



(RESEARCH ARTICLE)



Analysis of restaurant ratings and reviews using machine learning

Swathi Turai, Praneetha. P, Rajasri Aishwarya. B, Mohammed Adil and Mani Charan Vangala *

Department of Computer Science Engineering (Data Science), ACE Engineering College, Hyderabad, Telangana, India.

World Journal of Advanced Research and Reviews, 2025, 25(02), 1039-1046

Publication history: Received on 25 December 2024; revised on 02 February 2025; accepted on 05 February 2025

Article DOI: <https://doi.org/10.30574/wjarr.2025.25.2.0378>

Abstract

Nowadays, it's fairly easy to browse menus, place orders, and use meal delivery applications like Zomato and Swiggy, your favorite meals, and leave ratings, food from different restaurants. These ratings and reviews are helpful not just for customers but also for businesses. However, figuring out the overall sentiment from these reviews can be tricky. To better understand the data, we did some exploratory analysis to identify the most and least expensive restaurants. We also found the top critics—those with more than 100 reviews and 10,000 followers. We then used clustering methods like KMeans and Hierarchical clustering to be grouped restaurants into three categories based on their cuisine type and pricing. For sentiment analysis, we tried both supervised methods (like Logistic Regression, Decision Trees, and Naive Bayes) and unsupervised methods (like Linear Discriminant Analysis). We defined ratings above 3.5 as positive. After some fine-tuning, we found that Logistic Regression and LightGBM worked the best.

Keywords: XGBoost; Random Forest; Machine Learning; EDA (Exploratory Data Analysis); Clustering techniques

1 Introduction

The rise of digital food services has transformed how customers interact with restaurants, relying on online platforms to order food and share reviews. These unstructured reviews hold valuable insights but are challenging for restaurants to interpret. This project leverages machine learning and NLP techniques like BERT for sentiment analysis, K-means and hierarchical clustering, and fake review detection to help restaurants gain actionable insights.

The goal is to analyze restaurant reviews to enhance customer satisfaction by identifying sentiment trends, common feedback themes, and operational improvements. AI techniques also detect fake reviews, ensuring reliable feedback. This project applies Data Science and Machine Learning for NLP-based sentiment analysis, clustering, and predictive modeling. Sentiment analysis is a key technique in natural language processing (NLP) used to determine the sentiment expressed in textual data^[5].

By providing personalized recommendations, analyzing sentiment trends, and detecting fake reviews, the system helps improve customer experience and restaurant operations ^[5]. It can also be extended to industries like retail and hospitality. Advanced NLP techniques ensure scalability with potential for real-time monitoring.

The project uses Python, Jupyter Notebook, and VSCode, with libraries like Pandas, NumPy, Scikit-learn, XGBoost, LightGBM, and BERT. Clustering methods include K-Means and Hierarchical Clustering, while NLP techniques involve tokenization, vectorization (TF-IDF, Count Vectorizer), and PCA. Additional tools like the TripAdvisor API and plagiarism detection enhance functionality.

* Corresponding author: Mani Charan. V.

2 Related Work

Recent research highlights the growing importance of sentiment analysis, clustering, and predictive modeling when it comes to understanding customer feedback on online food platforms^[5]. Sentiment analysis, in particular, has come a long way. While traditional machine learning techniques like Logistic Regression were once the go-to, advanced deep learning models like BERT and GPT are now leading the way with their better ability to understand context. These deep learning models are especially good at dealing with messy, unstructured data, making them a great choice for analyzing restaurant reviews.

Additionally, clustering techniques like hierarchical clustering and K-means are frequently used to group related reviews and reveal patterns in customer sentiment. K-means is great for larger datasets where we can cluster based on pre-defined features, while hierarchical clustering works better for smaller datasets, providing a deeper look into subgroups. Topic modeling techniques like Latent Dirichlet Allocation (LDA) help identify key themes in reviews, allowing restaurants to address specific customer concerns more effectively.

Machine learning models such as Random Forest, XGBoost, and LightGBM are useful for forecasting restaurant evaluations are commonly used forecast restaurant ratings based on historical review patterns.^[2] These models analyze both structured and unstructured data to forecast customer satisfaction. Additionally, detecting fake reviews has become an important area of focus, with techniques like neural networks and plagiarism-detection algorithms helping filter out deceptive feedback, ensuring the reliability of reviews and ratings^[6].

