Sentiment Analysis of Hindi Tweets using Deep Learning

Abhijit Bera[1], Dr. Chiranjit Changdar[2], Dr. Kalyani Maity Das[3]

[1]*Natural Science Research Centre of Belda College under Vidyasagar University, Belda, Paschim Medinipur, WB, India*
[1]*Assistant Professor, Department of Computer Science & Engineering, Techno India Main, EM-4, Sector-V, Salt Lake, Kolkata-70009, India*
[2]*Assistant Professor, Department of Computer Science, Belda College, Belda, Paschim Medinipur, WB, India*
[3]*Associate Professor, Department of Mathematics, Belda College, Belda, Paschim Medinipur, WB, India*

Abstract-- **Sentiment Analysis is crucial in today's multilingual world. People converse in about 22 different languages on their blog posts, internet forums, product review websites, social networking accounts, and other online spaces just in Indian. Social networking services such as Facebook, Twitter, Google Plus, and others. This research aims to increase the overall sentiment analysis accuracy of Hindi language's tweets by employing a recurrent neural network (RNN) without considering the language's linguistic variability. It has also been explored to separate sentiments from Hindi-language tweets categorize them as positive or Negative or Neutral. For this study, a sample as 4300 Hindi Tweets as sample reviews was taken into consideration. The experimental outcomes of these realistic assessments are expected to be useful for future investigations.**

Keywords-- **Machine learning, Neural Network, Sentiment Analysis, Multilingual Tweets, Recurrent Neural Network (RNN)**

I. INTRODUCTION

People's methods of expressing their thoughts and opinions have altered due to the quick advancement of communication technology and the accessibility of high-speed Internet connections, even in remote locations. These new channels include social media, online forums, blog entries, and product review websites. Social networking services, such as Facebook, Twitter, Google Plus, and others, serve as forums for millions of people to communicate their thoughts and feelings. Sentiment analysis, which predicts the emotions in a phrase, sentence, or corpus of documents, is essential for comprehending the attitudes, opinions, and feelings represented within an online mention. More specifically, it's a framework for assigning labels as positive, negative, or neutral to discussions or content(Pang et al., 2002; Chen et al., 2016). Customers can utilize such a conviction analysis to learn more about a product or service, and marketers can use it to gauge how the public feels about their brand or offering. However, because of Twitter's tone, polarity, lexicon, and grammar, it can be difficult to analyze multilingual tweets that convey human emotions. however, there hasn't been much investigation into native or regional tongues yet.

Recurrent neural networks (RNNs) are used in this paper's analysis of Hindi tweets in an effort to maximize accuracy while minimizing processing time. The datasets from Twitter API v1.1 for popular Indian languages including Hindi tweets and CoRR by Hassan et al., (2016) have been integrated for the current study. A model of a recurrent neural network (RNN) has been developed and evaluated using Hindi, containing reviews of restaurants, movies on IMDB, books on Amazon, and cricket matches. An analysis that contrasts the results of the several models that were employed has been provided. It's noteworthy to note that when it comes to Hindi tweet sentiment categorization, the recurrent neural network (RNN) outperforms other more established techniques like the SVM and Naive Bayes approaches in terms of accuracy, performance, and lowest time complexity. This study describes similar work on recurrent neural network model-based Hindi tweet analysis. Algorithms are used to illustrate the approach, and the model's output, which includes around 4300 Hindi twitter samples, and varying training batch sizes, is provided. The suggested approach ascertains the emotion class of the tweets, signifying the categorization of the tweets into positively or negatively labeled tweets. This approach is universal and can be used with any type of data associated with any topic. The research issues listed below are addressed by the technique used to categorize Twitter users' critiques on Hindi movies.

