Sentiment Analysis on Multiple Indian Languages

B Ramya Asa Latha[1], Assistant Professor, Department of CSE,
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh.
Tagore Yuvaraj Singh[2], **Sanam Venkata Manoj Kumar**[3], **Turimella Deepthi Sai Sri**[4],
Tella Welson Raju[5]
[2,3,4,5] UG Students, Department of CSE,
Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh.
ramyaashalatha@gmail.com, yuvarajtagore@gmail.com,
svenkatamanojkumar@gmail.com, deepthisaisriturimella@gmail.com,
welsonraju@gmail.com

DOI:10.48047/IJFANS/V11/I12/209

Abstract

Sentiment analysis has become popular in the computer science community as it is essential for moderating and analyzing information across the internet. There are various applications for sentiment analysis, such as opinion mining, social media monitoring, and market research. Sentiment analysis in Indian languages is gaining importance due to the growth of content on social media, news articles, and other online platforms in Indian languages. Since India is a diversified country it has many languages that are used by millions of people but many Indian languages do not have enough resources for moderation on the internet to analyze the sentiment in the text to use them for either eradicating hate speech or to improve the productivity of companies by the understanding of customer needs from reviews. This paper explains an approach that helps in analyzing the sentiment which helps in content moderation and avoiding negativity on the internet. This approach uses the BERT algorithm for sentiment analysis in English. All the text in other languages will be translated into English and their sentiment is then analyzed. In this approach, we use the BERT algorithm for sentiment analysis on translated English text. This approach works well because sentiment analysis using BERT gives higher accuracy and the translation of text from Indian languages is made easy by the advent of natural language processing. By combining both the above-discussed processes we can analyze the sentiment in multiple Indian languages.

Keywords: Natural Language Processing, BERT, Translation, Indian Languages, Googletrans

Introduction

Sentiment analysis is the process of identifying, extracting, and quantifying subjective information from text data using natural language processing (NLP) and machine learning techniques. Sentiment analysis seeks to ascertain the emotional tone or attitude portrayed in a piece of text, such as a tweet, review, or news story. Sentiment analysis may be used to categorize text as positive, negative, or neutral, as well as more subtle categories like "Positive," and "Negative." This is accomplished by examining numerous aspects of the text, such as the words used, the sentence structure, and the context in which the text is produced. Sentiment analysis has numerous and diverse uses. It may be used to examine client comments in marketing. It can be used to moderate content over social media to reduce hate speech on the internet.

Natural language processing (NLP) is a type of artificial intelligence that helps computers understand and interpret human language. It's also used to help people communicate with computer systems using natural language. NLP is a type of computer science that helps us understand and process human language. NLP algorithms use statistical models and machine learning techniques to learn from large amounts of data, which helps them identify patterns and relationships in language use. It's used in chat-bots and virtual assistants to understand what people say, in sentiment analysis to understand how people feel about things, in translation to help people communicate with people who speak different languages, and in many other areas.

We perform sentiment analysis for Indian languages in this paper. The languages we use are Telugu, Hindi, Bengali, Tamil, Malayalam, and Marathi. The Indian languages have fewer resources compared to the English language because the number of people that use the language is comparatively lesser than English speakers. This paper proposes an approach that helps to create a sentiment analyzer with even fewer resources using the already available resources.

Multilingual sentiment analysis is extremely useful in today's globalized world, as businesses and organizations must comprehend client sentiment and opinion across several languages and cultures. It can analyze social media data, customer reviews, comments, and surveys in some languages to give insights into customer preferences, opinions, and attitudes. Sentiment Analysis in English is extremely popular and we use a similar approach for the Indian languages to obtain better results.

The sentiment analysis of English text is done using the BERT algorithm in this paper. BERT (Bidirectional Encoder Representations from Transformers) is a Google-developed pre-trained deep learning system for natural language processing (NLP). BERT is a sort of transformer-based neural network architecture that models language in both directions. Unlike prior NLP algorithms, BERT examines text in both ways (left-to-right or right-to-left), allowing it to acquire a more thorough grasp of context and meaning. The BERT algorithm is pre-trained on a huge corpus of text data (for example, Wikipedia), allowing it to discover broad patterns and correlations between words and sentences.

