International Journal of Applied Engineering & Technology

ASPECT TERM EXTRACTION AND SENTIMENT POLARITY ASSIGNMENT WITH LEXICAL RESOURCES IN ASPECT BASED SENTIMENT ANALYSIS

Vijay Kumar Soni [1] and Dr. Smita Selot[2]
[1,2]Department of CSE, SSTC, Chhattisgarh, India

ABSTRACT

The high volume of user-generated content on digital platforms has highlighted the necessity of extracting meaningful insights from various languages. In sentiment analysis, identifying aspect terms is crucial for capturing the emotions of user opinions. This research paper introduces a specialized Aspect Term Extraction Technique for Hindi text, addressing the unique linguistic challenges posed by the language. The proposed method combines natural language processing (NLP) and deep learning techniques to automatically identify aspect terms from Hindi text data. Additionally, by integrating a lexicon-based approach to set the polarity of Hindi sentences and BERT multi-class classification for Aspect-Based Sentiment Analysis (ABSA), we achieved an accuracy of 86.97% in classifying 50K Hindi reviews.

Keywords: Hindi, NLP, BERT, ABSA, Deep learning.

INTRODUCTION

In the era of information explosion, the analysis of user-generated content has become integral for interpreting public sentiment. Among the diverse range of languages contributing to this digital communication, Hindi stands out as one of the most widely spoken languages, reflecting a rich diversity of cultural expressions and linguistic details. Understanding the sentiment embedded in Hindi text is pivotal for numerous applications, ranging from customer feedback analysis to market research and beyond.

Sentiment analysis, a field at the intersection of natural language processing (NLP) and deep learning, is fundamental for discerning opinions expressed in textual data. Sentiment analysis is a fascinating area within text analytics that helps us understand a speaker's opinions about a person or an object. This analysis can be performed at a high level, focusing on the overall sentiment of the text. However, sometimes we need a more detailed understanding of emotions. To achieve this granularity, it is crucial to identify specific entities within the text and analyse the sentiments associated with them. In the field of Natural Language Processing (NLP), these entities are referred to as "aspects," and the process of analyzing their related sentiments is known as ABSA (Akhtar et al 2018).

This research paper introduces an Aspect Term Extraction Technique tailored explicitly for Hindi text. The objective is to address the linguistic subtitles of Hindi, such as nouns, compound words, postpositions, and verb forms, which influence the expression of opinions in this language. In addition to a polarity lexicon, the classification of opinionated social media texts involves considering the presence of conjunctions, negative words, positive words, and consecutive negative words. By developing a hybrid technique, the research aims to enhance the accuracy and relevance of aspect term extraction for Hindi, contributing to a better understanding of sentiments expressed in this linguistic domain.

CHALLENGES IN ABSA FOR INDIAN LANGUAGE

Most current methods aim to determine the total sentiment of a sentence or paragraph, despite the diverse topics (such as laptops or mobiles) and their specific aspects (like the keypad or charger for a laptop, and the battery or camera for a mobile). For instance, consider a review about a mobile: “□□□□□□ □□ □□□□□ □□ □□□□□ □□ □□□□□ □□□□□ □□□□□ □□.” The overall sentiment of the sentence appears to be conflict, as it mentions both positive and negative aspects of the mobile. However, this general sentiment does not provide much useful information. This is where ABSA becomes valuable. ABSA identifies specific aspects mentioned in the review, such as “□□□□□” and “□□□□□”. This process is called Aspect Term Extraction (ATE). Following ATE, Polarity Sentiment Classification (PSC) determines the sentiment associated with each aspect. For example, the sentiment for “□□□□□” might be positive, while the sentiment for “□□□□□” could be negative. Figure 1 shows the overall process of ATE with PSC.

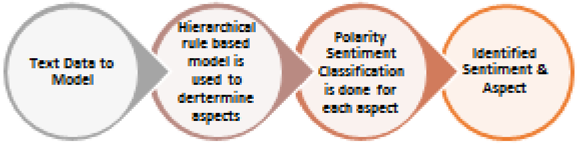
Text Data to Model
Hierarchical rule based model is used to dertermine aspects
Polarity Sentiment Classification is done for each aspect
Identified Sentiment & Aspect

Figure 1. The flow of ATE with PSC

This approach is particularly useful on social networking sites, review websites, blogs, and similar platforms. It allows us to understand a writer's or user's genuine sentiment at a more granular level, providing more detailed insights than just the overall sentiment.

