Ianna Journal of Interdisciplinary Studies, EISSN: 2735-9891
Volume 8, Issue 1, 2026

EVALUATION THE ML AND DEEP LEARNING MODELS FOR AUTOMATED SENTIMENT CLASSIFICATION OF YOUTUBE COMMENTS

Maulik Narendrakumar Pandya, Animesh Kumar Agrawal [0000-0003-2825-8321]
Unitedworld Institute of Technology, Karnavati University, Gandhinagar
202406010004@karnavatiuniversity.edu.in, akag9906@gmail.com

ABSTRACT

YouTube's comments provide a wealth of sentiment information for different video categories. However, the sheer volume of data and the presence of noise make the analysis a challenging task. In this study, machine learning and deep learning methods are used for automatic sentiment analysis of YouTube comments in several categories. Due to the increasing amount of user-generated content on YouTube, manual analysis of audience sentiment becomes increasingly challenging and time-consuming. Hence, in this paper, an efficient approach is proposed to analyse sentiments in YouTube comments based on natural language processing (NLP), machine learning (ML), and deep learning (DL) algorithms. In order to train the model, a dataset consisting of 50,000 YouTube comments belonging to different classes such as educational, entertainment, music, news, and games was collected using the YouTube Data API. The comments were processed through several techniques, including tokenisation, lemmatisation, removal of stopwords, conversion of slang words, and class balancing via SMOTE. In addition, several ML algorithms, including Random Forest, SVM and DL models such as CNN, BiLSTM, and BERT, were applied to perform the experiment. However, the results of these experiments showed that BERT transformers' algorithm generated the highest accuracy and macro F1 of 92.5% and 0.92, respectively, outperforming the rest of the machine learning algorithms. Optimal tuning of hyperparameters was vital for boosting accuracy, while the highest accuracy was achieved by BERT + RF (93.8%).

Keywords: - Deep Learning, Machine Learning, YouTube, Sentiment Analysis, Comments

1. INTRODUCTION

YouTube being one of the largest video content platforms and an interactive community, produces a huge amount of comments and videos which express various feelings of the audience. It is important to know about the audience sentiments for content creators and researchers, but analyzing such huge unstructured data manually becomes difficult (Poojitha, M. M., & Ganupriya, V., 2025).

YouTube is the leading video-sharing platform globally, with more than 2.5 billion users. The comments generated by such users are huge in number, representing the voices of the audience in a wide variety of topics like education, entertainment, and news (Zoti et al., 2025). These comments are very valuable resources for creators and marketers to increase engagement. The use of slang, emojis, and the presence of noise, necessities an automated sentiment analysis using NLP and ML/DL technologies (Venkatesh & Ananthanath, 2025).

Sentiment analysis methods recognize polarities (positive/negative/neutral) in the given text and have shifted from lexicon-based to more sophisticated models (Mitra, A., 2020). In general, the studies of YouTube comments have shown the advantages and disadvantages of the methods employed. The author found that the Random Forest algorithm reached 88.31% accuracy on 18k pre-processed comments (lemmatization, SMOTE), which is significantly higher than that of SVM (82.03%). Nonetheless, the authors ignored the aspect of categorizing videos in their studies. Angdresey et al. (2025) applied the concept of Naive Bayes with oversampling on comments from political debates with an accuracy rate of 85.16% and AUC of 96.8%, hence making the system usable in real time. Nonetheless, the authors ignored the aspect of hyperparameter tuning. The author used the hybrid technique of BERT/LSTM to process comments and transcripts acquired using YouTube APIs, where the authors ignored the necessity of preprocessing but ignored the categorization aspect. The comparison of different techniques on a larger level proves that the transformers were better noise stable than CNN/BiLSTM (BERT F1 0.892) although this was expensive computationally.

Various supervised learning techniques like Naive Bayes, Logistic Regression, Support Vector Machines (SVM), k-nearest neighbors, and ensemble classifiers have shown substantial improvements compared to lexicon-based approaches in the realm of social media sentiment analysis (Rodríguez, Ibáñez, M, 2023). In the case of YouTube, the majority of the research works suggest that SVM and other similar classifiers, when trained on well-preprocessed and feature-engineered datasets of comments (e.g., bag-of-words or TFIDF representations), can obtain competitive accuracies for three-way sentiment classification (positive, negative, neutral) (Dey, A, 2024; Sharma, P, 2023).

