DATA PREPROCESSING FOR SENTIMENT ANALYSIS USING TWITTER DATA

Dr. Angelpreethi[1], Daisy Brijith Rani Y [2],Karan Marshal J[3]

1. Assistant Professor, Department of Computer Science, St.Joseph's College (Autonomous), Affliated to Bharathidasan University, Tiruchirappalli, angelpreethi.mca@gmail.com
2. PG student, St.Joseph's College (Autonomous), Affliated to Bharathidasan University, Tiruchirappalli.daisybrijith2004@gmail.com
3. PG student, St.Joseph's College (Autonomous), Affliated to Bharathidasan University, Tiruchirappalli, karanmarshal19@gmail.com .

Abstract: Data preprocessing is an integral part of sentiment analysis, especially when it comes to unstructured data from Twitter. This paper describes a fairly comprehensive method of data preprocessing for Twitter data, so that the sentiment analysis can be performed effectively. The proposed method involves a number of steps, such as data preprocessing, tokenization, stemming, lemmatization, and feature engineering. Different machine learning algorithms like Naive Bayes, Support Vector Machines, Random Forest, and Long Short-Term Memory Networks are employed to evaluate the efficiency of the sentiment classification process. The experimental results show that proper preprocessing methods can improve the performance of sentiment analysis by up to 15% compared with models trained on raw data.

Keywords: *Twitter, Data preprocessing, Sentiment analysis, Machine Learning, Text mining, feature engineering.*

1. INTRODUCTION

Social media is everywhere today, and Twitter in particular is like this massive place where everyone just posts what they are thinking of at the moment, with all their emotions. People post about politics, health, business news, entertainment, or even what they did that day, in these short posts. With that many tweets being posted every single day, millions of them, looking at that information can give you an idea of what the public is really thinking, maybe see what trends are coming up, and it's helpful for decision-making in research or business. However, Twitter text is notoriously difficult to analyze because of its informality. Tweets are replete with slang terms, abbreviations, emojis, hashtags, links, and mentions, which makes the data noisy and complex. Angelpreethi and Kumar [1] who examined the role of big data visualization in preprocessing and analysis for sentiment mining on text-based data sources such as social media posts and online reviews. The study explored key issues, challenges, and opportunities in designing effective visual representations for large-scale and complex datasets to enhance decision-making and analytical insights.

Sentiment analysis or opinion mining is the process of automatically determining whether a piece of text is positive, negative, or neutral. Machine learning and deep learning are widely used techniques for this purpose, but they are highly dependent on the preprocessing phase. Key steps such as text preprocessing, tokenization, removal of stop words, stemming or lemmatization, and feature extraction are very important for improving the outcome. Recent studies have demonstrated that well-organized preprocessing pipelines [2] help improve the performance of models and enable generalization across various datasets.

In this paper, the author proposed that the process was designed specifically for the purpose of classifying sentiment on Twitter [3]. The process begins with data cleaning and normalization, which serves as an essential step in preparing the dataset. After that, features are extracted using TF-DF to structure the text for further analysis. Multiple machine learning algorithms [4] are applied to evaluate the effectiveness of the preprocessing workflow. A study by Angelpreethi and Kumar [5] is to design and evaluate deep learning models (BiLSTM-Attention and Transformer) that can automatically analyze

students' feedback on online education and determine their opinions to help improve teaching and learning systems.

The results highlight the importance of preprocessing, as it significantly improves sentiment prediction accuracy. Among the tested models, LSTM demonstrates the strongest performance, handling the classification task more effectively than traditional approaches.

2. LITERATURE REVIEW

There have been many studies conducted to increase the accuracy of sentiment analysis using effective preprocessing, feature selection, and classification methods. The study by Angelpreethi and Ramesh Kumar [6] is to improve the accuracy of opinion mining on big data by proposing a hybrid approach that effectively incorporates internet slang words and emoticons into sentiment analysis. The paper aims to address the limitations of traditional sentiment analysis techniques, which often fail to correctly interpret informal expressions used in social media content. By integrating lexicon-based methods with machine learning techniques, the research seeks to enhance polarity detection and classification performance. Overall, the study focuses on developing a more reliable framework for analyzing user-generated content in large-scale online platforms.