ABSA for the Hindi language faces several challenges that arise from the linguistic, cultural, and resource-specific characteristics of Hindi. These challenges impact the accuracy and applicability of sentiment analysis tools. Here are the key challenges in performing ABSA for Hindi.

Morphological Complexity: Hindi exhibits a complex morphological structure with compound words, inflections, and derivations. This complexity poses difficulties in accurately identifying and extracting aspect terms, affecting the precision of sentiment analysis.

Semantic Ambiguity: Hindi text often contains words and phrases with multiple meanings, leading to semantic ambiguity. Disambiguating the intended sentiment in context is a significant task for sentiment analysis models.

Limited Annotated Datasets: The availability of annotated datasets for training sentiment analysis models in Hindi is limited. Insufficient labelled data hampers the development of robust and accurate ABSA models for Hindi.

Lack of Standardization: The absence of standardized sentiment analysis guidelines for Hindi makes it challenging to develop universally applicable models. Variations in sentiment expression across different domains and contexts further complicate the standardization process.

Successfully overcoming these obstacles will lead to the development of robust and culturally sensitive ABSA tools for the Hindi language.

STATE-OF-THE-ART BERT MODELS

The challenge of achieving a universal representation of text data in NLP has long been a hurdle. However, BERT (Bidirectional Encoder Representations from Transformers) emerged as a groundbreaking solution to this issue. BERT, a sophisticated text embedding model, exhibits exceptional precision in handling various NLP tasks. The fundamental idea behind such applications is to establish an accurate technique for representing text in a manner comprehensible to machines, effectively converting natural language into machine-understandable instructions (H. Liu et al. 2020). Figure 2 shows the basic architecture of the BERT model.

In 2018, Devlin introduced BERT, a language model designed to deeply learn bidirectional text representation for subsequent utilization in ML models. BERT caters to both holistic and tokenized NLP tasks, operating at both sentence and individual text element levels, respectively. Leveraging pre-trained language models like BERT significantly reduces the time required for design, training, and testing, all while yielding accurate results. BERT’s training comprises two stages: pre-training on unlabelled data and subsequent training on application-specific labelled data (Pathak, A el at. 2021). BERT utilizes word embedding techniques to represent input sequences, considering an arbitrary continuous sequence of text tokens. This model requires retraining with data to grasp and represent text effectively, capable of handling nearly every language representation-related task.

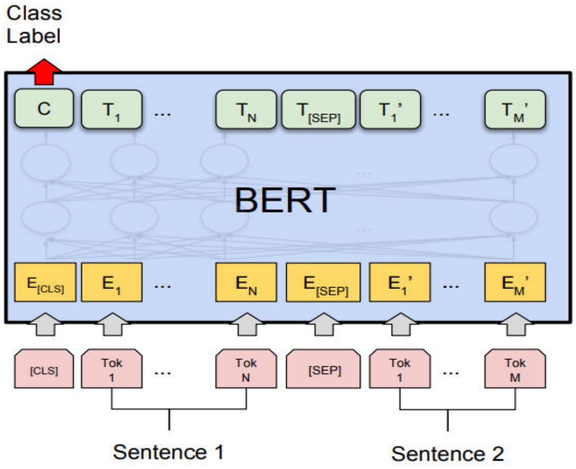

Figure 2. BRET Architecture

RELATED WORK

Akhtar et al. (2018) gathered and annotated 5,417 review sentences from various online sources across 12 different domains. They proposed a multiclass classification model that achieved F-measures of 46.46%, 56.63%, 30.97%, and 64.27% for aspect category detection in the domains of electronics, mobile apps, travel, and movies, respectively. For sentiment classification, the model attained accuracies of 54.48%, 47.95%, 65.20%, and 91.62% in the same domains. The primary contributions of this research are two-fold: establishing a benchmark for aspect category detection and sentiment classification.