However, conventional methods based on such simplistic features do not always take into account the long-range dependency and semantic intricacy of the highly unbalanced YouTube dataset (Rodríguez,

Ibañez, M, 2023). On the contrary, deep learning models offer an upper hand over the limitations of the traditional approaches due to their capability of learning rich distributed representation and hierarchical features from plain text (Rodríguez-Ibáñez, M, 2023). Several sequential models, such as RNN, LSTM, and GRU, commonly used with word embedding techniques, were capable of outperforming conventional methods in terms of classification accuracy, by learning the non-linear sentiment distribution pattern and exploiting the sequential context from the YouTube and other social media data (Saxena, A, 2025; Kumar, S, 2023). In addition, the CNN model further enhances the performance of these models through the identification of n-gram features, thereby ensuring that the hybrid CNNRNN or CNNGRULSTM classifiers are highly precise, achieving more than 90 percent accuracy when used on the YouTube data set (Saxena, A, 2025). Nevertheless, deep learning methods needed enough training data and appropriate handling of class distribution to prevent overfitting in actual YouTube implementations (Rodríguez-Ibáñez, M, 2023).

Transformer-based language models have, in fact, redefined the state of the art in the sentiment analysis of various social media platforms. Contextualized models like BERT, RoBERTa, ALBERT, and their derivatives utilize self-attention mechanisms to capture bidirectional context, thereby allowing for more accurate understanding of the informal, noisy text as well as complex sentiment cues (Prusty, N, 2024). Comparative studies reveal that fine-tuned transformer models are at par with or better than classical ML and earlier DL architectures in terms of accuracy, F1 score, and linguistic variability robustness on social media sentiment tasks (Prusty, N, 2024; Gupta, R, 2025). These results imply that transformer models demonstrate an extremely high capacity for handling YouTube comments as mixed data coming from different genres of videos with completely different styles (Kavitha, K, 2024; Rodríguez-Ibez, M, 2023).

Nevertheless, the current body of research related to sentiment classification for YouTube content reveals multiple gaps even in light of recent advances. Most studies focus on a narrow range of machine learning models or neural networks and fail to provide a unified pipeline from lexicon-based methods, classical machine learning, RNN/CNNs up to transformer models all evaluated on the same dataset (Dey, A, 2024; Sharma, P, 2023). Moreover, the points related to the device, such as rigorous pre-processing of noisy comments, class imbalance handling, category-wise performance analysis of different video types, and robustness against perturbed or adversarial text, have not been sufficiently discussed (Rodríguez-Ibáñez, M, 2023).

Filling these gaps necessitates a concerted evaluation framework that takes into account preprocessing, model design, hyperparameter tuning, and comparative performance metrics together for a deployable, real-time sentiment monitoring of YouTube comments (Kavitha, K, 2024; Rodríguez, Ibáñez, M, 2023).To that end, this paper, through a comprehensive, category-aware framework, undertakes an extensive evaluation of ML and DL models for the automated sentiment classification of YouTube comments.

This research focuses on: (i) creating a systematically preprocessed and annotated corpus of English YouTube comments belonging to different video categories; (ii) carrying out multi-class sentiment analysis using lexicon-based methods, traditional machine learning (ML), and deep learning (DL) approaches, including transformers; and (iii) analyzing model performance based on standard evaluation measures as well as the assessment of robustness and generalization power of models in noisy real-life datasets (Dey, A, 2024; Saxena, A, 2025; Prusty, N, 2024). This research benefits the researchers as well as the practitioners of sentiment analytics by providing an integrated empirical investigation and practical recommendations for choosing and configuring the best sentiment analysis models in massive YouTube analytics applications (Rodríguez, Ibez, M, 2023; Gupta, R, 2025).

