IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 1, January 2025

DOI: 10.17148/IJARCCE.2025.14113

Sentiment Analysis of Tweets Containing Fuzzy Sentiment

Harika R[1], Jahnavi Baddikuri[2], J Jhansi Sri Saranya [3], Joshna M[4],

Dr. Sudhakar Avareddy[5]

Department of Computer Science and Engineering, Ballari Institute of Technology & Management, India[1-5]

Abstract: Sentiment analysis involves determining the polarity of statements—classifying them as positive, neutral, negative, satisfied, or frustrated—based on opinions, feedback, and comments shared on social media or e-commerce platforms. It is a key area in text mining for analyzing and classifying user sentiments. This work focuses on identifying the sentiments of tweets and retweets on Reddit by searching for specific keywords and evaluating their polarity as positive or negative. Sentiment analysis is performed using feature selection for scoring words, with classification algorithms employed for training, testing, and sentiment evaluation. Performance metrics such as accuracy, precision, and time are compared across ANN, CNN, and LSTM models.

Keywords: Fuzzy sentiment, tweet sentiment analysis, sentiment polarity

1.INTRODUCTION

Emotions play a vital role in expressing human feelings. Sentiment analysis, or opinion mining, studies people's emotions, attitudes, and opinions using natural language processing and machine learning. It is a critical area in text mining, widely applied for decision-making by extracting and analysing online reviews. Social media platforms like Facebook, Reddit, and Instagram generate vast amounts of user opinions, which can serve as valuable business intelligence for applications like marketing and prediction. Sentiment analysis identifies the sentiment or attitude of an author on a specific topic, requiring efficient algorithms to process diverse opinions. Tasks include emotion recognition and polarity detection, applied to areas such as product reviews, political analysis, and entertainment. This work focuses on extracting opinions from Reddit using specific keywords and classifying tweet sentiments using the Naive Bayes Classifier (NBC). Tweets are pre-processed, trained, and analysed with NBC, which assumes feature independence. Naive Bayes is simple, efficient for large datasets, and often outperforms more complex methods. This approach provides a summary of opinions from Reddit, classifying sentiments as positive, neutral, or negative, making it a foundational method for text categorization.

The main objectives of this project are:

1. To design a robust architecture for sentiment analysis on tweets, dividing functionalities between the Admin module and the User module for efficient management and real-time analysis.
2. To implement machine learning classifiers in the Admin module for training and comparing algorithms based on accuracy scores.
3. To select the best-performing machine learning algorithm to ensure accurate sentiment classification of real-time tweets in the User module.
4. To facilitate the comparison of multiple machine learning algorithms for evaluating their performance in terms of accuracy and reliability.
5. To deploy the optimal machine learning model in the User module for real-time sentiment analysis of tweets.
6. To provide a user-friendly interface in the User module for seamless sentiment classification of tweets.
7. To enhance the efficiency of sentiment analysis by leveraging feature ensemble techniques for better model performance.

2.MOTIVATION

The exponential growth of social media platforms like Twitter, which generate a vast amount of user content daily. Tweets often encapsulate public opinions, emotions, and trends on a wide range of topics, from social and political issues to customer feedback. However, the informal language, brevity, and mixed sentiments commonly found in tweets make accurate sentiment analysis a complex task. This project seeks to address these challenges by leveraging advanced machine learning techniques and a feature ensemble approach to improve the accuracy and reliability of sentiment analysis. By employing a dual-module architecture—one for training and testing algorithms in the Admin module and

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 1, January 2025
DOI: 10.17148/IJARCCE.2025.14113

another for real-time sentiment analysis in the User module—the system ensures optimal performance in processing real-world data. The ultimate goal is to provide actionable insights that help businesses, organizations, and researchers make informed decisions based on public sentiment, while also demonstrating the practical application of cutting-edge machine learning techniques in sentiment analysis.

3.LITERATURE REVIEW

Sentiment analysis has been a subject of extensive research in recent years, with various machine learning algorithms and techniques being employed to enhance accuracy and efficiency. Below is an overview of relevant studies that form the foundation for this project:

1. **Littérateur survey on: Sentiment analysis algorithms and applications: A survey W. Medhat, A. Hassan, and H. Korashy**
   They provided a comprehensive overview of sentiment analysis (SA), exploring recent algorithmic enhancements and applications. The paper categorizes contributions to SA techniques and discusses related fields like transfer learning and emotion detection, offering insights into the latest trends and advancements**.**
2. **Littérateur survey on: Affective computing and sentiment analysis," IEEE Intel-ligent Systems E. Cambria,**
   Cambria highlights the significance of emotion processing in AI and polarity detection. The survey explores how affective computing and sentiment analysis analyze public sentiment across domains, leveraging human-computer interaction and multimodal signal processing for insights from social data
3. **Littérateur survey on: "An ensemble sentiment classification system of reddit data for airline services analysis Y. Wan and Q. Gao**
   Wan and Gao apply an ensemble classification strategy on Reddit data using multiple classifiers, including Naive Bayes, SVM, and Random Forest. Their approach, validated with 10-fold evaluation, outperforms individual classifiers, demonstrating improved accuracy for airline service sentiment analysis.
4. **Littérateur survey on: Comparison research on text pre- processing methods on reddit sentiment analysis Z. Jianqiang and G. Xiaolin**
   Jianqiang and Xiaolin examine the impact of pre-processing methods like expanding acronyms and replacing negations on classifier performance. Results indicate improved accuracy and F1-measure for Naive Bayes and Random Forest, with minimal effects from removing URLs, numbers, or stop words.
5. **Littérateur survey on: Energy-efficient wireless sensor networks using learning techniques S. Tokle, S. R. Bellipady, R. Ranjan, and S. Varma**
   Tokle et al. propose using Self-Organizing Maps (SOM) to reduce multimedia data in Wireless Multimedia Sensor Networks (WMSNs). Their design reduces energy consumption and increases network lifetime by 158.03% compared to JPEG.