Literature Survey

A. Multilingual Sentiment Analysis using RNN-LSTM and Neural Machine Translation

We convert the statements in Indian languages to English language using Neural Machine Translation. In NMT models, which employ deep learning techniques to train neural networks on enormous datasets of parallel texts, pairs of words that reflect the same concept in several languages are referred to as parallel texts. The translated text is sent to RNN+LSTM based hybrid model which gives out the required sentiment output.In a variety of NLP applications, RNN+LSTM has shown to be an efficient architecture for processing sequential input.

B. Sentimental analysis of Indian regional languages on social media

This study uses a variety of machine learning methods to analyse the sentiment of Indian languages. The author employs machine learning techniques including Deep Learning, KNN, Naive Bayes, Linear Support Vector, and Textblob. The algorithm in this case receives data collected from Twitter using the TwitterAPI. The Textblob technique is primarily suggested by the author because it offers greater accuracy than other models. This article analyses several machine learning techniques and demonstrates Textblob's functionality. With a 98% accuracy rate, it performs better than other algorithms in comparison.

C. Sentiment analysis of comments in social media

This study has shown the results of the analysis of Arabic text comments or tweets from Twitter which are scraped using the Twitter API. It verifies the relationship between emojis and text. It says that if emoji and text are used combined for sentiment analysis it would be difficult to analyze. The results of this study state that emoji and text in a statement have higher coherence which means they provide better reliable information when they are combined than when they are divided.

D. Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques

This analysis performed by the author is a comparison of working between multiple text analysis algorithms and analysing their performance. The comparison performed is independent which means that one approach doesn't coincide with other approach there by giving accurate results of each approach. Multiple languages like English, Russian, etc are analysed for sentiment in them. Many algorithms like SVM, KNN, and other classification algorithms are analyzed and accuracy results of them is taken into consideration.

E. Deep learning and multilingual sentiment analysis on social media data: An overview

This study is an overview of all the previous studies published from 2017-2020. This study analyzes all the approaches published and gives an overview of them. The results obtained in this study are

1. Multilingual sentiment analysis is a scarce research area in deep learning.
2. The available approaches are lacking complexity and they do not give expected outcomes.

3. There is a halt in deriving new algorithms for sentiment analysis.
4. The research has shifted from base language sentiment analysis to cross-lingual and code-switching text.

Problem Identification

There is extensive research done on sentiment analysis in the English language but sentiment analysis in Indian languages is scarce of such research and the available research lacks accuracy. There is a growth in the usage of the internet over the years. Indian languages are being used on the internet in social media and to give reviews. There is no moderation in this online text in Indian languages. Companies must assess the emotion of such messages to determine the demands of the user and provide those needs. This will improve the company's sales and increase the profits and trust of the company. [14-22]

Disadvantages

- Fewer resources and data lead to less accurate results.
- Understanding the language from its basic structure is difficult for the system.
- Understanding the context is difficult from the basic syntax of the Indian languages.
- Indian languages lack sufficient research and resources.
- The used approaches are less complex and are not suited for a sufficient understanding of Indian languages.

Methodology

The dataset we use is the Stanford IMDB dataset with around 25000 reviews with both positive and negative sentiments. We use google trans API for the translation of the text from the Indian language to English text. The obtained translated text is then sent as input to the pre-trained BERT model which gives the sentiment of the given Indian language text.

Googletrans API

Using Google's translation API, the Googletrans API Python package offers quick and precise translations across many languages. It gives users the option to translate words, phrases, or even whole documents from one language to another.

The library is based on the Google Cloud Translation API, which offers sophisticated translation features including support for more than 100 languages, intelligent language identification, and real-time text translation. This API is used by Googletrans to offer Python programmers translation services.

The UI of the Googletrans API is straightforward and user-friendly. It has a “Translator” class that lets you translate text, a “LangDetectException” class that deals with language detection issues, and a "LANGUAGES" dictionary that gives a mapping between language codes and language names.

Before using the Googletrans API, you must install the library using pip. By creating a Translator object and utilizing its translate method after downloading the library, you can translate text from one language to another. The translation process requires two inputs: the source text and the target language.

Transliteration, pronunciation, and usage examples are additional features that the Googletrans API supports. It can handle complex sentences and phrases, making it suitable for use in natural language processing (NLP) applications.

