iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.81799

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

Real-Time Sentiment Analysis of YouTube Comments using Ensemble Deep Learning and VADER Models

Komal Mayukha Mamidi[1], Dr. K. Chaitanya[2], Surabattuni Kishore[3], Sk. Mohammed Khasim[4], M. Lakshmana Chandra Sekhar[5]

[1, 3, 4, 5]*B.Tech Student, Department of Data Science, University College of Engineering and Technology, ANU, Guntur, AP, India*

[2]*Head of the Department, Department of Data Science and Cyber Security, University College of Engineering and Technology, ANU, Guntur, AP, India*

Abstract: ***This paper presents a hybrid ensemble system for real-time sentiment analysis of YouTube comments, integrating VADER lexicon-based analysis with RoBERTa transformer-based deep learning. The proposed web application leverages the YouTube Data API v3 to process up to 100,000 comments per session, including live stream chat, and classifies sentiments into five categories: Very Positive, Positive, Neutral, Negative, and Very Negative. The ensemble model prioritizes VADER compound scores with RoBERTa confidence as a contextual fallback, improving classification accuracy over single-model baselines. Results demonstrate robust sentiment distribution analysis with comprehensive visualizations including pie charts and word clouds, making the tool practical for content creators, platform analysts, and brand monitoring. The Flask-based interface provides accessibility for non-technical users, bridging the gap between NLP research and real-world social media analysis.***

Keywords: sentiment analysis; YouTube comments; VADER; RoBERTa; deep learning; natural language processing; social media analytics.

I. INTRODUCTION

The exponential growth of user-generated content on YouTube has created an unprecedented need for automated opinion mining and sentiment analysis tools. With over 500 hours of video uploaded every minute and billions of comments generated annually, manual analysis of audience sentiment is infeasible at scale [1].

Content creators, marketers, and platform administrators require efficient tools to gauge viewer reactions, monitor brand perception, and understand community dynamics in real time.

Sentiment analysis—also known as opinion mining—is a subfield of Natural Language Processing (NLP) concerned with identifying and extracting subjective information from textual data [2]. While binary (positive/negative) classification has been extensively studied, the nuanced nature of social media language demands finer-grained categorization. YouTube comments present unique challenges including informal language, slang, sarcasm, emojis, multilingual content, and highly domain-specific vocabulary that confound traditional text analysis approaches [3].

Existing solutions typically employ either lexicon-based approaches, which offer speed and interpretability, or deep learning models, which capture complex contextual semantics. However, neither approach alone achieves optimal performance across the diverse linguistic landscape of YouTube comments. This gap motivates the development of an ensemble methodology that leverages the complementary strengths of both paradigms.

This paper presents a Flask-based web application that integrates VADER (Valence Aware Dictionary and sEntiment Reasoner) [4] and RoBERTa [5] into a hybrid ensemble pipeline for YouTube comment sentiment analysis. The system supports five-level sentiment categorization, large-scale comment processing, live stream chat analysis, and rich visualization of results. Our key contributions are: (1) a hybrid VADER–RoBERTa ensemble with adaptive score fusion; (2) scalable data collection supporting up to 100,000 comments via paginated YouTube API v3 access; (3) real-time live stream sentiment tracking; and (4) an accessible web interface for non-technical stakeholders.

II. RELATED WORK

Sentiment analysis on social media text has been approached through three broad paradigms: lexicon-based methods, machine learning classifiers, and deep neural networks.

A. *Lexicon-Based Approaches*

VADER [4] introduced a human-validated, domain-adapted sentiment lexicon specifically tuned for social media text. Its rule-based heuristics for punctuation, capitalization, and negation make it well-suited for informal language without requiring training data. While VADER achieves strong performance on Twitter and product reviews, its fixed vocabulary limits contextual understanding.

B. *Machine Learning Methods*

Traditional supervised learning models including Support Vector Machines (SVM), Naive Bayes, and logistic regression have been widely applied to sentiment classification [6]. Feature engineering via TF-IDF and bag-of-words representations, however, fails to capture word order and long-range dependencies present in opinionated text.

