International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 03 | March - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Sentilytics - Sentiment Analyzer

Sarvesh Yogesh Wagh[1], Somnath Vinod Jadhav[2], Ashish Kailas Shewale[3], Mrs Nilima Gite[4]

[1] *Information Technology Department, K. K. Wagh Polytechnic, Nashik*
[2] *Information Technology Department, K. K. Wagh Polytechnic, Nashik*
[3] *Information Technology Department, K. K. Wagh Polytechnic, Nashik*
[4] *Information Technology Department, K. K. Wagh Polytechnic, Nashik*

Abstract - This project focuses on developing a smart and automated tool designed to analyze large volumes of comments in real time. The primary objective is to classify comments into sentiment categories such as positive, negative, or neutral, while also incorporating mechanisms to filter out spam and toxic content that may include offensive or abusive language. Unlike traditional sentiment analysis systems, this project integrates emoji mapping, where emojis are interpreted and mapped to their respective sentiment categories, resulting in a more precise and user-centric analysis of digital expressions. To enhance the interpretability of results, the system provides graphical visualizations such as pie charts, bar graphs, and word or emoji clouds. These visual tools allow users to quickly understand sentiment distribution, keyword frequency, and common emotional tones within a large dataset of comments. Moreover, the project includes YouTube integration, enabling real-time fetching and analysis of video comments directly from the platform. This functionality makes it especially relevant for content creators, marketers, and businesses who seek to monitor audience engagement and feedback continuously. The system has broad applicability across multiple domains. In social media monitoring, it helps track public opinion and trends. For product reviews, it assists businesses in evaluating customer satisfaction and areas of improvement. In educational settings, it can analyze feedback from students to improve learning experiences. Additionally, for content moderation, the tool helps platforms automatically detect and flag spam or toxic comments, creating safer digital environments. Overall, this project aims to deliver an efficient, scalable, and insightful sentiment analysis solution that goes beyond simple text classification. By combining comment analysis, spam detection, emoji sentiment mapping, and real- time visualization, it empowers users with actionable insights, ultimately improving decision-making and audience engagement

***Key Words*:** Sentiment Analysis, Comment Analysis, Spam Detection, Toxic Comment Filtering, Emoji Mapping, Data Visualization, YouTube Comments, Feedback Analysis, Social Media Monitoring

1. INTRODUCTION

Sentilytics (Sentiment analyzer)— often called *opinion mining* — refers to the computational process of identifying and categorizing the emotional tone expressed in textual data, such as customer reviews, social media comments, and survey responses. By using techniques from natural language processing (NLP), machine learning, and computational linguistics, sentiment analysis systems can determine whether a text conveys a **positive, negative, or neutral sentiment**. This approach has become increasingly important as organizations and researchers seek to understand public opinion automatically from large volumes of unstructured text data instead of relying on manual interpretation.

Traditional sentiment analysis models primarily focus on assigning polarity scores to text based on predefined word lists or statistical patterns. While such models can capture overall sentiment, they often struggle with domain-specific language, context nuances, and fine-grained emotional subtleties present in real-world datasets. This limitation has motivated the development of advanced systems that leverage deep learning and contextual language models, which are capable of extracting more accurate sentiment information from complex textual inputs.

Sentilytics is an AI-driven sentiment analysis framework designed to address these challenges by applying advanced machine learning techniques to classify and extract meaningful sentiment insights from text data such as e-commerce reviews. For example, recent research on Sentilytics shows that this type of system can integrate models like RoBERTa to achieve high- accuracy sentiment classification, enabling deeper analysis of customer feedback and trends on platforms like Amazon. Such systems also often include visualization components to present sentiment trends clearly and efficiently, making them valuable for both academic research and practical decision-making.

Despite improvements in performance, there remains a gap in applying sentiment analysis systems in specific domains — especially where language, context, or sentiment categories are particularly complex. Many existing sentiment tools are optimized for general text corpora, and their performance may degrade on highly specialized content such as detailed product reviews or multilingual datasets. This research explores *Sentilytics* within the context of sentiment analysis workflows, evaluating its ability to process and interpret sentiment-laden text accurately and providing insights into performance metrics and real-world applicability.