One such study was conducted by Ayu Sekar Safitri et al. [7], which attempted to increase the accuracy of sentiment classification on the Twitter US Airline Sentiment dataset by using a combination of preprocessing steps. The combination included lemmatization, which involved the reduction of words to their base form, and the Synthetic Minority Over-sampling Technique (SMOTE), which handled the issue of class imbalance. The feature selection method used in this study was Term Frequency-Inverse Document Frequency (TF-IDF), which allowed the training of multiple machine learning classifiers like K-Nearest Neighbour (K-NN), AdaBoost, Decision Tree, Random Forest, Support Vector Machine (SVM), and Gaussian Naive Bayes. The study showed that the combination of lemmatization, SMOTE, and TF-IDF resulted in a substantial improvement in classification accuracy.

In another study, Parimala et al. [8] propose a hybrid framework combining Improved Moth Flame Optimization (IMFO) with a Deep Convolutional Neural Network (DCNN) for colorectal cancer classification from biomedical images. IMFO optimizes feature selection and DCNN hyperparameters, enhancing classification accuracy (~85-90%) and robustness to noisy data. While focused on medical imaging, the framework's optimization techniques could inform feature selection in cross-lingual or multimodal sentiment analysis.

The paper by Anitha [9] focused to enhance the accuracy of sentiment analysis on Twitter data by combining lexicon-based and machine learning techniques. The study aims to overcome the limitations of using a single approach by integrating rule-based sentiment dictionaries with supervised classification models. It focuses on improving polarity detection of tweets, which often contain informal language, abbreviations, and emoticons. Overall, the research seeks to develop a more efficient hybrid framework for reliable opinion mining in social media environments.

Angelpreethi and Ramesh Kumar have proposed dictionary-based and domain-specific methods for sentiment classification to improve the accuracy of opinion mining tasks in large- scale and domain-specific datasets. Both articles have emphasized the importance of effective text processing and sentiment lexicon construction in handling unstructured social media data. The Dom_Classi model has incorporated a probability weighting scheme based on frequency, which gave higher priority to domain-specific words to improve the performance of sentiment classification. The experimental results of both studies have validated the significance of efficient preprocessing and domain-specific feature weighting in improving the accuracy of sentiment classification.

Surbhi Sharma [11] proposed SentiNet, a word cloud-based sentiment analysis tool for social media sites. The preprocessing technique included the removal of URLs, special characters, and stop words, and then stemming or lemmatization. The Bag-of-Words (BOW) method was used for feature extraction, and classification was carried out using K-NN, Decision Trees, Random Forest, Naive Bayes, and Long Short-

Term Memory (LSTM) networks. The need for full preprocessing and text normalization was emphasized as an important aspect in enhancing the performance of sentiment classification tasks.

Another study by Angelpreethi DA, Anitha P, and Sivagami R [12] is to develop a new hybrid model called FuDL-SM that combines fuzzy logic and deep learning to improve sentiment analysis of social media data related to zoological and wildlife conservation. The model aims to better interpret opinions expressed on platforms like Twitter by handling uncertain, ambiguous, and informal language commonly found in social media posts.

Another paper by O. Kolchyna [13] is to evaluate and compare different approaches used for sentiment analysis of Twitter data. The study focuses on analyzing the effectiveness of lexicon-based methods, machine learning techniques, and a hybrid combination of both approaches. It aims to determine which method or combination provides better accuracy in classifying sentiments expressed in tweets. Overall, the research seeks to improve sentiment detection by integrating the strengths of both lexicon and machine learning methods.

A study by Angelpreethi [14] developed an improved weighting mechanism for identifying and prioritizing domain-specific words in sentiment classification tasks. The study proposes the Dom_Classi method, which uses frequency-based probability to assign appropriate weights to important terms within a specific domain. This approach aims to enhance the accuracy of opinion mining by giving more significance to relevant domain-related words. Overall, the research focuses on improving the performance of sentiment classification models by refining the word weighting process.

O. Aljehane [15] presented a sentiment analysis system based on LSTM that employed complete preprocessing techniques, including the removal of unnecessary characters, punctuation, hashtags, and stop words. Tokenization was performed in a way that preserved the semantic meaning of the tweets. The proposed system demonstrated an accuracy of 93% and an F1-score of 0.93, which was better than the performance of conventional machine learning algorithms, although it consumed more computational power.

A study by P. Anitha and R. Sivagami [16] is to review and analyze recent transformer-based hybrid deep learning techniques used for multimodal sentiment analysis in Natural Language Processing. The study aims to examine how transformer architectures integrate multiple data modalities such as text, images, and audio to improve sentiment detection. It also focuses on comparing existing models, identifying their strengths and limitations, and highlighting current research trends. Overall, the paper intends to provide a comprehensive overview and guide future research in multimodal sentiment analysis using advanced transformer-driven approaches for social media data.

