

The Public Pulse: A Unified NLP Framework for Sentiment, Topic, and Toxicity Analysis of YouTube Discourse

¹ Akansha Dudpuri, ² Anshika Tyagi, ³ Aditya Bahuguna, ⁴ Janhvi Singh, ⁵ Ramveer Singh

^{1,2,3,4} B.Tech Student, Dept. of Computer Science & Engineering, Raj Kumar Goel Institute of Technology, Ghaziabad, UP, India

⁵ Assistant Professor, Dept. of Computer Science & Engineering, Raj Kumar Goel Institute of Technology, Ghaziabad, UP, India

Abstract - Online platforms such as YouTube generate vast volumes of user-generated textual content that reflect public opinion, emotions, and behavioural patterns. Analysing this content is essential for understanding audience sentiment, dominant discussion themes, and the presence of toxic or harmful speech. Existing Natural Language Processing (NLP) approaches typically address sentiment analysis, topic modeling, and toxicity detection as independent tasks, results in fragmented insights and increased system complexity. This paper presents The Public Pulse, an integrated and interpretable NLP framework that unifies sentiment analysis, topic modeling, and toxicity detection within a single analytical pipeline. The system processes YouTube comments using lexicon-based sentiment analysis, Latent Dirichlet Allocation (LDA) for topic extraction, and rule-based toxicity detection, with results visualised through an interactive Streamlit dashboard. Experimental results demonstrate that the proposed approach provides coherent insights into public discourse while remaining computationally efficient and suitable for academic and resource constrained environments.

Key Words: Sentiment Analysis, Topic Modeling, Toxicity Detection, YouTube Comments, NLP, LDA, TextBlob

1. INTRODUCTION

The rapid expansion of social media platforms has fundamentally changed how individuals express opinions, emotions, and attitudes toward digital content. Among these platforms, YouTube has emerged as one of the most influential spaces for public discourse, enabling users to engage through comments on news, educational, entertainment, and socio-political videos. These user-generated comments represent a valuable source of unstructured textual data that reflects collective sentiment, discussion themes, and patterns of online behaviour. However, conventional engagement indicators such as views, likes, and shares provide only surface-level insights and fail to capture the emotional tone, topical focus, and presence of harmful or toxic language within discussions. Recent advances in Natural Language Processing (NLP) have enabled automated analysis of large-scale social media text, leading to increased adoption of sentiment analysis, topic modeling, and toxicity detection techniques for understanding online discourse [1], [8], [11]. Sentiment analysis has been widely applied to YouTube comments to classify user opinions into positive, negative, and neutral categories. Lexicon-based approaches, such as those implemented using TextBlob and similar resources, have demonstrated effectiveness in extracting sentiment polarity while maintaining computational efficiency

and interpretability, making them suitable for large-scale and real-time analysis [5], [9], [10].

Beyond sentiment polarity, topic modeling techniques have been employed to uncover latent themes within YouTube comment datasets. Latent Dirichlet Allocation (LDA) remains one of the most commonly used unsupervised topic modeling methods due to its ability to identify coherent discussion topics from noisy and unstructured text without requiring labelled data [4], [6], [7]. Prior studies applying LDA to YouTube comments have successfully revealed dominant themes related to public opinion, trending content, tourism perception, and ideological discourse, highlighting its relevance for contextual analysis of user interactions.

In parallel, the increasing prevalence of abusive and harmful language on social media platforms has driven significant research into toxicity and hate speech detection. rule based keyword filtering methods provide transparent and interpretable mechanisms for identifying explicit toxic expressions, while empirical studies show that toxicity often intensifies within polarized or high engagement discussions [1], [3]. Large-scale analyses further indicate that toxic language is frequently associated with specific topics and emotional states, reinforcing the need to jointly analyze sentiment, topic structure, and toxicity rather than treating them as independent dimensions [8]. Despite these advancements, existing approaches often focus on sentiment analysis, topic modeling, or toxicity detection as isolated tasks. Such fragmented analysis limits the ability to understand how emotional tone, discussion context, and harmful behaviour interact within online discourse [3], [11]. Additionally, many existing systems prioritize model complexity over interpretability, reducing their accessibility for practical deployment and exploratory analysis.