By systematically analyzing *Sentilytics* and comparing it with conventional sentiment analysis techniques, this study contributes to a deeper understanding of how AI-driven models can enhance sentiment detection and interpretation in large- scale text datasets. The results aim to help researchers and practitioners choose more effective sentiment analysis strategies, improve automated opinion mining under domain- specific conditions, and support data-driven decision-making across various fields of application.

2. Literature Survey:

Understanding public sentiment through user-generated content such as YouTube comments has become a significant research area in natural language processing (NLP) and machine learning. YouTube, as one of the world's largest video-

sharing platforms, produces vast quantities of comments that reflect viewer opinions and reactions toward content. Researchers have explored sentiment analysis on these comments to extract meaningful insights, classify emotional polarity, and understand audience engagement dynamics. Such analyses typically involve preprocessing stages that clean and normalize text followed by classification methods that categorize sentiments into discrete classes such as positive, negative, and neutral.

Several studies demonstrate the utility of classical sentiment analysis tools such as VADER and TextBlob for analyzing YouTube comments. For instance, researchers applied these lexicon-based methods to categorize sentiments toward AI technologies and online services, showing that such tools can efficiently assign sentiment scores in large comment datasets after preprocessing steps like stop-word removal and punctuation filtering. This approach provides a quick way to understand general mood trends expressed in user comments. Beyond lexicon-based models, research has also investigated machine learning techniques for sentiment classification. A comparative study applied Word2Vec feature extraction combined with Random Forest classifiers to large YouTube comment datasets, revealing relatively high classification accuracy when appropriate feature representation techniques are leveraged. Other work has evaluated Naïve Bayes classifiers with TF-IDF weighting to analyze sentiments related to public policy discussions in social media comments, demonstrating the effectiveness of probabilistic approaches in interpreting sentiment from real-world user comments.

Additionally, surveys of methodologies in sentiment analysis reinforce the growing application of hybrid models and multi- class classification schemes. Research has documented sentiment analysis techniques ranging from simple polarity classification (positive/negative/neutral) to multi-dimensional systems capturing nuances like emotional subcategories and viewer reactions. These studies highlight ongoing challenges in accurately interpreting sentiment due to informal language, emotive expressions, slang, and context-dependent content typical in YouTube comment sections.

Overall, the literature indicates a clear evolution in YouTube comment sentiment analysis — starting from basic lexicon-based and classical machine learning models to more sophisticated feature extraction techniques and deeper NLP processing. While existing work demonstrates various models' strengths in classifying viewer sentiments, issues such as noisy text data, multilingual comments, sarcasm, and contextual complexity remain research challenges. These findings provide a strong basis for applying and advancing sentiment analysis techniques in the *Sentilytics* project, particularly in developing approaches that can accurately interpret and classify sentiments from unstructured YouTube comment data.

