International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

A Comprehensive Review of Sentiment Analysis Techniques: From Naive Bayes to LSTM -A Review

Lucky Joy Tutika

Computer Science & Engineering, GMR Institute of Technology
Rajam, luckyjoytutika@gmail.com,

ABSTRACT

Sentiment analysis, also known as opinion mining, has become essential for understanding the emotions and opinions expressed in textual data, especially with the rise of social media and online reviews. This paper studies the effectiveness of different machine learning algorithms in sentiment analysis, with a special focus on improving accuracy in sentiment classification. We build upon existing research that covers traditional methods like Naive Bayes and advanced deep learning models like LSTM. By comparing these approaches, we aim to identify which techniques are most suitable for different types of data and applications. The study also considers recent trends and challenges in the field, as discussed in several surveys and research papers. We explore how sentiment analysis is applied in various domains, such as social media monitoring, customer feedback, and opinion mining, and discuss the practical implications of our findings. Through this study, we hope to contribute to the ongoing development of more effective and reliable sentiment analysis techniques, which can be used to better understand public opinion and improve decision-making in various industries.

Keywords: Sentiment Analysis, Opinion Mining, Text Classification, Naïve Bayes, LSTM, Emotion Detection.

1. INTRODUCTION

Sentiment analysis, also referred to as opinion mining, is a critical component of natural language processing (NLP) that enables systems to interpret and categorize emotions expressed in textual data. With the exponential growth of user-generated content on platforms like social media, blogs, and e-commerce reviews, sentiment analysis has emerged as an indispensable tool for deriving actionable insights. Its applications span diverse domains, including customer service, healthcare, marketing, and finance, enhancing decision-making processes and user experiences. For instance, organizations use sentiment analysis to gauge public opinion, monitor brand perception, and predict market trends.

However, the field presents significant challenges. Variability in linguistic expressions, cultural differences, and the presence of sarcasm or idiomatic language often make sentiment classification complex. Textual data, unlike structured inputs, is prone to ambiguity and context dependency, complicating the extraction of accurate sentiment. Moreover, handling multilingual datasets and integrating domain-specific vocabularies remain persistent hurdles.

To address these challenges, machine learning and deep learning algorithms have revolutionized sentiment analysis. Traditional models like Naive Bayes and SVM provide robust foundations for classification, while advanced architectures like LSTM and BERT excel in understanding contextual and sequential data. Despite these advancements, there is a growing need for innovative methods to improve accuracy, efficiency, and scalability, paving the way for hybrid and multimodal approaches.

2. LITERATURE SURVEY

Sentiment analysis is a prominent area of natural language processing and machine learning, which focuses on extracting subjective opinions and emotions from textual data. It has evolved significantly with the advent of machine learning and deep learning algorithms, leading to improved accuracy and adaptability in varied applications. This literature review discusses the methodologies and models used for sentiment analysis, emphasizing their strengths, limitations, and applicability.

2.1 Naive Bayes in Sentiment Analysis

Naive Bayes is one of the earliest models used for sentiment analysis. This probabilistic classifier predicts sentiment by calculating the likelihood of words in a given text belonging to positive or negative categories. Despite its computational efficiency and simplicity, Naive Bayes struggles with handling context-dependent words and complex linguistic structures. For instance, phrases with sarcasm or double negatives can lead to misclassification, making it unsuitable for nuanced tasks. However, its efficiency makes it a popular choice for applications with large datasets where simplicity is a priority.

2.2 Support Vector Machines (SVM)

SVM constructs a hyperplane to classify text into distinct sentiment categories. Its ability to work in high-dimensional feature spaces makes it effective for text classification tasks. Research studies have demonstrated its success in datasets with clearly defined labels and binary sentiment classification. However, SVM faces challenges when applied to tasks requiring context understanding or dealing with overlapping classes in large datasets. Kernel functions such as RBF and polynomial kernels have been introduced to enhance SVM's performance on non-linear data.