Bidirectional Encoder Representations from Transformers Algorithm (BERT)

In 2018, Google created the pre-trained deep learning algorithm BERT (Bidirectional Encoder Representations from Transformers). It is intended to comprehend natural language processing tasks including text classification, question answering, and language translation.

The deep neural network with many layers of transformers makes up the BERT architecture. The transformers are a sort of attention mechanism that enables the network to concentrate on various input sequence segments while processing data.

Because the network is bidirectional, data from both sides of the input sequence are processed concurrently. As opposed to only processing words independently, this enables BERT to comprehend the context and meaning of each word inside a phrase.

We use Softmax as the activation function in the algorithm.

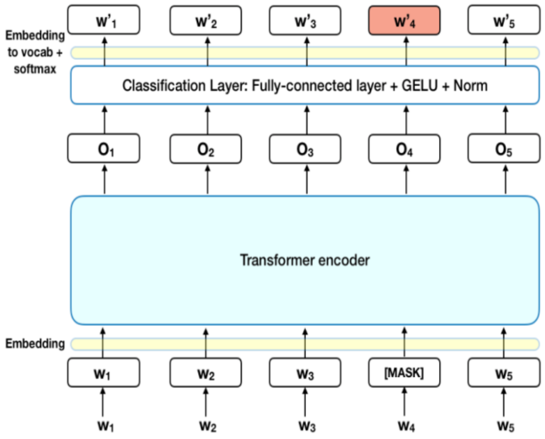


Fig 1: Architecture of BERT Algorithm

Implementation

The steps below are used to implement the method indicated above.

- We import the tensorflow, transformers, and other packages.
- We use the 'tf.keras.utils' library function to import the Stanford IMDB sentiment analysis dataset.

- We separated the text into train and test equivalents for preprocessing, using 20% test and 80% train.
- The preprocessed dataset is imported into a pandas dataframe, which is simple to use.
- We use the Python transformers package to instantiate the BERT model and BERT tokenizer.
- The BERT model doesn't take the pandas data frame as input so we convert the data frame to sequences of data using the InputExample Function.
- We then convert all the data into InputExample format and then tokenize it into objects from which we finally get the BERT acceptable input dataset which is done using two custom functions.
- Adamoptimizer, CategoricalCrossentropy, and SparseCategoricalAccuracy are the optimizer, loss function, and accuracy metric.
- The model is created using the transformers library.
- We use the compile function to compile the instantiated model.
- The instantiation model is put together, and the train data is fed into it.
- We then use the model to forecast the text's emotions.
- We take the test data and tokenize the data.
- The obtained output is fed as input to the BERT model.
- The BERT model analyses the given input tokenized data and produces the output accuracies of the sentiment being negative and positive.
- The maximum is considered as the required sentiment which is done using the np.argmax function.

Results

The below plot shows the dataset's value counts showing how many positive and negative reviews are given as input for training the BERT model.

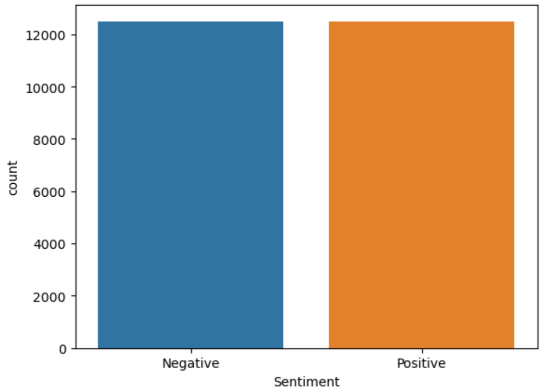
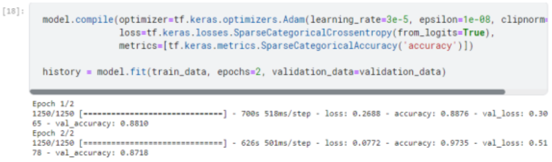