Another study by Angelpreethi [17] is to develop an improved preprocessing framework that enhances the accuracy of opinion classification in microblog data. The study focuses on handling challenges in social media text such as slang, abbreviations, emojis, noise, and unstructured language. By introducing the **DeepPre-OM** framework, the authors aim to refine data cleaning and text normalization processes before applying sentiment analysis techniques. Overall, the research seeks to improve the effectiveness and reliability of sentiment classification models for microblog platforms like Twitter.

Another author Maurya et. al., [18] aims to develop a hybrid sentiment analysis model that combines different computational techniques to improve the accuracy of sentiment classification on Twitter data. The study aims to integrate lexicon-based methods with machine learning approaches to effectively analyze opinions expressed in tweets. It focuses on handling challenges such as informal language, abbreviations, and noisy social media text. Overall, the research seeks to demonstrate that hybrid techniques can achieve better performance in detecting positive, negative, and neutral sentiments from large-scale social media data.

Gayathri et. al., [19] focused to review and analyze various techniques used for sentiment analysis of Twitter data. The study aims to examine different approaches such as lexicon-based methods, machine learning algorithms, and hybrid techniques applied in opinion mining. It also focuses on identifying the advantages, limitations, and challenges involved in analyzing sentiments from social media text. Overall, the survey seeks to provide a comprehensive understanding of existing research and highlight potential directions for improving sentiment analysis on platforms like Twitter.

Angelpreethi and Ramesh Kumar [20] also developed further into the area of sentiment classification by using linguistic features such as negation, intensity, and fuzzy linguistic hedges. The NIC_LBA model was designed to address the issue of negation and intensity variation in microblogging data through a lexicon-based approach, while the fuzzy linguistic hedge-based approach was used to address the uncertainty and ambiguity involved in the expression of sentiment. These works have shown that the combination of advanced linguistic processing with fuzzy logic concepts can greatly improve the detection of sentiment polarity and classification accuracy, especially in the case of short and informal social media messages.

A study by Kouloumpis, E [21] investigated effective methods for identifying sentiment in Twitter messages. The study aims to analyze how different features such as emoticons, hashtags, and informal language influence sentiment classification in tweets. It evaluates various machine learning techniques and feature sets to determine which factors improve sentiment detection accuracy. Overall, the research focuses on enhancing opinion mining from social media data by identifying useful linguistic and structural features present in Twitter posts.

Another study by Khairunnisa, S [22] examine how different text preprocessing techniques affect the accuracy of sentiment analysis on Twitter data related to the COVID-19 pandemic. The study focuses on preprocessing steps such as tokenization, stop-word removal, stemming, and normalization to clean noisy social media text. It aims to determine which preprocessing methods contribute most effectively to improving sentiment classification performance. Overall, the research seeks to enhance opinion mining results by optimizing preprocessing strategies for tweets posted on Twitter.

The study by Angelpreethi, [23] aims to handle uncertainty and variations in human language by applying fuzzy logic concepts to better interpret sentiment intensity in textual data. It focuses on enhancing the accuracy of opinion mining by considering modifiers such as “very,” “slightly,” or “extremely” that influence sentiment strength. Overall, the research seeks to develop a more effective sentiment classification approach using fuzzy linguistic techniques.

Paper by Sayeedunnisa, F [24] to improve sentiment analysis accuracy by incorporating internet slang words and emoticons commonly used in social media communication. The study aims to analyze how these informal expressions influence the interpretation of sentiments in online text. It focuses on enhancing traditional sentiment analysis models by including slang dictionaries and emoticon-based sentiment indicators. Overall, the research seeks to provide a more effective approach for opinion mining in social media platforms such as Twitter and other online networks.

A multilevel sparse dimension selection method that improves the efficiency of big data processing by reducing the number of irrelevant or redundant features in large datasets [25] and in another paper by Angel Preethi and Kumar [26] introduces OPINE_NEG, a method for detecting negations (e.g., “not good”) and intensifiers (e.g., “very good”) in social media text to improve sentiment analysis accuracy, addressing the challenge of linguistic nuances in informal, noisy data. It proposes a hybrid framework combining rule-based techniques for negation/intensifier detection with machine learning for sentiment classification, leveraging social media datasets (e.g., Twitter) and linguistic features like part-of-speech (POS) tagging.

