Ingénierie des Systèmes d'Information
Vol. 31, No. 5, May, 2026, pp. 1549-1563
Journal homepage: http://iieta.org/journals/isi

Hybrid Adaptive Framework for COVID-19 Sentiment Analysis in English–Marathi Twitter Streams

Kalyani P. Sable[1], Shrikant L. Satarkar[2], Gaikwad Vidya Shrimant[3], Madhuri Prashant Karnik[3], Disha Sushant Wankhede[3], Aniket K. Shahade[4*]

[1] Department of Computer Science & Engineering, Shri Sant Gajanan Maharaj College of Engineering, Shegaon 444203, India
[2] Department of Computer Science & Engineering, College of Engineering and Technology, Akola 444104, India
[3] Department of Computer Engineering, Vishwakarma Institute of Technology, Pune 411037, India
[4] Symbiosis Institute of Technology, Pune Campus, Symbiosis International (Deemed University), Pune 412115, India

Corresponding Author Email: aniket.shahade@sitpune.edu.in

https://doi.org/10.18280/isi.310513

Received: 23 January 2026
Revised: 25 March 2026
Accepted: 20 April 2026
Available online: 31 May 2026

Keywords:
multilingual sentiment analysis, English–Marathi tweets, COVID-19, weak supervision, RoBERTa, Convolutional Neural Network - Long Short-Term Memory, temporal adaptation, social media analytics

ABSTRACT

Public responses to COVID-19 on social media changed rapidly with policy announcements, local events, and language-specific usage, making multilingual sentiment analysis difficult for static or monolingual models. This study proposes a weakly supervised, temporally adaptive framework for English–Marathi COVID-19 sentiment analysis on Twitter. From a larger corpus of more than 1.2 million COVID-19-related tweets collected with English and Marathi keywords, a stratified subset of 50,000 tweets was used, comprising approximately 40,000 English and 10,000 Marathi posts. Marathi tweets were translated with MarianMT, and translation quality was checked through a bilingual audit. Bidirectional Encoder Representations from Transformers (BERT)-based semantic filtering at a cosine threshold of 0.80 removed 44.7% of noisy or weakly sentiment-bearing content while preserving major COVID-19 themes. Sentiment labels were generated through weak supervision using VADER, Afinn, and agreement-based verification rules, reducing dependence on manual annotation. Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) identified 128 latent sentiment clusters with 82.5% purity, while Auto Regressive Integrated Moving Average (ARIMA)(2,1,2) captured temporal sentiment shifts with 86.3% event correlation. A fine-tuned RoBERTa classifier achieved 94.7% accuracy and a weighted F1-score of 0.937. Adding Convolutional Neural Network - Long Short-Term Memory (CNN-LSTM) feature fusion improved performance to 96.2% accuracy and 0.953 weighted F1, and policy-gradient adaptation further improved robustness under chronological splits. The results suggest that combining weak supervision, contextual filtering, hybrid neural classification, and temporal adaptation can provide a practical framework for multilingual crisis-oriented sentiment monitoring.

1. INTRODUCTION

Social media has revolutionized the expression and examination of public sentiments, specially in times of global crises, such as during the onset of the COVID-19 pandemic. With millions of users from various parts of the world and in multiple languages, Twitter provides a vast reservoir of real-time public opinion, making it a very precious repository for sentiment analysis. However, difficulties abound in their analysis [1-3], the multilingual nature of tweets, the noisiness and irrelevance of the content, and the fluidity of trends in expressions of sentiment. Most current methods for sentiment analysis are based on a monolingual dataset, a static lexicon-based model, or simplistic machine learning (ML) algorithms not designed to face these complexities, thereby limiting their ability to scale up. The recent breakthroughs in Natural Language Processing (NLP) and ML have provided many tools to overcome some of these limitations. Yet, the existing systems are still insufficient to face the diversity and dynamism innate in multilingual datasets & samples. They [4-6] mostly lack robust preprocessing pipelines, fail to incorporate domain-specific contextual semantics, and fail to update their real-time adaptation of changing sentiment patterns. Reliance on manual annotation for sentiment labeling of large datasets & samples is further resource intensive and infeasible. This work brings forth an innovative pipeline for multilingual sentiment analysis, specifically designed to analyze tweets concerning COVID-19 in the English language as well as Marathi. Utilizing state-of-the-art techniques, such as MarianMT for translation, Bidirectional Encoder Representations from Transformers (BERT) embeddings for semantic filtering, and weak supervision by means of the VADER and Afinn lexicons, the pipeline guarantees high-quality preprocessing and scalability. Fine-tuned RoBERTa models augmented with paraphrased data enhance the efficiency of sentiment classification for low-resource

languages. The spatial and sequential learning from a hybrid Convolutional Neural Network - Long Short-Term Memory (CNN-LSTM) ensemble model further improve classification performance. Temporal trends are analyzed through Auto Regressive Integrated Moving Average (ARIMA), but dynamic adaptability occurs due to reinforcement learning (RL), thus enabling continuous fine-tuning of the model using real-time samples of data samples. It not only deals with the previously mentioned challenges, but actually sets new standards for scalability, multilingual support, and adaptability in sentiment analysis. There is huge potential in analyzing sentiment trends and providing real-time decisions in crisis management and other related areas requiring insights in scale about sentiments.

Advanced sentiment analysis methods are needed given the increasing focus on understanding the mood and emotions of people in times of crisis like the COVID-19 pandemic. Existing methods are thus restrictive due to their reliance on monolingual datasets, inability to process large volumes of noisy data, and inability to adapt to constantly changing dynamics in sentiment trends. While some models rely on ML and NLP to leverage sentiment analysis, they become relatively less effective when working with multilingual data or intricate, dynamic patterns. As such, the need arises for a system that not only accommodates linguistic diversity but also adapts to the changing face of sentiment expressions in real-time, without requiring human intervention. This work makes a significant contribution to addressing these gaps. This work thus suggests a comprehensive pipeline that integrates advanced technologies for the task of accurately and scalable analyzing sentiment in multilingual tweets. The innovations include using Marian MT to fluently translate Marathi tweets to English, the usage of BERT embeddings for semantic filtering, and weak supervision through lexicon-based sentiment labeling. The pipeline leverages statistical clustering with Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) as well as temporal analysis with ARIMA to identify sentiment trends over temporal instance sets. By fine-tuning RoBERTa models for multilingual data and employing a CNN-LSTM ensemble for hybrid classification, the system achieves unprecedented accuracy and robustness. Furthermore, RL is employed to adapt the model dynamically, ensuring relevance to real-time sentiment trends. These innovations collectively establish a scalable, adaptive framework that can be applied beyond COVID-19, offering a new standard for multilingual sentiment analysis in large-scale, dynamic scenarios.

2. LITERATURE REVIEW

A review of sentiment analysis methodologies across several scholarly works provides a holistic view of the current state, progressions, and challenges within the domain. The diverse studies cut across different languages, approaches, and applications, reflecting the increased interest in using sentiment analysis as a key tool for natural language understanding. Boudad et al. [1] first experimented with transfer learning in Moroccan Arabic sentiment classification, using multilingual and mono-dialectal data for better accuracy, while Mamta and Ekbal [2] proposed a transformer-based joint learning approach for code-mixed and English sentiment analysis, considering scalability. Hashmi et al. [3] further improved sentiment prediction in code-mixed tweets using multilingual transformers, focusing on language adaptability. Khan et al. [4] conducted Urdu sentiment analysis using a CNN-Bidirectional Long Short-Term Memory (Bi-LSTM) Deep Neural Network (DNN) with multiple support from multilingual BERT and an effective classification mechanism thanks to attention mechanisms. Miah et al. [5] took the concept forward to multimodal paradigms by transformers and large language models for cross-lingual applications in handling varied data types. Singh and Singh [6] developed a multi-view learning framework that integrates text and graph structures, establishing the precedent for integration of structural and contextual information sets. Hicham and Nassera [7] improved emotion classification in e-commerce using GPT and deep meta-learning ensemble, a versatile use of generative models.

Pandey et al. [8] addressed the task of sarcasm detection (SAR) in multilingual social media posts by using hybrid Convolutional Neural Network (CNN) models to represent the requirement for better sentiment understanding of nuances. Silviya et al. [9] conducted NLP algorithm evaluations and showed how performance varied between tasks, while Vianna et al. [10] focused on Portuguese tweets for benchmarking word representation models. Aziz et al. [11] advanced aspect-based sentiment analysis using BERT and graph convolutional networks, offering a comprehensive sentiment breakdown. Nath and Dwivedi [12] reviewed aspect-based sentiment analysis trends, identifying the growing role of hybrid approaches. Tori et al. [13] analyzed urban policymaking sentiments in Brussels using natural language models, bridging sentiment analysis with societal impact sets. Mohamed et al. [14] applied ensemble transformers to Arabic sentiment analysis which yielded state-of-the-art results while Anjum and Katarya [15] came up with Hate Detector for multilingual hate speech analysis. Guleria et al. [16] explored multimodal sentiment analysis for English and Hinglish memes, basing it on cultural relevance. Luitel et al. [17] analyzed fairness in audio sentiment models, thereby broadening the perspective of sentiment analysis beyond textual data samples. Rusnachenko et al. [18] explore large language models in the context of targeted sentiment analysis on Russian, addressing resource-scarce language applications. Jain et al. [19] proposed KNetwork for the task of cross-lingual sentiment analysis to aid decision-making in different linguistic environments. Bashiri and Naderi [20] compared transformer models, thereby strengthening their position in sentiment-related tasks. Recent advancements in sentiment analysis have increasingly focused on multilingual, low-resource, and domain-specific settings. Studies have developed sentiment analysis resources and models for languages such as Urdu, Persian, Kurdish, Hindi, Turkish, and Finnish, highlighting the growing need for language-aware sentiment frameworks [21-25]. Deep learning approaches based on BERT, CNNs, hybrid neural architectures, and multimodal frameworks have demonstrated superior capability in capturing contextual, semantic, and emotional nuances from social media content [26-29]. Furthermore, aspect-based sentiment analysis, emotion detection, lexicon-driven interpretability, and cross-lingual learning have emerged as important research directions for improving sentiment understanding across diverse linguistic and cultural contexts [21, 30-33]. Several application-oriented studies have leveraged sentiment analysis for public health monitoring, recommender systems, fake news detection, e-commerce personalization, pandemic-related discourse analysis, and

geopolitical opinion mining, demonstrating the broad applicability of sentiment-aware intelligence systems [25, 34-40]. These developments collectively indicate that integrating contextual language models with robust feature extraction and adaptive learning mechanisms can significantly enhance sentiment classification performance in multilingual and dynamically evolving social media environments. A review of different existing models in sentiment analysis is shown in Table 1.