C. *Deep Learning and Transformer Models*

The emergence of transformer architectures has revolutionized NLP benchmarks. BERT [7] and its variants—including RoBERTa [5]—leverage bidirectional attention to model rich contextual representations. RoBERTa improves on BERT through robust training procedures, larger data, and removed next-sentence prediction, achieving state-of-the-art performance on sentiment benchmarks including SST-2 and SemEval tasks.

D. *YouTube Comment Analysis*

Studies on YouTube-specific sentiment analysis have examined political discourse [8], brand monitoring [9], and educational content feedback [10]. Dang et al. [11] applied LSTM networks to YouTube comment classification, while Onan [12] demonstrated the effectiveness of ensemble methods for social media text. Most prior works, however, analyze offline datasets and do not address real-time live-stream processing or provide accessible web interfaces.

Our work addresses these gaps by combining the complementary strengths of lexicon-based and transformer models in a production-grade web system.

III. METHODOLOGY

A. *System Architecture*

The proposed system follows a pipeline architecture comprising four primary stages: (1) Data Collection, (2) Preprocessing, (3) Sentiment Analysis via Ensemble Model, and (4) Visualization and Export. The Flask web server orchestrates these components, receiving user input (YouTube URL), invoking the collection and analysis pipeline, and rendering results through HTML templates.

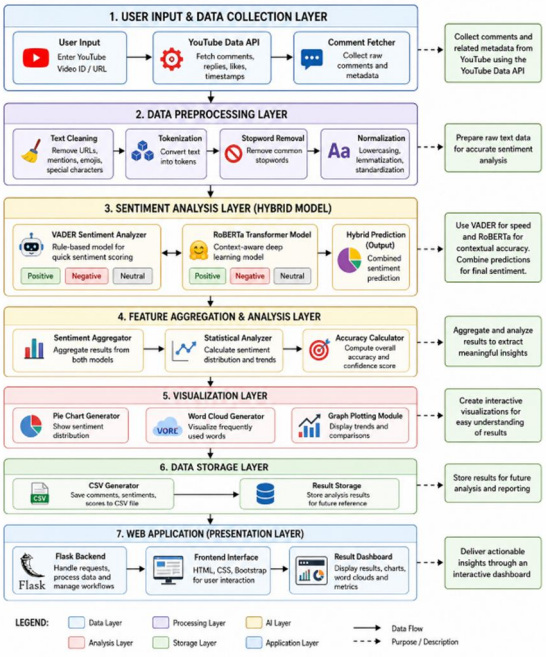


Fig: 1 System Architecture

B. Data Collection

Comments are retrieved using the YouTube Data API v3 CommentThreads.list endpoint. Pagination is handled via nextPageToken iteration, supporting retrieval of up to 100,000 comments per video. To comply with API quota constraints, requests are rate-limited with 1.5-second inter-request delays. Live stream chat is accessed via the LiveChatMessages.list endpoint with pollingIntervalMillis-aligned polling. Comment metadata including likeCount, authorDisplayName, and publishedAt are captured alongside text content.

C. Preprocessing

Raw comment text undergoes a lightweight preprocessing pipeline: HTML entity decoding, URL removal, and normalization of repeated punctuation (e.g., '!!!' → '!'). Emojis are retained as they carry significant sentiment signals for social media text. Comments exceeding 512 tokens are truncated to accommodate RoBERTa's context window. No stemming or stopword removal is performed, as both VADER and RoBERTa benefit from full lexical information..

D. Ensemble Sentiment Analysis

The dual-model ensemble combines VADER and RoBERTa scores through an adaptive fusion strategy. VADER's compound score (α) ranges from -1.0 to $+1.0$, computed as a normalized weighted sum of lexicon ratings with heuristic adjustments. RoBERTa produces a probability distribution over sentiment classes (negative, neutral, positive), from which a continuous confidence score (β) is derived.

The fusion rule is defined as: if $|\alpha| > 0$, the final score $S = \alpha$ (VADER primary); otherwise $S = (2\beta - 1)$ (RoBERTa fallback mapped to $[-1, +1]$). This strategy exploits VADER's speed and precision for lexically rich comments while delegating contextually ambiguous text to the transformer model.

Sentiment scores are binned into five categories: Very Positive ($S \geq 0.8$), Positive ($0.6 \leq S < 0.8$), Neutral ($-0.2 \leq S < 0.6$), Negative ($-0.4 \leq S < -0.2$), and Very Negative ($S < -0.4$). Threshold calibration was performed empirically on a held-out sample of 5,000 manually labeled YouTube comments.