Md Shad Akhtar et al. (2020) presented a multi-task end-to-end approach that demonstrates competitive performance in aspect term extraction tasks. Notably, this approach shows superior results in the laptop domain and a 2% improvement in Hindi datasets. It is important to highlight that, unlike existing single-task systems which address aspect term extraction and aspect sentiment classification independently by developing and evaluating separate models, the multi-task systems integrate these tasks.

Pathak A et al. (2021) proposed two ensemble models based on Multilingual BERT, named mBERT-E-MV and mBERT-E-AS. These models outperformed the existing state-of-the-art models on Hindi datasets. For the aspect polarity classification task, they reported accuracies of 69.95%, 51.22%, 75.47%, and 78.09% across four respective domains. Similarly, mBERT-E-AS achieved F1-scores of 73.38%, 52.31%, 59.65%, and 78.61% for aspect category detection in the same domains. Additionally, it reported accuracies of 70.49%, 48.78%, 75.47%, and 79.77% for aspect polarity classification across the four resources.

Sai Aparna et al. (2021) proposed a model utilizing two different word embedding algorithms, Word2Vec and fastText, for feature generation, alongside various machine learning (ML) and deep learning (DL) models for classification. In the aspect-based sentiment analysis (ABSA) task, the LSTM model outperformed other ML and DL models, achieving accuracies of 57.93% with Word2Vec features and 52.32% with fastText features. Generally, the classification models using Word2Vec embeddings performed better than those using fastText embeddings.

V. Yadav (2021) proposed a system for Aspect-Based Sentiment Analysis (ABSA) in the Hindi language using machine learning algorithms. The system utilized a dataset of Hindi product reviews collected from various online sources across 12 domains and explored additional review datasets, including those for vehicles and cars. Additionally, an unsupervised approach was found to have a significant impact on Hindi language ABSA, achieving an overall accuracy of 54.05%.

K. M. Kavitha et al. (2022) utilized a lexicon-based approach for sentiment classification in Hindi, achieving an 86.45% accuracy on 2,717 Hindi reviews.

METHODOLOGY

Sentiment Lexicon

A lexicon consists of a list of lexical features, each assigned a polarity score or semantic orientation such as positive, negative, or neutral. For Hindi sentiment analysis, we employ a rule-based approach that determines the polarity of a sentence based on this list of positive and negative lexical features.

- **Handling Consecutive word order***:* When two negative words appear consecutively in a sentence, they can result in a positive sentiment. For instance, consider the sentence "□□□□□□ □□ □□□□□ □□□□□ □□□□ □□". Here, the consecutive negative words "□□□□□ □□□□" indicate a positive overall sentiment. To handle such cases, we can establish a rule to identify these patterns and correctly assign the polarity.
- **Handling Negation:** Negation words significantly impact the polarity of a sentence. In English sentiment analysis, up to three words following a negation are checked against a lexicon to determine their impact. However, Hindi, which does not follow the Subject-Verb-Object (SVO) order and is a free word order language, requires a different approach. For handling negations in Hindi, we check up to two words preceding and following the negation words to see if they are present in the lexicon. In English, negation words such as "not," "no," "isn't," and "aren't" are typically considered during sentiment analysis. Similarly, in Hindi, negation words like "□□□□" and "न" are taken into account. Additionally, it was observed that sentences involving "कम" follow a similar pattern to those containing "□□□□," prompting the application of the same logic in the analysis.

Dataset Collection and Preprocessing

- **Collecting Data:** Collect a mobile review domain-specific dataset of Hindi text, including social media posts, reviews, and IIT Patna mobile review dataset, to ensure the model's applicability to specific domains.
- **Text Preprocessing:** Perform standard preprocessing steps, removal of symbols, characters, numbers, English words, stop words, duplicate words, and sentences.