To address these limitations, this paper presents The Public Pulse, a unified NLP-based framework for comprehensive analysis of YouTube comments. The proposed system integrates lexicon-based sentiment classification, LDA-based topic modeling, and rule-based toxicity detection within a single analytical pipeline. User generated comments undergo structured preprocessing, followed by sentiment scoring, topic extraction, and toxicity identification. The analytical outputs are presented through visualizations that enable intuitive exploration of sentiment distributions, topic trends, and toxicity patterns.

By consolidating multiple NLP techniques into a modular and scalable framework, The Public Pulse provides an interpretable and practical approach for analysing large-scale YouTube comment data. The system supports applications in media analytics, academic research, and online discourse monitoring, offering a balanced solution that emphasizes transparency,

efficiency, and integrated insight into public sentiment and behaviour.

2. LITERATURE REVIEW

This section presents a comprehensive review of existing research related to sentiment analysis, topic modeling, and toxicity detection in social media discourse, with a specific focus on YouTube comments. The objective is to highlight key methodological approaches, identify existing research gaps, and establish the foundation for the proposed Public Pulse framework.

Sentiment analysis of YouTube comments has been widely studied using lexicon-based and machine learning approaches. Several works demonstrate that lexicon-driven methods, such as TextBlob and VADER, are effective in classifying comments into positive, negative, and neutral categories while maintaining simplicity, transparency, and low computational overhead [5], [9], [10]. These characteristics make lexicon based techniques particularly suitable for large-scale social media analysis where real-time processing and interpretability are essential.

Beyond basic sentiment classification, multiple studies have explored automated sentiment feedback systems for YouTube videos. Approaches leveraging lexicon-based and lightweight machine learning models have shown promising results in capturing audience opinion trends without requiring extensive labeled datasets [9], [10]. However, while effective for polarity detection, such methods often lack contextual awareness and thematic understanding.

Topic modeling has been extensively employed to uncover latent discussion themes in YouTube comment datasets. Latent Dirichlet Allocation (LDA) remains one of the most widely adopted unsupervised topic modeling techniques due to its interpretability and robustness when handling noisy, unstructured text [4], [6], [7]. Studies applying LDA to YouTube comments demonstrate its ability to identify dominant themes across domains such as tourism, trending content, and public opinion, provided that adequate preprocessing and parameter tuning are applied.

Research focusing on harmful and sensitive content has combined topic modeling with sentiment analysis to identify patterns of abusive or extremist discourse. Topic-based clustering has been shown to reveal structured narratives surrounding hate speech, enabling a deeper understanding of how toxic content evolves within specific discussion themes [1], [4]. These studies emphasize that sentiment polarity alone is insufficient for identifying harmful behaviour without thematic context.

Toxicity and hate speech detection form another critical area of YouTube comment analysis. Rule-based approaches using predefined keyword dictionaries have been proposed as transparent and computationally efficient baselines for identifying explicit abusive language [1]. While effective for detecting overt toxicity, such methods often struggle with implicit abuse, sarcasm, and context-dependent expressions.

To overcome these limitations, large-scale empirical studies have employed deep learning models for toxicity detection. These approaches demonstrate significantly higher accuracy and recall when identifying abusive content and reveal strong correlations between certain discussion topics and elevated toxicity levels [3], [8]. However, deep learning models require substantial labeled data, computational resources, and complex training pipelines, limiting their suitability for lightweight and easily deployable systems.

Several studies have proposed integrated analytical pipelines that combine sentiment analysis, topic modeling, and toxicity detection to gain richer insights into online discourse. Unified frameworks have shown that toxic comments frequently cluster around specific topics and are often associated with negative sentiment, providing more meaningful interpretations than isolated analytical methods [1], [8]. These findings highlight the importance of multi dimensional analysis for understanding user behavior on social media platforms.

Visualization has also been identified as a crucial component for interpreting analytical results. Prior research demonstrates that bar charts, temporal plots, and word clouds are effective tools for representing sentiment distributions, topic prevalence, and discourse evolution over time [6], [7], [11]. Visualization driven analysis enhances interpretability and supports non technical users in understanding complex NLP outputs.

Despite significant advancements, existing studies reveal several limitations. Most systems focus on individual analytical tasks rather than providing a unified and user friendly framework. Additionally, many implementations are designed primarily for academic evaluation and lack interactive interfaces suitable for practical deployment. Limited emphasis is placed on integrating sentiment, thematic context, and toxicity analysis within a single scalable pipeline.