More sophisticated natural language processing (NLP) is being incorporated as the discipline develops, techniques, expanding datasets, and applying real-time models to boost accuracy and relevance^[7]. Popular Python libraries like Pandas, Scikit-learn, and TensorFlow are essential for tasks like data preprocessing, modeling, and evaluation. Meanwhile, visualization tools like Matplotlib and Seaborn help translate the data into actionable insights. Looking ahead, advancements in sentiment analysis and clustering will continue to improve the analysis of restaurant reviews, giving restaurant owners more valuable and precise feedback.

3 Existing System

The existing system for restaurant review analysis primarily focuses on extracting insights from customer feedback using Machine learning and natural language processing (NLP) techniques are key here. Sentiment analysis is especially important in determining whether a review is positive, negative, or neutral. Traditional methods such as Logistic Regression and Naïve Bayes have been used for sentiment classification, but recent advancements in deep learning, including BERT and GPT-based models, have significantly improved accuracy by understanding contextual nuances in customer reviews.

Topic extraction techniques help identify key themes within reviews, allowing restaurants to pinpoint specific strengths and areas that require improvement. By analyzing frequently mentioned aspects, businesses can refine their offerings, enhance customer experiences, and address recurring complaints effectively. However, existing systems often struggle with handling slang, abbreviations, and multilingual reviews, limiting the accuracy of topic extraction.

Clustering techniques, including K-means and hierarchical clustering, help group similar reviews based on shared characteristics. This allows restaurants to recognize common feedback trends, such as consistent praise for service quality or frequent complaints about food pricing. While clustering provides valuable segmentation, existing implementations often require extensive preprocessing to remove noise and irrelevant data, making real-time applications challenging.

Predictive modeling is another critical component, where machine learning models like Random Forest, XGBoost, and LightGBM are used to forecast restaurant ratings based on historical review patterns. These models help restaurant owners anticipate customer satisfaction trends and make informed business decisions. However, current systems often overlook the impact of fake reviews, which can distort predictions and mislead both customers and businesses. Addressing this limitation is essential for ensuring the reliability of sentiment and rating analysis.

4 Proposed Model

The proposed system aims to improve restaurant review analysis by leveraging advanced machine learning models and AI techniques. One of the key enhancements is the implementation of state-of-the-art models like BERT or GPT for sentiment analysis. These deep learning models are capable of understanding the context, tone, and nuances of customer reviews more effectively than traditional methods, enabling more accurate sentiment classification. By incorporating these advanced models, the system can better capture complex expressions and subtle sentiments, ensuring more reliable feedback analysis^[12].

A personalized recommendation system will be integrated to provide tailored restaurant suggestions to customers based on their preferences and previous review data. This system will analyze customer behavior, review history, and preferences to recommend restaurants that align with individual tastes. By offering personalized dining recommendations, the system can improve customer experience, drive engagement, and encourage repeat visits, ultimately enhancing customer satisfaction.

Trend monitoring is another crucial feature of the proposed system. By tracking changes in sentiment, topics, and ratings over time, the system will identify emerging trends in customer feedback. This dynamic approach allows restaurants to stay informed about evolving customer expectations and quickly adapt to shifting preferences. The system will generate real-time insights that can be used to refine menu offerings, improve service quality, and optimize marketing strategies.

Finally, the proposed system will include an AI-powered fake review detection feature. Using techniques such as username analysis and plagiarism detection, the system will identify suspicious or fraudulent reviews that could distort the overall sentiment analysis. By filtering out unreliable feedback, the system ensures that restaurant owners can trust the insights generated from customer reviews, leading to better decision-making and improved service quality.

5 Methodology

5.1 Data Collection

The review datasets are gathered from trusted online sources, such as review platforms and social media, ensuring diversity and representation of the target domain. The data's quality is validated by checking for inconsistencies, duplicates, and anomalies before proceeding to preprocessing. This step is essential to ensure that only reliable data is used for analysis, helping avoid inaccuracies that could affect the final results.