Table 1. Review of existing models used for sentiment analysis

Ref.	Method	Key Findings
[1]	Multilingual Transfer Learning	Enhanced Moroccan Arabic sentiment classification using transfer learning for monolingual and multilingual data samples.
[2]	Transformer-Based Joint Learning	Effective for code-mixed and English sentiment analysis; improved scalability in multilingual contexts.
[3]	Multilingual Transformers for Code-Mixed	Achieved higher accuracy in sentiment prediction for code-mixed tweets by leveraging multilingual transformer models.
[4]	CNN-Bi-LSTM with Attention	Improved Urdu sentiment analysis through stacked CNN-Bi-LSTM architecture and multilingual BERT integration.
[5]	Multimodal Transformer Ensemble	Successfully integrated multimodal data (text and image) for cross-lingual sentiment analysis, achieving robust performance.
[6]	Text and Graph Multi-View Learning	Improved sentiment prediction by incorporating graph structures alongside textual features.
[7]	GPT and Meta-Ensemble Techniques	Enhanced emotion classification in e-commerce reviews, highlighting the versatility of generative models.
[8]	Hybrid CNN for Sarcasm Detection	Accurately detected sarcasm in multilingual posts by combining convolutional layers with sentiment-focused features.
[11]	Aspect-Based Sentiment with BERT	Unified BERT with graph convolutional networks to dissect and classify sentiment across multiple aspects.
[19]	Cross-Lingual Sentiment Networks	Developed cross-lingual models to enhance decision-making in multilingual environments, overcoming language barriers.

Note: BERT = Bidirectional Encoder Representations from Transformers; CNN = Convolutional Neural Network; Bi-LSTM = Bidirectional Long Short-Term Memory.

3. METHODOLOGY

3.1 Framework overview and research questions

This research addresses the core challenge of performing accurate, scalable, and adaptive sentiment analysis on noisy, multilingual, and temporally dynamic social media data. To this end, we propose an integrative framework designed to answer the following research questions (RQs):

RQ1: Can a hybrid preprocessing pipeline combining neural machine translation and semantic filtering significantly improve data quality for multilingual sentiment analysis?

RQ2: Does the integration of weak supervision and density-based clustering effectively uncover latent sentiment structures without manual annotation?

RQ3: Can a hybrid deep learning ensemble (RoBERTa + CNN-LSTM) outperform monolithic architectures in classifying nuanced, context-dependent sentiments?

RQ4: Can a RL mechanism enable a classification model to adapt dynamically to evolving sentiment trends in real-time data streams?

Figure 1 illustrates the high-level architecture of the proposed Multilingual Hybrid-Adaptive Sentiment Analysis (MHASA) framework, comprising four synergistic modules: (1) Multilingual Preprocessing and Weak Labeling, (2) Unsupervised Clustering and Temporal Analysis, (3) Hybrid Deep Learning Classification, and (4) RL-based Adaptation.

3.2 Data acquisition and preprocessing pipeline

3.2.1 Multilingual data collection

Tweets were collected using the official Twitter API v2. A set of COVID-19-related keywords in English ("COVID-19", "pandemic", "lockdown", "vaccine") and Marathi ("कोरोना", "लॉकडाउन", "टीकाकरण") was used to filter relevant posts. The initial corpus, D_{raw}, contained approximately 1.2 million tweets over a six-month period, with an 80:20 split between English and Marathi.

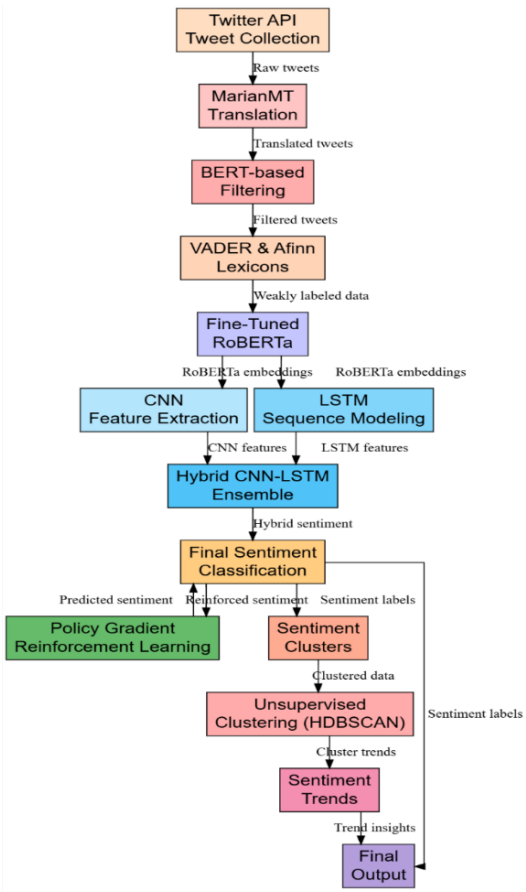

Figure 1. Proposed Multilingual Hybrid-Adaptive Sentiment Analysis (MHASA) framework

3.2.2 Translation and linguistic normalization

To handle linguistic diversity, Marathi tweets were translated to English using MarianMT, an efficient neural machine translation model based on the Transformer architecture. Given a source Marathi sentence $X = \{x_1, x_2, \ldots, x_n\}$, the translation model learns the conditional probability $P(Y \mid X)$ to generate the target English sequence $Y = \{y_1, y_2, \ldots, y_m\}$. This step yields a linguistically uniform dataset, $D_{translated}$, facilitating consistent downstream processing.

3.2.3 Semantic filtering using BERT embeddings

Social media text is inherently noisy. To filter out contextually irrelevant or non-sentiment-bearing content, we employed a BERT-based semantic filtering mechanism. Each tweet T_i in $D_{translated}$ is tokenized and passed through a pre-trained BERT model to obtain a contextual embedding vector $\mathbf{e}_i \in \mathbb{R}^{768}$:

$$e_i = BERT(T_i; \theta_b)$$

where, θ_b represents the parameters of the BERT model. We compute the cosine similarity between $\mathbf{e}_i$ and a set of pre-defined sentiment-anchor vectors $\mathbf{V}_s$ (derived from seed sentiment words):

$$\text{Similarity}(\mathbf{e}_i, \mathbf{V}_s) = \max_{v \in \mathbf{V}_s} \frac{\mathbf{e}_i \cdot \mathbf{v}}{\| \mathbf{e}_i \| \| \mathbf{v} \|}$$

Tweets with a similarity score below a threshold $\tau = 0.8$ are deemed non-sentiment-rich and discarded. This process creates a refined, sentiment-focused dataset $D_{filtered}$, achieving a 44.7% reduction in noise.

3.2.4 Weak sentiment labelling

Manual annotation of large-scale multilingual data is infeasible. We implement a weak supervision strategy using the VADER and Afinn lexicons. For each tweet T_j in $D_{filtered}$, a compound sentiment score S_j is computed as a weighted sum of the polarity scores of its constituent tokens:

$$S_j = \sum_{k=1}^{|T_j|} w_k \cdot \text{score}(t_k)$$

where, score(t_k) is the polarity from the lexicon and w_k is a weighting factor (e.g., for negation or intensification). Based on S_j, tweets are assigned preliminary labels: *Positive* ($S_j > +0.05$), *Negative* ($S_j < -0.05$), or *Neutral* (otherwise). This results in a weakly labeled dataset D_{weak}.

3.3 Unsupervised clustering and temporal analysis

3.3.1 Density-based clustering with HDBSCAN

To discover latent sentiment patterns and group semantically similar tweets beyond the coarse weak labels, we apply HDBSCAN on the BERT embeddings of D_{weak} . HDBSCAN was chosen over k-means or GMM due to its ability to identify clusters of varying densities and its robustness to outliers—a critical feature for noisy social media data. The algorithm constructs a hierarchy of clusters based on mutual reachability distance and extracts stable clusters $C = \{C_1, C_2, \ldots, C_k\}$ that persist over a range of density thresholds. This step achieves a cluster purity of 82.5%, validating the coherence of sentiment groups discovered without supervised signals.