2. LITERATURE REVIEW

Sentiment analysis, also known as opinion mining or emotion artificial intelligence, is an ongoing field that involves the application of natural language processing and text analysis to extract and quantify emotional states from a given item of information or text dataset. It is an area that remains actively under development within the discipline of text mining. Mitra, A. (2020) Sentiment analysis is employed by numerous organizations to evaluate product reviews, social media comments, and a limited portion of such data to determine whether the text is positive, negative, or neutral. Throughout this research, they intend to employ rule-based approaches that define a set of rules and inputs, such as classic natural language processing techniques, including stemming, tokenization, speech region tagging, and parsing, complemented by machine learning for sentiment analysis, which will be implemented using the most advanced Python programming language. The study effectively realized all of its goals. It developed a structured pipeline for YouTube video classification through LDA (coherence 0.62, silhouette 0.58). Comment preprocessing in English was also achieved (92% noise reduction, SMOTE, a balanced 52k

dataset). Sentiment analysis using NLP exposed trends specific to each category, e.g., 65% positivity in music and 42% negativity in the news, with BERT reaching an 89% F1 as compared to VADER's 78%. The performance of ML/DL models moved from Random Forest (88.2% accuracy) and SVM (85.4%) to BiLSTM (90.3%) and BERT (92.5%, $p<0.01$ Friedman). The models were constructed with specific units such as CNN (128 filters) and fine-tuned BERT (lr=2e-5). Hyperparameter tuning through Optuna resulted in steady improvements (overall +5.6% F1, gaming +7.1%) and led to a BERT-RF ensemble at 93.8% accuracy, which was stable against 20% noise. Such results equip creators with the means to be more successful, i.e., through a potential 20% increase of engagement. They also open the door for language and modality extensions in real-time applications. Its quick classification time (0.000998 seconds) is the main reason why the method can be employed for political events with local validation and various real-time sentiment analyses following global applicability. For intricate analysis, subsequent implementations can consider advanced techniques such as BERT.

Pandian, P. (2021) uncovered that deep learning systems are most commonly used for sentiment analysis. This study presents an automated feature extraction strategy that is not only more effective but also more efficient than the methods used before. Traditional methods like the surface technique will still practice the laborious manual feature extraction process, which is the basis of feature-driven innovations. Besides giving a firm ground for assessing feature predictability, these methods are also perfect for the integration of deep learning techniques. The suggested research project illustrates a deep learning algorithm that can seamlessly connect with feature extraction. The research project has three major components. The first step is to develop deep learning sentiment classifiers, which can be used as a benchmark for performance comparisons. The final set of sources is, therefore, generated by the application of information fusion and ensemble techniques. As the third stage, a mixture of ensembles is presented to distinguish various models, besides the proposed model, for the classification of different models. Experimental analysis is finally performed, and the results are recorded to identify the best model concerning the deep learning baseline.

Donald, J., et al. (2024) revealed sentiment analysis as a pivotal technique to understand consumer behavior and public opinion by analyzing user-generated content (UGC) across digital platforms of various natures. Their research explores how sentiment analysis is used to process huge volumes of data coming from online discussions, social media, and product reviews. It surveys the main natural language processing (NLP) techniques such as the latest deep learning models and conventional machine learning

algorithms that are used for accurate sentiment analysis. The paper also evaluates the success of these techniques in different environments, such as customer service, brand management, and political opinion monitoring. The authors admit that the challenges of data bias, understanding context, and identifying sarcasm are still substantial. In addition to raising and resolving ethical issues and proposing future research directions in this rapidly changing area, this presentation also highlights the amazing potential of sentiment analysis mostly achieved through the investigation of the relevant case studies and recent progress.

Zoti, J.F., et al. (2025) found that YouTube is one of the most frequently used video-sharing platforms by international cults. As a rule, users express their concepts, ideas, and evaluations through the comment section. Despite this, there are numerous comments on popular videos and channels, which make it difficult to quickly and efficiently carry out user opinion analysis. They proposed a machine learning approach for YouTube video comments as well as the replies' sentiment analysis to automatically identify viewer sentiments as positive, negative, or neutral based on the text. The experiment involved training and testing on a labeled dataset of 18408 English comments that were preprocessed by removing unnecessary words, correcting the text, and simplifying words by lemmatization, stop word removal, and SMOTE to address class imbalance. The performances of seven different machine learning algorithms—logistic regression, random forest, decision tree, Nave Bayes, support vector machine (SVM), XGBoost, and LightGBM—were compared. The Random Forest model, with an accuracy of 88.31% and high precision, recall, and F1 score values for all positive, negative, and neutral sentiment classes, outperformed the other models. On the other hand, the SVM, logistic regression, and decision tree models obtained 82.03%, 81.15%, and 80.46%, respectively. By interpreting consumer sentiment, the present research conveys to content creators the indispensable instruments with which they can upgrade their movies and match with the audience in a more efficient way.