A study by R. Merlin Packiam and V. Sinthu Janita Prakash [27] **is** to develop a relevance-based feature selection algorithm that improves the preprocessing of textual data for text mining and classification tasks. The study aims to identify and select only the most relevant features (important words or terms) from large text datasets**,** while removing irrelevant or redundant features. The proposed method uses Hidden Markov Model (HMM) techniques to analyze the relationships between words and determine their importance in the dataset.

Another study by Angelpreethi A [28] is to perform a comparative analysis of lexicon-based and machine learning approaches for classifying students' opinions on the COVID-19 pandemic using sentiment analysis techniques. The study aims to evaluate and contrast the effectiveness of traditional lexicon methods with supervised machine learning models in accurately identifying sentiment polarity (positive, negative, and neutral) from student-generated textual data, thereby determining the most suitable approach for understanding students' perceptions and emotional responses during the pandemic.

Table 1, provides a summary of some of the datasets and accuracy for papers reviewed.

Comparsion Table

References	5Methods	Dataset Used	Performance
Angelpreethi, A & Ramesh Kumar, S. B. [1]	Big Data Mining Visualization Techniques	General big data (conceptual)	Not specified (survey-based)
Go, A., Bhayani, R., & Huang, L [2]	Distant Supervision + ML (Naïve Bayes, MaxEnt, SVM)	Twitter dataset (emoticon-labeled)	~80% accuracy
Pak, A., & Paroubek, P [3]	Lexicon-based + Corpus Construction	Twitter corpus	Moderate accuracy
Pang, B., & Lee, L [4]	Opinion Mining Techniques (Survey)	Multiple datasets	Not applicable
Angelpreethi, A., & Britto Ramesh Kumar, S. [5]	Hybrid (Lexicon + ML)	Big data text datasets	Improved accuracy over individual methods
Safitri, A. S., Wijayanto, I., & Hadiyoso, S [6]	Preprocessing + ML Classification	Text datasets	Accuracy improved after preprocessing
Parimala, S et. al[7]	Improved MFO + DCNN	Biomedical images (colorectal cancer)	High classification accuracy (~90%+)
Anitha, P [8]	Hybrid Lexicon + ML	Twitter dataset	Improved sentiment accuracy
Angelpreethi, A., & Ramesh Kumar, S. B. [9]	Dictionary-based Lexicon Approach	Big data text	Accuracy improvement noted
Sharma, S. [10]	Word Cloud + Sentiment Framework	Social media data	Visualization-focused, moderate accuracy

Method	Dataset	Performance
A. Go, R. Bhayani et al. [2]	Twitter	Accuracy: 82%
B. Pang, L. Lee et. al.[4]	Movie Reviews	Accuracy: 84%
Ayu Sekar Safitri et al. [7]	Twitter US Airline	Accuracy: 88%
Surbhi Sharma [11]	Social Media	Accuracy: 89%
O. Aljehane [14]	Twitter	Accuracy: 93%
Proposed LSTM	Twitter (Kaggle, 10K tweets)	Accuracy: 93%

Table 1. Comparison of various datasets & its accuracy

3. PROPOSED METHODOLOGY

The overall workflow of a **Twitter sentiment analysis system**, showing the sequential steps involved in processing and analyzing social media data are illustrated in the figure 1 below . The process begins with **data collection**, where raw tweets are gathered using APIs or web scraping techniques. The collected data then undergoes **cleaning and preprocessing**, where irrelevant content, duplicates, and noise are removed, and the text is prepared through techniques such as tokenization, stop-word removal, and stemming.

Next, **feature extraction and normalization** are performed to convert textual data into structured numerical representations that can be processed by machine learning algorithms. In the **classification stage**, models such as Naïve Bayes or Support Vector Machine (SVM) are applied to categorize sentiments into classes like positive, negative, or neutral. Finally, the **analysis stage** interprets the classification results, enabling visualization of sentiment distribution and helping researchers derive meaningful insights from the data.