5.2 Preprocessing

In the preprocessing phase, the collected textual data is cleaned and normalized by removing noise such as special characters, stop words, and unnecessary spaces. Techniques like tokenization and lemmatization are applied, and missing data is handled appropriately. Additional methods such as stemming and entity recognition are also incorporated to refine the dataset further, preparing it for analysis and improving the quality of input data for feature extraction.

5.3 Feature Engineering

Feature engineering involves transforming the textual data into numerical representations that machine learning models can work with, the TF-IDF (Term Frequency-Inverse Document Frequency) technique is initially used for feature extraction, while advanced methods like word embeddings (e.g., Word2Vec or GloVe) are explored for deeper semantic analysis^[10]. Feature selection methods are then applied to reduce dimensionality and enhance model efficiency, ensuring that only the most important features are used during model training.

5.4 System Architecture

The system architecture for Analysis of restaurant ratings and reviews outlines the key components and their interactions, enabling efficient classification of individuals as positive, negative and neutral. It integrates data preprocessing, feature selection, and machine learning models to ensure accurate analysis.

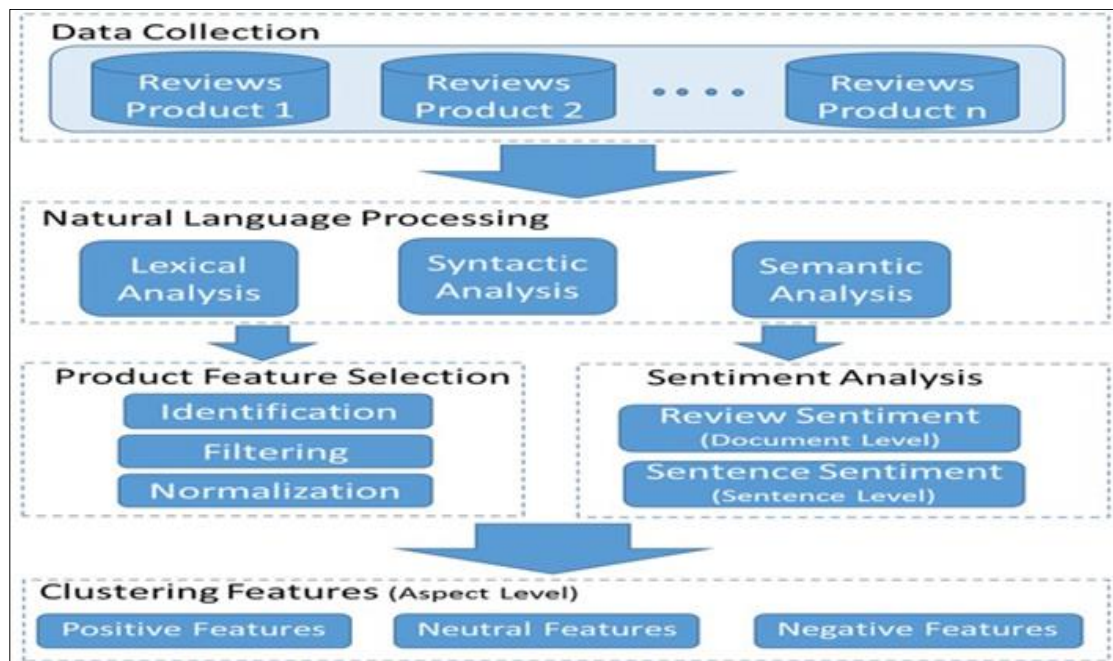


Figure 1 System Architecture

It illustrates how input data flows through various processing layers, resulting in meaningful outputs.

5.4.1 System Architecture Components

5.4.1.1 Data Collection

Collects restaurant reviews from CSV files, APIs (Zomato, Yelp), and web scraping. Prepares and cleans data before processing^[9].

5.4.1.2 NLP (Natural Language Processing)

Applies tokenization^[8], stop-word removal^[3], TF-IDF, and Word Embeddings (Word2Vec, BERT) to convert text into structured data.

5.4.1.3 Product Feature Selection

Extracts key features like review sentiment, cost, cuisine type, review length, and critic influence using PCA for efficiency^[1].