4.METHODOLOGY

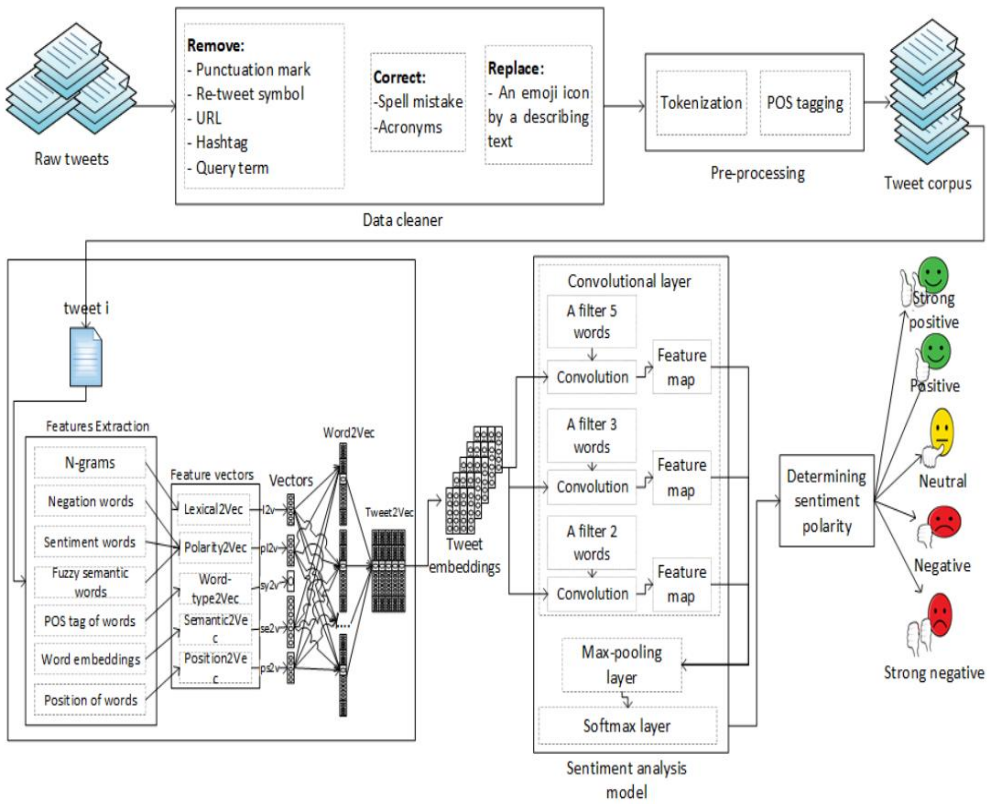
Building a machine learning system for sentiment analysis of fuzzy tweets involves multiple well-defined steps, each crucial to the success of the project. Below is a detailed description of the methodology followed in this project:

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 1, January 2025

DOI: 10.17148/IJARCCE.2025.14113



1.Algorithms used in Project

ANN

ANN is a set of connected input-output networks in which weight is associated with each connection. It consists of one input layer, one or more intermediate layers, and one output layer. Learning of neural networks is performed by adjusting the weight of the connection. By updating the weight iteratively performance of the network is improved. During training, the interconnection weights are optimized until the network reaches the specified level of accuracy. It has many advantages like less affected by noise and good learning ability.

CNN

CNN is one of the deep learning techniques that is used for the classification and detection of objects in mages. Here CNN will work differently where instead of neurons being connected to every neuron in the previous layer, they are instead only connected to neurons close to it and all have the same weight. The CNN is made up of multiple layers that make it unique like the convolutional layer, pooling layer, rectified linear unit layer (ReLU), and a fully connected layer. Here the convolutional layer will extract the feature map and ReLU layers act as an activation function for reducing the dataset dimensionality and the fully connected layer will perform the classification on the training dataset. So, in this system, the CNN algorithm is used for resolving the classification problems.

LSTM

LSTMs have been widely used in various applications, including machine translation, sentiment analysis, speech recognition, and generating captions for images. They have demonstrated superior performance compared to traditional RNNs in many sequence modeling tasks, especially when dealing with long sequences or data with long-range dependencies. Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture designed to overcome the vanishing gradient problem that can occur when training traditional RNNs. LSTMs are particularly effective for sequence prediction tasks, such as time series forecasting, natural language processing, and speech recognition. The key innovation of LSTM networks is the introduction of memory cells, which allow the network to remember information over long sequences. These memory cells have a gating mechanism composed of three gates: the input gate, the forget gate, and the output gate. These gates regulate the flow of information into and out of the cell, allowing the LSTM to selectively remember or forget information as needed during training.

Admin Module

In this admin module we train and test the dataset using three DL classifiers for accuracy calculation comparison.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 1, January 2025

DOI: 10.17148/IJARCCE.2025.14113

Upload Dataset

For the Sentiment analysis we have taken the dataset which consists of 156062 rows of labeled data. In this we have words and statement in one column and in another column we have label class of sentiment polarity, means positive, negative and neutral.

Stop words Removal

This step under process of the feature extraction, in every review statement of the training dataset we identified the stop words and remove it. In the API which we are using we have a list which consists of list of English stopwrods like 'a', 'about', 'according', 'across' etc. We can remove stop words because of the these words useless for any judgment of the text classification, and we can achieve the database space and computation cost.