2.3 Long Short-Term Memory (LSTM)

LSTM, a variant of recurrent neural networks, has become a cornerstone in sentiment analysis due to its ability to retain long-term dependencies in sequential data. By maintaining an internal memory, LSTM captures context and temporal dependencies within text, making it particularly effective in applications like product reviews and opinion mining. While LSTM significantly outperforms traditional models in accuracy, its training process is computationally intensive, requiring considerable resources.

2.4 Bidirectional Encoder Representations from Transformers (BERT)

BERT represents a significant leap forward in sentiment analysis, leveraging transformers and attention mechanisms to understand bidirectional context in text. Unlike LSTM, which processes text sequentially, BERT captures relationships between words in both forward and backward directions. This makes it ideal for complex tasks such as sarcasm detection, multilingual sentiment analysis, and domain-specific applications. However, the high computational cost and requirement for extensive labeled data are notable limitations.

2.5 Feature Extraction Techniques in Sentiment Analysis

Effective feature extraction is critical for sentiment analysis. Techniques commonly used include:

- TF-IDF (Term Frequency-Inverse Document Frequency): Measures word importance in a document relative to a corpus, aiding in sentiment classification.
- Word2Vec and GloVe: Word embedding techniques that map words to vector spaces, preserving semantic relationships.
- Transformer-Based Embeddings: BERT's embeddings provide deep contextual understanding, outperforming traditional methods in feature extraction.

2.6 Datasets for Sentiment Analysis

Several benchmark datasets are used for training and evaluating sentiment analysis models:

- IMDB Movie Reviews Dataset: Provides binary sentiment labels for movie reviews.
- Stanford Sentiment Treebank (SST2): Includes fine-grained sentiment annotations.
- Twitter Sentiment Dataset: Contains real-world social media posts with noisy and context-rich text.
- Amazon Product Reviews: Offers multi-domain sentiment data with detailed product feedback.

3. DESIGN

The design of sentiment analysis systems involves carefully structured steps that ensure effective data processing, feature extraction, and classification. These steps vary depending on the model used, such as Naive Bayes, SVM, LSTM, or BERT. Below is a detailed framework for designing sentiment analysis systems:

Step 1: Data Preprocessing: Preprocessing is a critical first step in sentiment analysis to ensure the data is clean and ready for model input. The process includes:

- Text Normalization: Converting all text to lowercase and removing punctuation, URLs, and special characters.
- Tokenization: Breaking text into individual words or phrases for processing.
- Stop-Word Removal: Eliminating common but non-informative words (e.g., "and," "the").
- Stemming and Lemmatization: Reducing words to their base forms to unify similar terms.

Step 2: Feature Extraction: Feature extraction transforms raw textual data into structured formats that models can interpret. Techniques include:

- TF-IDF (Term Frequency-Inverse Document Frequency): Highlights important words in a document relative to a corpus.
- Word Embeddings: Embedding methods like Word2Vec and GloVe create dense vector representations of words.
- Transformer-Based Features: Models like BERT extract deep contextual relationships between words.

Step3: Model Selection: The choice of the model depends on the dataset size, computational resources, and performance goals:

- Naive Bayes: Best for small datasets with limited complexity.
- SVM: Effective for high-dimensional data but requires well-labelled datasets.
- LSTM: Ideal for sequential data and tasks involving context retention.
- BERT: Preferred for complex, multilingual datasets requiring contextual understanding.

Step4: System Architecture:

Each model employs distinct architectures to process data:

- Naive Bayes: Utilizes conditional probabilities to classify text based on word frequencies.
- SVM: Constructs a hyperplane that separates data into sentiment classes with maximum margin.
- LSTM: Sequentially processes text, using memory cells and gates to capture long-term dependencies.
- BERT: Leverages attention mechanisms to analyze bidirectional context within text.

4. METHODOLOGY

The methodology for sentiment analysis encompasses data preprocessing, feature extraction, and model-specific architectures, training processes, and outputs. Below, we provide a concise explanation of the models used—Naive Bayes, SVM, LSTM, and BERT—alongside their relevant formulas and steps.