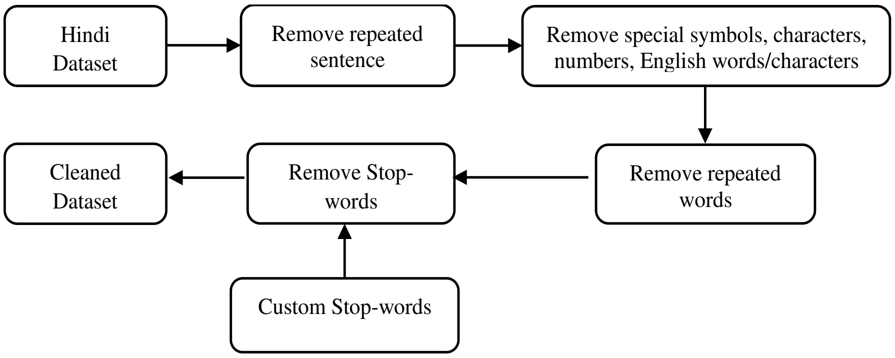

Figure 3. Data Preprocessing

Linguistic Analysis

Apply part-of-speech tagging to identify the grammatical category of each word, Consider the noun (NN) category as an aspect term extraction because in the mobile domain reviews maximum aspect is considered like a noun aiding in the recognition of potential aspect terms.

Table 1: Dataset with aspects

Final_Text	Aspect
□□□□□□ □□□ □□□□□□ □□□ □□□□□ □□□□□ और□□□□□ □□ □□□ □□□□ □□□□	□□□□□/ □□□□□
यह □□□□□□ □□□ एक□□□□□□ □□□□□□□□□□ □□ और□□□ □□□ □□□ □□ □□□□□□□□	□□□□□□□□□न
□□□□ □□□□ □□□□□ □□□□ □□□□ औरयह एक□□□□ □□□ □□□□□ □□ □□□ □□□ □□	□□□□□
□□□ यह □□□ □□□□ □□□ □□□ □□□□□ □□□□ □□ □□□□□□□ और □□□□□□□□ □□□ □□□□□□	□□□□□

FEATURE ENGINEERING

- **Lexicon-based Features:** Lexicon features refer to the predefined lists of words or phrases that are associated with specific sentiment polarities, such as positive, negative, and neutral. These lexicons serve as a reference to determine the sentiment expressed in a sentence. In this research a customized lexicon is tailored to the specific context of the domain like mobile phone reviews in Hindi, the lexicon would include words frequently used to describe mobile phones, categorized by their sentiment polarity. The rule-based algorithm was developed to identify the polarity of sentences using the customized lexicon-based approach. Proposed algorithms of lexicon-based polarity identification are:

Step 1: *Load positive, negative, and neutral lexicons containing words with their respective polarities.*

Step 2: *Tokenize the input sentence into words.*

Step 3: *Initialize counters for positive, negative, and neutral words.*

Step 4: *Repeat steps 5 to 12 until reach the last word of the sentence*

Step 5: *Check if the word is in the positive lexicon. If yes, increment the positive counter.*

Step 6: *Check if the word is in the negative lexicon. If yes, increment the negative counter.*

Step 7: *Check if the word is in the neutral lexicon. If yes, increment the neutral counter.*

Step 8: *Initialize a flag or counter to detect consecutive negative words.*

Step 9: *If the positive counter is highest, classify the sentence as positive.*

Step 10: *If the negative counter is highest, classify the sentence as negative.*

Step 11: *If the neutral counter is highest or all counters are equal, classify the sentence as neutral.*

Step 12: *If the next word is also in the negative lexicon, set the polarity to positive due to the double negation rule.*

- **Domain-Specific Features:** Integrate domain-specific features, such as keywords or terms relevant to the specific mobile domain like battery, screen, camera, etc., to enhance the model's ability to extract contextually important aspect terms.

Model Selection and Training

- **BERT model:** The BERT-multilingual model is a variant of the BERT (Bidirectional Encoder Representations from Transformers) model, specifically designed to support multiple languages. making it versatile for various Natural Language Processing (NLP) tasks across different languages. Some of the languages include English, Hindi, Spanish, Chinese, Arabic, and many more. Being a cased model means that it distinguishes between uppercase and lowercase characters. This feature is important for languages and contexts where the case of the letter carries meaning.
- **Performance Metrics:** Evaluate the model's performance using standard metrics such as precision, recall, and F1 score. Tailor the evaluation criteria to emphasize the importance of correctly identifying aspect terms.
- **Fine-Tuning and Optimization:** Fine-tuning model hyperparameters based on evaluation results is a critical step in optimizing the performance of machine learning models. This process involves adjusting various hyperparameters, which are external configurations set before training begins, to enhance the model's accuracy, efficiency, and overall effectiveness (Srivastav et al. 2023). Table 2 displays the specific values of hyperparameters employed during the training phase across the datasets.