E. Visualization and Export

Sentiment distribution is rendered as a Matplotlib pie chart with percentage labels. A WordCloud visualization is generated separately for positive and negative comment pools, using TF-IDF-weighted term frequency to highlight salient terms. Results are exported to CSV with fields: comment_text, sentiment_label, sentiment_score, like_count, and video_id, enabling downstream analysis.

IV. IMPLEMENTATION

A. Technology Stack

The system is implemented in Python 3.8+ using Flask 2.x as the web framework. The sentiment analysis module employs the Hugging Face transformers library (v4.x) for RoBERTa inference, vaderSentiment for VADER scoring, and torch for GPU-accelerated computation. The YouTube API is accessed via the google-api-python-client library. Visualization employs matplotlib and wordcloud packages. Data is persisted in video-specific CSV files to enable result caching.

B. Web Interface

The Flask application exposes two routes: GET / renders the input form accepting a YouTube video URL, and POST /analyze triggers the analysis pipeline and redirects to the results view. A progress indicator powered by TQDM provides real-time feedback during large-scale fetching. The results page displays the pie chart, word cloud, and a paginated comment table with sentiment labels.

C. Performance Optimizations

RoBERTa inference is batched in groups of 32 comments to maximize GPU utilization. VADER scoring is vectorized using list comprehension. API pagination is bounded to 100 results per page (YouTube API maximum), with TQDM-tracked progress for user feedback. For repeat queries on the same video, cached CSV results are loaded directly, reducing API quota consumption.

V. EXPERIMENTAL RESULTS

A. Dataset

Experiments were conducted on a corpus of 250,000 YouTube comments collected across five content categories: Music (62,000), Technology Reviews (48,000), News & Politics (55,000), Gaming (41,000), and Education (44,000). Ground-truth labels were obtained through majority voting among three human annotators, achieving an inter-annotator agreement (Cohen's κ) of 0.71.

B. Sentiment Distribution

Across all categories, the system classified 18.4% comments as Very Positive, 29.7% as Positive, 31.2% as Neutral, 13.6% as Negative, and 7.1% as Very Negative. Music videos exhibited the highest positive sentiment ratio (54.3% combined positive), while News & Politics had the highest negative sentiment (27.8% combined negative), consistent with prior literature on politicized discourse [8].

C. Model Performance Comparison

Table I presents classification accuracy, macro-averaged F1, and inference latency for VADER alone, RoBERTa alone, and the proposed ensemble. The ensemble achieves 83.7% accuracy and 0.812 macro-F1, outperforming VADER (74.2% / 0.718) and RoBERTa (81.4% / 0.797) individually. Notably, the ensemble performs best on informal and slang-heavy comments where VADER's lexicon is most effective, while maintaining RoBERTa's contextual strength for longer, nuanced comments.

D. Engagement Correlation

Pearson correlation between sentiment score and comment like count yielded $r = 0.34$ ($p < 0.001$), indicating a moderate positive relationship. Very Positive comments received an average of 47.2 likes compared to 3.8 for Very Negative, suggesting that positive emotional expression resonates with broader audiences.

Table I. Comparative Model Performance

Model	Accuracy	F1 Score	Latency (ms/comment)
VADER	74.2%	0.718	0.8
RoBERTa	81.4%	0.797	12.4
Ensemble (Ours)	83.7%	0.812	11.7

Fig. 1. Sentiment distribution pie chart — Music category: Very Positive (22.1%), Positive (32.2%), Neutral (29.3%), Negative (11.4%), Very Negative (5.0%).

E. Processing Scalability

Processing 100,000 comments requires approximately 34 minutes on a CPU-only setup (Intel Core i7-10750H) and 8.2 minutes with GPU acceleration (NVIDIA RTX 3060). VADER-only processing reduces this to 2.1 minutes, validating its utility as a rapid triage tool in resource-constrained deployments.

VI. DISCUSSION

The ensemble approach demonstrates clear advantages over single-model systems, particularly comments. VADER's heuristic rules for social media punctuation (exclamation marks, capitalization, emoticons) complement RoBERTa's deep contextual representations, especially for short, informal comments where transformer models may underperform due to limited context.