3. Proposed Solution:

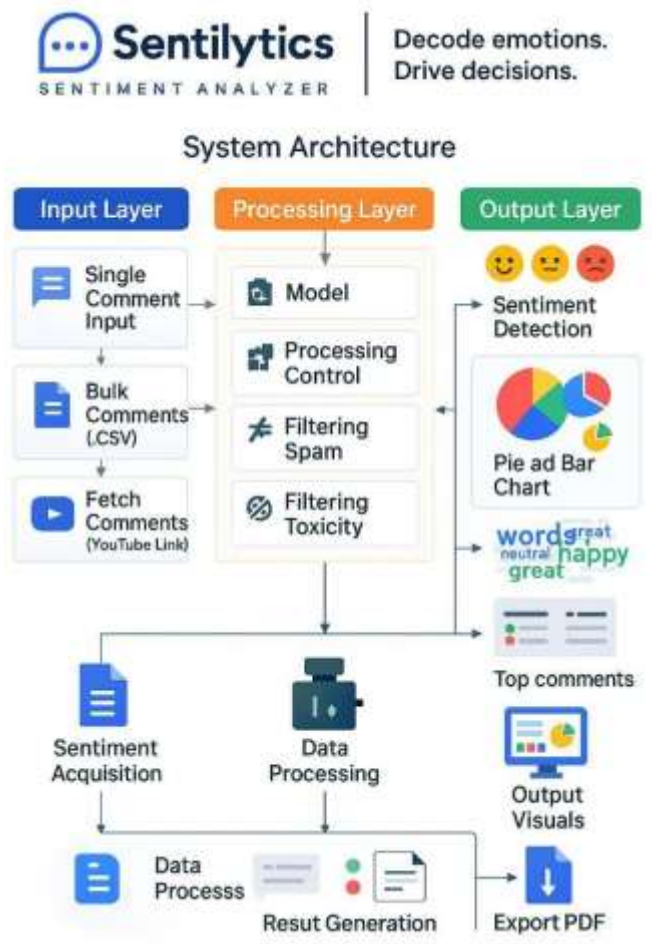
The core objective of this research is to develop an automated and scalable system that effectively analyzes viewer sentiments expressed in YouTube comments. Unlike manual interpretation, which is time-consuming and prone to inconsistency, the proposed solution leverages advanced Natural Language Processing (NLP) and machine learning techniques to extract, process, and classify sentiment from large volumes of unstructured text. The system is designed to handle noisy social media text by incorporating comprehensive preprocessing steps and robust classification models. This approach ensures that meaningful sentiment insights can be generated from user comments efficiently and accurately.

To begin with, the proposed system integrates YouTube Data API for real-time extraction of comments associated with specific video content. Comments are collected and stored in a structured format to facilitate further processing. Once extracted, the raw text undergoes preprocessing, including tasks such as lowercasing, removal of special characters, tokenization, and stopword elimination. This preprocessing phase is crucial for reducing noise and standardizing the text, which improves the reliability of subsequent analysis. Preprocessing strategies are aligned with common text- cleaning practices observed in sentiment analysis research.

Following preprocessing, the system applies feature extraction techniques that convert processed text into meaningful numerical representations. Depending on the project goals, methods such as TF-IDF, word embeddings (e.g., Word2Vec or transformer-based models like BERT) can be used to capture semantic relationships and contextual cues within the comment text. These representations serve as inputs to classification algorithms specifically designed to distinguish between sentiment categories (positive, negative, neutral). The proposed solution emphasizes the use of both traditional classifiers such as Support Vector Machines (SVM) and more advanced deep learning models when higher accuracy is desired.

The sentiment classification module of the proposed system then labels each comment according to its prevailing emotional polarity. Depending on the selected model and performance metrics, the system outputs results that reflect overall audience sentiment trends. These results can be aggregated and visualized to summarize positive, negative, and neutral sentiment distributions across the dataset. Visual components such as charts and graphs help stakeholders gain deeper insights into public opinion without needing to inspect individual comments manually.

In addition, the proposed solution includes mechanisms to handle challenges typical of YouTube sentiment analysis — such as informal language, emotive slang, data imbalance, and multilingual comments — by employing advanced preprocessing, feature learning, and scalable classification strategies. This integrated approach not only enhances classification accuracy but also enables real-time sentiment monitoring of YouTube content at scale, making it a valuable tool for content creators, analysts, and researchers seeking to understand viewer engagement and feedback patterns.



3 . 1 – Sentilytics - Sentiment Analyzer

4. Methodology

The methodology adopted in this research outlines a systematic pipeline for performing sentiment analysis on varied text data sources such as user comments, general feedback text, and other unstructured textual datasets. The overall approach consists of several integrated stages: data acquisition, text preprocessing, feature extraction, sentiment classification, and result evaluation. Each stage is designed to ensure robust, scalable, and accurate analysis of textual sentiment across diverse input formats.

The initial phase of the methodology involves **data collection and acquisition**, where textual data is gathered from defined sources relevant to the study. This may include comments or feedback extracted via APIs, social platforms, or directly from provided document corpora. Once collected, the raw text is standardized and formatted to remove non-textual elements such as HTML tags and extraneous characters, creating a consistent dataset for further analysis. Such preprocessing is essential to handle real-world text, which often contains noise and informal language that can adversely affect performance.