Apply Classification Algorithms

In this sub module after conversion of the dataset into the TfidfVectorizer we apply the classification algorithms of the following.

- CNN
- ANN
- LSTM

5.PROPOSED SYSTEM

The proposed system is designed to perform sentiment analysis on social media platforms, particularly focusing on extracting opinions from Reddit data. The architecture is divided into two primary modules: the **Admin Module** and the **User Module**, each addressing specific functionalities to ensure efficient processing and analysis of sentiment data. The following sections describe the proposed system in detail:

1. System Overview

The system processes large volumes of textual data to classify sentiments into categories like positive, neutral, and negative. It utilizes machine learning algorithms for training and testing, while leveraging deep learning (DL) techniques such as ANN, CNN, and LSTM for enhanced classification accuracy.

2. Admin Module

The Admin Module is responsible for training the system and ensuring the best-performing model is deployed for real-time sentiment analysis.

Key Features:

1. **Dataset Upload:**
   - Input is a dataset containing labeled data with sentiment categories (positive, negative, neutral, satisfied, frustrated).
   - Admin uploads datasets for processing.
2. **Data Pre-processing:**
   - Stop-word Removal: Eliminates words that do not contribute to sentiment classification.
   - TF-IDF Vectorization: Converts text data into numerical format using Term Frequency-Inverse Document Frequency for better machine learning compatibility.
3. **Model Training:**
   - Utilizes ANN, CNN, and LSTM classifiers.
   - The models are trained on the pre-processed dataset to learn sentiment classification.
4. **Model Testing:**
   - Conducts testing to compare performance metrics like accuracy, precision, recall, and F1-score across all classifiers.
   - The best-performing model is selected based on these metrics.

3. User Module

The User Module provides a user-friendly interface for real-time sentiment analysis.

Key Features:

1. **Real-Time Analysis:**
   - Fetches live tweets or Reddit posts based on user-provided keywords.
   - Applies the deployed machine learning model for sentiment classification.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 1, January 2025

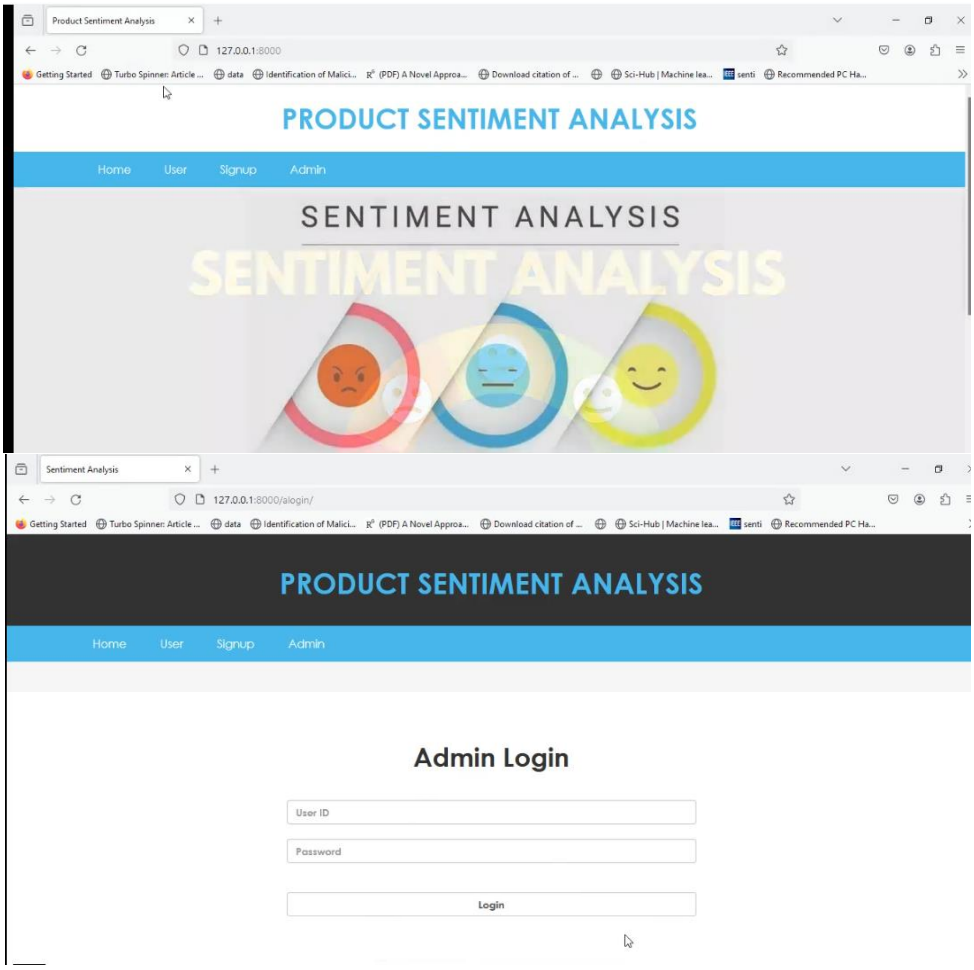
DOI: 10.17148/IJARCCE.2025.14113

2. **Sentiment Output:**
    - Classifies the sentiment of each post into positive, negative, or neutral categories.
    - Displays sentiment trends, offering insights into public opinions.
3. **User Interaction:**
    - Allows users to input specific keywords for targeted sentiment analysis.
    - Generates visual outputs like graphs for sentiment distribution.

