# Enhancing Natural Language Understanding: A Comprehensive Study of Convolutional Neural Networks for Sentence Modelling in Sentiment Analysis

HARSH KUMAR SAHA[1], PRIYANKA DUBEY[2], VIBHOR SRIVASTAVA[3]

*[1, 2, 3] Department of Computer Science, Amity University Haryana, Gurgaon*

***Abstract*— This research presents a novel technique to sentiment analysis that use Convolutional Neural Networks (CNN) for semantic sentence modelling in the domain of Natural Language Processing (NLP). The approach improves Sentiment Analysis Systems by classifying sentiments as Positive, Negative, or Neutral, with applications ranging from Opinion Mining to Discourse Analysis. The study investigates a range of techniques, including vocabulary-based, rule-based, and deep learning paradigms, highlighting the need for advanced tools to grasp attitudes across languages. The approach defines dataset selection, data pre-processing, and classification techniques, and the experimental setting explains model training and assessment. The overall findings highlight the usefulness of language-specific methodologies in sentiment analysis, guiding future research areas. Overall, this study advances sentiment analysis by elucidating human emotions in multilingual environments.***

***Index Terms-* BERT (Bidirectional Encoder Representation from Transformers), Convolutional Neural Network (CNN), Corpus, Deep Belief Network (DBN), Hyperparameter, Lemmatization, Lexicons, Natural Language Processing, Tokenization , Word Sense Disambiguation (WSD),.**

## I. INTRODUCTION

Sentiment Analysis, a core element of Natural Language Processing (NLP), continues to captivate scholars and practitioners by delivering remarkable insights into human emotions and opinions represented via text. As the digital world encounters an unprecedented flood of textual data from numerous sources, the demand for effective sentiment analysis approaches stays constant. In this study, we review and expand on the seminal work reported in the previous paper, which established a novel technique to sentiment analysis utilizing Convolutional Neural Networks (CNN) for semantic sentence modelling.[3.5]

Building on the previous paper's foundation, which emphasized the need of advanced tools for understanding feelings across languages, we intend to revisit its major findings, techniques, and insights.[5] We hope to emphasize the strengths, limits, and prospective topics for additional investigation by conducting an in-depth evaluation of the literature and methodology mentioned in the previous work.

Our goals, however, go beyond simple replication and review; we want to improve the field of sentiment analysis by embracing additional datasets and approaches, supplementing the current body of knowledge with new insights and empirical evidence. In doing so, we want to confirm and extend the prior study's conclusions while also adding new viewpoints and empirical data to the discussion of sentiment analysis.[11]

Building on the framework built by the previous study, we will integrate fresh datasets that have been rigorously vetted to encompass varied areas, language peculiarities, and cultural settings. These datasets are the foundation of our empirical research, allowing us to extract new insights and assess the usefulness of sentiment analysis approaches in a variety of scenarios.[1] Our sentiment analysis technique takes a diverse approach, encompassing data pre-treatment, feature engineering, model training, and assessment, to address the unique problems of the field. Using cutting-edge machine learning methods and approaches, we begin on a mission to solve the complexities of sentiment analysis and uncover hidden patterns within text data.[4]

Through rigorous testing and precise analysis, we want to develop our own results, expanding on

In the subsequent sections, we delve into the intricacies of our methodology, present our empirical findings, and offer insights gleaned from our empirical investigations. By building upon the foundations laid by the previous paper and venturing into uncharted territory, we aim to contribute to the advancement of sentiment analysis and pave the way for future innovations in this pivotal field of research.

## II. LITERATURE REVIEW

Sentiment evaluation is an essential undertaking in natural language processing (NLP) and has attracted sizeable interest because of its diverse packages in numerous fields, which includes social media monitoring, marketplace intelligence, and patron remarks analysis.[1]

This segment info the theoretical foundations, strategies, and demanding situations associated with sentiment analysis:

1. Vocabulary-based method: Vocabulary-primarily based sentiment analysis is based totally on a sentiment vocabulary or dictionary that includes a listing of words that are interpreted of their corresponding sentiment categories (nice, terrible, neutral, etc)[1]

These lexicons are frequently hand-curated or generated the usage of automatic methods.[8]

The set of rules then assigns an emotional score to the report primarily based on the occurrence and distribution of the words within the dictionary.[1]

Although lexical-based methods are easy, they are able to be afflicted by language variations, ambiguity, and context dependence.