Table 2: Hyperparameters Details

Dataset	Max Sequence Length	Batch Size	Learning rate	Training Epochs
Hindi: Mobile review	128	64	5×10^{-5}	7

By following this comprehensive methodology, the proposed aspect term extraction technique for Hindi text aims to achieve high accuracy, robustness, and applicability across diverse linguistic contexts. Continuous refinement and adaptation based on user feedback and emerging linguistic patterns contribute to the model's effectiveness in capturing nuanced sentiments in Hindi. Figure 4 shows the complete working of the proposed model.

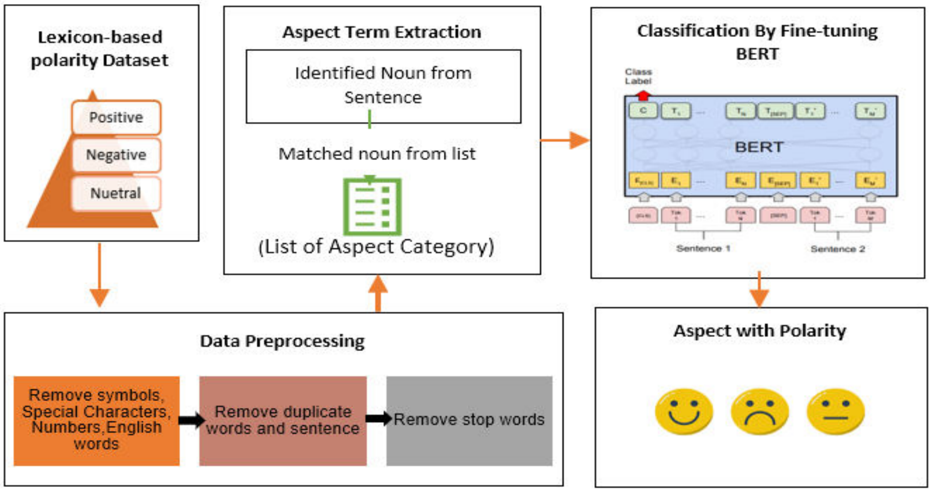
Lexicon-based polarity Dataset
Positive
Negative
Nuetral
Aspect Term Extraction
Identified Noun from Sentence
Matched noun from list
(List of Aspect Category)
Classification By Fine-tuning BERT
Class Label
BERT
Sentence 1
Sentence 2
Data Preprocessing
Remove symbols, Special Characters, Numbers,English words
Remove duplicate words and sentence
Remove stop words
Aspect with Polarity

Figure 4. Overall Process of Proposed Model

RESULTS

The proposed hybrid model combines a rules-based approach for aspect extraction with the use of a pre-trained multilingual BERT model for text classification. Implemented on a 50K text size mobile review dataset in Hindi, the model achieved an impressive accuracy score is 86.97%, F1 score is 87.00%, recall is 86.97% and precision is 87.12% for aspect-based sentiment analysis.

The novelty of this research lies in its multiple approach, first to set the polarity of sentences using a lexicon rule-based approach then POS tagger for aspect extraction, and the pre-trained multilingual fine-tuning BERT model for text classification. Table 3 shows the comparative different model results by authors. As an outcome, the model is capable of determining the aspect with sentiment for any given mobile review sentence, showcasing its effectiveness in understanding and analyzing sentiment towards various aspects of mobile devices. Test the model with real reviews, then display the output as follows. Figure 5 shows the sample output of the proposed model and Figure 6 shows the confusion matrix.

```
# Predict aspects and sentiments from sentences
new_sentences = ["यह □□□□ □□ □□□□□ □□□□ □□□□□ □□□"]
Predicted aspect: □□□□□
Predicted sentiment: POSITIVE
```