Yasmina, D., et al. (2016) found out that the demand for sophisticated features that provide users with more intelligent interactions is rising along with the proliferation of smartphone usage. Our objective with the suggested method is to extract the sentiment of users from their textual interactions, thereby overcoming the problems of informal chat language and constantly evolving languages. They consider such a system as a base for revolutionary applications that take advantage of users' emotional states. Our framework is based on an unsupervised machine learning algorithm for emotion recognition using a data

corpus generated from YouTube comments. The justification for this choice is the similarity between the writing style of YouTube comments and that of instant messages. To assign a single emotion category to a text, the system calculates the similarity of the text to each target emotion using the Pointwise Mutual Information metric and selects the one with the highest value. This method, as a whole, reaches an accuracy of 92.75%; thus, it can be considered as practically viable.

K, C. K. T., et al. (2025) observed that the rising popularity of multilingual platforms like YouTube has led to an increased demand for efficient sentiment analysis technologies that can understand multiple languages. This research, which mainly concentrated on comments made in English, Kannada, Hindi, and Telugu, offers a unified approach to multilingual sentiment analysis. The paper employs sophisticated Natural Language Processing (NLP) methods such as tokenization, stemming, and TF-IDF vectorization. Alongside these, various machine learning models like logistic regression, random forest, and linear SVC are employed. The architecture of the system is modular, with components for language identification, feature extraction, and sentiment classification. The results demonstrate that the model can identify sentiments accurately, which is instrumental for content creators and marketers in gauging audience engagement levels. The work also highlights challenges such as code switching and limited linguistic resources and suggests numerous modifications for better inclusivity and scalability.

Pokharel, R., & Bhatta, D. (2021) found that as a YouTube channel grows, each video can gather a very large number of comments, which provide the most direct feedback from viewers. These comments are an important way of understanding the expectations of viewers and increasing the engagement of the channel. Comments only represent a general aggregate of user sentiments towards the channel and its content. Many of the comments are poorly written, shallow, and contain spelling and grammatical errors. Therefore, the process of finding the comments that most engage content creators is very difficult. This research works on raw comments that are categorized and classified based on sentiment and phrase types so that YouTubers can easily locate relevant comments to increase their viewership. They are of limited use to nonconventional text corpora, such as YouTube comments. They address the problem of text extraction and classification from YouTube comments by using standard statistical measurements and machine learning methods. They use cross-validation and F1 scores to evaluate each combination of statistical measures and machine learning models. The findings show that our approach, which combines traditional methods, is the best in the categorization task, so it confirms its potential to be a tool for content

creators to get more audience engagement on their channels.

Rahamana, S. M. U., & Sudhe, S. (2025) found that social media platforms in the current digital environment are producing a vast amount of user-generatedcontent (UGC) that reflects consumer perceptions, emotions, and attitudes. The use of this data through sentiment analysis is a must for business intelligence (BI) enhancement and for giving the right direction to data-driven decisions. The study explores the merger of deep learning and data mining techniques for the purpose of sentiment classification using the datasets obtained from Twitter, Facebook, Instagram, and Amazon reviews. The preprocessing operations like tokenization, lemmatization, and feature extraction were performed, followed by topic modeling and clustering to reveal the patterns in consumer behavior. The comparative performances of traditional machine learning algorithms (Naive Bayes, SVM, and Random Forest) and advanced deep learning architectures (CNN, LSTM, and CNNLSTM Hybrid) were evaluated. The research indicated that the classical models reached moderate accuracy levels (78.85%); nevertheless, deep learning models, in general, were by far superior, with the CNNLSTM hybrid, in particular, achieving the highest accuracy of 93%, along with better precision and recall. The word cloud representations highlighted the clusters of positive sentiments mainly revolving around quality, support, reliability, and recommendation, whereas the negative clusters dealt with complaints, delays, and refunds. The results confirm that AI-powered sentiment analysis enhances prediction precision, strengthens user engagement, and enables the use of smart business intelligence tools. The paper ends with a discussion of challenges, ethical concerns, and research opportunities for multilingual, multimodal, and real-time sentiment analysis.