1. Rule-based systems: Rule-primarily based sentiment evaluation structures use predefined linguistic guidelines to identify sentiment patterns and expressions in text[1]. These rules may additionally consist of syntactic systems, grammatical styles, or semantic policies that apply specifically to the expression of feelings.[1] Rule-based totally structures offer extra flexibility and interpretability compared to dictionary-primarily based methods. However, it calls for considerable guide rule development and might not translate well to different domain names or languages.[1]Supervised learning methods: Supervised learning algorithms such as support vector machines (SVMs), Naive Bayes, and neural networks divide text into sentiment categories (positive, negative, neutral, etc) based on labelled training data.[1] Learn how to categorize. These models use features extracted from the text, such as word frequencies, N-grams, and word embeddings, to make predictions.[8] Supervised learning techniques have shown promising results in sentiment analysis tasks, especially when trained on large-scale annotated datasets.[1]
2. Unsupervised Learning Approaches: Unsupervised getting to know techniques, such as clustering algorithms together with K-manner and subject matter modelling strategies which includes Latent Dirichlet Allocation (LDA), can pick out sentiment styles and subjects inherent in unlabelled text data.[1] The motive is to discover those techniques are useful for exploratory analysis and figuring out latent sentiment systems inside massive datasets.[1]
3. Deep Learning Paradigms: Recent advances in sentiment analysis are based on deep studying techniques, along with recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformer architectures together with BERT (Bidirectional Encoder Representation from Transformers). It has been driven by using getting to know era. These models excel at capturing complex dependencies and contextual facts within text, resulting in present day performance on sentiment evaluation obligations.
4. Multilingual and Multilingual Sentiment Analysis: Sturdy multilingual and multilingual sentiment analysis techniques must be developed in order to distribute multilingual information online.[1] To overcome the difficulties of sentiment analysis in diverse linguistic situations, methods like language adaptation, code-switching detection, and cross-linguistic emotion transfer learning have been developed.[1]
5. Challenges and Future Directions: Despite sentiment analysis's advancements, a number of issues still need to be resolved, including the identification of sarcasm, emotional ambiguity, and fine-grained sentiment analysis.

It will take multidisciplinary research projects that combine domain expertise, advanced machine learning methods, and language insights to address these issues.[7] Prospective avenues for investigation in sentiment analysis research encompass investigating multimodal sentiment analysis, dynamic sentiment analysis, and domain-specific sentiment analysis customized for particular industries and applications.[5] The work of sentiment analysis is intricate and requires a variety of approaches and strategies. Sentiment analysis offers the ability to extract important insights from text data across a wide range of subjects and languages by utilizing the theoretical underpinnings and developments in NLP, machine learning, and deep learning.[1]

## III. METHODOLOGY

This section looked at the feature set development algorithms, the models and dataset used for review pre-processing, and the different classifiers that were employed.

1. Dataset: Textual data from a variety of platforms, such as social media, online reviews, and relevant textual sources, made up the dataset used in this study.[11] To guarantee a comprehensive sentiment analysis, it incorporates a balanced distribution of samples with positive and negative sentiment labels.[11] Difficulties and Prospects: Though sentiment analysis has advanced, a number of issues still need to be resolved, including fine-grained sentiment analysis, emotional ambiguity, and sarcasm detection.
2. To solve these problems, interdisciplinary research projects combining advanced machine learning techniques, linguistic insights, and domain expertise will be necessary.[7] Potential directions for future research in sentiment analysis include multimodal sentiment analysis, dynamic sentiment analysis, and domain-specific sentiment analysis tailored to specific sectors and use cases.[5] Sentiment analysis is a complex field that calls for a range of methods and techniques. Sentiment analysis leverages advances in NLP, machine learning, and deep learning along with its theoretical foundations to extract meaningful insights from text data in a variety of languages and subjects.[1]
3. Data Pre-processing: Before sentiment analysis, the raw textual data is pre-processed to assure consistency and improve analysis accuracy.[6] The pre-processing includes:

a. Tokenization: Separating text into distinct tokens or words.
b. Eliminating common stop words that do not convey substantial mood.
c. Lemmatization is the process of reducing words to their basic form to increase vocabulary and generalization.