4. System Workflow

1. **Data Collection:**
    - Extract posts from Reddit using APIs, filtered by specific keywords.
2. **Pre-processing:**
    - Normalize text (e.g., handle acronyms, remove special characters).
3. **Feature Extraction:**
    - Vectorize the cleaned text for machine learning compatibility.
4. **Model Training and Testing:**
    - Train ANN, CNN, and LSTM models.
    - Evaluate performance and deploy the best model.
5. **Real-Time Prediction:**
    - Input live text data into the deployed model for classification.
6. **Results Presentation:**
    - Sentiment results are presented to users via the graphical interface.

6.RESULT

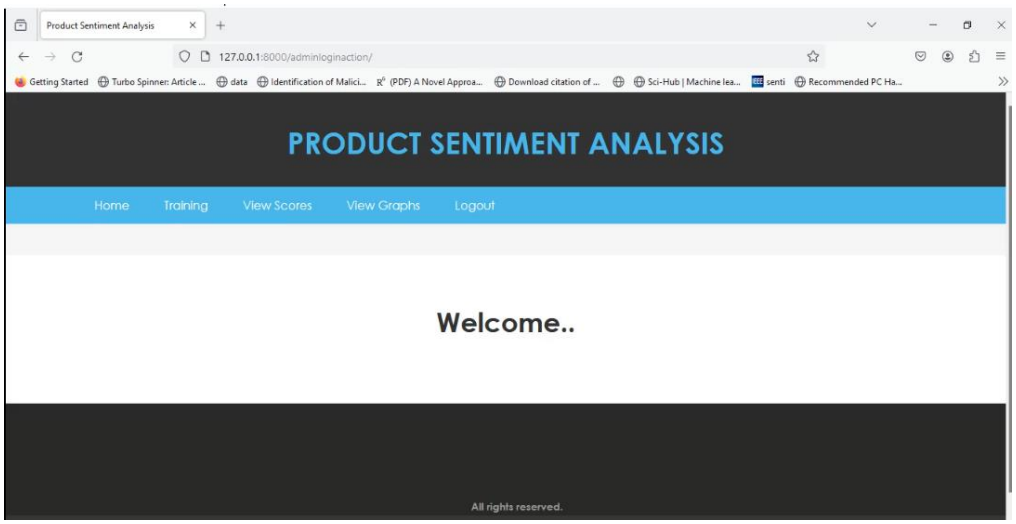


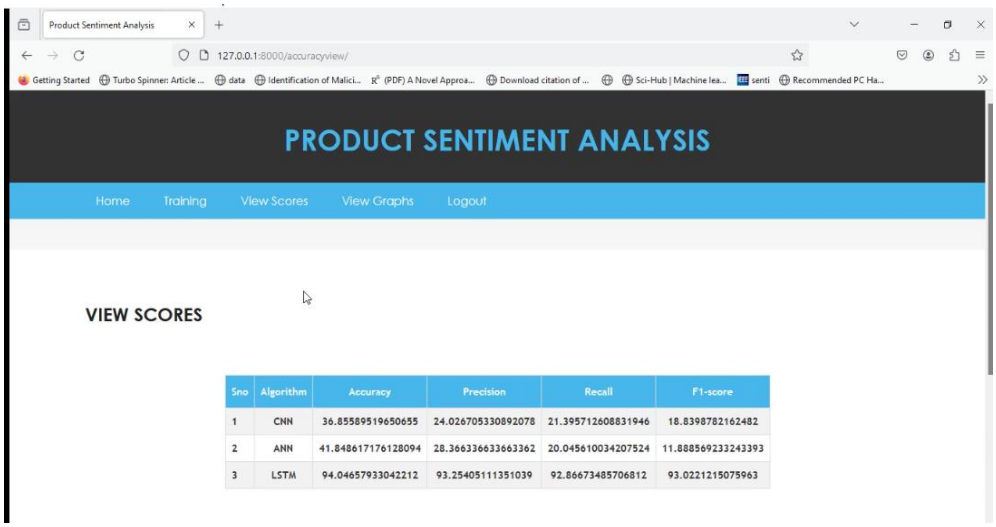
IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

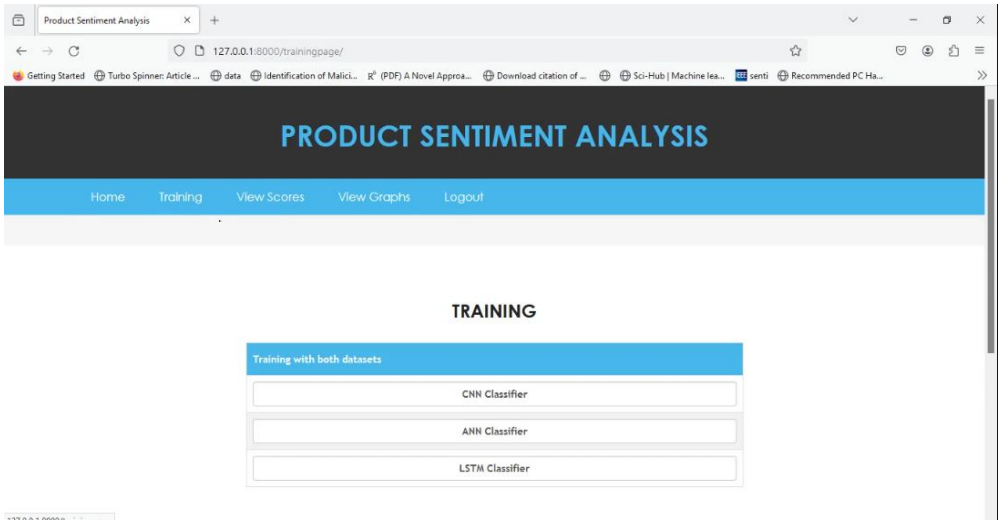
International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 1, January 2025