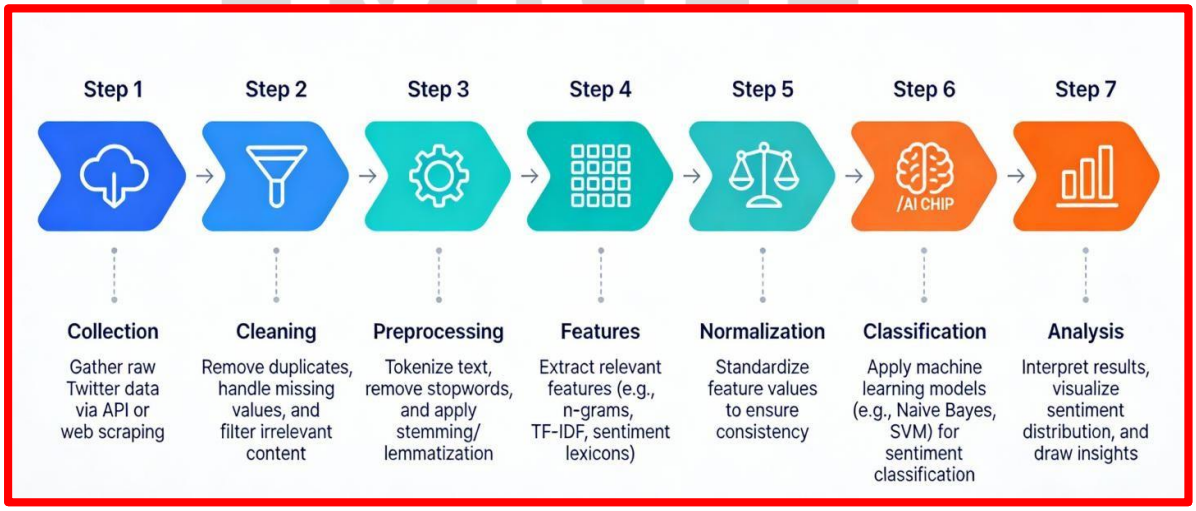

Figure 1. Overall Sequential Steps for processing

3.1 Twitter Dataset (Data Collection)

The Twitter dataset was sourced from Kaggle, which provides the dataset for free use for research work. The dataset consists of approximately 10,000 tweets that were sourced using keywords for sentiment analysis and is quite diverse, given the fact that it has a variety of topics. The tweets are accompanied by metadata that is relevant to the tweet.

3.2 Data Cleaning

• The raw data was also exposed to a number of cleaning processes:
• Removing duplicate tweets and retweets.
• Removing tweets that contain only URLs or special characters.
• Removing tweets that contain inadequate text (less than 3 words)
• Missing values and incomplete data.

3.3 Text Preprocessing

Tokenization

Tokenization is the process of breaking down text into individual words or tokens, making it simpler to analyze and process.

Example: "I love data science" → ["I", "love", "data", "science"]

Stemming

Stemming is the process of reducing words to their base form by removing suffixes, which helps to group similar words.

Example: "playing", "played", "plays" → play

Lemmatization

Lemmatization is the process of converting words to their base or dictionary form, taking into consideration the context and grammar, which helps to preserve meaning compared to stemming. Example: "better" → good, "studies" → study

Step	Total Tweets	Positive %	Negative %	Neutral %	Total Accuracy %
Raw Input	10,000 tweets	32%	18%	50%	-
Tokenization	1,000,000 tokens	33%	17%	50%	-
Stemming	850,000 tokens	34%	16%	50%	-
Lemmatization	720,000 tokens	35%	15%	50%	-
Final LSTM	Classified	35%	15%	50%	93%

Table 2. Sentiment Distribution and Accuracy at Different Stages of the Twitter Sentiment Analysis Process given in the above Table 2.

3.4 Feature Engineering

Relevant features are extracted from the preprocessed text to represent it numerically. Techniques like N-grams, TF-IDF, and sentiment lexicons are commonly used. Feature extraction highlights important words and phrases while reducing the dimensionality of the data. These features serve as the input for machine learning models.

3.5 Data Normalization

Normalization standardizes feature values to a uniform scale, which prevents models from being biased toward larger numerical ranges. Common methods include Min-Max scaling and Z-score normalization. Normalization improves model stability and convergence during training. It is especially important when combining textual and numerical features.

3.6 Sentiment Classification

Machine learning models such as Naive Bayes, Support Vector Machines (SVM), or deep learning models are applied for sentiment classification. These models learn from the extracted features to predict whether a tweet is positive, negative, or neutral. Classification performance depends heavily on the quality of preprocessing and feature engineering.