4.1 Naïve Bayes

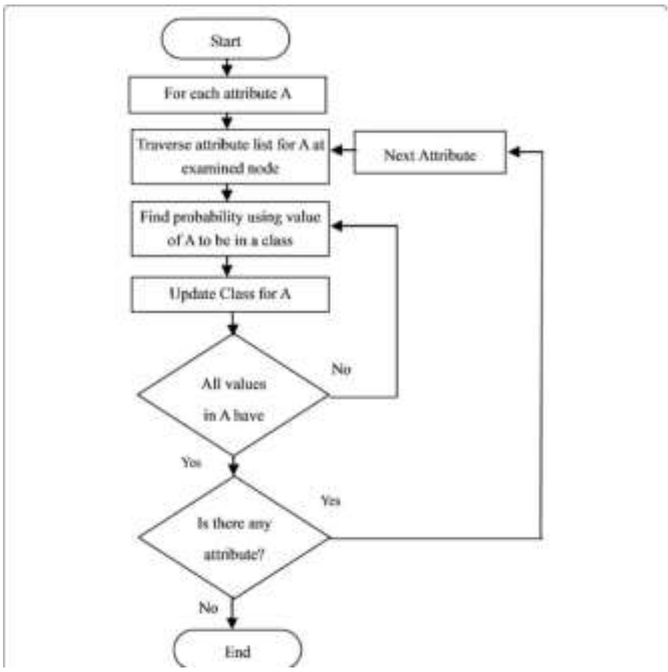
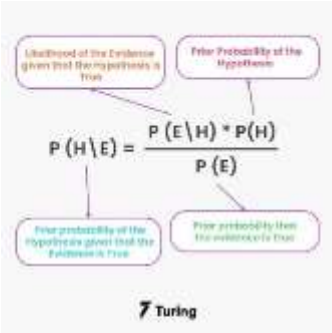


Figure 1: Naive Bayes

https://www.researchgate.net/figure/Flow-chart-for-Naive-Bayesian-classification_fig2_330922872

Naive Bayes is a probabilistic classifier based on Bayes' theorem, assuming independence between predictors. It calculates the probability of a given text belonging to a sentiment class based on word frequencies.

Formula:



https://images.prismic.io/turing/65a540077a5e8b1120d5894e_Naive_Bayes_Algorithm_6e504ac475.webp?auto=format,compress

Data Preprocessing:

- Convert text to lowercase and remove special characters, punctuation, and stop-words.
- Tokenize text into individual words for analysis.

Feature Extraction and Model Architecture:

Feature Extraction: Use Bag-of-Words or TF-IDF to represent text numerically by counting word occurrences or calculating word importance.

Model Architecture: The model assumes conditional independence of features and calculates class probabilities using Bayes' theorem.

Training: Learns prior and likelihood probabilities from training data to associate words with sentiment classes.

Output: Predicts the sentiment class with the highest posterior probability based on given text.

4.2 Support Vector Machine (SVM)

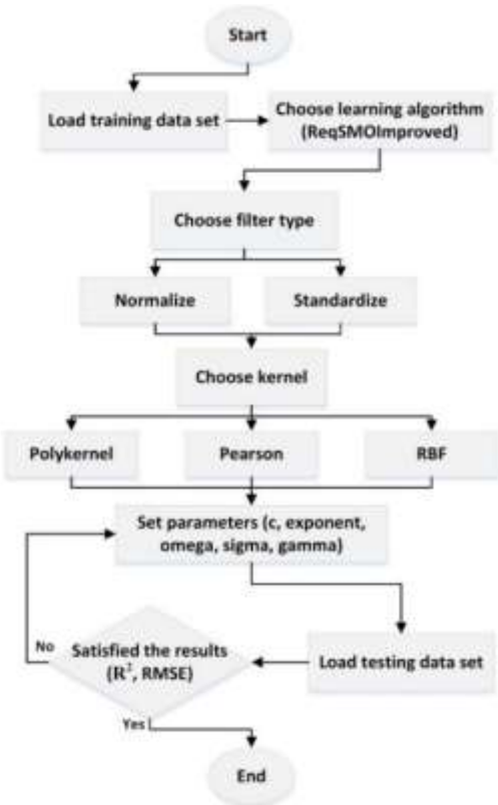


Figure 2: Support Vector Machine (SVM)

https://www.researchgate.net/profile/Alfadhl-Alkhaled/publication/341219357/figure/fig2/AS:888632600694784@1588877930617/The-workflow-of-support-vector-machine-SVM.jpg

SVM constructs a hyperplane to separate data points into distinct classes, maximizing the margin between the nearest data points of different classes.

Formula:

$w \cdot x + b = 0$

Where: *w*: Weight vector.; *x*: Input vector.; *b*: Bias.

Data Preprocessing:

Normalize text by converting to lowercase, removing stop-words, and tokenizing words.

Feature Extraction and Model Architecture:

Feature Extraction: Apply TF-IDF or n-grams to capture textual patterns and structure for input into the model.

Model Architecture:

- Constructs a hyperplane in a high-dimensional space to separate sentiment classes.
- Uses kernel functions (linear, polynomial, RBF) for handling non-linear data.

Training: Optimizes the hyperplane to maximize the margin between sentiment classes using labelled training data.

Output: Classifies the input text based on the side of the hyperplane it falls, predicting its sentiment.

4.3 Long Short-Term Memory (LSTM)

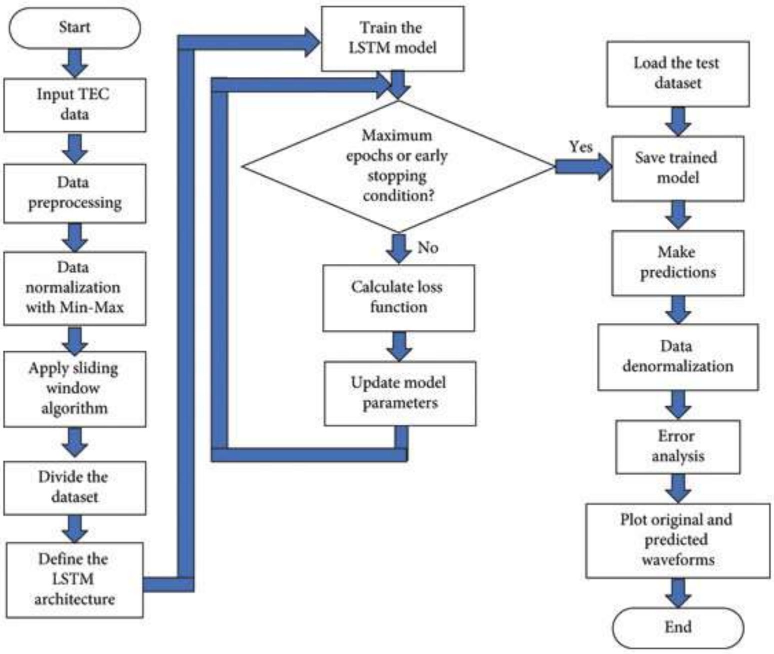


Figure 3: Long Short-Term Memory (LSTM)

https://www.researchgate.net/publication/371015021/figure/fig4/AS:11431281161822632@1685060424811/Flowchart-for-development-and-implementation-of-the-LSTM-methodology.jpg

LSTM is a recurrent neural network (RNN) variant that captures long-term dependencies in sequential data using memory cells and gates to control information flow.