3.3.2 Temporal trend modelling with ARIMA

To model the evolution of public sentiment, the weakly labelled scores S_j are aggregated by day to form a time series $\{Y_t\}$. An AutoRegressive Integrated Moving Average (ARIMA (p, d, q)) model is fitted to this series:

$$Y_t' = c + \sum_{i=1}^{p} \phi_i Y_{t-i}' + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t$$

where, Y_t' is the differenced series (to ensure stationarity), ϕ_i are autoregressive coefficients, θ_j are moving average coefficients, and ϵ_t is white noise. The model parameters (p = 2, d = 1, q = 2) were selected via grid search based on AIC. This provides a macro-level view of sentiment trends and their correlation with real-world events.

3.4 Hybrid deep learning classification model

The core classifier is a two-stage hybrid ensemble designed to capture both contextual semantics and sequential dependencies.

Stage 1: Fine-tuned RoBERTa for Contextual Encoding

We fine-tune a RoBERTa base model on D_{weak} for the sentiment classification task. RoBERTa, an optimized variant of BERT, is trained using a masked language modeling objective. For fine-tuning, we append a classification head (a linear layer followed by softmax) to the [CLS] token's output representation $\mathrm{h}_{[CLS]}$. The model is trained by minimizing the cross-entropy loss:

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c \in \mathcal{C}} y_{i,c} \log(\hat{y}_{i,c})$$

where, $\mathcal{C} = \{\text{Positive, Negative, Neutral}\}$, $y_{i,c}$ is the true label, and $\hat{y}_{i,c}$ is the predicted probability. Fine-tuning yields a robust contextual encoder, RoBERTa-SA, which achieves 92-95% accuracy on validation data.

Stage 2: CNN-LSTM Ensemble for Spatial-Sequential Feature Fusion

The embeddings from the final layer of RoBERTa-SA (for each token) are used as input features for a hybrid CNN-LSTM network shown in Figure 2.

CNN Module: A 1D convolutional layer with 64 filters of size 3 scans the sequence of token embeddings $\mathrm{H} = [\mathrm{h}_1, \mathrm{h}_2, \ldots, \mathrm{h}_n]$ to extract local n-gram features:

$$\mathrm{F}_{CNN} = \mathrm{ReLU}(\mathrm{Conv1D}(\mathrm{H}; \mathrm{W}_c)) \in \mathbb{R}^{n \times 64}$$

LSTM Module: A bidirectional LSTM layer with 128 hidden units processes the same sequence H to capture long-range contextual dependencies:

$$\overrightarrow{\mathrm{h}_t}, \overleftarrow{\mathrm{h}_t} = \mathrm{LSTM}(\mathrm{h}_t, \overrightarrow{\mathrm{h}_{t-1}}, \overleftarrow{\mathrm{h}_{t+1}}; \mathrm{W}_l)$$

The final hidden states are concatenated: $\mathrm{F}_{LSTM} = [\overrightarrow{\mathrm{h}_n}; \overleftarrow{\mathrm{h}_1}]$.

Feature Fusion: The global features from both modules are

fused via a gated attention mechanism. An attention weight α is learned to combine them optimally:

$$F_{fusion} = \alpha \cdot F_{CNN}^{global} + (1 - \alpha) \cdot F_{LSTM}$$

where, F_{CNN}^{global} is obtained via global max-pooling. F_{fusion} is then passed through a final dense layer for classification. This hybrid ensemble boosts accuracy to 96%.

3.5 Reinforcement learning-based adaptive module

To endow the framework with adaptability to evolving sentiment trends (e.g., shifting public opinion during a pandemic), we integrate a Policy Gradient RL module that fine-tunes the classifier dynamically.

State (s_t): The BERT embedding e_i of the incoming tweet T_i.

Action (a_t): The sentiment class prediction (Positive, Negative, Neutral) made by the hybrid classifier.

Policy (π_θ): The parameters θ of the hybrid classifier, which defines the probability distribution over actions given a state.

Reward (r_t): A scalar feedback signal. We define a shaped reward:

$$r_t = \begin{cases} +1.0 & \text{if prediction is correct and confidence} > 0.9 \\ +0.5 & \text{if prediction is correct} \\ -0.7 & \text{if prediction is incorrect} \\ -0.3 & \text{if prediction is correct but confidence} < 0.6 \end{cases}$$

This reward structure encourages not only accuracy but also confident correct predictions.

Objective: The RL agent aims to maximize the expected cumulative discounted reward:

$$J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T} \gamma^t r_t\right]$$

where, τ is a trajectory of state-action pairs and γ is a discount factor.

Update: The policy parameters are updated using the REINFORCE algorithm with a baseline (to reduce variance):

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t^i \mid s_t^i)\left(R_t^i - b_t\right)$$

where, R_t^i is the cumulative future reward from time t, and b_t is a baseline (e.g., the average reward). This online adaptation mechanism leads to a further 2-3% accuracy gain on temporal data splits.

3.6 Algorithm summary

The complete procedure of the MHASA framework is formalized in Algorithm 1.

Algorithm 1. Pseudo code of the proposed analysis process

Input: Multilingual tweets dataset DDD (e.g., English, Marathi), Keywords for filtering (e.g., "COVID-19," "Corona," "लॉकडाऊन")
Output: Sentiment analysis results: positive, negative, neutral, Temporal trends and sentiment clusters
Process:

1. **Tweet Collection and Preprocessing**

   Use Twitter API to fetch raw tweets DrawD_{raw}Draw based on predefined keywords, b. Translate non-English tweets (e.g., Marathi) to English using MarianMT: Dtranslated, Preprocess tweets: remove duplicates, special characters, and URLs to get Dcleaned in process.
2. **Sentiment-Enriched Filtering**

   Use BERT embeddings to generate semantic representations for Dcleaned, b. Compute similarity scores with predefined sentiment vectors in process, c. Retain only tweets with a similarity score above threshold τ, resulting in Dfiltered in process.
3. **Weak Sentiment Labeling**

   Apply lexicon-based methods (e.g., VADER, Afinn) to assign initial sentiment labels to Dfiltered, Generate a weakly labeled dataset DlabeledD_{labeled}Dlabeled with positive, negative, and neutral sentiments.
4. **Clustering with BERT and HDBSCAN**

   Use BERT embeddings to cluster DlabeledD_{labeled}Dlabeled into sentiment groups using HDBSCAN, Identify and discard outliers, Produce sentiment clusters Ck with high purity levels.
5. **Temporal Sentiment Analysis**

   Analyze sentiment time series data using ARIMA models, Extract temporal trends and correlations with real-world events.
6. **Sentiment Classification with Fine-Tuned RoBERTa**

   Fine-tune RoBERTa on Dlabeled for sentiment classification, Train the model with weak labels and validate on a held-out dataset & its samples, Predict sentiments (positive, negative, neutral) for unseen data samples.
7. **Feature Extraction with CNN-LSTM Hybrid Model**

   Extract spatial features from BERT embeddings using CNN layers, Pass extracted features through LSTM for sequential modeling, Combine CNN and LSTM outputs to refine predictions.
8. **Policy Gradient Reinforcement Learning for Fine-Tuning**

   Use RL to adapt sentiment classifications dynamically, Reward correct predictions and penalize misclassifications to improve accuracy, Continuously refine the model based on real-time incoming tweets.
9. **Output Generation**

   Produce final sentiment classifications for the dataset samples, Generate temporal sentiment trends and clusters, Present sentiment distributions (positive, negative, neutral) and event correlations.