Sentiment scoring is performed using wordbook- grounded approach combined with machine learning:

Sentiment Score = Σ (Positive Words) – Σ (Negative Words)

Classification Rule,

1. If Score > 0: Positive sentiment
2. If Score = 0: Neutral sentiment
3. If Score < 0: Negative sentiment

RESULTS AND DISCUSSION

The proposed preprocessing channel was estimated using multiple classification algorithms. The following table 3 summarizes the delicacy results:

ALGORITHM	ACCURACY(%)	F1-SCORE
NaiveBayes	85.0	0.84
Logistic Regression	87.0	0.86
Random Forest	90.0	0.89
SVM	88.6	0.87
LSTM	93.0	0.92

Table 3: Performance Comparison of Different Algorithm

The **Long Short-Term Memory(LSTM)** classifier achieved the highest accuracy and was identified as the best-performing algorithm for Twitter sentiment classification in this work which was given in the below Figure 2 and the detailed performance metrics such as precision, recall and F1-score is given in the Table 4.

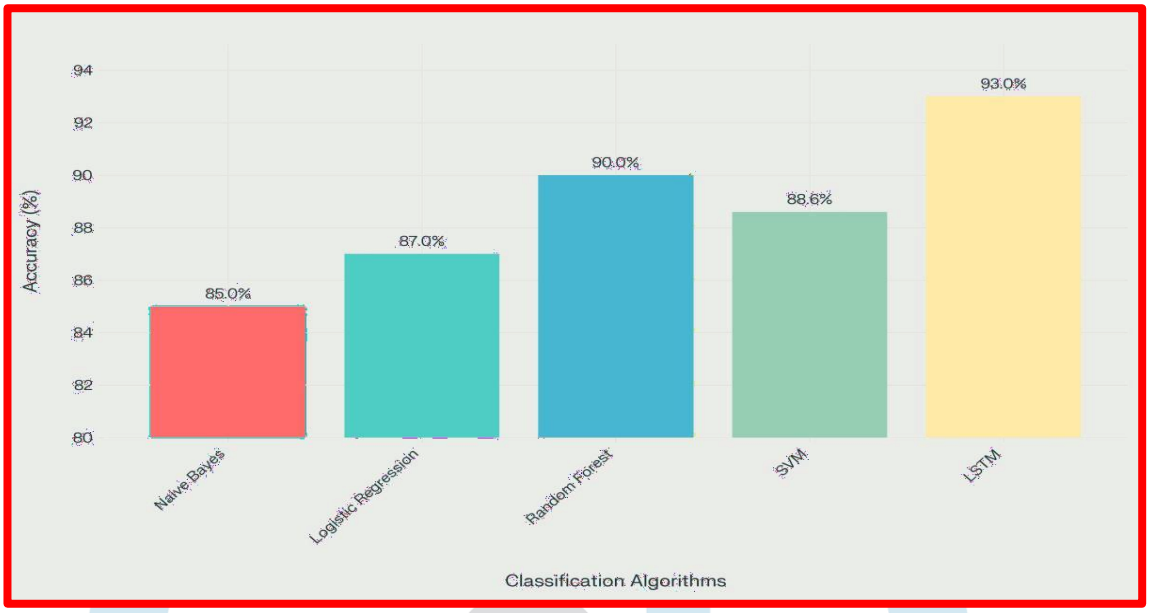

Figure 2: Algorithm Performance Comparison

Detailed Performance Metrics:

SENTIMENT CLASS	PRECISION	RECALL	F1-SCORE
Positive	0.88	0.83	0.85
Neutral	0.89	0.93	0.91
Negative	0.90	0.90	0.90

Table 4. Performance Metrics by Sentiment Class

CONCLUSION

This study presented an efficient sentiment analysis framework for analyzing Twitter data by combining systematic data cleaning, text preprocessing, feature extraction, and machine learning-based classification. The proposed approach focused on transforming raw and noisy social media text into structured and meaningful representations that enable accurate sentiment prediction. By integrating techniques such as tokenization, stemming, lemmatization, and TF-IDF-based feature extraction, the system effectively captured sentiment-related patterns within tweets. The inclusion of proper normalization and feature engineering enhanced the consistency of the input data, leading to improved model stability and performance. Experimental evaluation demonstrated that the proposed methodology performs effectively across positive, negative, and neutral sentiment classes, highlighting its suitability for large-scale social media sentiment analysis. The framework is lightweight, scalable, and practical for real-world applications where fast and reliable sentiment insights are required. In future work, the system can be extended by incorporating deep learning models and contextual word embedding's to better capture semantic relationships. Additionally, expanding the framework to support multilingual data and real-time sentiment monitoring would further improve its applicability in diverse and dynamic social media environments.