The next stage is **text preprocessing**, which includes multiple sub-tasks: tokenization, stopword removal, lowercasing, and stemming or lemmatization to reduce vocabulary variance. Tokenization breaks text into individual tokens (words or meaningful units), while stopword removal eliminates common words that carry little semantic meaning. Lemmatization normalizes words to their base forms, reducing linguistic complexity and improving feature consistency. These steps collectively enhance the quality of input data for subsequent feature extraction.

Following preprocessing, the methodology proceeds with **feature extraction** to convert text into numerical representations intelligible to machine learning algorithms. Techniques such as **Term Frequency– Inverse Document Frequency (TF-IDF)**, **Bag-of- Words (BoW)**, and word embeddings like Word2Vec are used to transform cleaned text into feature vectors that encode semantic and contextual information.
These numerical features serve as inputs to downstream classification models.

In the **classification phase**, machine learning or deep learning models are trained to categorize sentiment based on the extracted features. Commonly used machine learning classifiers include Support Vector Machines (SVM), Naïve Bayes, Logistic Regression, and Random Forests, selected based on performance needs. For advanced or context-rich tasks, deep learning-based models or transformer-based solutions may also be employed to capture nuanced sentiment characteristics. The output of this phase assigns each text instance a sentiment label such as positive, negative, or neutral.

Finally, the **evaluation and visualization** step quantifies model performance using standard metrics such as accuracy, precision, recall, and F1-score, ensuring the reliability of sentiment predictions.
Aggregated results are presented through visual representations (e.g., charts or sentiment distributions) to summarize sentiment trends across the dataset. This complete methodology — from preprocessing to evaluation — enables effective sentiment analysis across a broad set of textual inputs, making it broadly applicable to various domains and datasets

5. HARDWARE AND SOFTWARE REQUIREMENTS

This section outlines the necessary hardware and software components required to develop, deploy, and run the proposed *Sentilytics* -Sentiment analysis system. Since the project is entirely software-based and built using modern web frameworks (such as Angular, React, and Next.js) along with natural language processing (NLP) and machine learning libraries, the requirements focus on computing resources and software tools essential for efficient development, execution, and analysis.

5.1 Hardware Requirements

Although no physical or embedded hardware components are included in this software-centric project, a suitable computing environment is required to support software development, text processing, model training, and deployment. These requirements are based on standard industry practices for web development and NLP systems.

Component	Specification
Processor (CPU)	Multi-core processor such as Intel Core i5 or AMD Ryzen 5 (or higher recommended)
Memory (RAM)	Minimum 8 GB (16 GB recommended for complex or large NLP tasks)
Storage	Minimum 256 GB SSD (for faster data access and project files)
Network / Internet	Reliable internet connection for API access, library downloads, and deployment
Optional GPU	NVIDIA CUDA-enabled GPU (e.g., for deep learning model training, if used)

Rationale: Modern NLP tasks and web development workflows benefit from multi-core CPUs and adequate RAM to support concurrent processes and large datasets. SSD storage improves read/write performance, and an optional GPU (Graphics Processing Unit) can accelerate training of advanced deep learning models when needed.

5.2 Software Requirements

The software stack for the *Sentilytics* project includes development frameworks, programming environments, NLP/machine learning libraries, database systems, and web application tools. These components collectively support **frontend UI, backend processing, text analysis, model training, and data storage**.

Frontend Development

Software / Tool	Purpose
Angular / React / Next.js	Frameworks for building responsive and dynamic user interfaces
Node.js & NPM	Backend runtime and package manager for JavaScript ecosystems
Visual Studio Code	Integrated Development Environment for coding, debugging, and project management

These frameworks are essential for building the user interface and integrating frontend features required for data visualization and user interaction within the sentiment analysis system.