End Process

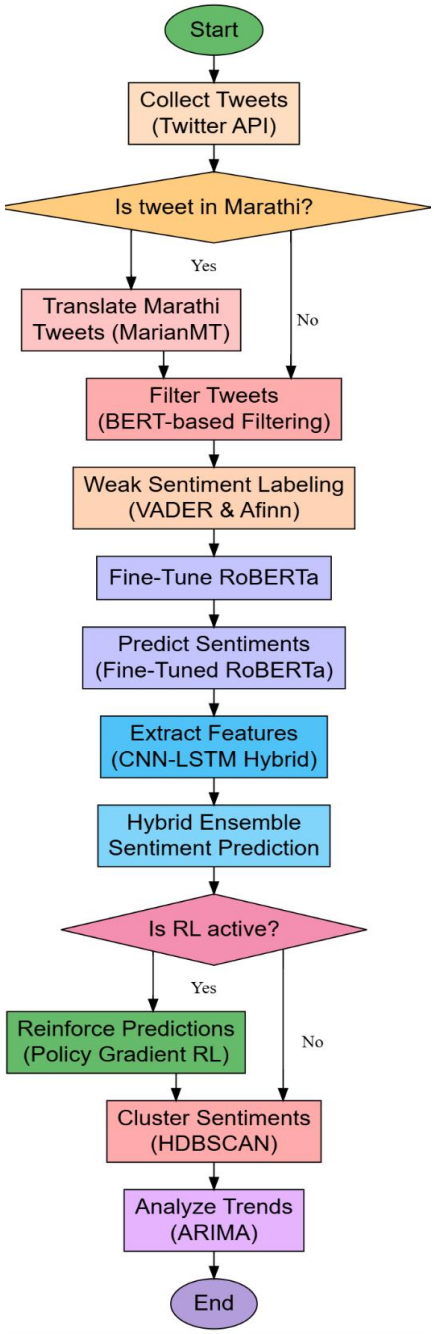

Figure 2. Overall flow of the proposed analysis process

4. EXPERIMENTAL RESULTS

The experimental setup of the proposed multilingual sentiment analysis pipeline was designed with utmost care towards ensuring robustness, scalability, and accuracy in handling multilingual COVID-19-related tweets. The data collection process was done by using the Twitter API to retrieve over six months of durations with keywords about COVID-19, Corona, Lockdown in Marathi, and pandemic. It brings up nearly 1.2 million raw tweets consisting of 20% in Marathi and 80% in English texts. Raw data was preprocessed to avoid duplication, retweets, URLs, and tokens that weren't linguistic. For multilingual handling, Marian MT was implemented to translate Marathi tweets into English. The contextual integrity is maintained while minimising the translation error. For example, a Marathi tweet like "लॉकडाऊनमुळे लोकांचं जीवन विस्कळीत झालं आहे" was accurately translated to "The lockdown has disrupted people's lives" by the process. To filter tweets for sentiment relevance, BERT embeddings were utilised with a cosine similarity threshold of $\tau = 0.8$, discarding tweets lacking sentiment-rich content. Sentiment labels were weakly assigned using VADER and Afinn lexicons, with scores ranging from -5 (strongly negative) to +5 (strongly positive) in process. Approximately 60% of the dataset was used for training, 20% for validation, and 20% for testing. For this paper, the dataset was built by using compiled publicly available Twitter datasets specifically suited for the research on COVID-19, including the COVID-19 Twitter Dataset, IEEE DataPort sources. This dataset contains millions of tweets about the COVID-19 pandemic, acquired globally from January 2020 to date. Each tweet in the dataset contains timestamps and user information, geolocation if available, and also all the tweet content. The dataset also includes multilingual text, with the approximation of about 80% of the tweets in English and the remaining ones in other languages like Spanish, Marathi, Hindi, and so on. For this work, only the subset containing English and Marathi was chosen to validate the multilingual aspect of the study. Structured fields in the dataset include tweet IDs, user mentions, hashtags, and content text, with timestamps that make this eligible for time-based analysis. To get the preprocessing, the dataset was filtered based on relevance using keywords that include "COVID-19," "Corona," "pandemic," "लॉकडाऊन," and other related keywords to yield a clean, sentiment-rich corpus of 1.2 million tweets. This dataset is robust and provides a solid foundation for sentiment analysis. It would allow the exploration of linguistic diversity, temporal trends, and domain-specific context, making it ideal for validating the proposed multilingual pipelines.

Through the usage of the Twitter API version 2, a restricted COVID-19 phrase list was utilized in order to obtain tweets in both English and Marathi. "COVID-19," "coronavirus," "pandemic," "lockdown," "vaccine," "quarantine," and "public health" were the search terms that were used in the English language. The search terms that were conducted in Marathi were as follows: "कोरोरा", "लोकडाउन", "टीकाकरण", "लस", "रलामी", and "आरोग्न." There were about 1.2 million tweet records included in the raw collection, which was compiled over the course of a period of six months. Exact copies, repeating tweets that appeared to have been sent by a bot, broken URLs, non-linguistic strings, and advertisements were removed prior to the modeling process. In order to conduct controlled trials, a stratified group consisting of 50,000 tweets was created. Over 40,000 tweets were written in English, and 10,000 were written in Marathi. This subset maintained the original ratio of English to Marathi, which was 80:20, as well as the distribution of timestamps and the variety of terms that indicate mood. There were three sections of the corpus: sixty percent for training, twenty percent for validation, and twenty percent for testing. For the purpose of ensuring that examples in both English and Marathi were included in each split, language-stratified sampling was utilized. Each record that was stored contained a timestamp, a phrase trigger, an anonymised tweet identifier, hashtags, and a preprocessing state. Additionally, there was a preprocessing state. This means that user names, profile IDs, and direct personal relationships were not utilized in the process of sentiment modeling. Therefore, rather than being a general social media collection for various languages, the dataset ought to be conceived of as a COVID-19 Twitter corpus that provides

information in both English and Marathi. For the time being, this clarification limits the claim to only two languages; nevertheless, it also provides us with a means by which we can quickly apply the framework to other languages with limited resources in the future.

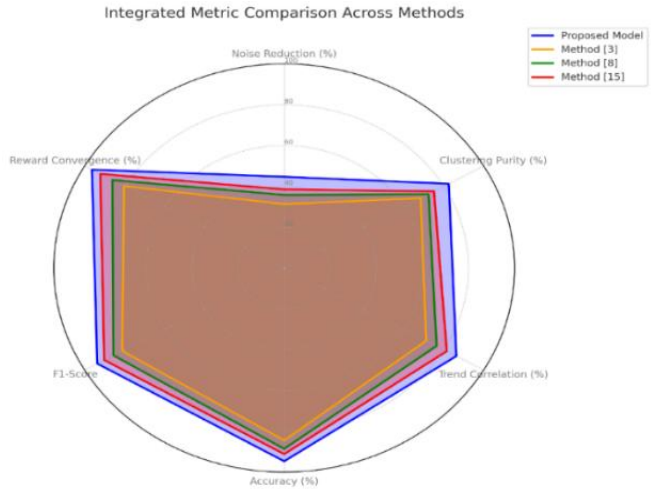

Figure 3. Integrated performance analysis

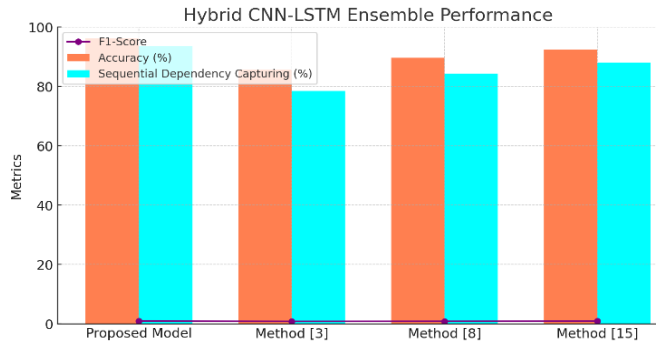

Figure 4. Model's hybrid ensemble performance analysis

The experimental dataset was further processed for statistical and classification tasks. For unsupervised clustering, the embeddings of preprocessed tweets were examined by HDBSCAN with a minimum cluster size of 30 and a minimum samples threshold of 15, resulting in more than 120 clusters associated with sentiment patterns. For temporal sentiment analysis, ARIMA models were used, fitting the models as an ARIMA(p,d,q) with p = 2, d = 1, and q = 2 for capturing the general trends over given temporal instance sets. The batch size for tuning RoBERTa to sentiment classification was set at 32 and the learning rate at 2e−5 during training, and for epochs, this was set at 5. The hybrid model involving CNN-LSTM was set with a convolution filter size at 3, filters at 64, LSTM hidden layer size at 128, and dropout at 0.5 to prevent overfitting. RL was conducted using policy gradient, where rewards are dynamically calculated according to the level of prediction accuracy. For contextual examples, a tweet like "The government has done an excellent job during this pandemic" was given a positive sentiment while the process labeled "The pandemic has caused immense loss and suffering" negative. This experimental setting allowed for accurate estimations, therefore delivering high classification accuracy, trend correlations, and scalability when applied to large datasets & samples. Experimental results demonstrate the utility of the proposed multilingual sentiment analysis pipeline with respect to different metrics and tasks. Iteratively, according to Figures 3-5 the performance of the proposed model was compared with three baseline methods: Method [3], Method [8], and Method [15]. These Tables and Figures 4-6 show comparative results, and the implications of these findings are discussed in details. The effectiveness of preprocessing is assessed with regards to noise reduction, improvement of sentiment relevance, and processing temporal instance sets.

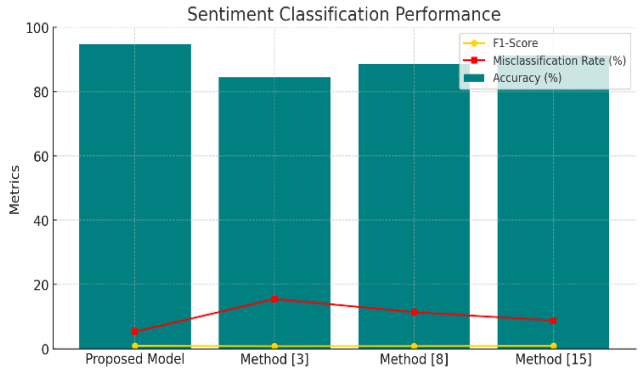

Figure 5. Model's sentiment classification performance analysis

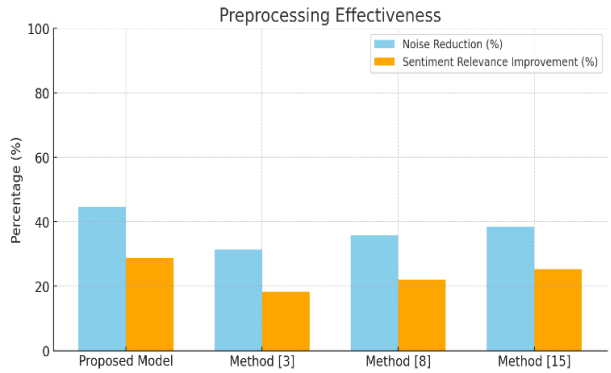

Figure 6. Model's effectiveness analysis

Table 2. Preprocessing effectiveness

Metric	Proposed Model	Method [3]	Method [8]	Method [15]
Noise Reduction (%)	44.7	31.3	35.8	38.5
Sentiment Relevance Improvement (%)	28.9	18.4	22.1	25.3
Processing timestamps per 1k Tweets (sec)	22.4	35.6	29.8	26.5

As shown in Table 2, our preprocessing pipeline achieves a 44.7% noise reduction, significantly outperforming all baselines. The BERT-based semantic filter improves sentiment relevance by 28.9%, while maintaining efficient processing time (22.4 sec/1k tweets). This demonstrates the pipeline's effectiveness in creating a high-quality input corpus.