In summary, while prior research confirms the effectiveness of sentiment analysis, topic modeling, and toxicity detection for YouTube comment analysis, there remains a clear gap in unified, interpretable, and accessible systems that integrate all three dimensions. The Public Pulse addresses this gap by combining lexicon-based sentiment analysis, LDA-based topic modeling, and rule-based toxicity detection into a modular framework supported by interactive visualizations, thereby extending existing research into a practical tool for large-scale public discourse analysis.

3. DATASET DESCRIPTION

The dataset used in this study was obtained from the Kaggle platform and consists of a large collection of publicly available YouTube comments related to high profile interview content. The dataset contains 50,000 textual comments extracted from two viral interviews published on the *Piers Morgan Uncensored* YouTube channel. These comments reflect a wide range of public reactions, opinions, and emotional expressions, making the dataset suitable for large scale textual analysis.

The dataset is designed to support multiple Natural Language Processing (NLP) tasks such as sentiment analysis, topic modeling, and toxicity detection. The diversity of opinions and linguistic styles present in the comments provides a realistic and challenging corpus for training and evaluating text-based analytical models.

The dataset includes both neutral and emotionally charged content, capturing positive, negative, and controversial reactions. Additionally, the dataset maintains anonymized user identifiers to preserve privacy while allowing behavioral analysis of engagement patterns. The balanced representation of comment polarity and engagement metrics helps in minimizing model bias and enables fair evaluation of classification performance.

DATASET ATTRIBUTES:

Each entry in the dataset consists of the following attributes:

TABLE -1: DATASET FEATURES

Attribute Name	Description
Comment	Raw user-generated text data
Anonymized Author	Privacy-preserved user identifier
Published At	Comment posting time
Likes	Audience engagement metric
Reply Count	Interaction level indicator

The dataset's scale, diversity, and real-world nature make it highly suitable for evaluating unified analytical frameworks that aim to extract sentiment trends, discussion topics, and toxic language patterns from online discourse. Its structured format and rich engagement metadata further enhance its applicability for visualization, temporal analysis, and interpretability of public opinion dynamics.

4. METHODOLOGY

The methodology adopted in The Public Pulse follows a systematic and modular Natural Language Processing (NLP) pipeline designed to analyse large-scale YouTube comment data. The overall approach integrates sentiment analysis, topic modeling, and toxicity detection into a unified framework, enabling comprehensive analysis of public discourse.

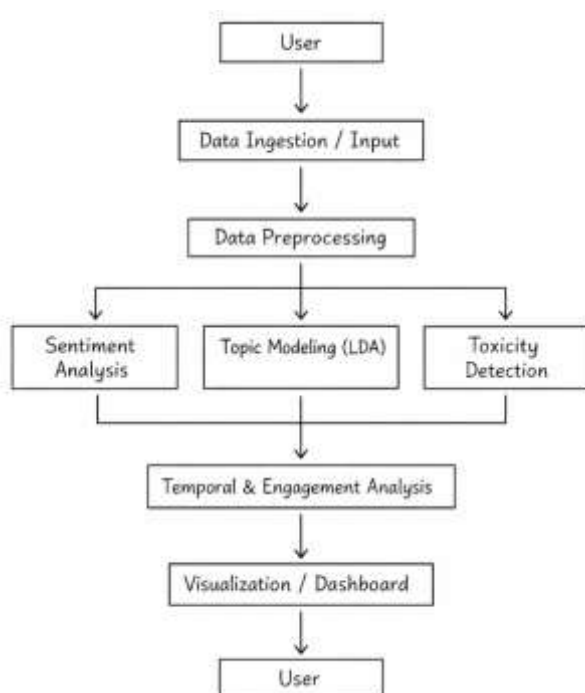


Fig. 1. Workflow of the proposed system

The methodology consists of multiple sequential stages, starting from raw data ingestion to visualization and interpretation of analytical results.

A. Data Collection and Input Handling

The dataset comprises 50,000 YouTube comments collected from two viral interviews hosted on the Piers Morgan Uncensored channel. The comments are stored in a structured CSV format and include textual content along with engagement related metadata such as timestamps, likes, and reply counts. The dataset is loaded into the system through a user-driven interface, allowing flexible selection and exploration of data.

B. Data Preprocessing

Raw YouTube comments often contain noise such as emojis, URLs, special characters, and inconsistent casing. To ensure analytical accuracy, the following preprocessing steps are applied:

- Conversion of text to lowercase
- Removal of URLs, HTML tags, and special characters
- Tokenization of text into individual words
- Stop word removal to eliminate non informative terms
- Lemmatization to normalize word forms

These preprocessing steps prepare the textual data for downstream NLP tasks while preserving meaningful linguistic patterns.