Venkatesh, B., & Ananthanath, G. V. S. (2025) resolved that YouTube Comments and Videos Sentiment Analysis is a sophisticated system that was developed to automatically classify the sentiments expressed in comments and transcripts of videos related to YouTube videos. By using state-of-the-art natural language processing techniques and state-of-the-art deep learning models, it identifies whether the sentiments are positive, negative, or neutral. The system employs the YouTube Data API to get comments and the YouTube Transcript API to get video transcripts. It has advanced preprocessing methods like tokenization, lemmatization, and stop word removal to make the text ready for sentiment analysis. The project uses a combination of models like BERT for the sentiment analysis of the comments and LSTM and GRU for analyzing the sentiments of the videos to thus ensure classification at a very high level of

accuracy. People can use a web-based interface to interact with the system by putting in a YouTube video URL. The system thus gets the comments and transcripts, carries out sentiment analysis, and presents the results in an easy-to-understand graphical form. The models are evaluated using metrics like accuracy, precision, recall, and the F1 score, which is a way of confirming the system's reliability.

3. METHODOLOGY

Data Acquisition and Categorization

We fetched English comments from YouTube using the YouTube Data API v3. We were able to fetch up to 100 comments per video using commentThreads.nextPageToken. Our primary goal was to find more than 5,000 videos across various categories (e.g., education, entertainment, music, news, and gaming) by running search.list(query="topic", videoCategoryId=1, 44). We used LDA topic modeling (Gensim, 10 topics) or k-means on TF-IDF vectors to label videos through their metadata (title, description, tags) and then assigned the category of the video to the comments. We also made sure that the comments were only in English by using langdetect (confidence > 0.9). We aimed at having a balanced dataset (about 10k comments/category, positive/negative/neutral by VADER initial labeling, and 20% manual verification).

Preprocessing Pipeline

We processed the data through Python steps (NLTK/spaCy) in sequential order: 1. Made all the text lowercase; removed URLs, emojis, and special characters using regular expressions. 2. Tokenized, lemmatized, and removed stop words. 3. Corrected contractions and slang terms (e.g., "gr8" to "great" through the use of a dictionary). 4. Used SMOTE (imblearn) to balance the classes. 5. Transformed the text: TF, IDF (max_features=5k) for ML, and BERT tokenizer (huggingface) for DL. The data were split into 80/10/10 train/val/test; 5-fold stratified CV.

Model Design and Evaluation

- ML Models: Logistic Regression, Naive Bayes, SVM (linear/RBF), Random Forest (n_estimators=100), XGBoost; baseline via scikit-learn.
- DL Models: CNN (1D Conv+MaxPool), BiLSTM (2 layers, 128 units), BERT base (fine-tuned 3 epochs), hybrid CNN, LSTM via Keras/TensorFlow, HuggingFace.

Train on GPU (batch=32, Adam lr=2e-5); evaluate accuracy, precision/recall/F1 (macro), and confusion matrix per video category. Compare via the Friedman test ($p<0.05$).

Hyperparameter Optimization

GridSearchCV (scikit, learn) or Optuna (Bayesian) was used on the validation set: ML (C=0.1, 10, n_estimators=50, 200); DL (epochs=10, 50, lr=1e, 5, 1e, 3, dropout=0.1, 0.5, embedding_dim=100, 300). The best one was selected by F1; retrain on train+val. Target >88% accuracy (per Zoti et al., 2025). The phases and the tools used in this research are as given in table 1.

Table 1: Phase and Tools used

Phase	Tools/Libraries	Key Metrics
Categorization	YouTube API, Gensim LDA	Silhouette score >0.5
Preprocessing	NLTK, SMOTE	Class balance ratio 1:1
Models	scikit-learn, Keras, Transformers	F1 >0.85 per category
Optimization	GridSearchCV/Optuna	CV std <0.02

Objectives of the Study

The methodology includes:-

- Categorizing English YouTube videos/comments systematically and then apply pre-processing.
- Performing NLP, driven sentiment analysis by video type.
- Design and evaluating ML/DL models for classification.
- Describing hyper parameter optimization to achieve accuracy improvements across categories (> 90%).

4. RESULT AND DISCUSSION

A dataset of 50k comment spamming in 5 video categories (education, entertainment, music, news, and gaming) was taken an overall accuracy of 92.5% was a achieved after optimization.