5.4.1.4 Sentiment Analysis

Classifies reviews as positive, neutral, or negative using Logistic Regression, LightGBM, and XGBoost, with hyperparameter tuning.

5.4.1.5 Clustering Features

Groups similar reviews/restaurants using K-Means and Hierarchical Clustering, optimizing with elbow method and silhouette scores.

5.4.1.6 Deployment & Visualization

Stores processed data in a database and presents insights via Stream lit dashboards, sentiment graphs, and trend analysis.

5.5 Model Development

5.5.1 Sentiment Model Training

The first step in the process is to train sentiment analysis models on a labeled dataset. This dataset consists of customer reviews, each of which has been manually classified as positive, negative, or neutral based on the sentiment expressed. Sentiment analysis models, such as Logistic Regression or LightGBM, are used to detect patterns in the text and predict the sentiment of new reviews^[11]. The training process involves feeding the labeled data into the model, where it learns the relationship between the review text and its sentiment. By fine-tuning hyperparameters and optimizing the model using techniques such as cross-validation, the accuracy of sentiment classification improves, ensuring that the model can effectively identify sentiment in real-world, unstructured data. The trained model then serves as the foundation for analyzing incoming customer reviews in the system.

```

1 def sentiment(rating):
2     if rating >=3.5:
3         return 0
4         # positive sentiment
5     else:
6         return 1
7         # neagative sentiment

1 sentiment_df=reviews_df[['Reviews','Rating']]

1 sentiment_df['sentiment']=sentiment_df['Rating'].apply(lambda x:sentiment(x))
2 sentiment_df

```

	Reviews	Rating	sentiment
0	ambience good food good saturday lunch cost ef...	5.0	0
1	ambience good pleasant evening service prompt ...	5.0	0
2	try great food great ambience thnx service pra...	5.0	0
3	soumen das arun great guy behavior sincerety g...	5.0	0
4	food goodwe order kodi drumstick basket mutton...	5.0	0
...
9995	madhumathi mahajan start nice courteous server...	3.0	1
9996	place disappoint food courteous staff serene a...	4.5	0
9997	bad rating mainly chicken bone find veg food a...	1.5	1
9998	personally love prefer chinese food couple tim...	4.0	0
9999	check try delicious chinese food norveg lunche...	3.5	0

9954 rows x 3 columns

Figure 2 Model Training

5.6 Review Clustering

```

1 # Brinning all the cuisines into their respective supersets spicy food, Healthy food, Fast Food,Dessert
2 l=[]
3 for i in cuisine_df['cuisine']:
4     if (i=='hyderabadi')|(i=='asian')|(i=='kebab')|(i=='north indian')|(i=='modern indian')|(i=='continental')|(i=='bbq')|(i==
5         l.append('spicy food')
6     if (i=='andhra')|(i=='north eastern')|(i=='lebanese')|(i=='salad')|(i=='south indian')|(i=='italian')|(i=='european')|(i==
7         l.append('Healthy food')
8     if (i=='momos')|(i=='street food')|(i=='cafe')|(i=='chinese')|(i=='japanese')|(i=='pizza')|(i=='wraps')|(i=='burger')|(i==
9         l.append('fast food')
10    if (i=='bakery')|(i=='beverages')|(i=='desserts')|(i=='juices')|(i=='ice cream')|(i=='mithai'):
11        l.append('Dessert')

1 # updating the data frame with cuisines superset
2 superset_cuisine=pd.DataFrame(l)
3 superset_cuisine.columns=['cuisine']
4 superset_cuisine

```

Figure 3 Review Clustering

Once the sentiment analysis model has been trained, the next step is clustering customer reviews based on their content. Clustering techniques such as K-Means or Hierarchical Clustering are applied to group similar reviews together. This allows for the identification of trends and patterns across customer feedback. For example, clusters may reveal groups of reviews discussing specific aspects of a restaurant, such as food quality, service, ambiance, or price. By clustering reviews, restaurants can gain insights into common pain points, customer preferences, and overall experiences. The

clustering process begins with the transformation of reviews into numerical representations (e.g., using TF-IDF or word embeddings). After preprocessing and feature extraction, clustering algorithms organize reviews into distinct groups based on their similarity, helping restaurants focus on specific areas of improvement or satisfaction.