C. Sentiment Analysis

Sentiment analysis is performed using the TextBlob library, which applies a lexicon-based approach to compute polarity scores for each comment. Based on the polarity value, comments are classified into three sentiment categories:

- Positive
- Negative
- Neutral

This approach is chosen for its interpretability, computational efficiency, and suitability for real-time analysis of informal social media text. Sentiment scores are stored for further aggregation and temporal trend analysis.

D. Topic Modeling

To identify dominant discussion themes, Latent Dirichlet Allocation (LDA) is employed for topic modeling. After preprocessing, a dictionary and document term matrix are constructed using the Gensim library. The LDA model is then trained to extract latent topics from the corpus.

Each topic is represented by a set of high probability keywords, enabling interpretation of prevalent discussion themes. The number of topics is selected empirically to balance topic coherence and interpretability. Topic distributions are computed for individual comments and aggregated to identify dominant themes.

E. Toxicity Detection

Toxicity detection is implemented using a rule-based keyword matching approach. A curated dictionary of offensive, abusive, and hate related terms is used to scan comment text. Comments containing one or more flagged terms are marked as toxic.

This transparent approach allows users to understand the basis of toxicity classification while maintaining low computational overhead. Toxicity scores are stored alongside sentiment and topic outputs for integrated analysis.

F. Temporal and Engagement Based Analysis

Using the comment timestamp metadata, sentiment and toxicity scores are aggregated over time to analyze trends and fluctuations in public opinion. Engagement metrics such as likes and reply counts are also correlated with sentiment and toxicity levels to study audience interaction patterns.

G. Visualization and Result Interpretation

Analytical results are visualized using Matplotlib, Seaborn, and WordCloud libraries. The visualization module generates:

- Bar charts for sentiment distribution
- Line graphs for sentiment trends over time
- Word clouds for dominant topic keywords
- Topic frequency and toxicity summaries

All visual outputs are presented through an interactive Streamlit-based dashboard, enabling users to explore results dynamically.

H. System Architecture Overview

The overall system follows a modular architecture where each analytical component : sentiment analysis, topic modeling, and toxicity detection, operates independently but contributes to a unified output pipeline. This design ensures scalability, flexibility, and ease of future enhancements, such as incorporating machine learning-based classifiers.

5. EXPERIMENTAL SETUP

The experimental setup of *The Public Pulse* system focuses on implementing sentiment analysis, topic modeling, and toxicity detection on a large-scale YouTube comment dataset. The experiments were conducted using Python-based Natural Language Processing (NLP) techniques in a modular pipeline. The dataset was first preprocessed to remove noise such as URLs, punctuation, emojis, and stopwords, followed by text normalization. Cleaned comments were then analyzed using three parallel modules. Sentiment analysis was performed using the TextBlob library, where polarity scores were used to classify comments into positive, negative, and neutral categories. Topic modeling was carried out using Latent Dirichlet Allocation (LDA) to extract dominant discussion themes from the comment corpus. Toxicity detection was implemented using a rule-based lexicon approach to identify offensive or abusive language.

The outputs from all analytical modules were aggregated and visualized through an interactive dashboard developed using Streamlit. The experiments were executed on a standard desktop environment using Python 3.x and common NLP libraries, ensuring computational efficiency and reproducibility. This setup demonstrates the feasibility of analyzing large volumes of social media text using lightweight and interpretable models.

6. RESULTS AND ANALYSIS

This section summarizes the experimental outcomes of *the Public Pulse* system, evaluating the performance of sentiment analysis, topic modeling, and toxicity detection on YouTube

comments. The sentiment analysis module categorizes comments into **positive, negative, and neutral** classes using TextBlob polarity scores. The results indicate a **higher proportion of neutral and negative sentiments**, reflecting the argumentative and opinion centric nature of discussions surrounding controversial interviews. The analysis demonstrates that lexicon based methods can effectively capture overall emotional trends in large scale social media data.