DOI: 10.17148/IJARCCE.2025.14113





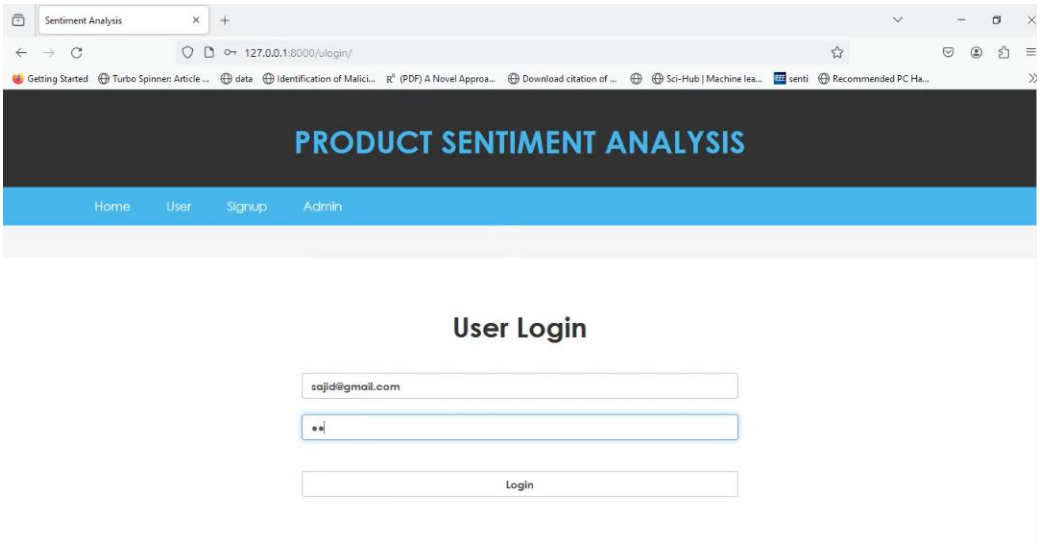


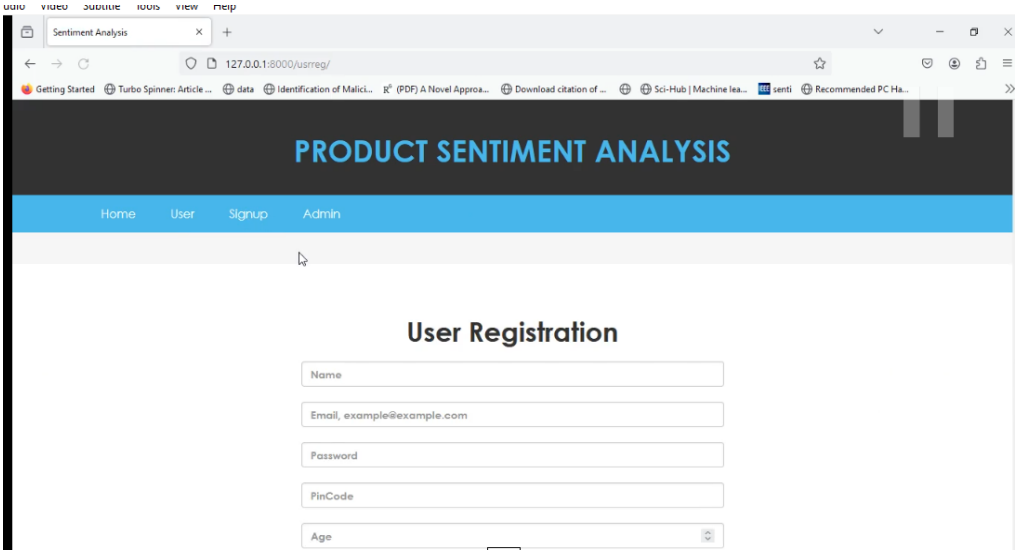
IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

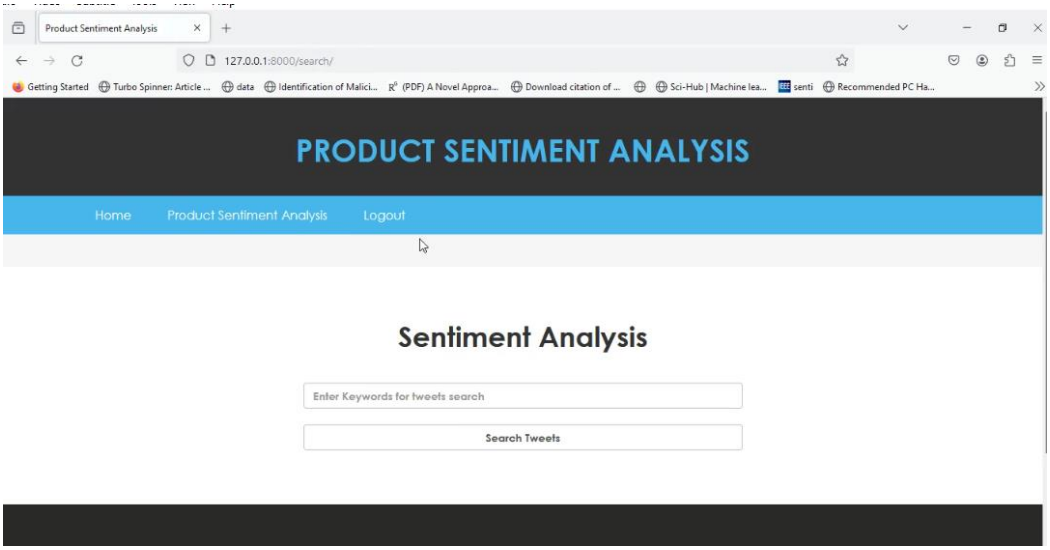
International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 1, January 2025