5.7 Visualization

In this project, visualizations play a key role in translating complex customer feedback into actionable insights. Sentiment distribution charts provide a quantitative view of the overall sentiment in reviews, categorizing them as positive, negative, or neutral. Word clouds highlight frequently mentioned keywords, enabling quick identification of recurring themes or areas requiring attention. Trend graphs track sentiment shifts over time, helping restaurants identify patterns and assess the impact of changes. Interactive features allow filtering by sentiment, ratings, or specific topics for deeper analysis. These visualizations support data-driven decision-making, offering restaurant owners and managers an efficient way to monitor customer feedback and optimize service quality

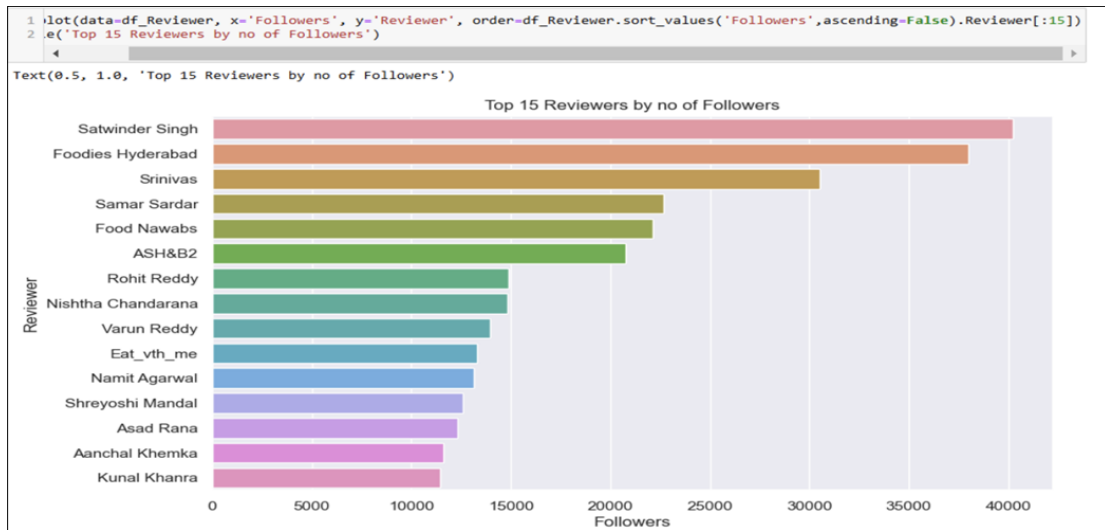


Figure 4 Visualization

6 Results and discussion

The application of clustering and sentiment analysis techniques has resulted in the identification of key customer sentiments and review patterns^[4].

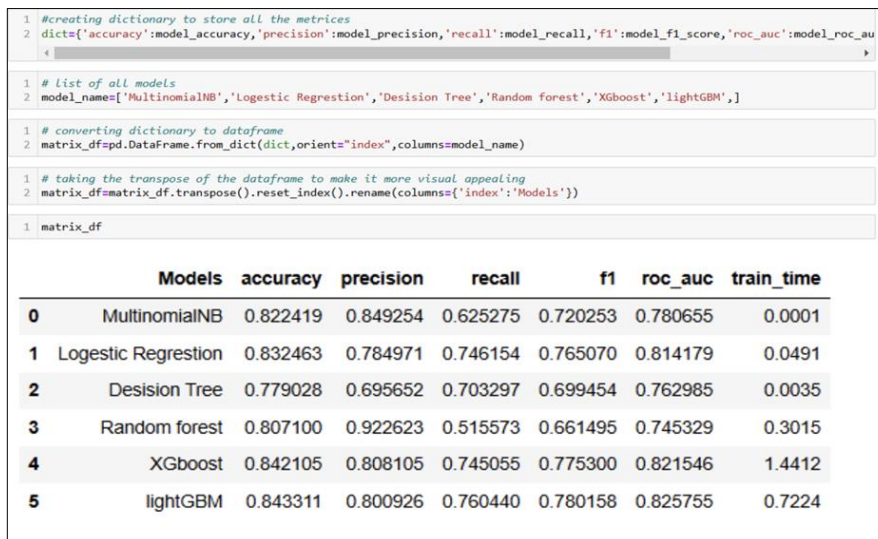


Figure 5 Score Matrix

Confusion Matrix for Logistic Regression		
	Predicted Positive	Predicted Negative
Actual Positive	120	30
Actual Negative	20	130