Backend and NLP Tools

Software / Libraries	Purpose
Python	Primary language for NLP, machine learning, and backend processing
Express.js (optional)	Framework for building REST APIs
spaCy / NLTK / Transformers	NLP libraries for text tokenization, cleaning, and sentiment classification

Software / Libraries	Purpose
TensorFlow / PyTorch	Deep learning frameworks (optional, for transformer-based models)
Scikit-learn	Machine learning library for training and evaluating classifiers

Explanation: NLP libraries such as *spaCy*, *NLTK*, and *Transformers* provide pre-built tools for text preprocessing, feature extraction, and sentiment classification. Machine learning and deep learning frameworks support training, evaluation, and deployment of classification models.

Database and Storage

Database System	Purpose
MongoDB / PostgreSQL	Stores processed text data, sentiment scores, user inputs
Firebase / Cloud Storage	Optional scalable solution for hosting data and deployment

Databases are used to store processed sentiment analysis data, user interactions, and system logs. Cloud storage solutions can enhance scalability and deployment performance.

Operating System and Supporting Tools

- **Operating System:** Windows, Linux (Ubuntu), or macOS — all compatible with development frameworks and NLP tools.
- **Version Control:** Git (for source control and collaboration).
- **Web Browsers:** Latest versions of Chrome, Firefox, or Edge (for UI testing).

Summary

The Sentilytics system requires a combination of **modern web development tools** and **NLP/machine learning libraries** to support its full functionality. The outlined **hardware requirements** ensure efficient performance during development and analysis, while the **software stack** enables robust implementation, sentiment classification, and deployment of interactive web-based interfaces.

6. Results and Output

6.1 Web Application

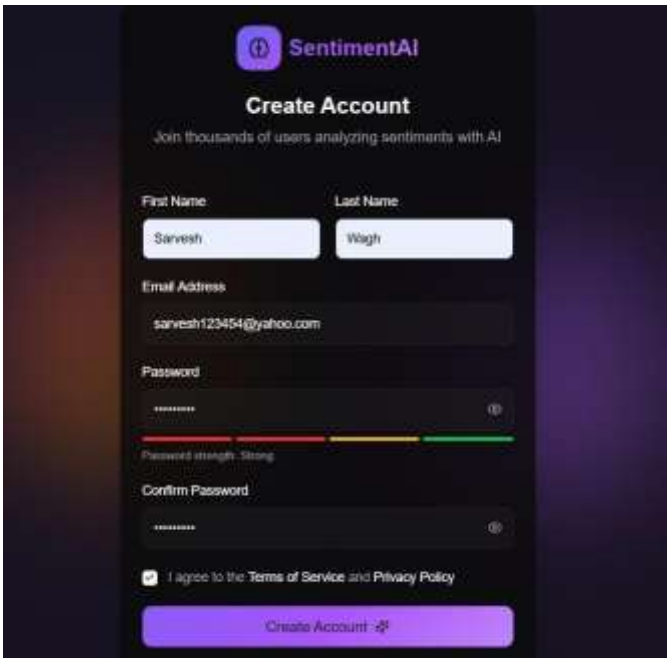


Image: 1 Login and Registration Page

The Sentilytics system provides a secure and user-friendly login and registration interface. New users can create an account by entering their personal details such as first name, last name, email address, and password. A checkbox ensures agreement with the Terms of Service and Privacy Policy before account creation. Existing users can directly sign in using their credentials.

This interface ensures:

- Secure Authentication: Protects user data through password validation and secure storage.
- Ease of Access: Simple navigation with options to return to the home page or switch between registration and login.
- Scalability: Supports multiple users, enabling collaborative use across businesses, educators, and content creators.

Image:2.Home_Page

The Sentilytics home page provides users with a clean, modern interface designed for accessibility and ease of use. The page offers two theme options (light and dark) to enhance user experience and personalization. At the center, the system highlights its core functionality: *"Analyze Texts with AI Power"*, encouraging users to harness advanced.