Figure 5. Aspect with a Sentiment

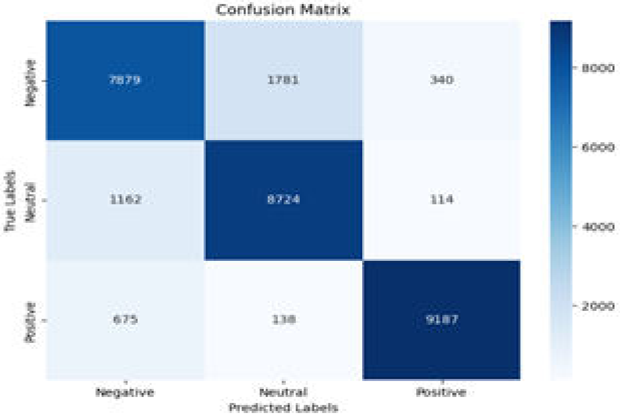

Figure 6. Confusion matrix of Mobile reviews

CONCLUSION

This research has presented a dedicated Aspect Term Extraction Technique tailored for the intricacies of the Hindi language. The methodology integrates linguistic analysis, feature engineering, and deep learning to achieve a robust and accurate model for identifying aspect terms in diverse Hindi textual data. The linguistic analysis component acknowledges the morphological complexities of Hindi. Part-of-speech tagging further refines the linguistic analysis, aiding in the identification of potential aspect terms.

The BERT model selection involves a thoughtful choice of algorithms, considering the characteristics of the Hindi language. Training the model on annotated data ensures that it generalizes well across diverse textual domains. The evaluation process, emphasizing precision, recall, and F1 score, provides a comprehensive assessment of the model's performance. Through iterations of fine-tuning, and optimization, the proposed methodology continually refines the model. The incorporation of contextual information, such as context window analysis and negation handling, enhances the model's adaptability to real-world scenarios. By bridging the gap in ABSA analysis for Hindi, this research contributes to the broader landscape of natural language processing, fostering a deeper understanding of user opinions in one of the world's most widely spoken languages.

REFERENCE

1) Akhtar, M.S., Ekbal, A., Bhattacharyya, P (2018). Aspect-Based Sentiment Analysis: Category Detection and Sentiment Classification for Hindi. In: Gelbukh, A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing Lecture Notes in Computer Science, vol 9624. Springer, Cham. https://doi.org/10.1007/978-3-319-75487-1.

2) H. Liu, I. Chatterjee, M. Zhou, X. S. Lu, and A. Abusorrah (2020). Aspect-Based Sentiment Analysis: A Survey of Deep Learning Methods, in IEEE Transactions on Computational Social Systems, vol. 7, no. 6, pp. 1358-1375, doi: 10.1109/TCSS.2020.3033302.

3) Akhtar, M., Garg, T., & Ekbal, A. (2020). Multi-task learning for aspect term extraction and aspect sentiment classification. Neurocomputing, 398, 247-256. https://doi.org/10.1016/j.neucom.2020.02.093.

4) Kumar, A., Sharan, A. (2020). Deep Learning-Based Frameworks for Aspect-Based Sentiment Analysis Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-15-1216-2_6.

5) Md Shad Akhtar, Tarun Garg, Asif Ekbal (2020). Multi-task learning for aspect term extraction and aspect sentiment classification, Neurocomputing, Volume 398, Pages 247-256, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2020.02.093.

6) Hetal Gandhi, Vahida Attar, Extracting Aspect Terms using CRF and Bi-LSTM Models, Procedia Computer Science, Volume 167,2020, Pages 2486-2495, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.03.301.

7) V. K. Soni and S. Selot (2021). A Comprehensive Study for the Hindi Language to Implement Supervised Text Classification Techniques, 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC), Solan, India, pp. 539-544, doi: 10.1109/ISPCC53510.2021.9609401.

8) Pathak, A.; Kumar, S.; Roy, P.P.; Kim, B.-G (2021). Aspect-Based Sentiment Analysis in Hindi Language by Ensembling Pre-Trained mBERT Models. Electronics, 10, 2641. https://doi.org/10.3390/electronics10212641.