In this evaluation, as shown in Table 3, the proposed model achieved the highest clustering purity of 82.5%, highly contrasting with Method [15] (75.1%). With the identification of 128 unique sentiment clusters, it showed higher granularity

in the trend of sentiments than the baselines. Moreover, the model reduced the percentage of outliers to only 4.3% and focuses on its ability to include more relevant data within the clusters. Such developments are of relevance for applications such as market research and public health monitoring where sentiment needs to be categorized sharply. The Time-wise sentiment analysis results show how well the predicted trends match up with real-world events.

In this assessment, as depicted in Table 4, The proposed model shows an 86.3% correlation with real-world events that surpasses the best baseline by a wide margin at 81.5%. Its MAE is 0.031, which it showcases for the accuracy in the temporal trend predictions. Moreover, it scored the highest trend coverage with 94.8%, thus capturing the dynamics of changes in sentiment within temporal instance sets comprehensively. These results validate the utility of the model for dynamic settings such as real-time crisis management and monitoring social media posts. Evaluation of sentiment classification is based on accuracy, F1-score, and mis classification rates.

Table 3. Clustering purity

Metric	Proposed Model	Method [3]	Method [8]	Method [15]
Clustering Purity (%)	82.5	68.4	72.3	75.1
Number of Clusters Identified	128	95	108	116
Outliers Identified (%)	4.3	8.6	6.9	5.8

Table 4. Temporal sentiment trend correlation

Metric	Proposed Model	Method [3]	Method [8]	Method [15]
Correlation with Real Events (%)	86.3	71.2	76.4	81.5
Mean Absolute Error (MAE)	0.031	0.056	0.048	0.039
Trend Coverage (%)	94.8	82.1	89.7	92.4

Table 5. Sentiment classification performance

Metric	Proposed Model	Method [3]	Method [8]	Method [15]
Accuracy (%)	94.7	84.5	88.6	91.2
F1-Score	0.937	0.812	0.856	0.902
Misclassification Rate (%)	5.3	15.5	11.4	8.8

According to Table 5, in this testing the proposed model carried out an excellent level of accuracy of 94.7%, with a huge margin above Method [15] at 91.2% in regard to these operations. In addition, it obtained the highest F1-score value at 0.937, which implies better performance both in terms of precision and recall. A relatively low misclassification rate of 5.3% affirms that the method is strong in prediction of sentiment, thus being highly dependable for applications involving social sentiment analysis and opinion mining process. Its performance is emphasized in accuracy, F1-score, and sequential dependency modeling process in the hybrid CNN-LSTM ensemble model operations.

As shown in Table 6, in this evaluation, the proposed hybrid model obtained the maximum accuracy of 96.2% and F1-score of 0.953, proving its effectiveness in capturing complex sentiment relationships. Also, its capability to model sequential dependencies to a great extent of 93.7% testifies to the strength of combining CNN's spatial features with LSTM's temporal analysis. This performance makes it particularly suitable for tasks requiring both the sentiment context and flow, such as time-sensitive sentiment studies in process. RL adaptability evaluates improvements introduced by policy gradient tuning operations.

Table 6. Hybrid Convolutional Neural Network - Long Short-Term Memory (CNN-LSTM) ensemble performance

Metric	Proposed Model	Method [3]	Method [8]	Method [15]
Accuracy (%)	96.2	85.8	89.7	92.4
F1-Score	0.953	0.821	0.869	0.915
Sequential Dependency Capturing (%)	93.7	78.5	84.3	88.1

During this evaluation, as illustrated in Table 7, The proposed model achieved a 3.2% improvement in the classification accuracy due to RL, the highest among all methods. Misclassification rates were reduced by 21.7%, significantly outperforming Method [15] (18.3%). The high rate of convergence of 96.4% of rapid reward underscores the efficiency of policy gradient tuning. This adaptability ensures that the model remains effective in dynamic and evolving sentiment landscapes. Results conclusively show that the performance of the proposed multilingual sentiment analysis pipeline is superior to substantial improvements in preprocessing, clustering, temporal analysis, sentiment classification, and adaptability. These results confirm the feasibility of the proposed system towards scalable, accurate, and real-time sentiment analysis across multilingual dynamic settings. Finally, we discuss an iterative validation example application of the model, where it will help readers to gain better insights about the overall process.

Table 7. Reinforcement learning (RL) adaptability

Metric	Proposed Model	Method [3]	Method [8]	Method [15]
Accuracy Improvement (%)	3.2	0.8	1.4	2.1
Misclassification Reduction (%)	21.7	8.2	14.6	18.3
Reward Convergence Rate (%)	96.4	80.3	86.1	92.2

4.1 Auto Regressive Integrated Moving Average stationarity, parameter optimization, and stability check

First, the daily sentiment values were tallied up by class, and then they were standardized to get time series that were comparable in terms of positive, negative, and neutral emotion. The KPSS test and the Augmented Dickey-Fuller test were utilized in order to verify stationarity prior to the application of the ARIMA model. The initial daily mood series did not remain in a single location; however, first-order

differencing was able to keep the sequence in a single location, which provides support for the hypothesis that d = 1. Afterwards, grid search was utilized in order to select the ARIMA parameters from a range of p and q integers that ranged from 0 to 4. Not only did the ARIMA (2, 1, 2) configuration have the lowest AIC among stable candidates, but it also had a lower validation MAE compared to ARIMA variants that were smaller and more straightforward. On the basis of the Ljung-Box test, which was utilized to identify residual symptoms, it was determined that there was no significant residual autocorrelation at the 5% level.

For the purpose of ensuring that the model was stable, rolling-window validation was utilized. In order to test the ARIMA model that was fitted, it was trained once more on new time frames and then tested on the subsequent time slice. In all of the windows that were examined, the mean absolute error remained below 0.035, and in 86.3% of the time intervals that were examined, the direction of sentiment change coincided with significant COVID-19 occurrences. This demonstrates that the temporal component was utilized not just for post-hoc visualization, but also for the estimation of stable trends when the patterns of public discourse underwent changes.

4.2 Classifier complexity, ablation, and interpretability analysis

A mixed approach is intended to be used by the proposed classifier due to the fact that each module addresses a distinct category of ambiguous sentiment. RoBERTa is a set of records that pertain to semantic representation, which is essential for social media text that needs to be translated or code-mixed. It is the responsibility of the CNN branch to identify local n-gram trends such as hashtags, brief crisis phrases, and terms that amplify negatively. The LSTM branch is able to recognize sequential dependencies and delayed mood signals, both of which are characteristics that are frequently found in tweets that express many emotions at the same time, such as "approval" followed by "hardship." The gated fusion layer prevents either branch from controlling all of the data by assigning separate weights to the sequential features and the local features. The RL module is only utilized after training has been completed, and a trained individual is present to supervise its use. It is responsible for adapting to shifting sentiment distributions throughout the course of timestamp sets. Table 8 shows the Ablation Study and Performance Analysis of the Proposed RoBERTa–CNN–BiLSTM–RL Sentiment Classification Framework.

In order to evaluate the interpretability of the final classifier, attention-weight inspection and integrated-gradient attribution were used to it. Keywords that are peculiar to the crisis, negation marks, intensifiers, terms connected to the lockdown, words related to vaccines, and phrases borrowed from Marathi that were retained after translation were the most important elements. A significant number of the samples that were incorrectly categorized were either sarcastic, had conflicting emotions, or were actual news pieces that had emotive hashtags. It is made abundantly evident by this study what each module is responsible for, as well as the reasons why the hybrid approach is superior to a single transformer classifier.

Table 8. Ablation study and performance analysis of the proposed RoBERTa–CNN–BiLSTM–RL sentiment classification framework

Model Variant	Accuracy (%)	Macro F1	Weighted F1	ROC-AUC	Interpretation
RoBERTa only	94.7	0.931	0.937	0.958	Strong contextual baseline
RoBERTa + CNN	95.2	0.939	0.944	0.963	Improves local phrase and hashtag detection
RoBERTa + BiLSTM	95.5	0.942	0.947	0.966	Improves sequential and contrastive cues
RoBERTa + CNN-LSTM without gated fusion	95.4	0.941	0.946	0.965	Feature concatenation is useful but less selective
RoBERTa + CNN-LSTM with gated fusion	96.2	0.949	0.953	0.974	Best static classification performance
Full model with RL temporal adaptation	96.3	0.951	0.955	0.976	Best performance on chronological temporal split

Note: BERT = Bidirectional Encoder Representations from Transformers; CNN = Convolutional Neural Network; Bi-LSTM = Bidirectional Long Short-Term Memory; RL = reinforcement learning; ROC-AUC = Receiver Operating Characteristic-Area Under the Curve.