- Base the research on the Hindi language which is morphologically richer and more free-order than the English language.
- Identifying the opinions that the movie reviewers tweeted. As a result, a technique has been developed in the following sections to address the issue raised in the problem Statement. The goal of the study is to accurately label and classify the movie tweets. The study also compares the performance of the two approaches when it comes to the accuracy attained while employing dual methodologies for sentiment analysis

II. LITERATURE REVIEW

While more linguistically motivated approaches are used for more fine-grained analysis (Riloff et al., 2003; Wiegand et al., 2010; Narayanan et al., 2009), a variety of machine-learning techniques have been used up to this point for document-level sentiment analysis detection (Madan et al., 2021; Pang et al., 2002; Zhao et al., 2008; Dave et al., 2003). Scholars have also investigated sentiment analysis of review documents in the following genres: blogs by Ku et al., (2006), discussion boards or forums by Kim et al., (2006), user reviews by Pang et al., (2004), expert reviews by Zhu et al., (2010), and Hindi tweets by Madan et al., (2021). In 2021, Madan et al., (2021) conducted a comparative evaluation of the Lexicon-Based Sentiment Analysis (LBA) using an upgraded Hindi Twitter dataset, and proposed the Lexicon-Based Approach (LBA) and the Hybrid Approach (HBA) for sentiment classification of movie tweets. For pos_hindi.txt and neg_hindi.txt, the SentiWordNet vocabulary resource yields accuracy values of 60.31% and 62.78%, respectively. Relatively good results have been obtained by the HBA that employs supervised machine learning methods after LBA. With an accuracy of 92.97%, the Decision Tree (DT) classifier is the most accurate of all the classifiers.

Rejanimol et al. (2017) conducted sentiment analysis on movie reviews using a combination of rule-based and lexicon-based techniques. They used movie reviews posted on Twitter to build a dataset. His tweets are gathered into a file, and phrases are taken out of them using punctuation. By tokenizing sentences based on word boundaries and applying grammatical rules to match each token against tokens in the positive and negative word corpuses, the sense and context of each sentence were ascertained. A matching algorithm was created to determine a word's polarity. The count determines the sentence's final polarity.

In order to help researchers, locate the resources they need, like annotated datasets, pre-process linguistic and lexical resources, and select the appropriate sentiment analysis technique in Indian languages by Razia et al., (2018) presented a systematic review of sentiment analysis in Indian languages, focusing on the taxonomy of Indian languages in sentiment analysis based on techniques, domains, sentiment levels, and classes. Sharma et al. (2018) used machine learning and data mining approaches to analyze sentiment. The task of sentiment classification in Hindihas been undertaken with the provision of resources. The advantages and disadvantages of the aforementioned methods for the sentiment categorization task have also been examined across the various genres.

The challenges and issues pertaining to user-generated content in Hindi ,.Using the data sets gathered from SAIL MIKE2015, In 2015 , Patra et al., (2015) employed a machine learning technique called Multinomial Naïve Bayes trained using n-gram and Senti Wordnet Features by Trivedi et al., (2011) with an accuracy of 45%. by Chowdhury et al., (2014) have analyzed the effectiveness of these two machine learning methods and tried to analyze sentiment from Hindi tweets using support vector machines (SVM) and maximum entropy (MaxEnt) on a combination of different feature sets. As a result, they used SVM against a feature vector and emoticons (emotional symbols) and achieved a high accuracy of 93%.

In 2018, Soman et al., (2018) had created a framework by conducting a systematic review to systematically identify the difficulties in comparing the sentiment analysis of tweets in English to Indian regional languages, such as Tamil, Malayalam, Telugu, Hindi, and Bengali.

However, various dependence problems, such the availability of certain elements like emotion symbols and its unsuitability for analyzing complicated phrases, limit the performance of the suggested study.

III. PROPOSED ALGORITHMS AND IMPLEMENTATIONS

In my current work, a straightforward but effective method for sentiment analysis of 4300 Hindi Tweets based on natural language processing (NLP) distinguishes between positive, negative and neutral comments by using a classifier of a neural network model, such as a Recurrent Neural Network (RNN), to improve the dataset's classification accuracy. Recurrent Neural Networks (RNNs) are used to identify the dataset's positive and negative remarks for sentiment analysis of Hindi tweets