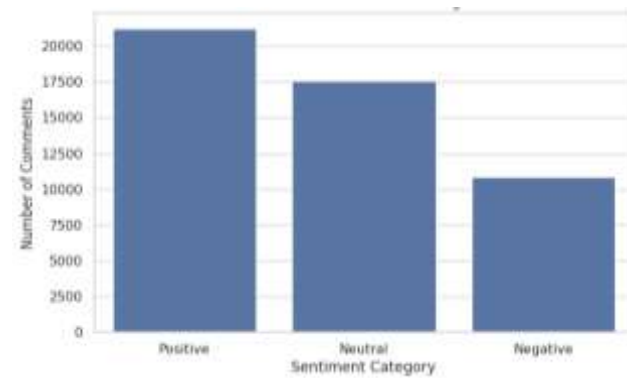


Fig. 1. Sentiment distribution of YouTube comments

A. Topic Modeling

Latent Dirichlet Allocation (LDA) was applied to uncover dominant discussion themes from the dataset. The model successfully extracted **coherent topics** representing public opinion, media discourse, ideological debate, and personal commentary.

TABLE II. EXTRACTED TOPICS WITH TOP KEYWORDS.

Topic ID	Top Keywords
T1	Israel, Hamas, Palestinians, conflict, people
T2	tate, Andrew, piers, interview, media
T3	women, men, gender, work, debate
T4	free, palestine, Zionist, UK
T5	people, opinion, discussion, public

B. Toxicity Detection

The toxicity detection module identified comments containing abusive or harmful language using keyword based indicators. A significant fraction of comments was flagged as toxic, particularly within emotionally charged topics

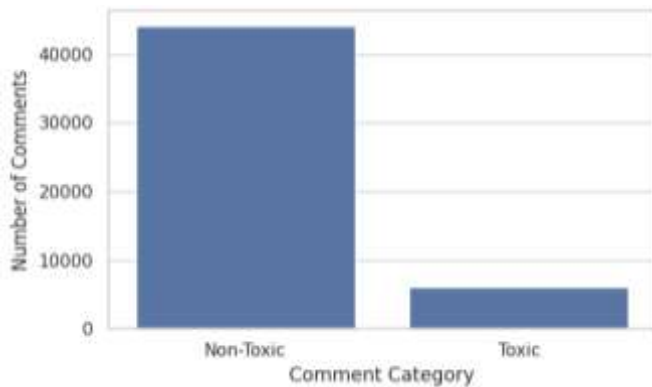


Fig. 2. Distribution of toxic and non-toxic comments.

C. Integrated Analysis

This presents a **topic-wise comparison of average sentiment polarity and toxicity levels**, integrating outputs from all analytical modules. The results show that topics associated with political conflict and ideological debate exhibit **lower sentiment scores and higher toxicity rates**, whereas informational or personality-focused topics tend to maintain **neutral sentiment with comparatively lower toxicity**. This demonstrates a clear inverse relationship between sentiment polarity and toxicity intensity, highlighting the importance of joint analysis.

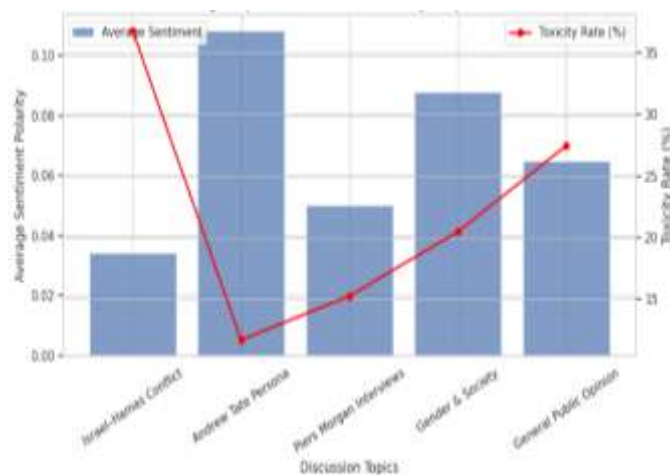


Fig. 3. Topic-wise Sentiment and Toxicity Comparison.

D. Discussion

The experimental results confirm that The Public Pulse provides a coherent and interpretable framework for online discourse analysis. While lexicon-based sentiment analysis and LDA topic modeling have inherent limitations in handling sarcasm and contextual nuance, their integration with toxicity detection enables deeper behavioural insights. The proposed system achieves an effective balance between **computational efficiency, interpretability, and analytical depth**, making it suitable for large-scale social media monitoring and research applications.