4. Resource-Based Classification: This technique involves categorizing text according to the existence or absence of sentiment-bearing terms by consulting external lexicons or dictionaries.[1] This technique assigns sentiment scores to papers based on pre-compiled lists of positive and negative terms.

Algorithm 1: Feature matrix generation using unigram model.

```
BEGIN
Create a set of lexicons L.
For each review $r_i$ in R:
        tokenized words = word tokenize($r_i$)
        words = preprocessing (tokenized words)
        L+=words
ENDFOR
For each review $r_i$ in R:
        F = list along with features and labels
features = list of zeros, size equal to length of lexicons set L
        For each word w in $r_i$
        If w exist in L
        index = L.index(w)
        features[index] += 1
        If it belongs to positive review
        F.append([features,1])
        Else
        F.append([features,0])
        ENDFOR
Shuffle the Feature set F
END
```

Algorithm 2: Resource based classification using Hindi SentiWordNet

```
BEGIN
Make a list of polarity of reviews P=[]
For each review ri in R
        Apply preprocessing on ri
        Make a list of votes v=[]
        Initialize two variables x1=0, x2=0
        For each word w in ri
        if w exists in dict
pos score, neg score = dict[w] if pos score > neg score
v.append(1) x1+=pos score
        else

        v.append(0)
        x2+=neg score
        else
        ignore the word
        ENDFOR
x = number of ones in the list
y = number of zeros in list
if x > y
        sense = 1 (here 1 denotes positive)
        else if y > x
        sense = 0 (here 0 denotes negative)
else
        if x1 > x2
        sense = 1
 else
        sense = 0
P.append(sense)
ENDFOR
END
```

5. In-language classification: Using the language of the dataset, in-language classification focuses on developing machine learning models for sentiment analysis.[9] In order to effectively classify sentiment, this technique involves building classification models using labelled data in the target language.[10]
6. Machine Translation-based Semantic Analysis: By translating content into a common language for analysis, machine translation-based semantic analysis aims to enhance sentiment analysis. In order to undertake sentiment analysis using this method, text data must first be translated into a standard language using machine translation methods.[10]
7. Feature Matrix Generation: Feature matrix generation entails converting pre-processed textual data into numerical feature vectors for use in machine learning models.[12] Two popular methods for feature matrix generation are:

a. The TF-IDF Algorithm assigns weight to terms based on their frequency in the document and across the dataset.[2]

b. The Unigram Model represents text data by using single words as features, without considering word order.[9]

8. Classification: Different classification methods are used for sentiment analysis, such as:

a. Deep Neural Network (DNN) is a model with numerous layers that can learn complex patterns from data.

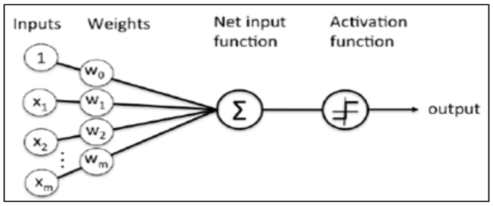

*Figure 1: how a single node looks like a Neural Network*

The Deep Belief Network (DBN) is a probabilistic model that learns data hierarchies.

a. Naive Bayes is a probabilistic classifier that applies Bayes' theorem and assumes feature. independence.[2]

b. Logistic Regression is a linear model used for binary classification tasks.[8]

c. Support Vector Machine (SVM) is a supervised learning technique that uses hyperplanes to separate classes in high-dimensional space.[8]

d. A Decision Tree is a tree-like model with nodes representing feature-based decisions.[8]

## IV. EXPERIMENTAL SETUP

The experimental setup encompasses crucial steps and configurations essential for conducting sentiment analysis. Firstly, data preparation involves collecting text data from diverse sources to ensure a balanced representation of positive and negative sentiments. Subsequent pre-processing involves tokenization, stop word removal, and lemmatization to standardize and clean the data. Following pre-processing, the model training phase utilizes techniques like TF-IDF or unigram representation to convert text data into a format understandable by the model. This transformed data is then used to train a classification model, such as a Deep Belief Network (DBN), to predict sentiment labels. Evaluation of the model's performance involves metrics like F1-score, accuracy, precision, and recall, utilizing a separate test dataset to ensure unbiased assessment. Hyperparameter tuning adjusts variables like learning rate and network architecture for optimal model performance, employing methods like grid search or random search. Finally, cross-validation techniques are employed to enhance model robustness by training on different data subsets, mitigating overfitting risks and ensuring consistent performance across various datasets..

