757-771.

- [10] Lo, Siaw Ling, Erik Cambria, Raymond Chiong, and David Cornforth. "Multilingual sentiment analysis: from formal to informal and scarce resource languages." *Artificial Intelligence Review* 48, no. 4 (2017): 499-527.
- [11] De Vries, E., Schoonvelde, M., Schumacher, G. (2018). No Longer Lost in Translation: Evidence that Google Translate Works for Comparative Bag-of-Words Text Applications. *Political Analysis*, 26(4), 417-430. doi:10.1017/pan.2018.26
- [12] Nemes, L. and Kiss, A., 2021. Social media sentiment analysis based on COVID-19. *Journal of Information and Telecommunication*, 5(1), pp.1-15.
- [13] Turney, Peter. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *Computing Research Repository - CORR*. 417-424. 10.3115/1073083.1073153.
- [14] Divate, M.S., 2021. Sentiment analysis of Marathi news using LSTM. *International Journal of Information Technology*, 13(5), pp.2069-2074.
- [15] Minaee, S., Azimi, E. and Abdolrashidi, A., 2019. Deep-sentiment: Sentiment analysis using ensemble of cnn and bi-lstm models. *arXiv preprint arXiv:1904.04206*.
- [16] Yu, Y., Si, X., Hu, C. and Zhang, J., 2019. A review of recurrent neural networks: LSTM cells and network architectures. *Neural computation*, 31(7), pp.1235-1270.
- [17] Huang, Zhiheng, Wei Xu, and Kai Yu. "Bidirectional LSTM-CRF models for sequence tagging." *arXiv preprint arXiv:1508.01991* (2015).
- [18] Kulkarni, Dhanashree Rodd, Sunil. (2022). Sentiment Analysis in Hindi—A Survey on the State-of-the-art Techniques. *ACM Transactions on Asian and Low-Resource Language Information Processing*. 21. 1-46. 10.1145/3469722.
- [19] Shah, Sonali Kaushik, Abhishek. (2019). Sentiment Analysis On Indian Indigenous Languages: A Review On Multilingual Opinion Mining. 10.20944/preprints201911.0338.v1.
- [20] Dashtipour, K., Poria, S., Hussain, A., Cambria, E., Hawalah, A.Y., Gelbukh, A. and Zhou, Q., 2016. Multilingual sentiment analysis: state of the art and independent comparison of techniques. *Cognitive computation*, 8(4), pp.757-771.

- [21] *Santhosh, N.M., Cheriyan, J. and Sindhu, M., 2021, September. An Intelligent Exploratory Approach for Product Recommendation Using Collaborative Filtering. In 2021 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS) (pp. 232-237). IEEE.*
- [22] *Samih, Younes, Suraj Maharjan, Mohammed Attia, Laura Kallmeyer, and Thamar Solorio. "Multilingual code-switching identification via lstm recurrent neural networks." In Proceedings of the second workshop on computational approaches to code switching, pp. 50-59. 2016.*
- [23] *Nair, A.J., Veena, G. and Vinayak, A., 2021, April. Comparative study of twitter sentiment on covid-19 tweets. In 2021 5th International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1773-1778). IEEE.*
- [24] *Li, Bo, and Heiga Zen. "Multi-language multi-speaker acoustic modeling for LSTM-RNN based statistical parametric speech synthesis." (2016).*
- [25] *Samih, Younes, Suraj Maharjan, Mohammed Attia, Laura Kallmeyer, and Thamar Solorio. "Multilingual code-switching identification via lstm recurrent neural networks." In Proceedings of the second workshop on computational approaches to code switching, pp. 50-59. 2016.*