An approach for selecting and classifying positive and negative comments using a Recurrent Neural network (RNN) model has been described. The most significant technique for categorization is a neural network. It uses a variety of classification techniques. Back propagation in neural networks is accomplished by turning on the neurons in the hidden layer. There are two stages to a neural network model: training and testing. Using a training dataset, a neural network model updates its model weights to provide mappings from inputs to outputs. The entire dataset's positive and negative comments are trained and given weights during the training phase. The dataset's weighted dictionaries are used for testing. We have employed various classifiers based on neural network models during the testing phase to boost

Step 1: Preparation of the Dataset: The dataset used in this study consists of a selection of tweets written in Hindi that were obtained using various Twitter APIs version 1.1. A variety of Twitter APIs offer language filtering on received posts. The Twitter Search URL's language option was configured to look for tweets in Hindi. The setting is "hin." Additionally, some datasets from Twitter data are merged with reviews from IMDB and Amazon by Hassan et al., 2017). The entire collection of datasets is divided into two groups: a training dataset, which makes up 80% of the datasets, and a test dataset, which makes up the remaining 20%.

Step 2: Preprocessing the Texts: Tokenization is the process of taking text from a corpus and breaking it up into smaller textual units called tokens (words or keywords). The tokens facilitate the process of eliminating superfluous tokens, such reviews into words and documents into paragraphs or phrases. Tokens and their index inside the corpus have been obtained by their utilization. Normalization: Recognizing Hindi words elongated words, and punctuation. Labeling: After classifying the tweets as "positive" or "negative," the positive and negative ones are merged into a single set.

Step 3: Build up the neural network model, such as the recurrent neural network (RNN) model algorithm.

Step 4: Train the model on the training dataset and label it.

Step 5: The Recurrent Neural Network (RNN) algorithm is compiled using the Python compiler.

Step 6: The data sets are used to obtain a final model fit to the training data set and test dataset and the accuracy of sentiment analysis.

Step-7: End

IV. EXPERIMENTAL RESULTS

The aforesaid models are tested in the experiments using datasets of varying sample sizes of Hindi-language tweets, retrieved using various Twitter APIs v1.1, as well as some datasets from IMDB movie reviews and Amazon reviews provided in Hassan et al. (2017). Tokenization, normalization, and labeling of tweets are the preprocessing tasks that are completed. The models for RNN have been constructed for testing with training data sets by utilizing the Python compiler, while the tokenization procedure is executed using Python's Keras packages. After evaluating the model using 128 batches of training data, the following result was discovered: The accuracy score and the execution time (measured in seconds) from model creation to test data validation are displayed in Table 1. Achievement

Table 1.

Comparison of the resultant methodology with already existing results

Ref. No.	Dataset Used	Approach	Algorithms Used	Accuracy
[9]	Internet Movie Database (IMDB)	Machine Learning (ML) algorithms and n-gram models have been applied to classify the dataset.	Maximum Entropy(ME), Support Vector Machine(SVM), Naïve Bayes (NB)	SVM using Unigram:(82.9),ME using Bigram: (77.4),SVM along with a combination of Unigramand Bigram models:(82.7)
[11]	Stock message textual data	Used Ekman Paul's[21] methodology to convert the text into low dimensional emotional space and then machine learning algorithms are Applied for classification.	Decision Tree (DT),Naïve Bayes (NB), Support Vector Machine (SVM)	Unigram:(83.7),Bigram:(80.4), Bigram+Unigram:(84.6)