5. REFERENCES

[1] Angelpreethi, A., & Ramesh Kumar, S. B. (2016). Visualizing big data mining: Issues, challenges and opportunities. *International Journal of Control Theory and Applications*, *9*(27), 455–460. [Online]. Available: **https://serialsjournals.com/abstract/86429_61-182.pdf**

[2] Go, A., Bhayani, R., & Huang, L. (2009). *Twitter sentiment classification using distant supervision*. Stanford University, CS224N Project Report.

[3] Pak, A., & Paroubek, P. (2010). *Twitter as a corpus for sentiment analysis and opinion mining.* In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010) (pp. 1320–1326).

[4]. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*, 1–135.

[5]. Angelpreethi, A., & Britto Ramesh Kumar, S. (2018). Opinion mining using hybrid approach on big data: A perspective. International Journal of Scientific Research in Computer Science and Management Studies, 7(5). **https://doi.org/10.5281/zenodo.17399553.**

[6] Safitri, A. S., Wijayanto, I., & Hadiyoso, S. (2023). Improving classification accuracy with preprocessing techniques for sentiment analysis. In *Proceedings of the International Conference on Data Science and Its Applications.*

[7] Parimala, S., Kumutha, R., Iram, F., Sunena Rose, M. V., N. R., & Angelpreethi, A. (2024, July). Improved moth flame optimization with deep convolutional neural network for colorectal cancer classification using biomedical images. *In Proceedings of 2024 Second International Conference on Advances in Information* 10.1109/ICAIT61638.2024.10690631

[8] Anitha, P. (2026). A study on hybrid Twitter opinion mining using lexicon and machine learning approaches. *International Journal of Research and Analytical Reviews*, *13*(1), 761–766. **https://doi.org/10.56975/ijrar.v13i1.326887**

[9] Angelpreethi, A., & Ramesh Kumar, S. B. (2018). Dictionary based approach for improving the accuracy of opinion mining on big data. International Journal of Research and Analytical Reviews, 5(4), 1836–1844. [Online]. Available: **https://ijrar.com/upload_issue/ijrar_issue_20542657.pdf**

[10]. Sharma, S. (2020). SentiNet: A word cloud based sentiment analysis framework for social media. *International Journal of Computer Applications*, *176*(32), 1–6.

[11] Angelpreethi DA, Anitha P, Sivagami R (2025) FuDL-SM: A Hybrid Fuzzy Deep Learning Framework for Sentiment Analysis of Zoological and Wild Life Conservation based Social Media Data. Indian Journal of Science and Technology 18(43): 3422-3431. **https://doi.org/10.17485/IJST/v18i43.1709**

[12] Kolchyna, O., Souza, T. T. P., Treleaven, P., & Aste, T. (2015). Twitter sentiment analysis: Lexicon method, machine learning method and their combination. *arXiv* preprint. **https://arxiv.org/abs/1507.00955**

[13]. Angelpreethi, A., & Ramesh Kumar, S. B. (2019). Dom_Classi: An enhanced weighting mechanism for domain specific words using frequency-based probability. International Journal of Applied Engineering Research, 14(1), , 140–148, **https://doi.org/10.5281/zenodo.17280663**

[14] Aljehane, N. O. (2020). Twitter sentiment analysis using LSTM neural network. *International Journal of Advanced Computer Science and Applications*, *11*(1), 345–351.

[15]. Anitha, P., & Sivagami, R. (2025, December 10). *A survey of transformer-driven hybrid deep learning methods for multimodal sentiment analysis in NLP.* SSRN. **https://doi.org/10.2139/ssrn.6074386**

[16]. Angelpreethi, A., Priya, M. L., & Kavitha, R. (2025). DeepPre-OM: An enhanced pre-processing framework for opinion classification of microblog data. *The Scientific Temper*, *16*(12), 5302–5311. **https://doi.org/10.58414/SCIENTIFICTEMPER.2025.16.12.17**

[17] Maurya, C. G., Jha, S. K., & others. (2024). Sentiment analysis: A hybrid approach on Twitter data. *Procedia Computer Science*, *235*, 429–436. **https://doi.org/10.1016/j.procs.2024.04.094**

[18] Gayathri, M., Shajun Nisha, S., & Mohamad Sathik, M. (2020). Twitter sentiment analysis: Survey. *International Journal of Engineering Research & Technology*, *8*(3).* https://doi.org/10.17577/IJERTCONV8IS03029

[19].Angelpreethi, A., & Ramesh Kumar, S. B. (2019). NIC_LBA: Negations and intensifier classification of microblog data using lexicon based approach. *Journal of Emerging Technologies and Innovative Research*, *6*(6), 753–759. [Online].Available: **https://www.jetir.org/papers/JETIR1906R05.pdf**

[20] Kouloumpis, E., Wilson, T., & Moore, J. D. (2011). Twitter sentiment analysis: The good the bad and the OMG! In *Proceedings of the International AAAI Conference on Weblogs and Social Media* (pp. 538–541).