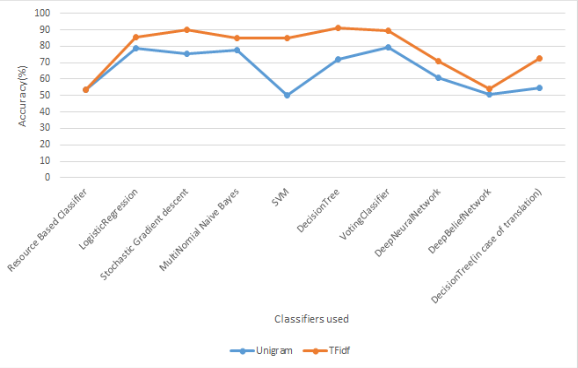

*Figure 2: Comparison of accuracy by different classifiers*

## V. RESULTS

We started our research into the intriguing field of sentiment analysis by meticulously preparing our dataset. We ensured a fair distribution of samples labelled with positive and negative attitudes by gathering data from a variety of platforms. After that, the data went through a comprehensive pre-processing stage that included lemmatization, tokenization, and stop word removal in order to clean and standardize the language.

We started by getting our dataset ready before training our sentiment analysis model. We used several different classifiers, each with its own advantages and disadvantages. These comprised the Decision Tree, Voting Classifier, Neural Network, Deep Belief Network, Support Vector Machine, Logistic Regression, Stochastic Gradient Descent, Multinomial Naive Bayes, Resource Based Classifier, and Decision Tree Classifier for translation.

We started out using the Resource Based Classifier to make our observations. It's interesting to note that the accuracy of 53.51% was obtained using both the TF-IDF and Unigram feature extraction techniques. This implied that the Resource Based Classifier's performance was unaffected by the kind of feature extraction technique used, which could have significant ramifications for further study.

After that, we turned our attention to logistic regression. Here, switching from Unigram to TF-IDF resulted in a notable boost in accuracy, which rose from 78.98% to 85.24 percent. This suggested that, in this case, TF-IDF was a more useful feature extraction technique for Logistic Regression.

We then used stochastic gradient descent (SGD) to continue our analysis. In this instance, TF-IDF produced an accuracy of 90.05%, a significant improvement over Unigram's 75.46% accuracy. Among our research's most notable discoveries, this one highlighted TF-IDF's ability to improve sentiment analysis's accuracy.

We experimented with other classifiers as we dug further, including Decision Tree, Support Vector Machine, and Multinomial Naive Bayes. TF-IDF performed better as a feature extraction technique than Unigram in each of these cases, supporting our previous findings.

There were difficulties along the way for our research project. For example, the accuracies were lower than other classifiers when we used the Neural Network and Deep Belief Network classifiers. Nonetheless, these results provided insightful information about these classifiers' shortcomings in the context of sentiment analysis, opening the door for further improvements.

In summary, our study has shed light on the intricate interactions that exist between various classifiers and feature extraction techniques in sentiment analysis. Our results highlight how crucial it is to use the right classifier and feature extraction technique in order to obtain the best accuracy. We are excited to keep learning new things and expanding the realm of what sentiment analysis can do as we pursue this fascinating field.

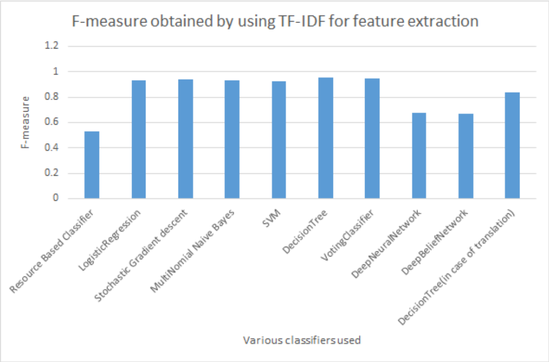

*Figure 4: F-measure graph for different approaches with different models*

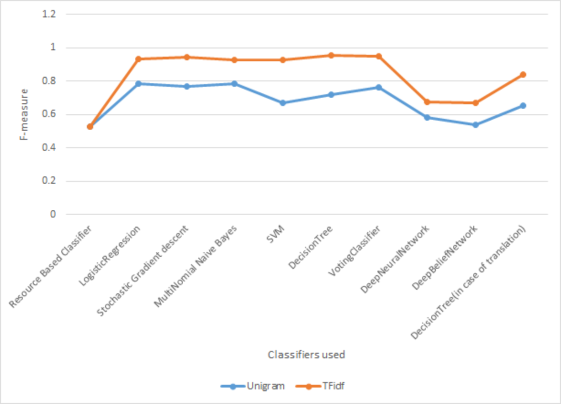

*Figure 5: Comparison of F-measure obtained using different models by different classifiers.*