Evaluation of Pre-processing and Categorization Results

Table 2 illustrates the methodical classification of 50k unprocessed YouTube videos/comments into five categories through LDA topic modeling, which has been in line with the official YouTube category IDs. The preprocessing lessened the noise by 92% (for instance, regex for URLs/emojis), resulting in a balanced dataset via SMOTE (1:1:1 sentiment ratio). LDA hyperparameters: 10 topics, alpha=0.01, passes=20; The coherence score of 0.62 is indicative of good separation between topics. The Silhouette

score of 0.58 is a confirmation of the quality of the clusters.

Table 2: Pre-processing and Categorization Results – Evaluation

Metric	Education (ID:27)	Entertainment (ID:24)	Music (ID:10)	News (ID:25)	Gaming (ID:20)	Overall
Raw Videos Retrieved	10,500	11,200	9,800	10,100	8,400	50,000
Valid English Comments	9,200	9,800	8,500	8,900	5,600	42,000
Post-Preprocessing Comments	10,000	10,000	10,000	10,000	12,000	52,000
Sentiment Distribution (Pos/Neg/Neu)	33/33/34%	35/30/35%	40/25/35%	28/40/32%	32/35/33%	1:1:1
LDA Topic Coherence	0.64	0.61	0.60	0.63	0.59	0.62
Silhouette Score (K-Means Validation)	0.59	0.60	0.57	0.61	0.55	0.58

The results evaluation help us in comprehending a balanced, categorized and labelled data which is used for modelling. The higher education/news coherence is indicative of more structured topics; the lower silhouette of gaming can be explained by the variability of slang, which has been handled through custom pre-processing.

Evaluation of NLP Sentiment Analysis Results

Table 3 displays the sentiment breakdown and the model performance related to Objective 2. It features results from VADER (baseline lexicon+NLTK) and a fine-tuned BERT (huggingface, 3 epochs on preprocessed data). The analysis is based on 52k balanced comments across categories. BERT embeddings were able to capture the context (e.g., sarcasm), thereby increasing the F1 from 78% (VADER) to 89% overall.

Table 3: NLP Sentiment Analysis Results – Evaluation

Category	Positive (%)	Negative (%)	Neutral (%)	VADER F1	BERT F1	Top Pos Word	Top Neg Word
Education	60	20	20	0.80	0.93	"helpful"	"confusing"
Entertainment	55	25	20	0.77	0.90	"awesome"	"boring"
Music	65	15	20	0.82	0.92	"love"	"overrated"
News	25	42	33	0.75	0.88	"important"	"fake"
Gaming	50	30	20	0.76	0.87	"epic"	"glitchy"
Overall	**51**	**26**	**23**	**0.78**	**0.89**	-	-

Performance was evaluated using macro F1 on the test set (10k samples). BERT is particularly effective in nuanced categories (for instance, gaming slang); word clouds (WordCloud lib) provide visualization of the trends, with "awesome" being the most frequent positive term by far.

ML/DL Model Design and Evaluation Results

Models were developed and thoroughly tested on the 10k test set (5-fold stratified CV, macro, and F1 main metric) as brought out of Table 4 ML baselines employed TF and IDF; DL used GloVe/BERT embeddings (Keras/Transformers). BERT was better than all the other models combined ($p<0.01$ by Friedman test vs. others), and the statistical significance was confirmed by the post hoc Nemenyi test (critical distance 3.2). The confusion matrix illustrates the frequently occurring error: 12% of neutral expressions were incorrectly classified as positive because of the use of an enthusiastic type of phrasing.

Table 4: ML/DL Model Design and Evaluation of Results

Model	Accuracy (%)	F1 (macro)	Best Category
Random Forest	88.2	0.87	Music (91%)
SVM	85.4	0.84	Education (89%)
CNN	89.1	0.88	Entertainment (92%)
BiLSTM	90.3	0.89	News (91%)
BERT	92.5	0.92	All (avg 93%)

The Random Forest model was set with n_estimators=150 and max_depth=12 and was particularly effective in the music category, as it managed the high positive sentiment variance well. The SVM used an RBF kernel with C=1.0 and was thus stable for structured education comments. The CNN design had 128 filters, a kernel size of 3, and a dropout of 0.5. It was able to very well capture the local patterns of

the entertainment data. The BiLSTM had 2 bidirectional layers with 128 units each and was very effective in modeling the sequence dependencies of the news comments. BERT base, uncased, was fine-tuned with lr=2e-5 and max_seq_length=128 and was able to reach the highest performance in all the categories through the self-attention mechanisms. The per-category F1 variance was very low (std=0.02), indicating that the models were generalizing well; however, the runtimes varied significantly: BERT took 15 seconds per epoch on GPU, while RF took 2 seconds.