DOI: 10.17148/IJARCCE.2025.14113





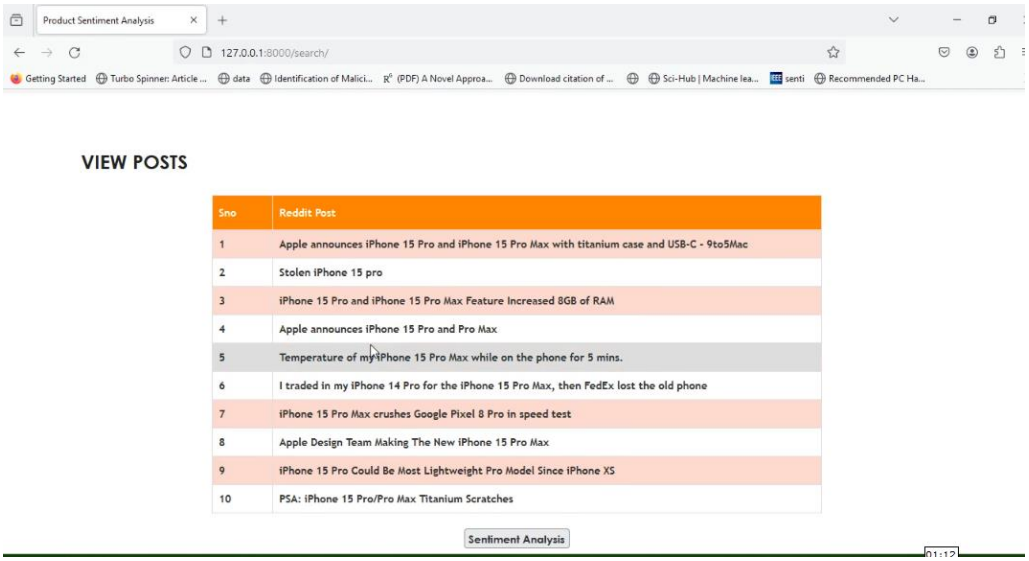


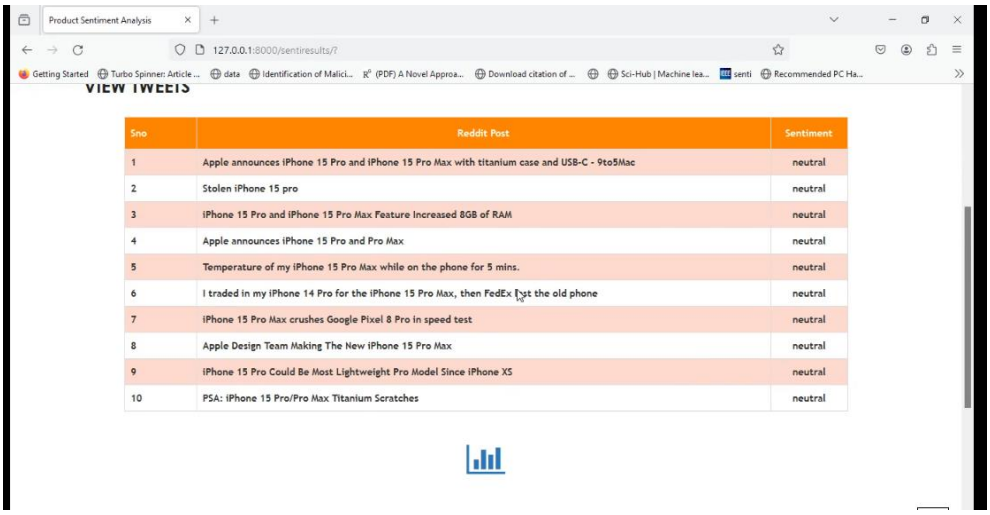
IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

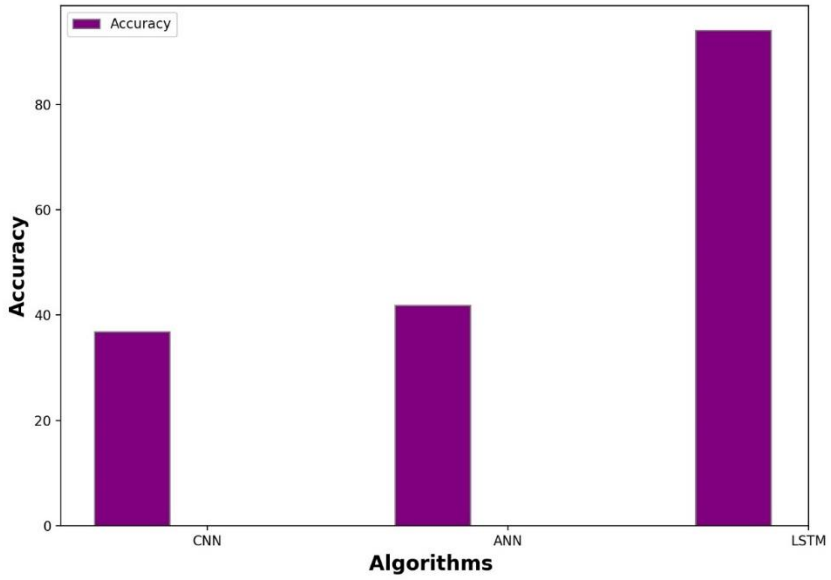
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 1, January 2025