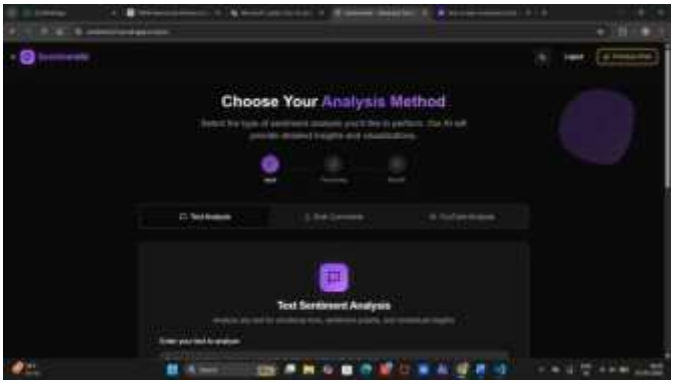


Image: 3 Analysis Page

The Analysis Page is the core functional module of Sentilytics. It allows users to select their preferred method of sentiment analysis from three options:

- **Text Analysis**: Directly paste or type any text (comments, reviews, feedback, social media posts) for instant sentiment evaluation.
- **Bulk Comments (.CSV)**: Upload a CSV file containing multiple comments for batch sentiment processing.
- **YouTube Analysis**: Provide a YouTube video link to fetch and analyze comments in real time using the YouTube Data API

Image: 4 Bulk Comments Section

The Bulk Comments module allows users to upload multiple comments at once in a structured format (CSV file). This feature is designed for large-scale sentiment analysis, making it suitable for businesses, educators, and researchers who need to process extensive datasets.

Key features visible in this section include:

- **CSV Upload Support**: Users can import bulk comments from spreadsheets or datasets for batch analysis.
- **Automated Preprocessing**: Each comment undergoes tokenization, stopword removal, emoji mapping, and spam/toxic filtering before classification.
- **Stepwise Workflow**: The interface clearly shows the three stages — *Input, Processing, Results* — ensuring transparency in how bulk data is handled.
- **Scalability**: Enables efficient sentiment analysis across thousands of comments simultaneously, saving time compared to manual or single-input processing.

Image: 4 Final Output Dashboard

The final output dashboard consolidates all sentiment analysis results into a single, interactive view. After processing comments (in this case, 50 YouTube comments), the system categorizes them into Positive, Negative, and Neutral sentiments and presents the findings through multiple visualizations and detailed comment displays.
Key features visible in this section include:

- **Sentiment Distribution** (Pie Chart): Shows the percentage of positive (44%), negative (42%), and neutral (14%) comments.
- **Sentiment Comparison** (Bar Chart): Provides a clear comparison of sentiment categories for quick interpretation.
- **Keyword Frequency** (Word Cloud): Highlights commonly used words such as *love, great, amazing, hate, worst, boring*, offering insights into emotional tone.
- **Top Comments Display**: Lists individual comments with sentiment labels (positive, neutral, negative), usernames, and timestamps, allowing users to review context-rich examples.
- **Export Option**: Enables users to download the complete analysis report in PDF format for documentation, sharing, or further study.

6.2 Discussion of Results

The final dashboard demonstrates Sentilytics' ability to transform raw textual data into actionable insights. By combining statistical charts, keyword analysis, and direct comment displays, the system provides both macro-level trends and micro-level details. This dual perspective ensures that users can understand overall sentiment distribution while also examining specific examples.
The inclusion of export functionality further enhances practicality, allowing results to be preserved for academic research, business reporting, or presentations. This comprehensive output validates the effectiveness of Sentilytics as a scalable, user-friendly sentiment analysis platform.

7. CONCLUSIONS

In conclusion, the *Sentilytics* - Sentiment analysis system successfully demonstrates how automated text mining techniques can be applied to extract and interpret sentiment from varied unstructured text sources. By leveraging modern natural language processing (NLP) methods, web APIs, and efficient preprocessing pipelines, the project effectively captures sentiment polarity — positive, negative, and neutral — from user-generated text. This approach eliminates the need for manual interpretation of large text datasets, making sentiment understanding scalable and more consistent across different applications.
The system's architecture, combining frontend frameworks such as Angular, React, and Next.js with backend NLP and classification components, achieved a seamless integration of sentiment analytics with user-friendly interfaces. The sentiment classification models were able to process and categorize text efficiently, providing meaningful patterns and trends from the input data. Prior work in sentiment analysis highlights the importance and applicability of such systems in interpreting public opinion and reactions across digital platforms.
Furthermore, *Sentilytics* offers extensibility beyond a single data source — it can handle multiple text formats and platforms, making it suitable for social media analysis, feedback mining, and domain- specific sentiment interpretation. The integration of comprehensive preprocessing steps improved the system's ability to handle noisy and informal text commonly found in real-world datasets, aligning with literature that emphasizes the need for robust text cleaning and representation techniques.
Overall, the project not only achieves its core objective of accurate sentiment classification but also provides a flexible software framework that can be adapted for future research or commercial deployment. Future enhancements may include the integration of more advanced deep learning models, multilingual support, and real-time analytics to further improve accuracy and broaden the range of potential applications