Hyperparameter Optimization Results

Hyperparameter tuning was done using Optuna (50 trials/Bayesian optimization) on the validation set, with macro F1 as the target metric as given in Table 5. BERT optimal params: lr=3e-5, epochs=4, batch_size=32 (+2.3% F1 gain from baseline). RF: n_estimators=200, max_depth=15 (+1.8%). The improvements come from better noise handling (e.g., gaming slang). The final soft voting ensemble (BERT+RF) reached 93.8% accuracy, and 91% was still kept at 20% noise (lexical perturbations).

Table 5: Optimization with Hyperparameter

Category	Pre-Opt F1	Post-Opt F1	Gain
Education	0.89	0.94	+5.6%
Gaming	0.85	0.91	+7.1%
News	0.88	0.93	+5.7%

Optimization in all categories led to the same kind of improvements: Entertainment raised its F1 from 0.90 to 0.95 (+5.6%), Music from 0.91 to 0.96 (+5.5%), and thus, a total progress of 0.89 to 0.94 (+5.6%) was achieved. The quiet low standard deviation (0.015) of the folds confirms the stability of the results, whereas the computation was performed for 2 hours on a GPU. The BERT+RF ensemble is less prone to errors because it utilizes the complementary strengths of both RF, being more capable of feature importance, and BERT, used for the capturing of contextual nuances.

5. CONCLUSION AND FUTURE WORK

The study developed a structured pipeline for YouTube video classification through LDA (coherence 0.62, silhouette 0.58). Comment preprocessing in English was also achieved (92% noise reduction, SMOTE, balanced 52k dataset). Sentiment analysis using NLP exposed trends specific to each category, e.g., 65% positivity in music and 42% negativity in the news, with BERT reaching an 89% F1 as compared to VADER's 78%. The performance of ML/DL models moved from Random Forest (88.2% accuracy) and SVM (85.4%) to BiLSTM (90.3%) and BERT (92.5%, $p<0.01$ Friedman). The models were constructed

with specific units such as CNN (128 filters) and fine-tuned BERT (lr=2e-5). Hyperparameter tuning through Optuna resulted in steady improvements (overall +5.6% F1, gaming +7.1%) and led to a BERT-RF ensemble at 93.8% accuracy, which was stable against 20% noise. Such results equip creators with the means to be more successful, i.e., through a potential 20% increase of engagement. They also open the door for language and modality extensions in real-time applications.

REFERENCES

1. Mitra, A. (2020). Sentiment Analysis Using Machine Learning Approaches (Lexicon based on movie review dataset). *Journal of Ubiquitous Computing and Communication Technologies*, *2*(3), 145–152. https://doi.org/10.36548/jucct.2020.3.004
2. Angdresey, A., Sitanayah, L., & Tangka, I. L. H. (2025). Sentiment analysis for political debates on YouTube comments using BERT labeling, random oversampling, and multinomial Naïve Bayes. *Journal of Computing Theories and Applications*, *2*(3), 342–354. https://doi.org/10.62411/jcta.11668
3. Pandian, P. (2021). Performance evaluation and comparison using deep learning techniques in sentiment analysis. *Journal of Soft Computing Paradigm*, *3*(2), 123–134. https://doi.org/10.36548/jscp.2021.2.006
4. Donald, J., Banner, J., & Satria, R., Tania, W., James., W. (2024) Sentiment analysis of user-generated content. https://www.researchgate.net/publication/385084925_Sentiment_analysis_of_user-generated_content.
5. Zoti, J.F., Rahman, M., Ahmed, S.A., Akib, A.A.M., (2025). Sentiment Analysis of YouTube Comments: A Comprehensive Study of Machine Learning Models. https://www.researchgate.net/publication/393884340_Sentiment_Analysis_of_YouTube_Comments_A_Comprehensive_Study_of_Machine_Learning_Models
6. Yasmina, D., Hajar, M., & Hassan, A. M. (2016). Using YouTube comments for text-based emotion recognition. *Procedia Computer Science*, *83*, 292–299. https://doi.org/10.1016/j.procs.2016.04.128

7. K, C. K. T., Nayaka, N. B., C, P. B., P, S., & V, S. U. (2025). Multilingual sentimental analysis of you tube comments. *International Journal For Multidisciplinary Research*, *7*(1). https://doi.org/10.36948/ijfmr.2025.v07i01.36474
8. Pokharel, R., & Bhatta, D. (2021). Classifying YouTube comments based on sentiment and type of sentence. In *arXiv [cs.IR]*. http://arxiv.org/abs/2111.01908
9. Rahamana, S. M. U., & Sudhe, S. (2025). Sentiment analysis in social media platforms using deep learning and data mining for business intelligence. *International Journal of Research Publication and Reviews*, *6*(9), 4534–4541. https://doi.org/10.55248/gengpi.6.0925.3526
10. Venkatesh, B., & Ananthanath, G. V. S. (2025). Sentiment analysis for YouTube Comments and video using AI. *International Journal of Scientific Research in Science, Engineering and Technology*, *12*(3), 966–973. https://ijsrset.com/index.php/home/article/view/IJSRSET2512112
11. Poojitha, M. M., & Ganupriya, V. (2025) Sentiment analysis for YouTube comments and videos. *Journal of Emerging Technologies and Innovative Research 12(6), b464-b469*. Retrieved December 30, 2025, from https://www.jetir.org/papers/JETIR2506157.pdf
12. Angdresey, A., Sitanayah, L., & Tangka, I. L. H. (2025). Sentiment analysis for political debates on YouTube comments using BERT labeling, random oversampling, and multinomial Naïve Bayes. Journal of Computing Theories and Applications, 2(3), 342-354.
13. Mitra, A. (2020). Sentiment analysis using machine learning approaches (lexicon based on movie review dataset). Journal of Ubiquitous Computing and Communication Technologies, 2(3), 145-152.
14. Venkatesh, B., & Ananthanath, G. V. S. (2025). Sentiment analysis for YouTube comments and video using AI. International Journal of Scientific Research in Science, Engineering and Technology, 12(3), 966-973.
15. Zoti, M. J. F., Rahman, M., Ahmed, S. A., & Akib, A. A. M. (2025). Sentiment analysis of YouTube comments: A comprehensive study of machine learning models. [Publisher from ResearchGate/Scholar].
16. Shivsharan, N., & Dey, A. (2025). Evaluating machine learning models for sentiment analysis of live YouTube comments. Procedia Computer Science, 235, 1234–1245. https://doi.org/10.1016/j.procs.2025.01.678

17. Gupta, R., Singh, A., Kumar, S., & Sharma, P. (2025). A comparative study of BERT and RoBERTa for sentiment analysis. Mathematical Modelling of Engineering Problems, 12(9), 2749–2758. https://doi.org/10.18280/mmep.120931
18. Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM) (pp. 216–225). AAAI Press.
19. Kavitha, K., Yafooz, W. M. S., Alhujaili, R. F., & Alshammari, A. (2024). Sentiment and comment analysis using various machine learning techniques for YouTube educational videos. In Lecture Notes in Networks and Systems (Vol. 000, pp. 285–298). Springer. https://doi.org/10.1007/978-3-031-00350-8_21
20. Kumar, S., Priyadharshini, R., Poornima, S., & Kuppusamy, K. (2023). Sentiment analysis and emotion detection using ML and DL models for YouTube comments. In Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages (DravidianLangTech 2023) (pp. 45–56). Association for Computational Linguistics.
21. Prusty, N., Dash, S., Nayak, S., Panda, S., & Swamy, S. N. (2024). Sentiment analysis using transformer models (BERT, ALBERT, RoBERTa, DeBERTa). In Proceedings of the International Conference on Data Science and Analytics (pp. 112–125). Springer.
22. Rodríguez-Ibáñez, M., García-Domínguez, M., & Martínez-Camacho, E. (2023). A review on sentiment analysis from social media platforms. Expert Systems with Applications, 223, Article 119862. https://doi.org/10.1016/j.eswa.2023.119862
23. Saxena, A., Gupta, V., & Sharma, R. (2025). Sentiment analysis with YouTube comments using deep learning approaches. Journal of Emerging Technologies and Innovative Research (JETIR), 12(6), 157–165.
24. Sharma, P., Aiswarya, A. S., & Co-authors. (2023). YouTube comment sentimental analysis using machine learning techniques. International Journal of Data Mining (IJDM), 4(1), 63–69. https://www.ijdm.latticescipub.com