DOI: 10.17148/IJARCCE.2025.14113





Graphs of the project



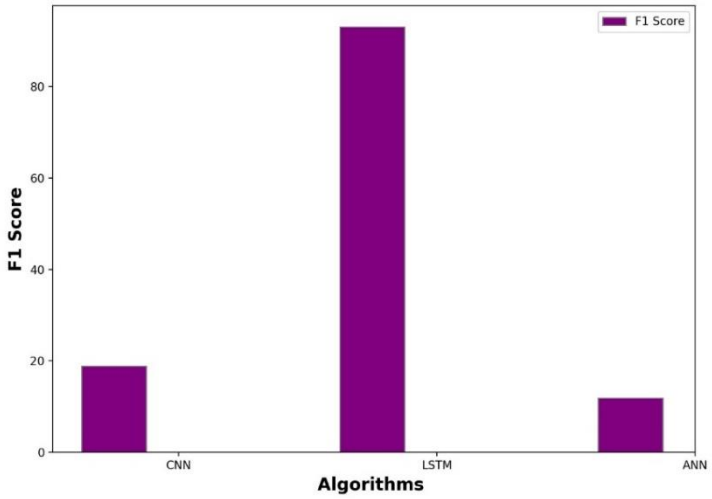
IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

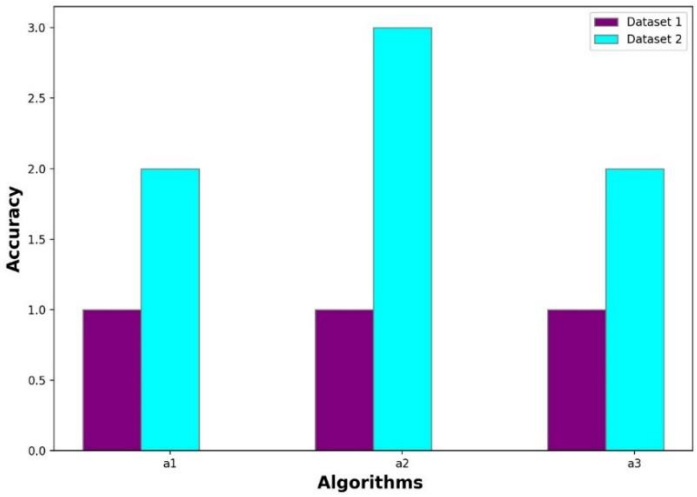
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 1, January 2025

DOI: 10.17148/IJARCCE.2025.14113

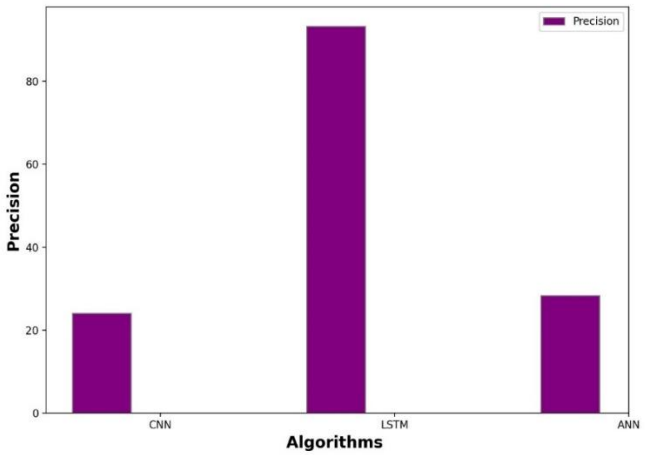
Accuracy of the Algorithms



F1 Score of the Algorithms



Accuracy of datasets on Algorithms



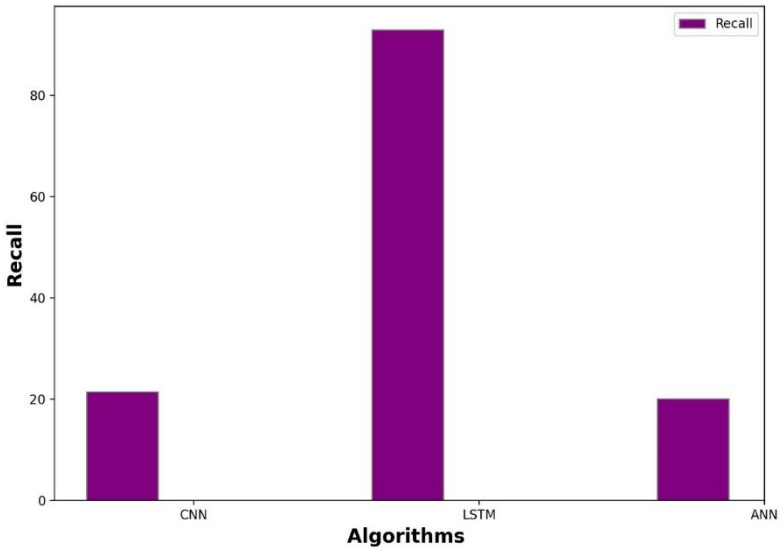
Precision on Algorithms

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 1, January 2025

DOI: 10.17148/IJARCCE.2025.14113



Recall on Algorithms

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 1, January 2025

DOI: 10.17148/IJARCCE.2025.14113

6.FUTURE SCOPE

The **Fine-Grained Sentiment Analysis** system has immense potential for expansion and refinement to address evolving challenges and leverage advancements in technology. Here are some key areas for future scope:

1. Enhanced Data Sources

- **Inclusion of Additional Platforms:** Extend the system to include other social media platforms like Facebook, Instagram, LinkedIn, and YouTube.
- **Multilingual Sentiment Analysis:** Develop capabilities to analyze sentiments in multiple languages using advanced translation models and natural language processing (NLP) techniques.