[8]	Internet Movie Database(IMDB)	Using count vectorizer and Tf-Idf for converting movie reviews into a numeric format and further classifying them using machine learning algorithms.	Support Vector Machine(SVM), Naive Bayes (NB), Maximum Entropy (ME), Stochastic Gradient Descent (SGD), N-gram	NB:(86.23),ME:(88.48),SVM: (88.94),SGD:(85.11)
[6]	Movie Review tweets, Us Election tweets, Food Reviews in the form of tweets	Applied machine learning algorithms along with various term weighing schemes: Tf-Idf, Term Frequency Inverse Class Frequency(TFICF), Mutual Information(MI),X2Statistic (CHI).	Logistic Regression (LR),Support Vector Machine(SVM), Naïve Bayes (NB)	SVMwithTf-Idf:83.75,LRwith MI:(93.2), LR with Tf-Idf: 82.35 on Movie Review Dataset
[34]	Manually annotated Twitter data	LBA has been used for analysis to transformdatain to the required format and then further classifying the tweets Using statistical methods.	BOW approach has been with Tf and Tf-Idf models for feature generation,	Text:(71.9),Visual Feature:(68.7),Multi-view:(75.2)
[21]	Topic-based Twitter tweets	Employed dictionary based experimental research for classification of tweets.	Support Vector Machine(SVM), Naive Bayes (NB)	NB:89%,SVM:76% (Using 250 tweets as the training set out of 350 tweets). NB:84%, SVM:79% (Using350 Tweets as the training set out of500tweets)
[2]	Manually annotated Twitter data	Two approaches are suggested: One using an enhanced Senti Word Net dictionary and another based on Tf-Idf and Unigram models for converting raw tweets to numeric matrices and further classifying them.	Used Senti Word Net Dictionary as a lexical resource.	Positive text document: (60.31),Negative text document:(62.71)SVM: (74.9),NB:(77.2)using Unigram model.

			Applied Unigram and Tf-Idf Models for feature extraction and NB, DT, SVM, LR, and Nearest Neighbor (NN) algorithms for tweet classification.	NB:(84.83),SVM:(89.65),DT: (92.97),NN:(87.34)usingTf-Idf Model.
Proposed Approach				
	Reviews of Amazon, IMDB Movie, Restaurants and Cricket matches of Hindi Tweets	Extracting sentiments out of the Hindi language tweet and an approach towards classifying a tweet into positive, or negative using 4300 sample Tweets in Hindi	Recurrent Neural Network(RNN) for tweet classification	The Accuracy of Classification using RNN algorithm of positive and negative tweets in the Sentiment Analysis dataset is 99.98%

It has been noted that, regardless of the quantity of sample data sets and training batch sizes, good training accuracy is obtained from sentiment analysis of Hindi tweets, based on natural language processing (NLP) using the classifier of a neural network model, such as recurrent neural network (RNN), to identify positive, negative and Neutral comments. This also increases the classification accuracy of the dataset. The figures in Figures 1 and 2 below illustrate the degree of testing and training accuracy and losses of Hindi Tweets and Figure 3: Bar Graph representing the accuracy of classification the number of Positive and Negative of Hindi Tweets.

The following results are also obtained the accuracy of classification the number of Positive, Negative of Hindi Tweets

Accuracy: 99.98%

Training Accuracy: 0.9998

Test Score: 0.00158
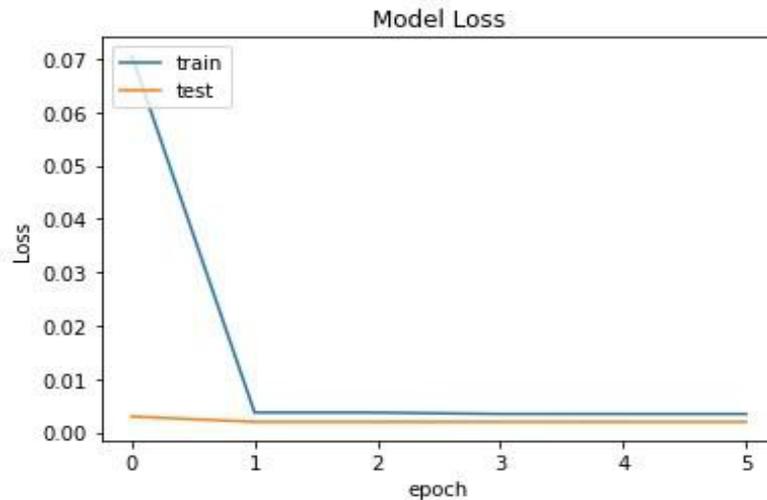
The execution time is 31.38s