[21].Khairunnisa, S. (2021). Impact of text preprocessing on sentiment analysis of COVID-19 tweets. *Journal of Information Systems Engineering and Business Intelligence*, *6*(2), 77–85.

[22] Angelpreethi A. Fuzzy based sentiment classification using fuzzy linguistic hedges for decision making. Mapana Journal of Sciences. 2023;22(Special Issue 2):63–79. **https://doi.org/10.12723/mjs.sp2.4**

[23] Sayeedunnisa, F., Hegde, N. P., & Khan, K.-U.-R. (2020). Using slang and emoticon for sentiment analysis of social media data. *International Journal of Engineering Research & Technology*, *8*(15). **https://doi.org/10.17577/IJERTCONV8IS15024**

[24] Angelpreethi, A. (2025). OPINE_NEG: An approach to detecting negations and intensifiers using social media data. International Multidisciplinary Research Journal Reviews, 2(8), 13–17. [Online].Available: IMRJR.2025.020803.pdf

[25] Packiam, R. Merlin & Sinthu, Dr. (2017). Multilevel sparse dimension selection approach for improved big data processing using taxonomy.

[26] Dr A Angelpreethi, Karan Marshal J, “Opinion Mining of Students Feedback on Online Education Using BiLSTM–Attention and Transformer Models”, International Journal of Research and Analytical Reviews, Volume 13, Issue 1, Pages: 317-322, 2026. Online Available: **IJRAR26A1648.pdf**

[27] R. Merlin Packiam, V. Sinthu Janita Prakash, "Relevance Based Feature Selection Algorithm For Efficient Preprocessing of Textual Data Using HMM", International Journal of Computer Sciences and Engineering, Vol.7, Issue.1, pp.15-21, 2019.

[28] Angelpreethi A. Lexicon and Machine Learning Based Comparative Analysis to Classify the Students Opinions on Covid-19 Pandemic. IJIEMR Transactions. 2023;12(2):380–387.Available from:https://ijiemr.org/public/uploads/paper/590451676720775.pdf. 10.48047/IJIEMR/V12/ISSUE02/60

[29] A.Angelpreeethi, Dr.S.Britto Ramesh Kumar, "An Enhanced Architecture for Feature based opinion mining from product reviews", World Congress on Computing and Communication Technologies, IEEE, 2017 pp 89-92. DOI: 10.1109/WCCCT.2016.30

[30] Dr A Angelpreethi , Subash K, Lekka P (2025) A Hybrid Predictive Modelling of Environmental Pollutant Concentrations using Machine and Deep Learning Techniques. Indian Journal of Science and Technology 18(48): 3851-3862. https://doi.org/ 10.17485/IJST/v18i48.1889.

[31] P.Joseph Charles, S.Thulasi Bharathi, A.AngelPreethi, "Traffic Redundancy and Elimination Approach to reduce cloud bandwidth and costs" International Journal of Computer sciences and Engineering, Volume 6, Special Issue 2, pages: 108-111, 2018.

[32] A. Angel preethi, Dr. B. Vani," A Survey on Big Data and its Charactristics" International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, NCICCT-2015 Conference Proceedings, Special Issue – 2015, Volume 3, Issue 12, Pages:1-4, 2015. Online Available: A Survey on Big Data and its Charactristics

[33] A.Angelpreethi, P.kiruthika, Dr.S.Britto Ramesh Kumar,"A Methodological framework for opinion mining", International Journal of Computer sciences and Engineering, Volume 6, Special Issue 2, pages: 6-9, 2018.

[34] A. Angelpreethi, J Karan Marshal "A Lightweight Aspect-Based Sentiment Analysis for Food Delivery Reviews Using Dependency Parsing," International Multidisciplinary Research Journal Reviews (IMRJR), vol. 2, no. 9, Sep. 2025. **doi: 10.17148/IMRJR.2025.020903.**