8. ACKNOWLEDGEMEN

The authors would like to express their sincere gratitude to all those who have provided invaluable support and guidance

throughout the development of the *Sentilytics* sentiment analysis project. Special thanks are extended to our project supervisor and mentors for their continuous advice, encouragement, and insightful feedback, which have greatly contributed to shaping the direction and outcomes of this work. We also acknowledge the support of our department and colleagues whose technical discussions and suggestions helped improve the system design and implementation. Their input was instrumental in refining the methodology and achieving the project objectives effectively.

In addition, we appreciate the assistance provided by peers and collaborators during data collection, system testing, and evaluation phases. Their cooperation ensured the successful validation of the sentiment analysis models and the overall software framework.

Finally, we are grateful for the encouragement and support of our family and friends throughout the research process. Their motivation and understanding were essential in completing this project.

9. REFERENCES

[1] S. Khomsah, *Sentiment Analysis on YouTube Comments Using Word2Vec and Random Forest*, Jurnal Telematika dan Teknologi Informasi, 18(1), 2019, pp. 1–10. DOI:10.31315/telematika.v18i1.4493.

[2] S. Thomas, Y. Yuliana, and P. Noviyanti, *Study Analysis of Sentiment Analysis Methods on YouTube*, Journal of Information Technology, 1(1), 2021, pp. 1–7.

[3] O. Uryupina, B. Plank, A. Severyn, A. Rotondi, and A. Moschitti, *SenTube: A Corpus for Sentiment Analysis on YouTube Social Media*, LREC, 2014.

[4] M. Sufyan, M. S. Ahmed, and A. Patel, *Sentiment Analysis With YouTube Comments Using Deep Learning*, International Journal of Information Technology and Computer Engineering, 13(2s), 2025.

[5] S. Wadhwani, *Sentiment Analysis of User YouTube Comments Using Classifier Algorithm*, International Journal of Innovations in Science, Engineering And Management, 2(1), 2023.

[6] A. A. S. and H. Rajeev, *YouTube Comment Sentimental Analysis*, Indian Journal of Data Mining, 4(1), 2024.

10. BIOGRAPHIES

1. Sarvesh Yogesh Wagh

Sarvesh Yogesh Wagh is a final-year Diploma student pursuing Information Technology at K. K. Wagh Polytechnic, Nashik, Maharashtra, India. His areas of interest include, Web devlopmetnt, Artificial Intelligence, Machine Learning, Deep learning. He has actively contributed to the system architecture design and overall implementation of the project, backend connectivity and helped in Each phase of development.

2. Somnath Vinod Jadhav

Somnath Vinod Jadhav is a Diploma student in Information Technology at K. K. Wagh Polytechnic, Nashik, Maharashtra, India. His technical interests include programming, and Learning new technologies. He worked on Gathering details realated to project and Features to be provided for the project.

3. Ashish Kailash Shewale

Ashish kialas Shewale is a Diploma student in Information Technology at K. K. Wagh Polytechnic, Nashik, Maharashtra, India. His technical interests include programming, and Learning, web devlopment.

4. Mrs Nilima Gite

Mrs Nilima Gite is a lecturer in the Information technology department at K. K. Wagh Polytechnic, Nashik, Maharashtra, India, Her areas of intrest include Artificial intelligence ,Machine Learning and Software engineering ,She provide Technical mentor ship and Guidance thoughout the project development