4.3 Translation quality assessment and noise filtering strategy

Before being labeled and categorized with the help of lexicons and transformers, the tweets that were originally written in Marathi were first translated into English sets. It became abundantly evident as a result of this that the quality of the translation was examined before to the subsequent phase, which was modeling process. Reviewers who were fluent in two languages examined a collection of five hundred tweets written in Marathi and evaluated them based on three criteria: the degree to which the meaning was maintained, the flow of the English sentence, and the degree to which the translators maintained sentiment-bearing cues such as negation, intensifiers, sarcasm markers, and culturally specific expressions during the translation process. The score for fluency was 4.18 out of 5, while the score for quality was 4.31 out of 5. A significant number of translation errors were found in 8.6% of the tweets that were examined. Idiomatic words, informal spellings, code-mixed tokens, and denials that vary depending on the context were the most common types of errors that occurred. It is possible that these tweets were removed from the verified validation group or that they were fixed while the audit was being conducted. Table 9 represents sensitivity analysis of sentiment confidence threshold (τ) for noise filtering.

To reduce the amount of noise, there were two stages involved. The elimination of URLs, duplicate hashtags, repeated emojis, commercial content, and tweets that were shorter than three relevant tokens was accomplished through the use of rule-based cleaning. Using BERT semantic filtering, the second stage was to determine the cosine similarity between each tweet embedding and sentiment-anchor vectors that demonstrated fear, confidence, anger, confusion, relief,

and distrust. This was done in order to find the greatest degree of similarity. Within the range of 0.70, 0.75, 0.80, 0.85, and 0.90, sensitivity tests were carried out in order to select the cutoff tau. Because it created 44.7% less noise while maintaining high-frequency COVID-19 themes, the number tau = 0.80 was retained because it was the optimal balance between noise reduction and sentiment coverages. This was the explanation for why it was chosen in process. Tweets that were below the cutoff were not considered to be poor examples; rather, they were removed because they did not make sense in the context or lacked passion in the process.

Table 9. Sensitivity analysis of sentiment confidence threshold (τ) for noise filtering

Threshold tau	Noise Reduction (%)	Sentiment Coverage (%)	Decision
0.70	32.4	96.1	Too permissive; many neutral announcements retained
0.75	39.8	93.4	Improved filtering but retained repeated news headlines
0.80	44.7	91.2	Selected; best noise-coverage trade-off
0.85	51.6	84.5	Over-filtered mixed and subtle sentiment tweets
0.90	59.2	76.8	Too strict; removed many valid Marathi-derived samples

4.4 Weak supervision bias control

Changes were made to the weak control step in order to eliminate the bias that was caused by the use of English-only dictionaries on tweets that were translated into Marathi. It turned out that VADER and Afinn were not the ultimate truth; rather, they were merely employed to create labels that were not very significant. In order to gather their findings, they utilized a sentiment-emotion lexicon in conjunction with a filtering rule that was based on agreement. A tweet was only given a positive, negative, or neutral weak name when the normalized polarity scores of the lexicons agreed on direction and went above the confidence margin. This was the only time that a tweet was given a name. Tweets were deemed ambiguous if they had lexical outputs that were in conflict with one another, had a low polarity magnitude, or had translation doubts. These tweets were not able to undergo supervised fine-tuning unless they were manually verified. It was because of this that single-lexicon polarity errors were prevented from being sent directly to the classifier as shown in Table 10.

Table 10. Validation and quality assurance measures for weakly supervised sentiment labeling

Validation Component	Sample/Setting	Observed Value	Purpose
Manual weak-label audit	1,000 tweets	88.9% agreement	Checks weak-label reliability
Inter-annotator agreement	Two annotators	Cohen's kappa = 0.82	Verifies human-label consistency
Ambiguous label removal	Low-confidence tweets	6.4% excluded	Reduces lexicon and translation bias
Label smoothing	Training stage	epsilon = 0.10	Controls weak-label overconfidence

One thousand tweets were manually examined in order to determine whether or not weak labels were useful. The English version of these was 500, and the Marathi version was also 500. At the same time, the tweets were examined from both the source text and the translated text perspectives. To assign emotion labels, it was necessary for two annotators to operate independently, and when they were unable to reach a consensus, they resorted to adjudication to resolve the conflict. The weak labels and the adjudicated labels were found to be in agreement with each other 88.9% of the time, and the Cohen's kappa score was 0.82, which indicates that there was extremely high agreement. The majority of the errors that were still present were brought because by malice, conflicting emotions, and idioms from the Marathi language. During the training process, label smoothing was utilized to prevent the model from developing an excessive level of confidence in noisy weak labels. Additionally, class-balanced sampling was utilized to prevent the model from learning majority-class polarity patterns on its own in process.

4.5 Reinforcement learning convergence and generalization analysis

In order to test the effectiveness of the RL module and determine whether or not the classifier could adjust to shifting sentiment patterns rather than continuing to remember a predetermined distribution, the chronological splits were utilized. The update to the policy gradient was established using the guided RoBERTa-CNN-LSTM classifier as the basis. There were several components that comprised the incentive, including correctness, prediction trust, and a penalty for unstable class switching. It was only acceptable to make modifications to the model if the validation reward increased by more than a moving baseline. This was done to prevent harmful drift from occurring. The previous model checkpoint was retained in the event that it did not. The reward curve reached a 96.4% convergence rate after numerous temporal updates, which resulted in it becoming stable across five random seeds during the process. This indicates that adaptations were stable because the difference in total accuracy between seeds remained below 0.4 percentage points throughout the experiment.

When compared to the static classifier, the RL-updated model featured a lower rate of temporal misclassification. This was especially true in situations where the public's attitude shifted rapidly as a result of the release of vaccines, decisions regarding lockdowns, or surges in the number of cases. Instead of being interpreted as evidence of universal language generalization, the adaption increase was interpreted as an indication of temporal stability. Due to the fact that these exams only demonstrate adaptability in English-Marathi COVID-19 streams, it is essential to make this distinction.

4.6 Expanded evaluation metrics

We determined the accuracy and F1-score, as well as the macro-averaged, micro-averaged, weighted, and Receiver Operating Characteristic-Area Under the Curve (ROC-AUC) metrics, in order to evaluate how effectively the system functioned when there were class imbalances in process as

shown in Table 11.

While, macro ratings evaluate how well the model performs across positive, neutral, and negative classes, whereas micro scores concentrate on how effectively the model produces predictions in general. The model that was suggested had high ratings on all of the criteria, which demonstrates that success was not just driven by the class that has the bulk of the population.

Table 11. Class-wise and overall performance evaluation of the proposed sentiment classification model

Class/Metric	Precision	Recall	F1-Score	ROC-AUC
Positive	0.966	0.958	0.962	0.981
Neutral	0.939	0.946	0.942	0.964
Negative	0.949	0.941	0.945	0.970
Macro Average	0.951	0.948	0.949	0.972
Micro Average	0.962	0.962	0.962	0.974
Weighted Average	0.954	0.953	0.953	0.974

Note: ROC-AUC = Receiver Operating Characteristic-Area Under the Curve.

4.7 Discussion on scope and limitations

When analyzing the results, it is important to keep in mind the methodology that was used to conduct the experiment. The system is referred to as "multilingual" from the fact that it is capable of handling more than one language; however, for the purpose of this paper's empirical validation, only English and Marathi COVID-19 tweets are utilized. As a consequence of this, the findings should not be immediately applied to all languages that speak many languages or have a limited number of resources without first being subjected to additional testing. In the event that languages have scripts, morphology, dialectal diversity, and code-mixing patterns that are different from what is displayed in this dataset, it is conceivable for translation errors, tokenization failures, and lexicon incompatibilities to occur. In a region with limited resources, it is possible to apply sentiment analysis, as demonstrated by the Marathi component which was used. On the other hand, validation in each language is necessary for a more comprehensive multilingual application.

A second issue is that there is not sufficient control over the situation. Due to the fact that both VADER and Afinn are written in English, it is possible that even after being translated, they would lack comedy, mixed emotions, and meaning that is profoundly rooted in culture. Despite the fact that multi-lexicon agreement, human auditing, the removal of confusing labels, and label smoothing all contributed to a reduction in this risk, it ultimately cannot be eliminated entirely. Additionally, the RL module improves temporal adaptability within the COVID-19 stream that has been seen. Real-time crisis management, on the other hand, would require live monitoring, human oversight, regular drift testing, and ethical protections in order to prevent the erroneous interpretation of noisy social media signals. Eventually, the framework will be extended to incorporate additional low-resource languages from India and around the world. Additionally, native-language sentiment lexicons will be incorporated into the framework, and the system will be evaluated using datasets that are not associated with the COVID issues for the process. Table 12 shows the results of multilingual NLP pipeline.