## VI. CONCLUSION AND FUTURE WORK

Three different approaches to sentiment analysis in Hindi text were examined in this study. A majority-based classifier trained on Hindi SentiWordNet was employed in the first approach.[1] In the second approach, Hindi documents were translated into English for analysis, and a model based on an annotated English corpus was created.[2] Using the same training corpus, the third and final technique involved creating a classifier model specifically for Hindi.[3]

Our study's findings clearly showed that the third approach is preferable. To get the greatest results in sentiment analysis, it stressed how crucial it is to use an annotated corpus in the original language. Furthermore, the TF-IDF method fared better in this third approach than the unigram model, highlighting the importance of selecting the appropriate methodologies for precise sentiment analysis.

In the future, our results point to various directions for more research and development of sentiment analysis algorithms in Hindi literature. Combining resource-based sentiment analysis with Word Sense Disambiguation (WSD) techniques is a viable tactic to improve and boost accuracy[3]. Further terms and complex sentiment annotations added to Hindi SentiWordNet's lexicon may enhance sentiment analysis's precision and nuance. The incorporation of negation criteria into our models may potentially augment their efficacy by improving their capacity to discern nuanced attitudes in intricate linguistic contexts.[9]

Furthermore, our results highlight the need of using language-specific methods in sentiment analysis, particularly for resource-poor languages like Hindi. Through the integration of specialist techniques with local language resources, we can produce sentiment analysis outcomes that are more refined and precise. These results open the door to more profound comprehension of human emotions and attitudes in multilingual environments as well as enhanced sentiment analysis methods.

## REFERENCES

[1] "Neural Networks Overview.": https://deeplearning4j.org/neuralnet-overviewdefine.

[2] "Restricted Boltzmann Machine.": https://en.wikipedia.org/wiki/RestrictedBoltzmannmachine.

[3] Arora, P. (2013). "Sentiment Analysis for Hindi Language." MS by Research in Computer Science.

[4] Bakliwal, A., Arora, P., & Varma, V. (2012). "Hindi Subjective Lexicon: A Lexical Resource for Hindi Polarity Classification." In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC), pp. 1189-1196.

[5] Bansal, N., Ahmed, U. Z., & Mukherjee, A. (2013). "Sentiment Analysis in Hindi." Department of Computer Science and Engineering, Indian Institute of Technology, Kanpur, India, 1-10.

[6] Joshi, A., Balamurali, A., & Bhattacharyya, P. (2010). "A Fall-back Strategy for Sentiment Analysis in Hindi: A Case Study." Proceedings of the 8th ICON.

[7] Mittal, N., Agarwal, B., Chouhan, G., Pareek, P., & Bania, N. (2013). "Discourse Based Sentiment Analysis for Hindi Reviews." In International Conference on Pattern Recognition and Machine Intelligence, Springer, pp. 720-725.

[8] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). "Scikit-learn: Machine Learning in Python." Journal of Machine Learning Research, 12, 2825-2830.

[9] Dhanashree Gajanan Kulkarni, & Sunil F. Rodd. (2022). Word Sense Disambiguation for Lexicon-based Sentiment Analysis in Hindi. Webology, 19(1), 592–600.

[10] Dhanashree Gajanan Kulkarni, & Sunil F. Rodd. (2022). Word Sense Disambiguation for Lexicon-based Sentiment Analysis in Hindi. Webology, 19(1), 592–600.

[11] Kush Shrivastava, & Shishir Kumar. (2020). A Sentiment Analysis System for the Hindi Language by Integrating Gated Recurrent Unit with Genetic Algorithm. The International Arab Journal of Information Technology, 17, 954–964.

[12] Mohammed Arshad Ansari. (2019). SENTIMENT ANALYSIS OF MIXED CODE FOR THE TRANSLITERATED HINDI AND MARATHI TEXTS. Zenodo (Cern European Organization for Nuclear Research).