Fig 2: The Dataset division of sentiment values



```
[18]:
model.compile(optimizer=tf.keras.optimizers.Adam(learning_rate=3e-5, epsilon=1e-08, clipnorm=
              loss=tf.keras.losses.SparseCategoricalCrossentropy(from_logits=True),
              metrics=[tf.keras.metrics.SparseCategoricalAccuracy('accuracy')])

history = model.fit(train_data, epochs=2, validation_data=validation_data)
```

```
Epoch 1/2
1250/1250 [==============================] - 700s 518ms/step - loss: 0.2688 - accuracy: 0.8876 - val_loss: 0.30
65 - val_accuracy: 0.8810
Epoch 2/2
1250/1250 [==============================] - 626s 501ms/step - loss: 0.0772 - accuracy: 0.9735 - val_loss: 0.51
78 - val_accuracy: 0.8718
```

Fig 3: Fitting the model

The above image shows the model being compiled and fit using the train data and the validation data.

The model has translated the given multiple language texts to English language using Googletrans API and tokenized the given text and fed the tokenized data to the trained BERT model which gives the accuracy of positive and negative sentiment and we take a maximum of the accuracies given and the maximum accuracy is considered as the main sentiment of the text which is shown in the below image.

Language	Text	Sentiment
Telugu	ఈ ఉత్పత్తిని కొనుగోలు చేయడంలో అర్థం లేదు	Negative
Telugu-English	There is no point in buying this product	
Hindi	रामू बहुत बुरा इसान है और उस पर भरोसा नही करना चाहिए।	Negative
Hindi-English	Ramu is a very bad person and should not be trusted.	
English	human extinction is near	Negative
Bengali	আমি এই ভাষা ভালোবাসি	Positive
Bengali-English	I love this language	
Gujarati	ગુજરાતઝડપથી વિકાસ કરી રહ્યું છે	Positive
Gujarathi-English	Gujarat is developing rapidly	
Malayalam	കേരളം അതിന്റെ സംസ്കാരത്തിനും പ്രകൃതി സൗന്ദര്യത്തിനും പേരുകേട്ടതാണ്	Positive
Malayalam-English	Kerala is famous for its culture and natural beauty	
Marathi	उत्पादन अपेक्षेपर्यंतपोहोचले नाही	Negative
Marathi-English	The product did not live up to expectations	
Tamil	இதுவரை என்னுடையஅனுபவம் அருமையாக இருந்தது	Positive
Tamil-English	My experience so far has been fantastic	

Fig 4: Sentiment analysis on some statements in Indian languages

Conclusion

In this paper, we present a method for implementing a multilingual sentiment analyzer by using a combination of GoogleTrans API and the BERT algorithm. Because the translation accuracy for GoogleTrans API is more than 94% and the better performance of the BERT algorithm in the English language this analyzer can assess sentiment in multiple languages, including Indian languages, with greater precision.

Limitations and Future Scope

This approach fails when there is a discrepancy in the google trans API because Google Trans API sometimes doesn't work as intended. The future scope of this study is as follows:

- We can improve the translation by implementing a translation API for each Indian language separately.
- We try hyperparameter tuning on the model to improve the accuracy even more than obtained.
- We will try increasing the number of layers in the BERT architecture to obtain more features from the algorithm there by improving the accuracy.

References

[1] Anupam Baliyan, Akshit Batra , Sunil Pratap Singh "Multilingual Sentiment Analysis using RNN-LSTM and Neural Machine Translation", IEEE, 2021 8th International Conference on Computing for Sustainable Global Development (INDIACom), March 2021.

[2] Kakuthota Rakshitha, Ramalingam H M, M Pavithra, Advi H D, Maithri Hegde"Sentimental analysis of Indian regional languages on social media", International Conference on Computing System and its Applications (ICCSA- 2021), Vol. 2, Issue 2, November 2021.

[3] Kia Dashtipour, Soujanya Poria, Amir Hussain, Erik Cambria, Ahmad Y. A. Hawalah, Alexander Gelbukh, Qiang Zhou, "Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques", Journal Name, Vol. 8, Issue 4, June 2016.

[4] Marvin M. Agüero-Torales, José I. Abreu Salas, Antonio G. López-Herrera,"Deep learning and multilingual sentiment analysis on social media data: An overview", Applied Soft Computing,, Vol. 107, August 2021.

[5] Adaikkan Kalaivani, Durairaj Thenmozhi "Multilingual Sentiment Analysis in Tamil, Malayalam, and Kannada code-mixed social media posts using MBERT", CEUR-WS, Vol. 11, December 2021.