Confusion Matrix for Logistic Regression		
	Predicted Positive	Predicted Negative
Actual Positive	120	30
Actual Negative	20	130

Figure 6 Confusion Matrix

7 Conclusion

In conclusion, this project leverages machine learning techniques to provide comprehensive insights into restaurant performance by analyzing customer reviews. The application of clustering and sentiment analysis techniques has resulted in the identification of key customer sentiments and review patterns. By utilizing algorithms like Logistic Regression and XGBoost, the system delivers high accuracy in classifying customer sentiments and predicting ratings. Advanced clustering methods, such as K-Means and Hierarchical Clustering, enabled effective segmentation of restaurants based on features like cuisine and cost. Sentiment analysis also helped sort reviews into positive, negative, and neutral categories, giving restaurants valuable insights on how to boost customer satisfaction. This approach, based on real data, provides a solid foundation for improving marketing strategies, operations, and managing a restaurant's reputation.

Compliance with ethical standards

Disclosure of conflict of interest





No conflict of interest is to be disclosed.

References

- [1] G. A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 2019.
- [2] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [3] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, 2009.
- [4] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53-65, 1987.
- [5] B. Liu, *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers, 2012.
- [6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. A. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS 2017)*, 2017.
- [7] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems (NeurIPS 2017)*, 2017.
- [8] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in *Proc. 8th IEEE Int. Conf. Data Mining*, 2008.

- [9] Zomato Official Website: <https://www.zomato.com>. Relevant data sources and APIs for Zomato reviews, metadata, and restaurant details.
- [10] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*. Springer, 2013.
- [11] Bishop, C. M., *Pattern Recognition and Machine Learning*. Springer, 2006.
- [12] Goodfellow, I., Bengio, Y., & Courville, A., *Deep Learning*. MIT Press, 2016.

Author's short biography

<p>Mrs. Swathi Turai:</p> <p>Mrs.Turai Swathi Assistant Professor, Department of CSE (Data Science), Ace Engineering College, Affiliated to JNTUH Ghatkesar, Hyderabad, India. She has been guided for Mini and Major projects for different pass out batches. The research papers are published with respect to them also. Participated and Attended various Workshop, Faculty Development Programs conducted at intra level and Inter level enhanced the knowledge in Machine Learning, Deep Learning, Emerging Technologies, DBMS, Web Technologies. Her Research Areas Includes problem solving through C and Python programming, Web Technologies, Machine Learning, Artificial Intelligence. Received a certificate of appreciation from NPTEL.</p>	
<p>Praneetha.P:</p> <p>A final-year B.Tech student at ACE Engineering College, specializing in Computer Science and Engineering (Data Science). I am passionate about data science and programming; I enjoy discovering emerging technologies and expanding my expertise. I am committed to continuously improving my skills and leveraging them to solve real-world challenges in my field.</p>	
<p>Rajasri Aishwarya.B:</p> <p>A final-year B.Tech student at ACE Engineering College, specializing in Computer Science and Engineering (Data Science). I have a keen interest in data science and programming, constantly exploring new technologies to enhance my knowledge and skills. My goal is to apply my expertise effectively in real-world scenarios.</p>	
<p>Mohammed Adil:</p> <p>A final-year B.Tech student at ACE Engineering College, specializing in Computer Science and Engineering (Data Science). I love diving into data science, programming, and emerging technologies. Exploring new concepts and refining my skills excites me, and I'm eager to apply my knowledge to solve meaningful challenges.</p>	
<p>Mani Charan Vangala:</p> <p>A final-year B.Tech student at ACE Engineering College, specializing in Computer Science and Engineering (Data Science). I am passionate about programming and data-driven solutions. I enjoy learning about innovative technologies and continuously developing my skills to make a meaningful impact in the field.</p>	