9) Sai Aparna, T., Simran, K., Premjith, B., Soman, K.P. (2021). Aspect-Based Sentiment Analysis in Hindi: Comparison of Machine/Deep Learning Algorithms. In: Smys, S., Balas, V.E., Kamel, K.A., Lafata, P. (eds) Inventive Computation and Information Technologies. Lecture Notes in Networks and Systems, vol 173. Springer, Singapore. https://doi.org/10.1007/978-981-33-4305-4_7.

10) Kumar, A., Dahiya, V., Sharan, A. (2021). ACP: A Deep Learning Approach for Aspect-category Sentiment Polarity Detection. In: Bhattacharyya, D., Thirupathi Rao, N. (eds) Machine Intelligence and Soft Computing. Advances in Intelligent Systems and Computing, vol 1280. Springer, Singapore. https://doi.org/10.1007/978-981-15-9516-5_14.

11) M.P. Geetha, D. Karthika Renuka (2021). Improving the performance of aspect-based sentiment analysis using fine-tuned Bert Base Uncased model, International Journal of Intelligent Networks, Volume 2, Pages 64-69, ISSN 2666-6030, https://doi.org/10.1016/j.ijin.2021.06.005.

12) V. Yadav, P. Verma and V. Katiyar (2021). "E-Commerce Product Reviews Using Aspect Based Hindi Sentiment Analysis," 2021 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, pp. 1-8, doi: 10.1109/ICCCI50826.2021.9402365.

13) Rani, S., & Kumar, P. (2021). Aspect-based Sentiment Analysis using Dependency Parsing. Transactions on Asian and Low-Resource Language Information Processing, 21, 1 - 19. https://doi.org/10.1145/3485243.

14) K. M. Kavitha, A. Nishmitha, G. K. Balgopal, K. K. Naik and M. G. Gaonkar (2022). "Aspect-based Sentiment Analysis of English and Hindi Opinionated Social Media Texts," 21st IEEE International Conference on Machine Learning and Applications (ICMLA), Nassau, Bahamas, pp. 1498-1503, doi: 10.1109/ICMLA55696.2022.00235.

15) V. K. Soni and D. Srivastava (2022). The Use of Supervised Text Classification Techniques: A Comprehensive Study,2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, pp. 422-425, doi: 10.1109/ICACITE53722.2022.9823416.

16) M. Priadarsini and J. Akilandeswari(2022).Recent Approaches on Aspect Based Sentiment Analysis incorporating Deep Learning techniques,13th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, pp. 1-7, doi: 10.1109/ICCCNT54827.2022.9984451.

17) Trisna , Komang Wahyu, Jie ,Huang Jin (2022). Deep Learning Approach for Aspect-Based Sentiment Classification: A Comparative Review, doi: 10.1080/08839514.2021.2014186.

18) Soni, Vijay Kumar, and Smita Selot (2022). A Survey of Deep Learning Techniques in the Field of Sentiment Analysis for the Hindi Language." i-Manager's Journal on Computer Science 10.1.

19) Sujata Rani and Parteek Kumar (2022). Aspect-based Sentiment Analysis using Dependency Parsing. ACM Trans. Asian Low-Resour. Lang. Inf. Process. 21, 3, Article 56, pages. https://doi.org/10.1145/3485243.

20) Nayak, A., & Agrawal, R. K. (2023). Aspect-based sentiment analysis of Odia hotel reviews using LSTM neural network. International Journal of Advanced Intelligence Paradigms, 15(1/2), 111-122.

21) K. Zarandi, S. Mirzaei and H. Talebi, "Aspect-base Sentiment Analysis with Dual Contrastive Learning," 2023 9th International Conference on Web Research (ICWR), Tehran, Iran, Islamic Republic of, 2023, pp. 276-282, doi: 10.1109/ICWR57742.2023.10139077.

22) Srivastav G., Kant, S., & Srivastava, D. (2023). An Efficient Sentiment Analysis Technique for Virtual Learning Environments using Deep Learning model and Fine-Tuned EdBERT. International Journal of Intelligent Systems and Applications in Engineering, 11(5s), 468–476. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/2808.

23) Kaur, G., Sharma, A (2023). A deep learning-based model using a hybrid feature extraction approach for consumer sentiment analysis. J Big Data 10, https://doi.org/10.1186/s40537-022-00680-6.