Figure 1: Graph representing the classification loss in Hindi Tweets in the dataset
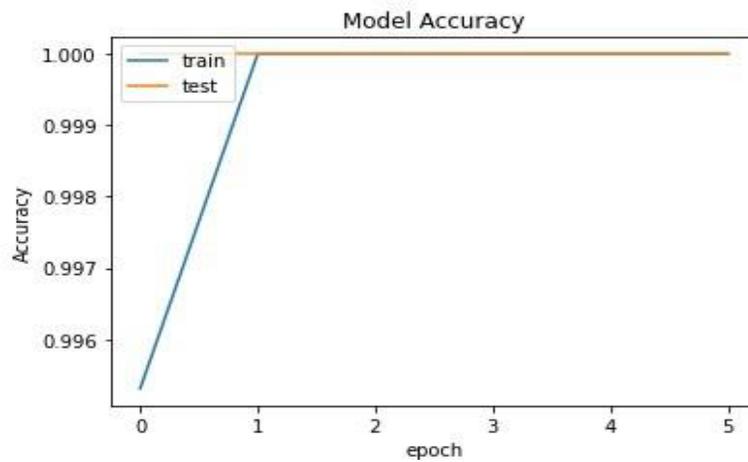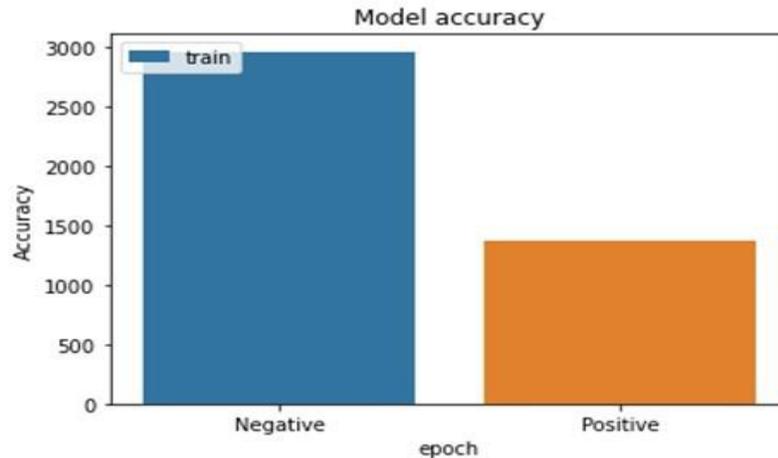

Figure 2: Graph representing the classification Accuracyin Hindi Tweets in the dataset

The accuracy of classification of the number of Positive and Negative Sentences of Hindi Tweets in the data set is shown as bar graph in Figure 3 below

Figure 3: Bar Graph representing the accuracy of classification the number of Positive andNegative Hindi Tweets

V. CONCLUSIONS

As the aforementioned outcome analysis make clear. Recurrent neural networks (RNNs) have been found to detect positive and negative and neutral comments with the highest accuracy among all the models currently in use, as indicated in Table 1. RNNs have also been demonstrated to improve the dataset's classification accuracy. Recurrent Neural Networks (RNNs) yield the best overall accuracy. In the Sentiment Analysis dataset, the accuracy of classifying the number of positive and negative sentences of Hindi Tweets is **99.98%.** It is expected that in the future, more advanced models with lower execution times will be created to effectively manage human vision.

REFERENCES

[1] B. Pang, L. Lee ,S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques", in Proceedings of the Conference on Empirical Methods in Natural Language Processing, Philadelphia, 79–86, 2002.

[2] B. Pang , L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," in A sentimental education. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL, 271–278, 2004

[3] Chen H, Sun M, Tu C, Lin Y, Liu Z ,"Neural sentiment classification with user and product attention", In: Proceedings of the conference on empirical methods in natural language processing, pp 1650–1659, 2016

[4] Chowdhury, S., Chowdhury, W. ," Performing sentiment analysis in bangla microblog posts. International Conference on Informatics", Electronics & Vision (ICIEV),pp 1–6, 2014

[5] [5]Dave, K., Lawrence, S., Pennock, D. M., " Mining the peanut gallery: Opinion extraction and semantic classification of product reviews",. Proceedings of the 12th international conference on World Wide Web, 519–528, 2003