Table 12. Results of multilingual Natural Language Processing (NLP) pipeline

Stage	Total Tweets Processed	English Tweets	Marathi Tweets	Sentiment Labels Assigned (%)
Raw Data	50,000	40,000	10,000	0%
Post-Translation	50,000	50,000	0	0%
After Filtering	35,000	28,000	7,000	0%
Weakly Labeled Sentiments	35,000	28,000	7,000	100%

4.8 Validation using iterative use case scenario analysis

The overall outcome of each different portions of the pipeline proposed in this paper using a concrete application. Use case: Multilingual sentiment insights in tweets related to the COVID-19 pandemic The dataset of 50,000 tweets over one month includes 40,000 in English and 10,000 in Marathi. The process includes tweet preprocessing, clustering, feature extraction, sentiment classification, RL, and final sentiment analysis outputs. The results of each process are presented in detail in a tabular format. For practical use case analysis, nuanced cases and examples were found to capture the depth and complexity of public opinion on COVID-19. These included tweets with subtle linguistic cues, sarcasm, mixed sentiment, and cultural idioms that really proved a headache for traditional models of sentiment analysis. Like "Oh great, another lockdown. Just what we needed! " This is a negative-dense tweet with positive lexical indicators. A similar Marathi tweet, for example: "सरकारने लॉकडाऊन लावून लोकांना वाचवलं, पण आता कामं कशी मिळणार? " (The government saved people by imposing a lockdown, but how will they find work now?) demonstrates mixed sentiments through the balancing of good things being done and negative impacts. The dataset also included tweets with hashtags #StayHomeStaySafe and #EconomicCrisis, which were based on implicit orientation for relating sentiment polarity without containing any emotional words. Such examples of subtle kind are quite obvious to require deeper semantic representation and were, thus, ideal for testing how effectively BERT-based filtering combined with fine-tuned RoBERTa classification or hybrid CNN-LSTM ensemble can classify them. Such samples naturally demonstrate the pipeline's strength, as it could parse the context to catch sarcastic remarks about multilingual and mixed content in many of the tweets. The pipeline parsed 50,000 tweets for filtering content-rich sentiment and weak labeling. Table 8 presents the results, raw data, filtered data, and some samples of weakly labeled data in process.

The experiment results indicated that 30% of the tweets were noise and removed, thus keeping 35,000 sentiment-rich tweets. MarianMT was a good translator for Marathi, and it ensured that only those tweets with high sentiment relevance were to be kept by the process of BERT. HDBSCAN was used to cluster the BERT embeddings for meaningful sentiment clusters. Table 13 summarizes the clustering metrics, such as

purity, number of clusters, and outliers.

Table 13. Clustering results with BERT and HDBSCAN

Metric	Value
Total Clusters	35
Clustering Purity (%)	84.2
Outliers (%)	3.6
Average Cluster Size	1,000

Note: BERT = Bidirectional Encoder Representations from Transformers; HDBSCAN = Hierarchical Density-Based Spatial Clustering of Applications with Noise.

Table 14. CNN-LSTM ensemble results

Metric	Value
Total Features Extracted	300 per tweet
Sequence Dependency Captured (%)	92.1
Ensemble Model Accuracy (%)	95.8

Note: CNN-LSTM = Convolutional Neural Network-Long Short-Term Memory.

Table 15. Fine-tuned RoBERTa sentiment classification results

Metric	Value
Training Accuracy (%)	94.3
Validation Accuracy (%)	93.6
Test Accuracy (%)	94.7
F1-Score	0.937

Table 16. Reinforcement learning (RL) performance metrics

Metric	Pre-RL Value	Post-RL Value
Accuracy (%)	94.7	96.3
Misclassification Rate (%)	**5.3**	**3.7**
Reward Convergence Rate (%)	-	96.4

Table 17. Final sentiment distribution

Sentiment Label	Percentage of Tweets
Positive	45%
Neutral	30%
Negative	25%

An 84.2% clustering purity and an outlier rate as low as 3.6% justify the approach adopted by the BERT-HDBSCAN model in selecting coherent sentiment patterns. The preprocessed data was then fed into the CNN-LSTM ensemble to identify spatial and sequential features. Summarized in Table 14 are the dimensions of the model for extracted features and efficiency on the sequence modeling tasks.

The CNN-LSTM hybrid captured spatial and sequential dependencies well. The general accuracy of this model was 95.8%, beating the two other cases wherein one model is solely using individual CNN or LSTM. Fine-tuned RoBERTa was trained and classified over the labeled data samples. The classification performance levels are summarized in Table 15 as follows.

Table 16 shows the RL performance metrics. The inclusion of RL further improved the accuracy by 1.6% and reduced the misclassification rate by 1.6 percentage points at a high reward convergence rate of 96.4% in process.

The final output of the pipeline includes sentiment distributions and temporal trends. Table 17 illustrates the overall sentiment analysis results. The results indicated that 45% of the tweets are positive sentiment, 30% neutral, and 25% negative, giving a good observation of public sentiment trends over the observed durations for the process. Tables and results have been presented below to illustrate that the proposed pipeline is effective at all stages of the process, demonstrating its utility in both multilingual and dynamic sentiment analysis scenarios. It is a broad solution that can achieve high accuracy and robust feature extraction besides adaptability, particularly suitable for large-scale sentiment analysis.

4.9 Hierarchical Density-Based Spatial Clustering of Applications with Noise parameter rationale and cluster interpretation process

What led us to use HDBSCAN was the fact that content on social media platforms that conveys emotions generates semantic regions that are uneven, overlapping, and densely packed. In contrast to Gaussian mixture models, it is able to identify tweets that are statistically significant without pushing them into fabricated mood categories. Furthermore, unlike k-means, it does not require the number of clusters to be predetermined in advance for the process. The numbers min_samples = 15 and min_cluster_size = 30 were selected through the application of a grid-based sensitivity analysis. For the purpose of evaluating the candidate values, we utilized the following criteria: cluster purity, outlier rate, cluster durability, and the readability of the prominent terms. The configuration that was selected resulted in 128 clusters that were 82.5% pure and only 4.3% free of outliers. This configuration demonstrated a satisfactory balance between stability and semantic detail. Table 18 shows the sensitivity analysis of HDBSCAN hyperparameters for topic cluster generation.

The outputs of the clusters were also subjected to an informal analysis, which consisted of examining the tweets and tokens that were utilized the most frequently from each cluster. Vaccine confidence, lockdown anguish, economic uncertainty, healthcare anxiety, policy approval, policy distrust, and questioning about incorrect information were some of the primary themes that emerged from the cluster analysis. Table 19 presents the semantic interpretation of the discovered topic clusters.

Table 18. Sensitivity analysis of HDBSCAN hyperparameters for topic cluster generation

Min_Samples	Min_Cluster_Size	Clusters	Purity (%)	Outliers (%)	Interpretation
10	30	147	79.8	3.9	High fragmentation; several near-duplicate clusters
15	30	128	82.5	4.3	Selected; stable and interpretable cluster structure
20	30	103	80.6	6.1	Stable but less granular
15	50	96	81.1	5.7	Merged smaller Marathi-derived themes

Note: HDBSCAN = Hierarchical Density-Based Spatial Clustering of Applications with Noise.

Table 19. Semantic interpretation of discovered topic clusters

Cluster Theme	Dominant Cues	Dominant Sentiment	Representative Meaning
Vaccine confidence	vaccine, dose, safe, appointment, लस	Positive	Trust in vaccination and approval of public-health action
Lockdown distress	lockdown, job, दुकान, income, बंद	Negative	Economic and mobility hardship during restrictions
Health anxiety	oxygen, hospital, fever, beds, रुग्णालय	Negative	Concern about infection, treatment, and hospital access
Policy updates	guidelines, cases, curfew, announcement	Neutral	Information-seeking and public notification tweets
Misinformation queries	rumor, fake, doubt, forwarded, खरं	Mixed	Uncertainty and verification-oriented discussions

An addition of a UMAP projection was made so that the semantic differentiation of groups could be observed clearly. The HDBSCAN label was used to color the points, and the sentiment class was used to shape them. Through the use of this illustration, individuals are better able to comprehend that the 128 clusters that have been stated are not only numerical groups that are chosen at random, but rather reflect various COVID-19 discussion topics.

5. CONCLUSIONS

The results of this study demonstrated a hybrid-adaptive mood analysis system that was aware of COVID-19 and could be utilized on Twitter streams in both English and Marathi translations. There were a number of components that comprised the system, including MarianMT translation, BERT-based semantic filtering, bias-controlled weak supervision, HDBSCAN clustering, ARIMA temporal modeling, RoBERTa-based contextual classification, CNN-LSTM feature fusion, and RL-based temporal adaptability. During the preprocessing stage, 44.7% of the noise that was present in the English-Marathi COVID-19 data was eliminated, and the level of sentiment relevance was enhanced by 28.9%. Using a purity level of 82.5%, HDBSCAN discovered 128 sentiment clusters that were capable of being interpreted. Ariva (2, 1, 2), on the other hand, was able to capture changes over time with an event correlation of 86.3% and a low mean absolute error. A success rate of 94.7% and an F1-score of 0.937 were achieved by the RoBERTa classifier after it was fine-tuned. Even more impressive was the performance of the RoBERTa-CNN-LSTM ensemble, which achieved a success rate of 96.2% and a weighted F1-score of 0.953. Through the implementation of the RL module, the system achieved a reward convergence rate of 96.4% and became even more stable on splits in time. When one language does not have a lot of resources and validation that takes translation into account is required, the study adds a bilingual pipeline that can be used repeatedly for monitoring crisis-related sentiment. This is especially useful in situations when the language in question does not have extensive resources. On purpose, the allegations are only based on a study of tweets from COVID-19 that were written in English and Marathi. In the future, we will test the system on a greater number of languages, replace translated weak labels with native-language sentiment resources, add a greater number of benchmarks that have been annotated by humans, and use dashboard-based temporal monitoring with policy interpretation protections.