7. CONCLUSION

This paper presented *The Public Pulse*, a unified Natural Language Processing framework for analyzing large-scale YouTube comment data. By integrating sentiment analysis, topic modeling, and toxicity detection into a single analytical pipeline, the system provides a structured and multidimensional understanding of online public discourse. Experimental results demonstrate that lexicon-based sentiment analysis combined with LDA-based topic modeling and rule-based toxicity detection can effectively capture emotional trends, dominant discussion themes, and patterns of hostile language in high-engagement digital environments. The findings highlight strong associations between negative sentiment and toxic expressions, emphasizing the need for integrated analytical approaches when studying polarized online discussions. Overall, *The Public Pulse* offers an efficient, interpretable, and scalable solution for social media discourse analysis.

REFERENCES

- [1] S. Shekhar and A. Saini, "Utilizing topic modelling to identify abusive comments on YouTube," in Proc. 2021 Int. Conf. Intelligent Technologies (CONIT), Hubli, India, 2021, pp. 1–5, doi: 10.1109/CONIT51480.2021.9498368.
- [2] H. M. Alawadh, A. Alabrah, T. Meraj, H. T. Rauf, A. S. Alghamdi, S. S. Alshamrani, and H. A. Alshamrani, "English language learning via YouTube: An NLP based analysis of users' comments," Computers, vol. 12, no. 2, p. 24, 2023.
- [3] U. Ghosh, S. Dutta, I. Mondal, A. Nandi, and A. K. Das, "Tracking the tides of toxicity: Temporal shifts in public emotion and discourse during the COVID 19 crisis," in Proc. IEEE Int. Conf. Advances in Computing Research on Science Engineering and Technology (ACROSET), Indore, India, 2025, pp. 1–6, doi: 10.1109/ACROSET66531.2025.11280719.
- [4] J. P. Latorre and J. J. Amores, "Topic modelling of racist and xenophobic YouTube comments: Analyzing hate speech against migrants and refugees," in Proc. 5th Int. Conf. Natural Language and Speech Processing, 2021, pp. 456–460, doi: 10.1145/3486011.3486494.
- [5] R. Pradhan, "Extracting sentiments from YouTube comments," in Proc. 6th Int. Conf. Image Information Processing (ICIIP), Shimla, India, 2021, pp. 1–4, doi: 10.1109/ICIIP53038.2021.9702561.
- [6] K. R. R. S. Thilakarathne, B. T. G. S. Kumara, and B. Kuhaneswaran, "Analyzing tourists' perceptions of tourism destinations using YouTube comments," in Proc. Int. Conf. Data Analytics for Business and Industry (ICDABI), Sakheer, Bahrain, 2021, pp. 301–305, doi: 10.1109/ICDABI53623.2021.9655833.

[7] O. Endait, M. Iyer, U. Somashetti, K. Zoting, and D. Bhagat, "YouTube trending video analysis using natural language processing and machine learning," in Proc. Int. Conf. Sustainability, Innovation & Technology (ICSIT), Nagpur, India, 2025, pp. 1–5, doi: 10.1109/ICSIT65336.2025.11295060.

[8] V. Su and N. Thakur, "COVID-19 on YouTube: A data driven analysis of sentiment, toxicity, and content recommendations," in Proc. IEEE 15th Annual Computing and Communication Workshop and Conf. (CCWC), Las Vegas, NV, USA, 2025, pp. 715–723, doi: 10.1109/CCWC62904.2025.10903713.

[9] P. Dhanalakshmi, B. A. Raj, A. R. Ashreen, B. B. Kalyan, A. A. Kumar, and N. Balakrishna, "Automated sentiment analysis for instant feedback on YouTube videos through comments," in Proc. Control Instrumentation System Conf. (CISCON), Manipal, India, 2024, pp. 1–6, doi: 10.1109/CISCON62171.2024.10696587.

[10] A. Dumpawar, A. Raut, A. Bais, R. Agrawal, C. Dhule, and P. K. Dhankar, "Real time sentiment analysis of YouTube comments using NLP and TextBlob via YouTube Data API," in Proc. Int. Conf. Engineering Innovations and Technologies (ICoEIT), Bhopal, India, 2025, pp. 325–330, doi: 10.1109/ICoEIT63558.2025.11211539.

[11] P. Mironela, P. Iustin, P. C. Mihaela, P. Claudiu, and G. L. Daniel, "Analysis of YouTube video comments with NLP methods," in Proc. 16th Int. Conf. Electronics, Computers and Artificial Intelligence (ECAI), Iasi, Romania, 2024, pp. 1–6, doi: 10.1109/ECAI61503.2024.10607575.