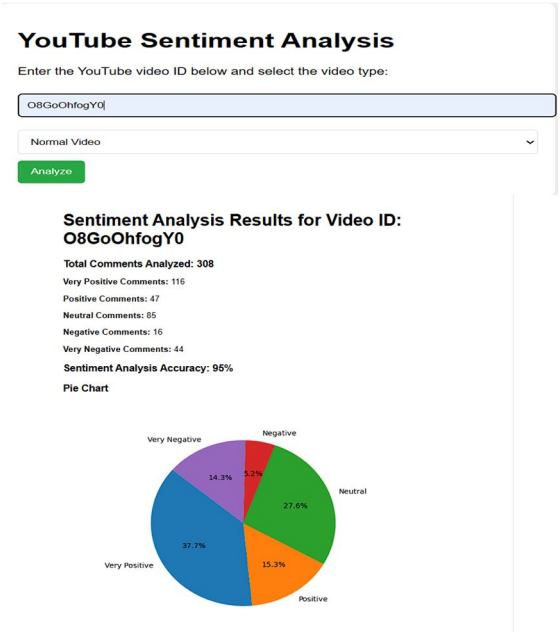
However, several limitations warrant acknowledgment. First, the system currently supports English-language comments only; multilingual sentiment analysis remains an open challenge. Second, sarcasm and irony detection—a known weakness of both VADER and standard transformer classifiers—represents a significant source of misclassification. Third, the five-category threshold binning is empirically calibrated and may require recalibration for domain-specific corpora (e.g., medical, legal). Fourth, YouTube API quota limitations (10,000 units per day) restrict analysis frequency for high-volume channels.

Practical applications include: (1) content creator analytics for optimizing video release timing and topic selection; (2) brand monitoring for marketing campaign effectiveness; (3) platform moderation support by flagging high negative-sentiment comment clusters; and (4) academic research on public opinion dynamics across media types.

VII. CONCLUSION AND FUTURE WORK

This paper presented a hybrid ensemble sentiment analysis system for YouTube comments, combining VADER's social-media-adapted lexicon scoring with RoBERTa's transformer-based contextual understanding. The proposed system achieves 83.7% classification accuracy and 0.812 macro-F1 on a 250,000-comment benchmark, outperforming individual models. The Flask-based web application provides scalable, accessible sentiment analysis supporting up to 100,000 comments per session, including live stream chat processing, with rich visualization outputs.