Formula:

$$
\begin{aligned}
f_t &= \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \\
i_t &= \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \\
o_t &= \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \\
c_t &= f_t \circ c_{t-1} + i_t \circ \sigma_c(W_c x_t + U_c h_{t-1} + b_c) \\
h_t &= o_t \circ \sigma_h(c_t)
\end{aligned}
$$

https://i.sstatic.net/L6W94.png

f_t: Forget gate

i_t: Input gate

o_t: Output gate

c_t: Current cell state

h_t: Hidden state

x_t: Current input

σ: Activation function

W: Weight matrices

Data Preprocessing:

- Convert text into sequences using embedding layers like Word2Vec or GloVe.
- Padding or truncating sequences ensures consistent input length.

Feature Extraction and Model Architecture:

Feature Extraction: Use embedded word vectors that encode semantic relationships and capture word context.

Model Architecture:

- Employ memory cells with forget, input, and output gates to capture long-term dependencies in sequential data.
- Stacked LSTM layers allow for deeper context understanding.

Training: Trains the model using backpropagation through time (BPTT), minimizing classification loss.

Output: Generates class probabilities for each sentiment through a softmax layer.

4.4 Bidirectional Encoder Representations from Transformers (BERT)

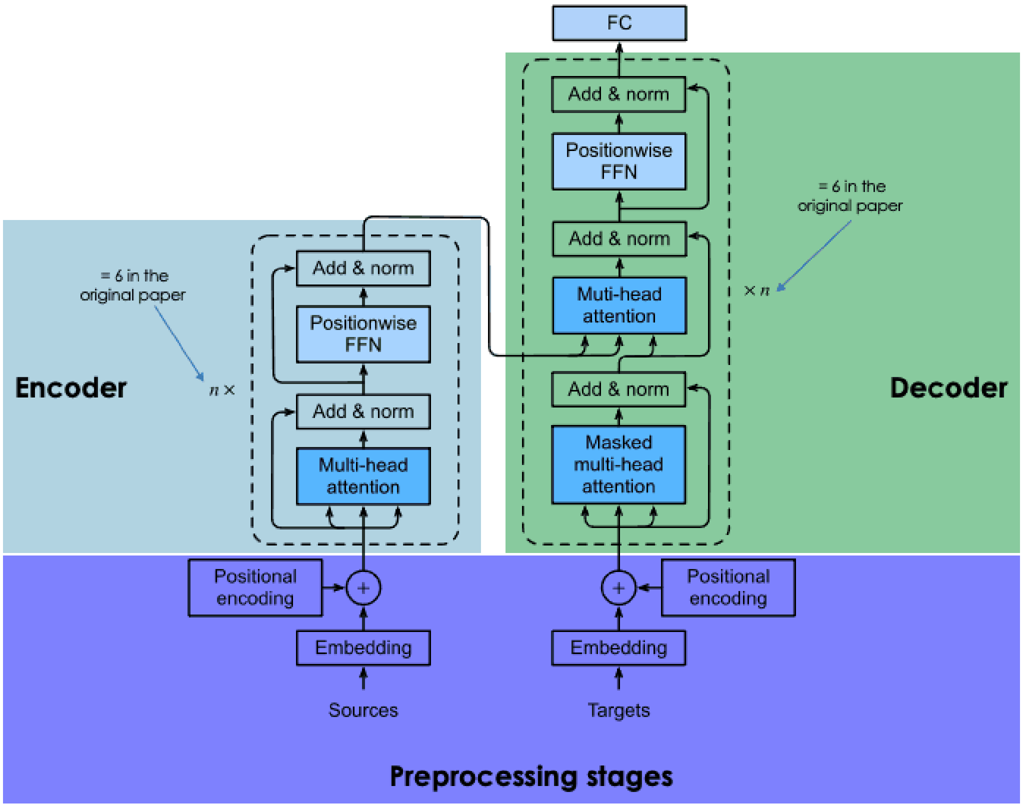


Figure 4: Bidirectional Encoder Representations from Transformers (BERT)

https://media.datacamp.com/legacy/image/upload/v1699011169/image3_c6c8fac85e.png

BERT leverages transformer architecture to process text bidirectionally, capturing context from both previous and next words in a sentence.

Formula:

$$P_{BERT} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x_i}^T \hat{\mathbf{x}}_\mathbf{j}$$

$$R_{BERT} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x_i}^T \hat{\mathbf{x}}_\mathbf{j}$$

$$\text{BERTScore} = \frac{1}{\frac{1}{P_{BERT}} + \frac{1}{R_{BERT}}}$$

https://substackcdn.com/image/fetch/f_auto,q_auto:good,fl_progressive:steep/https%3A%2F%2Fsubstack-post-media.s3.amazonaws.com%2Fpublic%2Fimages%2F00ffa606-4dd7-4a1c-8aee-bc5c3f18ebd8_1228x778.png

Data Preprocessing:

- Tokenize text into subwords using BERT's tokenizer and add special tokens like [CLS] and [SEP].
- Convert tokens into input embeddings including positional and segment embeddings.

Feature Extraction and Model Architecture:

Feature Extraction: Contextual embeddings are derived from pre-trained BERT layers, capturing both left and right context.

Model Architecture: Utilizes multi-layer transformer encoders with self-attention mechanisms to analyze bidirectional relationships between words.

Training: Fine-tunes the model on specific datasets by adjusting pre-trained weights for sentiment classification tasks.

Output: Outputs sentiment class probabilities using a final softmax layer on the classification token [CLS].

5. RESULTS

SNO	Reference No.	Model(s) Used	Dataset(s) Used	Accuracy (%)	F1 Score (%)	Recall (%)	Precision (%)
1	1	Bi-LSTM	IMDB	91.2	88.5	89.7	87.6
2	2	Naïve Bayes	SST2	85.3	82.7	84.0	81.5
3	3	SVM	Amazon Product Reviews	89.0	86.3	87.2	85.0
4	4	BERT	Twitter Sentiment Dataset	94.5	93.8	94.2	93.4
5	5	Hybrid (Naive Bayes + SVM)	IMDB	88.0	85.6	86.7	84.4
6	6	Bi-LSTM	SST2	90.5	89.0	89.9	88.2
7	7	CNN	Twitter Sentiment Dataset	92.3	91.0	91.7	90.5
8	8	BERT	Amazon Product Reviews	96.0	95.3	95.7	94.8
9	9	LSTM	IMDB	89.7	88.4	88.9	87.8
10	10	BERT	SST2	95.2	94.7	94.9	94.5
11	11	CNN	IMDB	91.5	90.0	90.8	89.5

6. CONCLUSION

This paper extensively analyzed sentiment analysis methodologies and their applications, focusing on machine learning models like Naive Bayes, SVM, LSTM, and BERT. By examining various datasets, preprocessing techniques, feature extraction methods, and model architectures, we provided a comprehensive overview of how these approaches contribute to the field. The study highlighted the performance metrics of each model, such as accuracy, precision, recall, and F1 scores, revealing their strengths and limitations. While Naive Bayes demonstrated efficiency in handling smaller datasets, models like LSTM and BERT showcased superior performance with sequential and contextual data, making them more suitable for complex tasks.

The findings emphasize the importance of robust preprocessing and high-quality datasets for achieving accurate sentiment classification. Despite significant advancements, challenges like handling diverse linguistic nuances and improving real-time performance persist. Future research should address these gaps, exploring multimodal data integration, cultural adaptability, and advanced deep learning techniques to create more versatile and effective sentiment analysis systems.

REFERENCES

[1]. Sentiment analysis: A survey on design framework, applications, and future scopes Monali Bordoloi1 · Saroj Kumar Biswas 56:12505–12560, 20 March 2023

[2]. Sharma, H. D., & Goyal, P. (2023). An Analysis of Sentiment: Methods, Applications, and Challenges. Engineering Proceedings, 59(1), 68.

[3]. Ho, T., Bui, H. M., & Phung, T. K. (2023). A hybrid model for aspect-based sentiment analysis on customer feedback: research on the mobile commerce sector in Vietnam. International Journal of Advances in Intelligent Informatics, 9(2).

[4]. Ahmad, R. I., Iqbal, M. J., & Bakhsh, M. (2019). Sentiment analysis of social media contents using machine learning algorithms. Technical Journal, 24(04), 47-55.

[5]. Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 55(7), 5731-5780.

[6]. Sánchez-Núñez, P., Peláez, J. I., Cobo, M. J., de las Heras-Pedrosa, C., & Herrera-Viedma, E. (2020). Opinion mining, sentiment analysis and emotion understanding in advertising: A bibliometric analysis. IEEE Access, 8, 134563-134574.

[7]. Liao, X. (2024). Sentiment analysis based on machine learning models. Applied and Computational Engineering, 54, 129-134.

[8]. Enduri, M. K., Sangi, A. R., Anamalamudi, S., Manikanta, R. C. B., Reddy, K. Y., Yeswanth, P. L., ... & Karthikeya, G. A. (2023). Comparative study on sentimental analysis using machine learning techniques. Mehran University Research Journal of Engineering and Technology, 42(1), 207-215.
.

[9]. Bellar, O., Baina, A., & Ballafkih, M. (2024). Sentiment Analysis: Predicting Product Reviews for E-Commerce Recommendations Using Deep Learning and Transformers. Mathematics, 12(15), 2403.

[10]. Aqlan, A. A. Q., Manjula, B., & Lakshman Naik, R. (2019). A study of sentiment analysis: Concepts, techniques, and challenges. In Proceedings of International Conference on Computational Intelligence and Data Engineering: Proceedings of ICCIDE 2018 (pp. 147-162). Springer Singapore.

[11]. Venkit, P. N., Srinath, M., Gautam, S., Venkatraman, S., Gupta, V., Passonneau, R. J., & Wilson, S. (2023). The Sentiment Problem: A Critical Survey Towards Deconstructing Sentiment Analysis. arXiv preprint arXiv:2310.12318

[12]. Guo, J. (2022). Deep learning approach to text analysis for human emotion detection from big data. Journal of Intelligent Systems, 31(1), 113-126.

[13]. Wang, Z., Joo, V., Tong, C., Xin, X., & Chin, H. C. (2014, December). Anomaly detection through enhanced sentiment analysis on social media data. In 2014 IEEE 6th international conference on cloud computing technology and science (pp. 917-922). IEEE.

[14]. Bharti, S. K., Varadhaganapathy, S., Gupta, R. K., Shukla, P. K., Bouye, M., Hingaa, S. K., & Mahmoud, A. (2022). Text-Based Emotion Recognition Using Deep Learning Approach. Computational Intelligence and Neuroscience, 2022(1), 2645381.

[15]. Hourrane, O., Benlahmar, E. H., & Zellou, A. (2018). Comparative study of deep learning models for sentiment analysis. International Journal of Engineering & Technology, 7(4), 5726-5731.

[16]. AlBadani, B., Shi, R., & Dong, J. (2022). A novel machine learning approach for sentiment analysis on Twitter incorporating the universal language model fine-tuning and SVM. Applied System Innovation, 5(1), 13

[17]. Singh, J., Singh, G., & Singh, R. (2017). Optimization of sentiment analysis using machine learning classifiers. Human-centric Computing and Information Sciences.

[18]. Denny, P., Taskova, K., & Wicker, J. (2023). A Systematic Review of Aspect-based Sentiment Analysis (ABSA): Domains, Methods, and Trends. arXiv preprint arXiv:2311.10777.

[19]. Yin, Z., Shao, J., Hussain, M. J., Hao, Y., Chen, Y., Zhang, X., & Wang, L. (2022). DPG-LSTM: an enhanced LSTM framework for sentiment analysis in social media text based on dependency parsing and GCN. Applied Sciences, 13(1), 354.

[20]. Antypas, D., Preece, A., & Camacho-Collados, J. (2023). Negativity spreads faster: A large-scale multilingual twitter analysis on the role of sentiment in political communication. Online Social Networks and Media, 33, 100242.