2. Improved Machine Learning Techniques

- **Integration of Transformers:** Incorporate state-of-the-art models like BERT, RoBERTa, and GPT for more accurate and context-aware sentiment classification.
- **Transfer Learning:** Utilize transfer learning to adapt pre-trained models to new datasets for domain-specific sentiment analysis.
- **Hybrid Models:** Combine traditional machine learning methods with deep learning techniques to improve accuracy and efficiency.

3. Advanced Sentiment Categorization

- **Emotion Detection:** Expand beyond positive, neutral, and negative sentiments to include fine-grained emotions like joy, anger, fear, sadness, and surprise.
- **Aspect-Based Sentiment Analysis:** Analyze sentiments based on specific aspects or features, such as service quality, product usability, or pricing.

4. Real-Time Capabilities

- **Streaming Data Analysis:** Implement real-time sentiment analysis for live streams or continuous data feeds, enabling instant insights and rapid decision-making.
- **Scalability:** Use distributed systems like Apache Kafka or AWS Lambda to process large volumes of real-time data efficiently.

5. Advanced Visualization

- **Interactive Dashboards:** Develop interactive dashboards with visual analytics to present sentiment trends, comparisons, and predictions dynamically.
- **Geographical Mapping:** Integrate geolocation data to visualize sentiments across different regions or demographics.

6. Domain-Specific Customization

- **E-commerce Feedback:** Tailor the system to analyze reviews and ratings on e-commerce platforms for improved customer insights.
- **Healthcare Sentiment:** Use sentiment analysis to understand patient feedback, public health opinions, or reactions to medical services.
- **Financial Markets:** Analyze sentiment trends in financial news and social media to predict market movements.

7. Ethical and Privacy Enhancements

- **Bias Mitigation:** Address algorithmic biases to ensure fair and unbiased sentiment analysis across diverse datasets.
- **Privacy Compliance:** Incorporate data anonymization and encryption techniques to adhere to privacy regulations like GDPR and CCPA.

8. Integration with Other Systems

- **Recommendation Systems:** Combine sentiment analysis with recommendation engines to suggest products or services based on user sentiments.
- **Chatbots:** Enhance chatbot interactions by integrating sentiment analysis to provide empathetic and context-aware responses.

9. Automation and RPA

- **Automated Reporting:** Enable automatic generation of sentiment analysis reports and summaries for stakeholders.
- **Robotic Process Automation (RPA):** Use RPA tools to streamline data collection, preprocessing, and sentiment analysis workflows.

10. Research Contributions

- **Benchmarking:** Establish benchmarks for sentiment analysis models and contribute datasets or methodologies to the research community.
- **Cross-Domain Learning:** Explore how sentiment analysis techniques can be applied to new domains like education, law, or entertainment.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 1, January 2025
DOI: 10.17148/IJARCCE.2025.14113

7.CONCLUSION

Sentiment analysis is used to find the sentiment of author behind his/her comment. In this work, tweets are extracted using a particular string search and these tweets are subjected to sentiment analysis using ANN,CNN,LSTM classifiers in order to classify them into positive, neutral and negative. As a part of analysis, the DL classifiers ANN, CNN are considered and their accuracy is estimated considering three features and accordingly increasing the number of tweets.

REFERENCES

[1] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams Engineering Journal, vol. 5, no. 4, pp. 1093–1113, 2014.
[2] E. Cambria, "Affective computing and sentiment analysis," IEEE Intel- ligent Systems, vol. 31, no. 2, pp. 102–107, 2016.
[3] S. Rosenthal, P. Nakov, S. Kiritchenko, S. Mohammad, A. Ritter, and V. Stoyanov, "Semeval-2015 task 10: Sentiment analysis in reddit," in Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015), pp. 451–463, 2015.
[4] Y. Wan and Q. Gao, "An ensemble sentiment classification system of reddit data for airline services analysis," in Data Mining Workshop (ICDMW), 2015 IEEE International Conference on, pp. 1318–1325, IEEE, 2015.
[5] Z. Jianqiang and G. Xiaolin, "Comparison research on text pre- processing methods on reddit sentiment analysis," IEEE Access, vol. 5, pp. 2870–2879, 2017.
[6] S. Tokle, S. R. Bellipady, R. Ranjan, and S. Varma, "Energy-efficient wireless sensor networks using learning techniques," Case Studies in Intelligent Computing: Achievements and Trends, pp. 407–426, 2014.
[7] M. Bouazizi and T. Ohtsuki, "A pattern-based approach for multiclass sentiment analysis in reddit," IEEE Access, vol. 5, pp. 20617–20639, 2017.
[8] W. Hua, Z. Wang, H. Wang, K. Zheng, and X. Zhou, "Understand short texts by harvesting and analyzing semantic knowledge," IEEE transactions on Knowledge and data Engineering, vol. 29, no. 3, pp. 499–512, 2017.
[9] D. Vilares, C. Go ´ mez-Rodr´ıguez, and M. A. Alonso, "Universal, unsupervised (rule-based), uncovered sentiment analysis," Knowledge-Based Systems, vol. 118, pp. 45–55, 2017. [10] D. Thomas, "Sentiment analysis of the conflict of ages series," in International Forum, vol. 20, pp. 31–43, 2017.