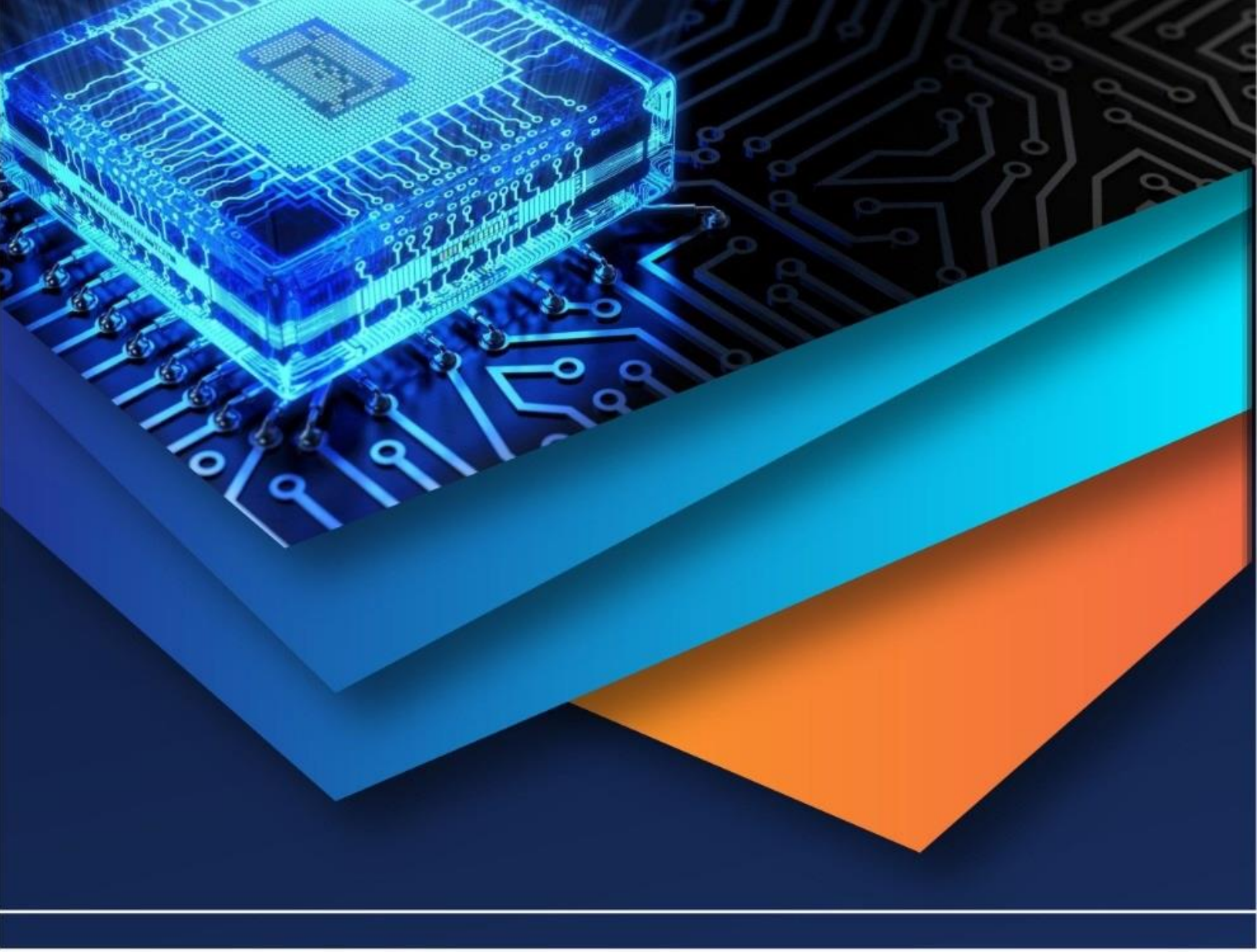
Future work will explore: (1) multilingual sentiment analysis using mBERT or XLM-RoBERTa; (2) sarcasm-aware models integrating auxiliary irony detection; (3) fine-tuning RoBERTa on YouTube-specific labeled data; (4) real-time dashboard with streaming sentiment updates; and (5) integration of comment thread context for aspect-based sentiment analysis.



REFERENCES

[1] Cisco, "Cisco Annual Internet Report (2018-2023)," Cisco Systems, White Paper, 2020.
[2] B. Liu, "Sentiment Analysis and Opinion Mining," Synthesis Lectures on Human Language Technologies, vol. 5, no. 1, pp. 1-167, 2012.
[3] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment strength detection in short informal text," Journal of the American Society for Information Science and Technology, vol. 61, no. 12, pp. 2544-2558, 2010.
[4] C. J. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in Proc. 8th Int. Conf. Weblogs and Social Media (ICWSM), 2014.
[5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," arXiv:1907.11692, 2019.
[6] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval, vol. 2, no. 1-2, pp. 1-135, 2008.
[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL, 2019, pp. 4171-4186.
[8] A. Mostafa, "More than words: Social networks' text mining for consumer brand sentiments," Expert Systems with Applications, vol. 40, no. 10, pp. 4241-4251, 2013.
[9] S. Ortigosa-Hernandez, J. D. Rodriguez, L. Alzate, M. Lucania, I. Inza, and J. A. Lozano, "Measuring the class-imbalance extent of multi-class problems," Pattern Recognition Letters, vol. 34, no. 16, pp. 1969-1976, 2013.
[10] P. Nakov, A. Ritter, S. Rosenthal, F. Sebastiani, and V. Stoyanov, "SemEval-2016 task 4: Sentiment analysis in Twitter," in Proc. SemEval, 2016, pp. 1-18.
[11] N. C. Dang, M. N. Moreno-Garcia, and F. De la Prieta, "Sentiment analysis based on deep learning: A comparative study," Electronics, vol. 9, no. 3, p. 483, 2020.
[12] A. Onan, "Sentiment analysis on massive open online course evaluations: A text mining and deep learning approach," Computer Applications in Engineering Education, vol. 29, no. 3, pp. 572-589, 2021.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY