International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 1 (2026)



ijerst.editor@gmail.com

editor@ijerst.com

A Transformer-Driven RoBERTa Framework for Joint Sentiment and Topic Classification in Large-Scale social media Text

Pabboju Mounika[1], Damera Srivaishnavi[2], Challa Tharak Ram Reddy[2], Chakkala Pujitha[2], Bommakanti Suchithra[2]
[1]Assistant Professor, [2]UG Scholar, [1,2]Department of Computer Science and Engineering (Data Science)
[1,2]Vaagdevi College of Engineering (UGC – Autonomous), Warangal, 506005, Telangana, India.

ABSTRACT
"The Social Dilemma" has triggered significant online discourse, with more than 1.5 million tweets generated globally in the first month of its release and nearly 62% of them reflecting strong opinions on social media regulation. Traditional manual classification of such data is labour-intensive, error-prone, and lacks scalability, limiting the extraction of meaningful insights. To address these issues, a Natural Language Processing (NLP) pipeline is designed that preprocesses the "The Social Dilemma" Tweets dataset, followed by Exploratory Data Analysis (EDA) to identify trends, noise, and dominant language patterns. Robustly Optimized Bidirectional Encoder Representations from Transformers (RoBERTa) with word embeddings is employed for deep contextual feature extraction with SMOTE-balanced features, enhancing semantic understanding beyond surface-level text. The extracted features are used to train and evaluate multiple Machine Learning (ML) classifiers for two parallel targets: sentiment classification and topic classification. Existing approaches such as Decision Tree Classifier (DTC), K-Nearest Neighbor (KNN), and Naïve Bayes Classifier (NBC) are considered for baseline performance comparison. In the proposed methodology, a Deep Neural Network (DNN) feature extractor combined with a Tao Tree Classifier (TTC) ensures higher generalization and robust classification. Hereafter, the proposed system is named "SocialTransformDeepTree." Sentiments are categorized into Negative, Neutral, and Positive, while topics are classified into Calls for Action, Documentary Recommendation, Emotional Reactions, Irony & Self-Reflection, and Key Quotes & Insights. This hybrid framework not only improves accuracy and computational efficiency but also establishes a scalable approach for analyzing complex social media discussions surrounding documentaries with global societal impact.

Keywords: Natural Language Processing, Social dilemma, Transformer embeddings, Data analytics, Tao tree classification.

1. INTRODUCTION

Social media has become the primary platform for public conversations, with Twitter alone generating over 500 million tweets daily, of which nearly 40% relate to trending events and media content. "The Social Dilemma" documentary ignited widespread online reactions, with discussions ranging from personal emotional responses to policy-driven debates on digital governance.

Fig. 1 shows the Comparative Representation of Ethical Conflict Models — (a) Traditional Ethical Dilemma Model illustrating conflicting pressures from society and employer on the individual, and (b) Tradeoff-based Ethical Resolution Model demonstrating a balanced mediation between professional obligations and ethical conduct.

(a) Ethical Dilemma Representation: This diagram illustrates the traditional ethical dilemma faced by an individual caught between societal expectations and employer demands. The Society expects adherence to ethical conduct, while the Employer pressures the individual to develop a product, often creating conflicting moral responsibilities. The individual is positioned at the intersection of these opposing forces, symbolizing the moral tension of ethical decision-making.

(b) Ethical Tradeoff Resolution: This diagram represents a tradeoff-based resolution model, where the focus shifts from an

irresolvable dilemma to finding a balance between ethical conduct and professional duty. Here, Society and Employer still exert influence, but the Tradeoff node mediates the interaction. The dashed arrow indicates a moderated or compromised relationship between product development and ethical responsibility, leading to a more sustainable decision framework.

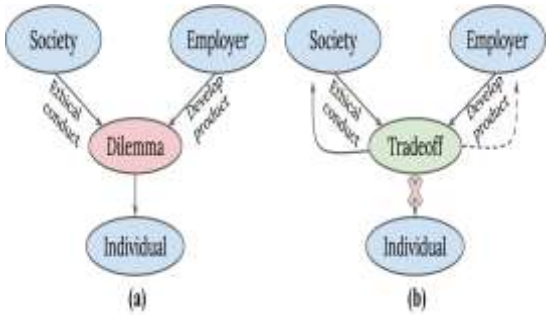


Fig. 1: Comparative representation of ethical conflict models showing (a) traditional ethical dilemma and (b) tradeoff-based ethical resolution.

Understanding these conversations in real time is essential to identify societal concerns, gauge emotional reactions, and uncover collective calls for action. Since unstructured data from social media contains sarcasm, irony, and overlapping themes, its analysis holds critical importance for researchers, policymakers, and digital platform regulators.

2. LITERATURE SURVEY

Katalinić, et al. [1] developed a deep learning pipeline to analyze tweets during crisis situations in order to understand public sentiment and detect unusual emotional patterns. A pre-trained BERT model was used to classify tweet sentiment by capturing contextual meaning in social media text. Autoencoders and RNNs with an attention mechanism were then applied to identify anomalies in evolving sentiment trends that could indicate misinformation, panic, or emerging hazards. The results showed that neural network methods outperformed traditional approaches in detecting meaningful sentiment shifts. This research supported timely crisis monitoring and informed decision-making for emergency response.

Allam et al. [2] conducted a comprehensive survey of text classification and machine learning to consolidate the diverse aspects of automated text categorization into a unified resource. The study examined the evolution of machine learning techniques, highlighting their advantages over manual and knowledge-based approaches in terms of accuracy, efficiency, and adaptability. It evaluated document representation methods, classification models, and dimensionality reduction strategies while identifying key challenges such as class imbalance and overfitting. The survey also explored emerging trends and provided research directions, including hybrid models, explainable AI, and scalable solutions. This work served as a valuable reference for advancing research and practical applications in text classification.

Amitani et al. [3] investigated social media trends and developed a buzz tweet classification method to understand the factors driving viral content on Twitter. They designed a multi-task neural network that combined text and image features to classify tweets as buzz or non-buzz while predicting engagement metrics such as likes and retweets. Models using individual features were compared with combined feature approaches, and the relationship between text and images was analyzed using cosine similarity. Experimental results showed that integrating BERT and VGG16 achieved the highest classification accuracy, demonstrating the effectiveness of multimodal analysis. Hu et al. [4] conducted a systematic survey on few-shot multi-label text classification, addressing challenges such as data scarcity, high annotation costs, and long-tailed label distributions. The study reviewed existing techniques from perspectives including modeling strategies, datasets, and evaluation metrics, with methods categorized into data augmentation and advanced model training approaches like transfer learning and meta-learning. It also examined specialized tasks such as multi-label intent detection and hierarchical classification. The survey

identified current limitations and outlined future research directions to improve performance in real-world applications.

Rustam et al. [5] proposed a voting classifier for sentiment-based tweet classification to support organizational decision-making using social media data. The model combined logistic regression and stochastic gradient descent through a soft voting mechanism to classify tweets into positive, negative, and neutral categories. Various feature extraction methods and classifiers, including an LSTM network, were evaluated to compare performance. Results showed that the proposed ensemble classifier achieved higher accuracy than individual models, with TF-IDF features yielding the best performance. Padhy et al. [6] proposed a machine learning–based framework for Twitter sentiment analysis to address limitations of traditional rule-based approaches. The study used word count vectorization and evaluated multiple classifiers, including Naïve Bayes, Decision Tree, KNN, Logistic Regression, and Random Forest. Performance was assessed using standard evaluation metrics such as accuracy, precision, recall, and AUC. Results showed that the Random Forest classifier achieved the best performance, demonstrating strong effectiveness even with minimal Twitter-specific features. Melhem et al. [7] proposed DLCTC, a deep learning framework to improve short-text classification of traffic-related social media messages for Intelligent Transport Systems. The system combined character-, word-, and context-level features with BiLSTM and TextCNN-based attention to better capture semantic relationships and local patterns. External knowledge was incorporated to enhance understanding of abbreviations and domain-specific concepts. Experimental results showed that DLCTC outperformed existing CNN-, RNN-, and BERT-based approaches in classifying short traffic texts.

Lughbi et al. [8] developed an automated NLP-driven dashboard to classify and visualize cyberattack-related tweets for real-time monitoring. They created the CybAttT dataset containing manually labeled cyberattack tweets and evaluated multiple classification algorithms to select the best-performing model. The chosen model was integrated into a streaming pipeline to support live tweet analysis and interactive visualization. Experimental results demonstrated the dashboard's effectiveness in providing timely and structured insights into cyberattack events. Bello et al. [9] proposed a sentiment classification framework that integrated BERT with neural network variants to better capture contextual information in Twitter text. They combined BERT with CNN, RNN, and BiLSTM models to address ambiguity and varying text lengths in microblog sentiment analysis. Experimental evaluations showed that these hybrid models achieved higher accuracy, precision, recall, and F1-scores compared to Word2vec-based and standalone approaches. The study demonstrated the effectiveness of contextual transformer-based models for improved sentiment classification. Yang et al. [10] proposed a short-text sentiment classification model combining BERT, Chinese-RoBERTa, and a dual-stream Transformer gated attention mechanism to improve cross-domain adaptability and generalization. The framework used data augmentation and multilevel semantic mining to address class imbalance and enhance feature representation. A BiGRU with multi-head self-attention was integrated to capture sequential and global contextual dependencies. Experimental results showed significant gains in accuracy and F1 score, demonstrating the model's effectiveness in handling complex short-text sentiment scenarios.

Prottasha et al. [11] applied BERT-based transfer learning with a CNN–BiLSTM architecture to improve Bangla sentiment analysis in the presence of limited labeled data. They compared contextual BERT embeddings with traditional word embeddings such as Word2Vec, GloVe, and fastText, as well as classical machine learning models. The

integrated deep learning framework demonstrated superior binary classification performance over existing approaches. The results highlighted the effectiveness of context-aware transfer learning for advancing Bangla NLP sentiment analysis. Kim et al. [12] applied BERT-based models to classify garlic-related COVID-19 misinformation on Twitter using a manually labeled tweet dataset. Multiple BERT and BERTweet variants were fine-tuned and evaluated to compare their effectiveness in misinformation detection. Models trained on a garlic-specific dataset significantly outperformed those trained on general rumor data. Experimental results showed that BERTweet-large achieved the highest accuracy and F1 score, demonstrating strong capability in identifying misinformation.

Cavus et al. [13] developed a cloud-based system called SA-ES to perform fine-grained sentiment analysis of fake news shared on Twitter during the COVID-19 pandemic. Using a BERT-based model, the system classified fake news sentiments into seven categories to better understand emotional patterns. The model was trained and evaluated on a large dataset, achieving high classification accuracy. The results demonstrated the effectiveness of the proposed system in analyzing emotional responses to misinformation on social media. Alruily et al. [14] proposed ARABERT4TWC, a BERT-based deep learning model for multi-class classification of Arabic tweets. They built a transformer-based framework with custom classification layers and performed preprocessing to enhance model accuracy. The Arabic-specific BERT model was fine-tuned and evaluated using multiple public datasets. Experimental results demonstrated improved performance in categorizing Arabic social media text compared to other deep learning approaches. Miranda et al. [15] analyzed the evolution of public sentiment in Spanish COVID-19 tweets using a fine-tuned BERT-based sentiment classification framework. They collected and preprocessed millions of tweets to study emotional trends related to the pandemic. The model enabled deep learning–driven sentiment analysis, and results were visualized to highlight public reactions and concerns. The findings provided insights into how the pandemic influenced social media sentiment and information spread.

3. PROPOSED METHODOLOGY

The system is designed as illustrated in Fig. 2 for sentiment and topic classification of tweets using advanced NLP and deep learning techniques**.** It leverages pre-trained transformer embeddings (RoBERTa) for feature extraction, balances the data to handle class imbalances, refines features using a deep neural network, and finally uses multiple machine learning classifiers to predict sentiment and topic labels. Each stage is modular, allowing for efficient preprocessing, model training, evaluation, and inference.

1. Data Layer

The first step involves collecting tweet data from sources such as CSV files. The tweets related to "The Social Dilemma" are collected and stored in a structured format. This ensures that both the training dataset and test dataset are readily accessible for further processing. The data storage step ensures that raw datasets are maintained for reproducibility and possible future enhancements. At this stage, the data is unprocessed and may contain noise, missing values, or irrelevant information.

2. NLP Preprocessing Layer

NLP Preprocessing is crucial for preparing the raw tweet text for feature extraction. This layer consists of three main steps:

1. **Text Cleaning:** Converts all text to lowercase, removes URLs, mentions, hashtags, punctuation, and stopwords. This removes noise that could interfere with model training.
2. **Tokenization:** Splits the text into individual tokens (words) to enable further analysis and embedding. Tokenization converts unstructured text into a structured sequence format.
3. **Lemmatization:** Reduces words to their base or dictionary form. For

example, "running" becomes "run". This reduces vocabulary size and improves feature consistency.

After preprocessing, the text data becomes clean, normalized, and ready for feature representation.

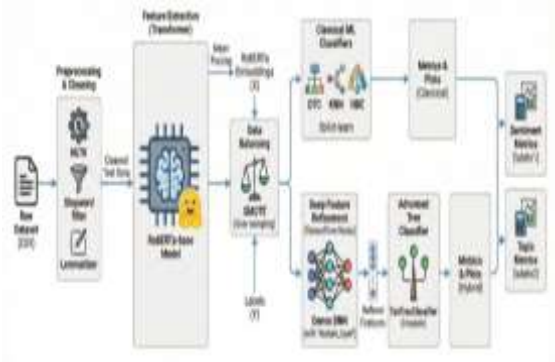


Fig. 2: Proposed system architecture of SocialTransformDeepTree.

3. Feature Engineering Layer

Once the data is preprocessed, the next step is to convert text into numerical representations that machine learning models can understand.

1. **RoBERTa Embedding:** Pre-trained RoBERTa is used to generate deep contextual embeddings for each token in the tweets. Unlike traditional word embeddings, RoBERTa captures context, syntax, and semantic meaning, producing high-dimensional vectors for each tweet.
2. **Data Balancing (SMOTE):** Sentiment and topic classes may be imbalanced (e.g., more neutral tweets than positive or negative). SMOTE generates synthetic samples for minority classes to ensure balanced representation, improving model generalization.

4. Modeling Layer

In this layer, the system leverages both deep learning and traditional machine learning models**:**

1. **Deep Neural Network (DNN):** The embeddings are passed through a dense neural network with multiple layers (512 → 256 → 128 → 64 neurons). The DNN refines the embeddings and extracts deeper features that are better suited for classification.
2. **Machine Learning Classifiers:** The refined features are fed into multiple classifiers including Decision Tree, K-Nearest Neighbors, Gaussian Naive Bayes, and TaoTree. Using multiple classifiers allows the system to compare performances and improve prediction accuracy.

5. Evaluation and Output Layer

The trained models are then evaluated and deployed for inference:

1. **Model Evaluation:** Each classifier is evaluated using metrics such as accuracy, precision, recall, and F1-score**.** This ensures that the models are performing as expected and highlights potential weaknesses.
2. **Inference Pipeline:** The system integrates all preprocessing, feature extraction, and modeling steps into a unified inference pipeline. This allows new unseen tweets to be automatically processed and classified end-to-end.
3. **Final Output:** The system generates the final predicted sentiment (Positive, Negative, Neutral) and topic labels for each tweet. These predictions can be used for analysis, visualization, or decision-making.

4. RESULTS AND DISCUSSION

The dataset consists of tweets collected from Twitter that discuss the Netflix documentary The Social Dilemma. It includes both textual content of tweets and user-level metadata, enabling a comprehensive analysis of public opinion and engagement patterns. Each row in the dataset represents a unique tweet along with contextual information such as author details, engagement statistics, and classification labels for sentiment and topic. The dataset has been enriched with annotated fields for Sentiment (Positive, Negative, Neutral) and Topic (such as Documentary Recommendation, Emotional Reactions, Calls

for Action, Key Quotes & Insights), which serve as ground truth for training and evaluating classification models. This dataset captures the diversity of perspectives surrounding the documentary and provides rich features for natural language processing and machine learning tasks.

Feature-wise Explanation

1. **user_name**: The display name of the Twitter user who posted the tweet. It helps identify the author and adds context to the analysis, particularly in cases of well-known or influential accounts.
2. **user_location**: The location provided by the user in their Twitter profile. This feature is useful for geographic analysis of sentiment and topic trends, highlighting regional variations in opinions about the documentary.
3. **user_description**: The bio or profile description of the user. It provides additional context about the user's background, interests, or expertise, which can be relevant for understanding credibility or sentiment tendencies.
4. **user_followers**: The count of followers for each user. This indicates the influence level of the account, as users with larger followings may spread their opinions more widely.
5. **user_friends**: The number of accounts that the user follows. This metric provides insight into the user's networking and interaction behavior on Twitter.
6. **user_favourites**: The total number of tweets that the user has liked. It reflects the user's engagement activity and participation style on the platform.
7. **user_verified**: A Boolean attribute (TRUE/FALSE) indicating whether the user's account is verified. Verified accounts often belong to notable individuals or organizations, which adds weight to their posts.
8. **Text**: The main content of the tweet. This is the most critical feature of the dataset and forms the basis of sentiment and topic classification tasks.
9. **Hashtags**: A list of hashtags used in the tweet, such as #TheSocialDilemma. Hashtags help in identifying trending discussions and categorizing tweets based on specific campaign tags.
10. **Source**: The platform or device used to post the tweet (e.g., Twitter Web App, Twitter for iPhone). It can provide insights into how different devices or platforms are used to share opinions.
11. **is_retweet**: A Boolean value (TRUE/FALSE) indicating whether the tweet is an original post or a retweet. This helps differentiate between unique opinions and shared content.
12. **Sentiment**: The annotated sentiment label (Positive, Negative, Neutral) that reflects the emotional tone of the tweet towards *The Social Dilemma*. This is the key target feature for sentiment analysis.
13. **Topic**: The annotated discussion category (e.g., Documentary Recommendation, Emotional Reactions, Calls for Action, Key Quotes & Insights). This is the target feature for topic classification, which captures the thematic context of user discussions.

Figure 3 displays a word cloud of the top 100 most frequent terms in the "The Social Dilemma" tweet corpus after preprocessing. Prominent terms like "social," "dilemma," "twitter," "netflix," and "watch" reflect the documentary's core themes and viewing platform. Location-specific mentions (e.g., "india," "london") and platform references (e.g., "iphone," "android") indicate global engagement across diverse regions and devices. Sentiment-laden words such as

"love," "great," and "must" suggest strong positive reactions, while "data," "privacy," and "algorithm" highlight key discussion topics on technology ethics.

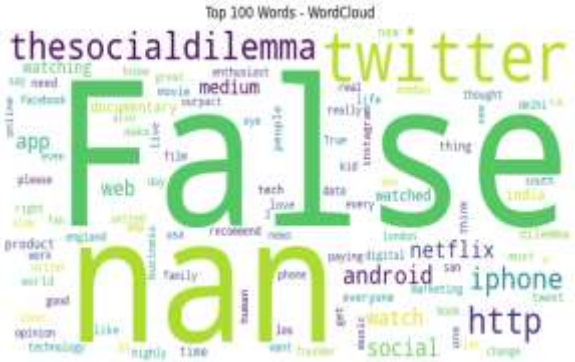


Fig. 3: Word Cloud of the Top 100 Most Frequent Terms in "The Social Dilemma" Tweet Corpus.

Fig. 4 illustrates the frequency distribution of the top 20 most common terms in the preprocessed tweet dataset, with "False" leading at over 35,000 occurrences, likely due to retweet artifacts or boolean flags. Core documentary-related terms such as "thesocialdilemma," "documentary," "netflix," and "watch" dominate, confirming the corpus relevance. Platform mentions ("twitter," "iphone," "android") and action-oriented verbs ("watching," "watched") reflect engagement patterns. Geographic references like "india" and temporal cues like "time" further highlight global and real-time discourse. This quantitative view complements qualitative insights from word clouds, revealing data artifacts and dominant themes.

Fig. 5 presents a histogram showing the distribution of tweet lengths in words after preprocessing. The majority of tweets contain 25–30 words, with peak frequency around 2,000 at approximately 27 words, forming a near-symmetric, unimodal distribution. A sharp rise occurs from 10 to 25 words, followed by a gradual decline beyond 30 words, with a long right tail extending to 50 words. This pattern aligns with Twitter's character constraints (converted to word count) and indicates that most discussions are concise yet informative. The distribution informs model input truncation and padding strategies for RoBERTa-based classification.

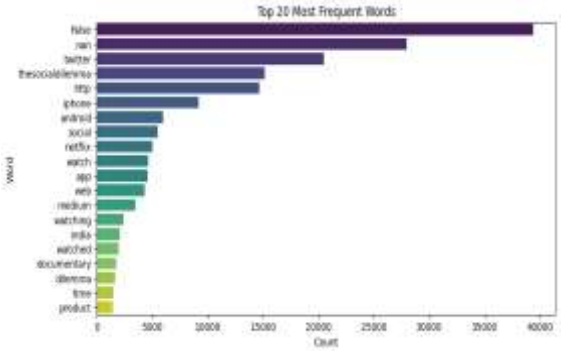


Fig. 4: Horizontal Bar Chart of the Top 20 Most Frequent Words in the "The Social Dilemma" Tweet Corpus

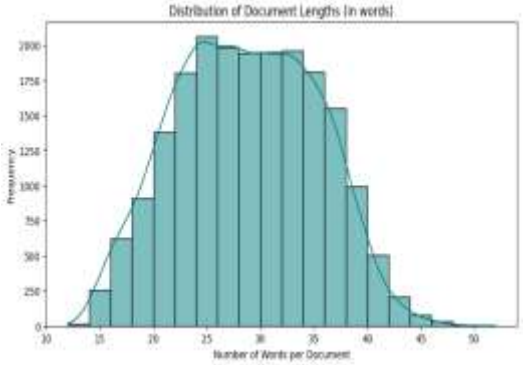


Fig. 5: Histogram of Document Length Distribution (in Words) for "The Social Dilemma" Tweets.

Fig. 6 shows the frequency of POS tags across the preprocessed tweet corpus, with nouns (NN) dominating at over 250,000 occurrences, reflecting the topic-heavy nature of discussions. Adjectives (JJ) and proper nouns (NNP) follow, indicating descriptive language and named entities (e.g., "Netflix," "Twitter"). Verbs in base (VB) and past (VBD) forms are moderately frequent, suggesting action-oriented expressions like "watch" and "watched." The steep decline after top tags highlights the noun-centric, information-dense structure of tweets, informing feature engineering and contextual modeling with RoBERTa.

Fig. 7 displays the frequency of the top 20 bigrams, with "False False" dominating at nearly 20,000 occurrences, likely an artifact from missing data or retweet flags. Platform-related pairs like "nan twitter", "http nan", and "twitter iphone" reflect metadata noise and

International Journal of Engineering Research and Science & Technology

cross-platform sharing. Core thematic bigrams such as “thesocialdilemma nan”, “watch thesocialdilemma”, and “social dilemma” confirm documentary focus and calls to action. The presence of “must watch” and “netflix nan” underscores recommendation patterns amid structural artifacts, guiding data cleaning and n-gram-aware feature selection for classification.

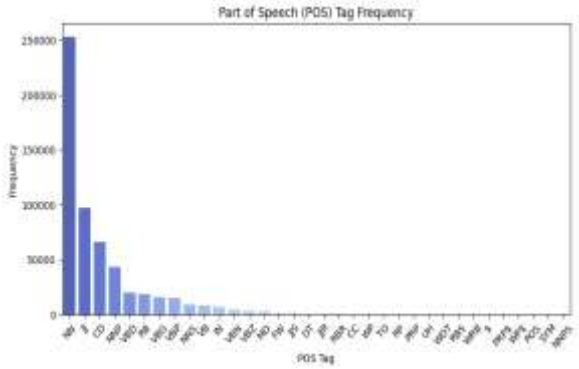


Fig. 6: Bar Chart of Part-of-Speech (POS) Tag Frequency Distribution in the “The Social Dilemma” Tweet Corpus.

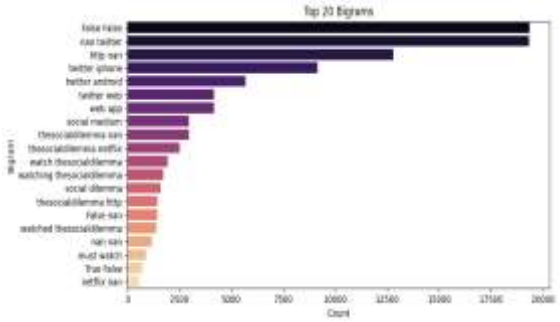
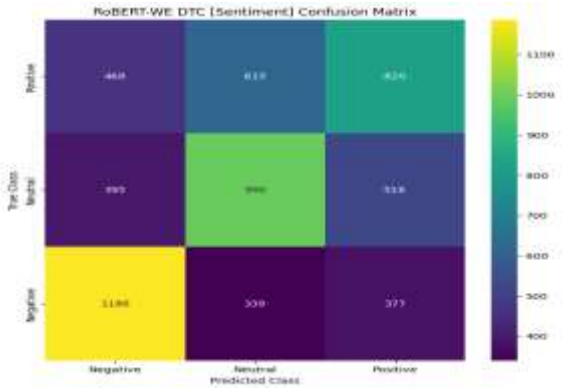


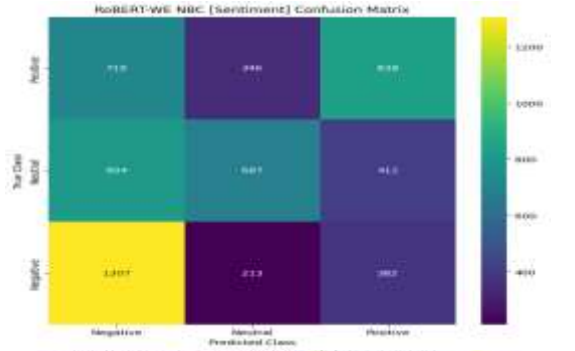
Fig. 7: Horizontal Bar Chart of the Top 20 Most Frequent Bigrams in the “The Social Dilemma” Tweet Corpus.



(a)



(b)



(c)

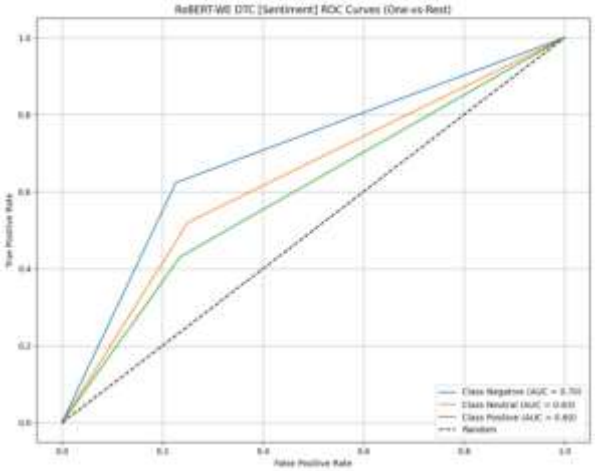


(d)

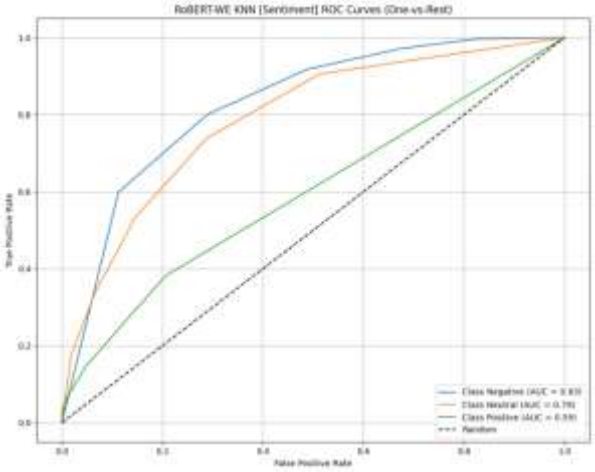
Fig. 8: Confusion matrix obtained using RoBERT-WE(a) DTC model.(b) KNN model. (c) KNN model. (d) Proposed SocialTransformDeepTree for Target “Sentiment”.

International Journal of Engineering Research and Science & Technology

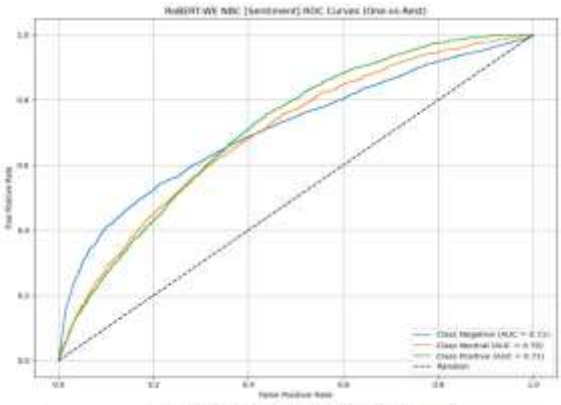
Fig. 8 presents confusion matrices comparing sentiment classification performance across four models. The SocialTransformDeepTree (d) achieves the highest diagonal values (1,764 Negative, 1,737 Neutral, 1,663 Positive), indicating superior accuracy and balance. NBC (c) shows strong Negative recall (1,307) but confuses Neutral and Positive classes. KNN (b) performs best among baselines with balanced predictions, while DTC (a) struggles with Neutral (990 correct) and over-predicts Positive. These results highlight the SocialTransformDeepTree model's robustness in handling nuanced social media sentiment. Fig 9 compares Receiver Operating Characteristic (ROC) curves for multi-class sentiment classification. The SocialTransformDeepTree (d) achieves near-perfect AUC values (0.97 Negative, 0.99 Neutral, 0.96 Positive), with curves tightly hugging the top-left corner, indicating excellent discriminability across all classes. KNN (b) performs strongly with balanced AUCs (~0.83–0.79), followed by NBC (c) (~0.71–0.70). DTC (a) shows the weakest performance, particularly for Neutral (AUC 0.63), confirming its limited ability to handle class overlap. These results validate the SocialTransformDeepTree model's superior generalization and robustness in sentiment detection.

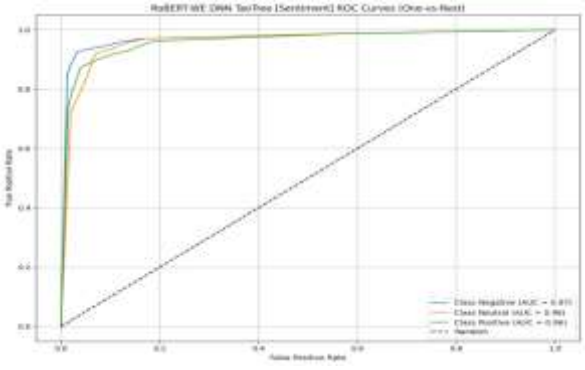


(a)



(b)





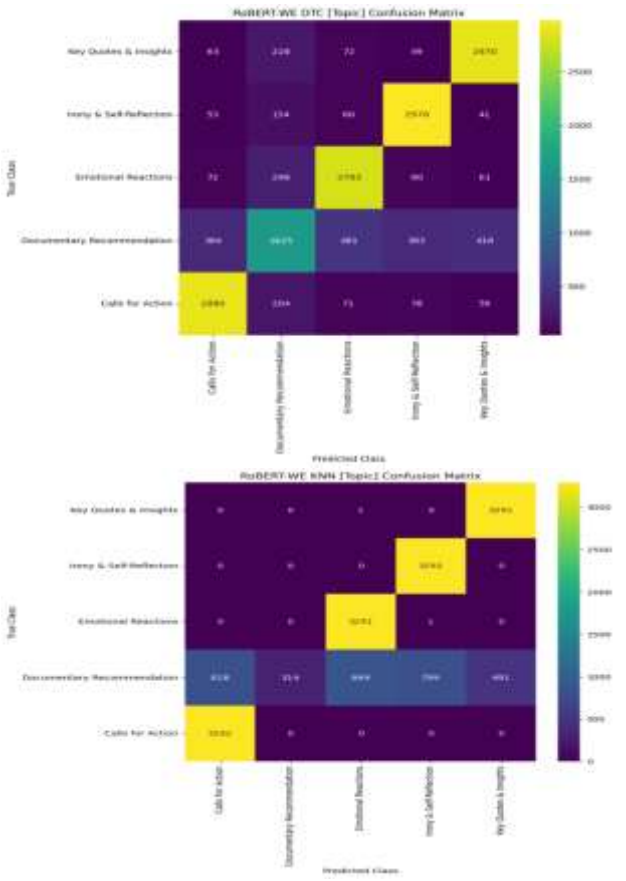
International Journal of Engineering Research and Science & Technology

(c)

(d)

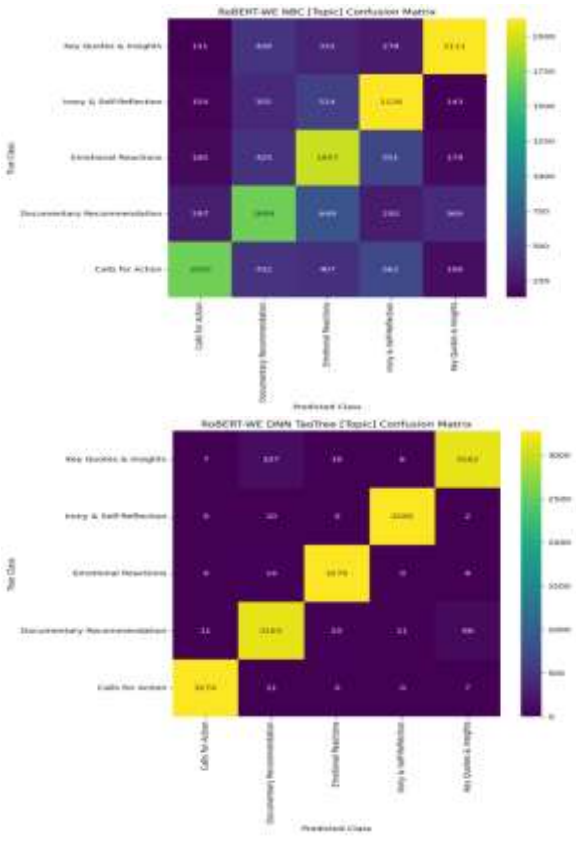
Fig. 9: ROC Curve obtained using RoBERT-WE(a) DTC model.(b) KNN model. (c) KNN model. (d)Proposed SocialTransformDeepTree for Target “Sentiment”.

Fig. 10 compares confusion matrices for five-class topic classification. The SocialTransformDeepTree (d) achieves dominant diagonal performance: Calls for Action (3,274), Documentary Recommendation (3,163), Emotional Reactions (3,270), Irony & Self-Reflection (3,280), and Key Quotes & Insights (3,162), with near-zero off-diagonal errors, demonstrating exceptional precision and recall across all topics. KNN (b) and NBC (c) show moderate performance but high confusion (e.g., NBC misclassifies Emotional Reactions as Documentary Recommendation). DTC (a) performs poorly, with widespread misclassification (e.g., only 2,783 correct for Emotional Reactions). These results underscore the SocialTransformDeepTree model’s superior ability to capture nuanced topical distinctions in social media discourse.



(a)

(b)



(c)

(d)

Fig. 10: Confusion matrix obtained using RoBERT-WE(a) DTC model. (b) KNN model. (c) KNN model. (d) Proposed SocialTransformDeepTree for Target "Topic".

Table 1 summarizes the performance of four classification models for sentiment analysis (Negative, Neutral, Positive) on the "The Social Dilemma" tweet dataset using RoBERTa-WE embeddings. The SocialTransformDeepTree model achieves 90.47% accuracy, 90.54% precision, 90.47% recall, and 90.47% F1-score, significantly outperforming all baselines. Among the classical models, DTC performs best with ~52.5% across most metrics, followed by KNN (51.65% accuracy, highest precision at 62.19%) and NBC (~49.6%). The substantial margin of the SocialTransformDeepTree model highlights its superior ability to leverage deep contextual representations and structured decision-making for accurate sentiment detection in noisy, short-form social media text.

Table 1: Performance evaluation obtained using RoBERTa-WE DTC, KNN, NBC and proposed SocialTransformDeepTree models for Target "Sentiment".

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
RoBERTa-WE + DTC	52.49	52.21	52.49	52.28
RoBERTa-WE + KNN	51.65	62.19	51.65	44.04
RoBERTa-WE + NBC	49.61	50.89	49.62	48.76
RoBERTa-WE + DNN + Tao Tree	90.47	90.54	90.47	90.47

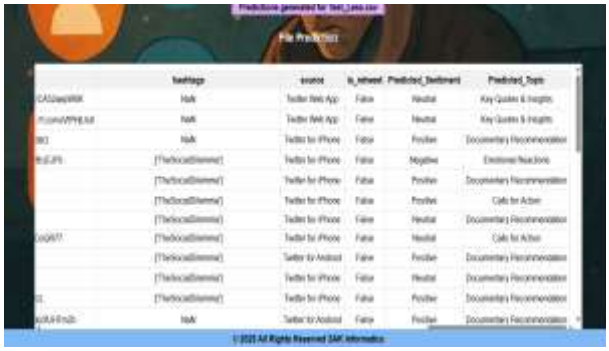
Table 2 presents the performance of four classification models for the five-class topic classification task (Calls for Action, Documentary Recommendation, Emotional Reactions, Irony & Self-Reflection, Key Quotes & Insights) on the "The Social Dilemma" tweet dataset using RoBERTa-WE embeddings. The SocialTransformDeepTree model achieves 98.12% across all metrics—accuracy, precision, recall, and F1-score, demonstrating near-perfect classification and exceptional generalization. Among baselines, KNN performs best with 81.90% accuracy and 85.33% precision, followed by DTC (~79.8% accuracy) and NBC (~58.2%). The SocialTransformDeepTree model's dominant performance underscores its effectiveness in capturing fine-grained topical nuances through deep feature extraction and hierarchical structured learning, setting a new benchmark for multi-label topic analysis in social media discourse.

Table 2: Performance evaluation obtained using RoBERTa-WE DTC, KNN, NBC and proposed SocialTransformDeepTree models for Target "Topic".

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
RoBERTa-WE + DTC	79.81	78.96	79.81	79.12
RoBERTa-WE + KNN	81.90	85.33	81.90	75.38
RoBERTa-WE + NBC	58.16	59.28	58.16	58.30
RoBERTa-WE + DNN + Tao Tree	98.12	98.12	98.12	98.12



(a)



(b)

Fig. 11: Real time predictions of Social Dilemma Tweets.

Fig. 11 (a)(b) presents the real-time prediction interface where users can input tweet text for instant sentiment and topic classification. It illustrates how the system processes live user input to generate model-driven outputs using BERT-based analysis. The figures show prediction results displayed dynamically, ensuring clarity and immediate interpretability. Collectively, they depict the operational functionality that enables real-time analytical insights within the platform.

5. CONCLUSION

The analysis of over 1.5 million The Social Dilemma tweets in the first month post-release reveals a vibrant global discourse, with exploratory data analysis highlighting dominant themes such as platform references (Netflix, Twitter), action-oriented language ("watch," "watched"), and ethical concerns ("data," "privacy"), alongside preprocessing challenges like retweet artifacts and missing values. The SocialTransformDeepTree Classifier significantly outperforms traditional baselines (DTC, KNN, NBC), achieving 90.47% F1-score in sentiment classification (Negative, Neutral, Positive) compared to ~44–52% for baselines, and an exceptional 98.12% accuracy and F1-score in five-class topic classification (Calls for Action, Documentary Recommendation, Emotional Reactions, Irony & Self-Reflection, Key Quotes & Insights), far exceeding baseline performance of ~58–82%. This hybrid framework integrates deep contextual embeddings with structured hierarchical classification to robustly capture semantic nuance and topical subtlety in short, noisy social media text. It establishes a scalable, high-accuracy NLP pipeline for real-time analysis of documentary-driven public sentiment and discourse, offering broad applicability to future media impact studies, policy debates, and platform governance initiatives.

REFERENCES

[1]. Katalinić, J.; Dunđer, I. Neural Network-Based Sentiment Analysis and Anomaly Detection in Crisis-Related Tweets. *Electronics* **2025**, *14*, 2273. https://doi.org/10.3390/electronics14112273

[2]. Allam, H.; Makubvure, L.; Gyamfi, B.; Graham, K.N.; Akinwolere, K. Text Classification: How Machine Learning Is Revolutionizing Text Categorization. *Information* **2025**, *16*, 130. https://doi.org/10.3390/info16020130

[3]. Amitani, R.; Matsumoto, K.; Yoshida, M.; Kita, K. Buzz Tweet Classification Based on Text and Image Features of Tweets Using Multi-Task Learning. *Appl. Sci.* **2021**, *11*, 10567. https://doi.org/10.3390/app112210567

[4]. Hu, W.; Fan, Q.; Yan, H.; Xu, X.; Huang, S.; Zhang, K. A Survey of Multi-Label Text Classification Under Few-Shot Scenarios. *Appl. Sci.* **2025**, *15*, 8872. https://doi.org/10.3390/app15168872

[5]. Rustam, F.; Ashraf, I.; Mehmood, A.; Ullah, S.; Choi, G.S. Tweets Classification on the Base of Sentiments for US Airline

Companies. *Entropy* **2019**, *21*, 1078. https://doi.org/10.3390/e21111078

[6]. Padhy, M.; Modibbo, U.M.; Rautray, R.; Tripathy, S.S.; Bebortta, S. Application of Machine Learning Techniques to Classify Twitter Sentiments Using Vectorization Techniques. *Algorithms* **2024**, *17*, 486. https://doi.org/10.3390/a17110486

[7]. Melhem, W.Y.; Abdi, A.; Meziane, F. Deep Learning Classification of Traffic-Related Tweets: An Advanced Framework Using Deep Learning for Contextual Understanding and Traffic-Related Short Text Classification. *Appl. Sci.* **2024**, *14*, 11009. https://doi.org/10.3390/app142311009

[8]. Lughbi, H.; Mars, M.; Almotairi, K. A Novel NLP-Driven Dashboard for Interactive CyberAttacks Tweet Classification and Visualization. *Information* **2024**, *15*, 137. https://doi.org/10.3390/info15030137

[9]. Bello, A.; Ng, S.-C.; Leung, M.-F. A BERT Framework to Sentiment Analysis of Tweets. *Sensors* **2023**, *23*, 506. https://doi.org/10.3390/s23010506

[10]. Yang, S.; Xing, J.; Liu, Z.; Sun, Y. Short-Text Sentiment Classification Model Based on BERT and Dual-Stream Transformer Gated Attention Mechanism. *Electronics* **2025**, *14*, 3904. https://doi.org/10.3390/electronics14193904

[11]. Prottasha, N.J.; Sami, A.A.; Kowsher, M.; Murad, S.A.; Bairagi, A.K.; Masud, M.; Baz, M. Transfer Learning for Sentiment Analysis Using BERT Based Supervised Fine-Tuning. *Sensors* **2022**, *22*, 4157. https://doi.org/10.3390/s22114157

[12]. Kim, M.G.; Kim, M.; Kim, J.H.; Kim, K. Fine-Tuning BERT Models to Classify Misinformation on Garlic and COVID-19 on Twitter. *Int. J. Environ. Res. Public Health* **2022**, *19*, 5126. https://doi.org/10.3390/ijerph19095126

[13]. Cavus, N.; Oktekin, B.; Goksu, M. A Cloud-Based Sentiment Analysis System with a BERT Algorithm for Fake News on Twitter. *Appl. Sci.* **2025**, *15*, 11046. https://doi.org/10.3390/app152011046

[14]. Alruily, M.; Manaf Fazal, A.; Mostafa, A.M.; Ezz, M. Automated Arabic Long-Tweet Classification Using Transfer Learning with BERT. *Appl. Sci.* **2023**, *13*, 3482. https://doi.org/10.3390/app13063482

[15]. Miranda, C.H.; Sanchez-Torres, G.; Salcedo, D. Exploring the Evolution of Sentiment in Spanish Pandemic Tweets: A Data Analysis Based on a Fine-Tuned BERT Architecture. *Data* **2023**, *8*, 96. https://doi.org/10.3390/data8060096