[6] Dattu, B. S," A Survey on Sentiment Analysis on Twitter Data Using Different Techniques", International Journal of Computer Science and Information Technologies, 6(6), 5358–5362,2015

[7] Gupta, U., Chatterjee, A., Srikanth, R., Agrawal, P, "A sentiment-and-semantics- based approach for emotion detection in textual conversations", arXiv:1707.06996, 2017

[8] Hassan, A., Mohammed, N., Azad, A. K. A. ," Sentiment analysis on Bangla and Romanized Bangla text (BRBT) using deep recurrent models", International Workshop on Computational Intelligence (IWCI), 51-56, 2016

[9] Hannah, Kim ,Young-SeobJeong, "Sentiment classification using convolutional neural networks." ,Applied Sciences 9, no. 11 , 2019

[10] Kaur, W., Balakrishnan, V., Singh, B." Social Media Sentiment Analysis of Thermal Engineering Students for Continuous Quality Improvement in Engineering Education", Journal of Mechanical Engineering, Special Edition(4(1)), 263-272, 2017

[11] Ku, L. W., Liang, Y. T., Chen, H. H. ,"Opinion Extraction, Summarization and Tracking in News and Blog Corpora. Proceedings of AAAI Spring Symposium on Computational Approaches to analyzing Weblogs", 2006

[12] Madan. A, Ghose. U, "Sentiment Analysis for Twitter Data in the Hindi Language",11th International Conference on Cloud Computing, Data Science & Engineering, 784-789 , 2021

[13] Narayanan, R., Liu, B., Choudhary, A. "Sentiment analysis of conditional sentences", Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 1,180-189, 2009

[14] Patra, B. G., Das, D., Das, A., Prasath, R. "Shared Task on Sentiment Analysis in Indian Languages (SAIL) Tweets - An Overview",. In Proceedings of the International Conference on Mining Intelligence and Knowledge Exploration, (pp. 650-655). Springer International Publishing, 2015

[15] Razia Sulthana, A., Jaithunbi, A. K., Sai Ramesh, L. " Sentiment analysis in twitter data using data analytic techniques for Predictive Modeling", In National Conference on Mathematical Techniques and its Applications (NCMTA 18). IOP Publishing, 2018

[16] Rejanimol, P. R., Sharika, M., Arun Kumar, N,"Sentimental Analysis of Movie Reviews Based on Twitter Data", International Journal of Pure and Applied Mathematics, 114(11), 137–146, 2017

[17] Riloff, E., Wiebe, J.,"Learning extraction patterns for subjective expressions",. In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP-03). Association for Computational Linguistics, 105–112 , 2003

[18] Shankar, Shilpa, Sridhar, Suma. " A Survey on Sentimental Analysis in Different Indian Dialects. International Journal of Advanced Research in Computer and Communication Engineering, 5(4), 1072–1076., 2016

[19] Sharma, A.K., Chaurasia, S. Srivastava , D.K.,"Sentimental short sentences classification by using CNN deep learning model with fine tuned Word2Vec", Procedia Comput. Sci., 167 , pp. 1139-1147, 2020

[20] Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., Potts, C."Recursive deep models for semantic compositionality over a sentiment Treebank", Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 1642, 1631–1642, 2013

[21] Trivedi, A., Srivastava, A. I., Singh, K., & Gupta, S. K," Literature Survey on Design and Implementation of Processing Model for Polarity Identification on Textual Data of English", International Journal of Computer Science Issues, 8(6), 309-311, 2011.

[22] Wiegand, M., Balahur, A., Roth, B., Klakow, D., Montoyo, A. " A survey on the role ofnegation in sentiment analysis. NeSp-NLP '10", Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, 60 – 68, 2010.

[23] Zhao, J., Liu, K., & Wang, G. " Adding redundant features for CRFs based sentence sentiment classification", Proceedings of 2008 Conference on Empirical Methods in Natural Language Processing, 117–126, 2008

[24] Zhu, F.,& Zhang, X., " Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. Impact of Online Consumer Reviews on Sales: The Moderating Role of Product and Consumer Characteristics. Journal of Marketing", 74(2), 133– 148, 2010