[6] F. Liu, Y. Liu, and R. Yang, "Cross-lingual sentiment analysis using bilingual embeddings", Springer, Vol. 7, June 2022.

[7] Anita Saroj, Sukomal Pal "Sentiment Analysis on Multilingual Code Mixing Text Using BERT-BASE", FIRE 2020: Forum for Information Retrieval Evaluation, Vol. 2826, December 2020.

[8] Vinothina V, Vigneshwaran R, Karthik V, Saravanan N, "An empirical study of transfer learning for multilingual sentiment analysis in Indian languages", International Conference on Advanced Computing and Communication Systems (ICACCS), 2020.

[9] A. Al Shamsi, Arwa & Bayari, Reem & Salloum, “Sentiment Analysis in English Texts”, Advances in Science Technology and Engineering Systems Journal,Vol. 5, 2021.

[10] P. Badjatiya, S. Gupta, M. Gupta and V. Varma, “Deep learning for hate speech detection in tweets”, WWW '17 Companion: Proceedings of the 26th International Conference on World Wide Web Companion, pp. 759-760, 2017.

[11] O. Hosam, “Toxic comments identification in arabic social media”, International Journal of Computer Information Systems and Industrial Management Applications, vol. 11, pp. 219-226, 2019.

[12] R. Naidu, S. K. Bharti, K. S. Babu and R. K. Mohapatra, “Sentiment analysis using Telugu SentiWordNet”, 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), pp. 666-670, 2017.

[13] Y. Sharma, V. Mangat and M. Kaur, “A practical approach to Sentiment Analysis of hindi tweets”, 2015 1st International Conference on Next Generation Computing Technologies (NGCT), Dehradun, India, pp. 677-680, 2015.

[14] Sri Hari Nallamala, et al., “A Literature Survey on Data Mining Approach to Effectively Handle Cancer Treatment”, (IJET) (UAE), ISSN: 2227 – 524X, Vol. 7, No 2.7, SI 7, Page No: 729 – 732, March 2018.

[15] Sri Hari Nallamala, et.al., “An Appraisal on Recurrent Pattern Analysis Algorithm from the Net Monitor Records”, (IJET) (UAE), ISSN: 2227 – 524X, Vol. 7, No 2.7, SI 7, Page No: 542 – 545, March 2018.

[16] Sri Hari Nallamala, et.al, “Qualitative Metrics on Breast Cancer Diagnosis with Neuro Fuzzy Inference Systems”, International Journal of Advanced Trends in Computer Science and Engineering, (IJATCSE), ISSN (ONLINE): 2278 – 3091, Vol. 8 No. 2, Page No: 259 – 264, March / April 2019.

[17] Sri Hari Nallamala, et.al, “Breast Cancer Detection using Machine Learning Way”, International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-8, Issue-2S3, Page No: 1402 – 1405, July 2019.

[18] Sri Hari Nallamala, et.al, “Pedagogy and Reduction of K-nn Algorithm for Filtering Samples in the Breast Cancer Treatment”, International Journal of Scientific and Technology Research, (IJSTR), ISSN: 2277-8616, Vol. 8, Issue 11, Page No: 2168 – 2173, November 2019.

[19] Kolla Bhanu Prakash, Sri Hari Nallamala, et al., “Accurate Hand Gesture Recognition using CNN and RNN Approaches” International Journal of Advanced Trends in Computer Science and Engineering, 9(3), May – June 2020, 3216 – 3222.

[20] Sri Hari Nallamala, et al., “A Review on ‘Applications, Early Successes & Challenges of Big Data in Modern Healthcare Management’”, Vol.83, May - June 2020 ISSN: 0193-4120 Page No. 11117 – 11121.

[21] Nallamala, S.H., et al., "A Brief Analysis of Collaborative and Content Based Filtering Algorithms used in Recommender Systems", IOP Conference Series: Materials Science and Engineering, 2020, 981(2), 022008.

[22] Nallamala, S.H., Mishra, P., Koneru, S.V., "Breast cancer detection using machine learning approaches", International Journal of Recent Technology and Engineering, 2019, 7(5), pp. 478–481.