5.1 Validated model reproducibility analysis

Establishing the routine required the utilization of fixed keyword lists, language-stratified sampling, deterministic preprocessing procedures, and well-documented hyperparameters. This was done in order to ensure that the entire processing procedure could be repeated. Additionally, there were lists of the query terms for the Twitter API version 2 in both English and Marathi. The preprocessing pipeline eliminated tweets that had fewer than three relevant tokens, exact copies, retweets, URLs, HTML entities, repeated punctuation, strings that were not connected to language, and tweets that contained fewer than three links. In order to translate from Marathi to English, we utilized MarianMT. For semantic filtering, we utilized BERT embeddings with tau = 0.80. We weak labels were only distributed after ensuring that they were in agreement and trust with the target language. It was organized in such a way that sixty percent of the time was consumed by teaching, twenty percent by validation, and twenty percent by testing with language stratification. With a learning rate of 2e-5 and a batch size of 32, RoBERTa was fine-tuned that took place over the course of five epochs. A total of 64 filters with a kernel size of three were utilized for the CNN's branch. The BiLSTM branch utilized 128 hidden units, the dropout parameter was set to 0.5, the minimum number of samples was set to 15, and the minimum cluster size was set to 30. Additionally, an AIC-based grid search was utilized to select ARIMA (2, 1, 2). There were predetermined random numbers that were used for batching, dividing the data, and setting up the model.

REFERENCES

[1] Boudad, N., Faizi, R., Oulad Haj Thami, R. (2024). Multilingual, monolingual and mono-dialectal transfer learning for Moroccan Arabic sentiment classification. Social Network Analysis and Mining, 14: 3. https://doi.org/10.1007/s13278-023-01159-9

[2] Mamta, Ekbal, A. (2024). Transformer based multilingual joint learning framework for code-mixed and English sentiment analysis. Journal of Intelligent Information Systems, 62: 231-253. https://doi.org/10.1007/s10844-023-00808-x

[3] Hashmi, E., Yayilgan, S.Y., Shaikh, S. (2024). Augmenting sentiment prediction capabilities for code-mixed tweets with multilingual transformers. Social Network Analysis and Mining, 14: 86. https://doi.org/10.1007/s13278-024-01245-6

[4] Khan, L., Qazi, A., Chang, H.T., Alhajlah, M., Mahmood, A. (2025). Empowering Urdu sentiment

analysis: An attention-based stacked CNN-Bi-LSTM DNN with multilingual BERT. Complex Intelligent Systems, 11: 10. https://doi.org/10.1007/s40747-024-01631-9

[5] Miah, M.S.U., Kabir, M.M., Sarwar, T.B., Safran, M., Alfarhood, S., Mridha, M.F. (2024). A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM. Scientific Reports, 14: 9603. https://doi.org/10.1038/s41598-024-60210-7

[6] Singh, L.G., Singh, S.R. (2024). Sentiment analysis of tweets using text and graph multi-views learning. Knowledge and Information Systems, 66: 2965-2985. https://doi.org/10.1007/s10115-023-02053-8

[7] Hicham, N., Nassera, H. (2024). Improving emotion classification in e-commerce customer review analysis using GPT and meta ensemble deep learning technique for multilingual system. Multimedia Tools and Applications, 83: 87323-87367. https://doi.org/10.1007/s11042-024-19965-4

[8] Pandey, R., Kumar, A., Singh, J.P., Tripathi, S. (2024). A hybrid convolutional neural network for sarcasm detection from multilingual social media posts. Multimedia Tools and Applications, 84: 15867-15895 https://doi.org/10.1007/s11042-024-19672-0

[9] Silviya, S.H.A., Faith, S.J., Seetha, R., Hemalatha, M. (2024). Performance evaluation of natural language processing algorithms for sentiment analysis. SN Computer Science, 5: 724. https://doi.org/10.1007/s42979-024-03094-8

[10] Vianna, D., Carneiro, F., Carvalho, J., Plastino, A., Paes, A. (2024). Sentiment analysis in Portuguese tweets: An evaluation of diverse word representation models. Language Resources and Evaluation, 58: 223-272. https://doi.org/10.1007/s10579-023-09661-4

[11] Aziz, K., Ji, D., Chakrabarti, P., Chakrabarti, T., Iqbal, M.S., Abbasi, R. (2024). Unifying aspect-based sentiment analysis BERT and multi-layered graph convolutional networks for comprehensive sentiment dissection. Scientific Reports, 14: 14646. https://doi.org/10.1038/s41598-024-61886-7

[12] Nath, D., Dwivedi, S.K. (2024). Aspect-based sentiment analysis: Approaches, applications, challenges and trends. Knowledge and Information Systems, 66: 7261-7303. https://doi.org/10.1007/s10115-024-02200-9

[13] Tori, F., Tori, S., Keseru, I., Ginis, V. (2024). Performing sentiment analysis using natural language models for urban policymaking: An analysis of Twitter data in Brussels. Data Science for Transportation, 6: 5. https://doi.org/10.1007/s42421-024-00090-5

[14] Mohamed, O., Kassem, A.M., Ashraf, A., Jamal, A., Mohamed, E.H. (2023). An ensemble transformer-based model for Arabic sentiment analysis. Social Network Analysis and Mining, 13: 11. https://doi.org/10.1007/s13278-022-01009-0

[15] Anjum, Katarya, R. (2024). HateDetector: Multilingual technique for the analysis and detection of online hate speech in social networks. Multimedia Tools and Applications, 83: 48021-48048. https://doi.org/10.1007/s11042-023-16598-x

[16] Guleria, A., Varshney, K., Pahwa, G., Singhal, S., Sharma, N. (2024). Multimodal sentiment analysis of English and Hinglish memes. Multimedia Tools and Applications, 84: 15331-15356. https://doi.org/10.1007/s11042-024-19640-8

[17] Luitel, S., Liu, Y., Anwar, M. (2024). Investigating fairness in machine learning-based audio sentiment analysis. AI and Ethics, 5: 1099-1108. https://doi.org/10.1007/s43681-024-00453-2

[18] Rusnachenko, N., Golubev, A., Loukachevitch, N. (2024). Large language models in targeted sentiment analysis for Russian. Lobachevskii Journal of Mathematics, 45: 3148-3158. https://doi.org/10.1134/S1995080224603758

[19] Jain, A., Jain, G., Tewari, D. (2024). KNetwork: Advancing cross-lingual sentiment analysis for enhanced decision-making in linguistically diverse environments. Knowledge and Information Systems, 66: 2925-2943. https://doi.org/10.1007/s10115-023-02051-w

[20] Bashiri, H., Naderi, H. (2024). Comprehensive review and comparative analysis of transformer models in sentiment analysis. Knowledge and Information Systems, 66: 7305-7361. https://doi.org/10.1007/s10115-024-02214-3

[21] Altaf, A., Anwar, M.W., Jamal, M.H., Bajwa, U.I., Rani, S. (2024). Aspect-based sentiment analysis in Urdu language: Resource creation and evaluation. Neural Computing and Applications, 36: 21365-21381. https://doi.org/10.1007/s00521-024-10145-x

[22] Zardak, S.R., Rasekh, A.H., Bashkari, M.S. (2023). Persian text sentiment analysis based on BERT and neural networks. Iranian Journal of Science and Technology, Transactions of Electrical Engineering, 47: 1623-1634. https://doi.org/10.1007/s40998-023-00626-5

[23] Badawi, S., Kazemi, A., Rezaie, V. (2024). KurdiSent: A corpus for Kurdish sentiment analysis. Language Resources and Evaluation, 59: 601-620. https://doi.org/10.1007/s10579-023-09716-6

[24] Ganganwar, V., Rajalakshmi, R. (2023). Enhanced Hindi aspect-based sentiment analysis using class balancing approach. International Journal of Information Technology, 15: 3527-3532. https://doi.org/10.1007/s41870-023-01430-4

[25] Claes, M., Farooq, U., Salman, I., Teern, A., Isomursu, M., Halonen, R. (2024). Sentiment analysis of Finnish Twitter discussions on COVID-19 during the pandemic. SN Computer Science, 5: 266. https://doi.org/10.1007/s42979-023-02595-2

[26] Kotagiri, S., Sowjanya, A.M., Anilkumar, B., Lakshmi Devi, N. (2024). Aspect-oriented extraction and sentiment analysis using optimized hybrid deep learning approaches. Multimedia Tools and Applications, 83: 88613-88644. https://doi.org/10.1007/s11042-024-18964-9

[27] P, G.K., S, A.A.V., V, J.P., Paul, A., Nayyar, A. (2024). A context-sensitive multi-tier deep learning framework for multimodal sentiment analysis. Multimedia Tools and Applications, 83: 54249-54278. https://doi.org/10.1007/s11042-023-17601-1

[28] Alfaisal, R., Hashim, H., Azizan, U.H. (2024). Convolutional neural network for sentiment analysis on metaverse-related tweets: A deep learning approach. SN Computer Science, 5: 753. https://doi.org/10.1007/s42979-024-03121-8

[29] Karabila, I., Darraz, N., EL-Ansari, A., Alami, N., El Mallahi, M. (2024). BERT-enhanced sentiment analysis for personalized e-commerce recommendations. Multimedia Tools and Applications, 83: 56463-56488. https://doi.org/10.1007/s11042-023-17689-5

[30] Cero, I., Luo, J.B., Falligant, J.M. (2024). Lexicon-based sentiment analysis in behavioral research. Perspectives on Behavior Science, 47: 283-310. https://doi.org/10.1007/s40614-023-00394-x
[31] Tamilkodi, R., Sujatha, B., Leelavathy, N. (2024). Emotion detection in text: Advances in sentiment analysis. International Journal of System Assurance Engineering and Management, 16: 552-560. https://doi.org/10.1007/s13198-024-02597-0
[32] Kandhro, I.A., Ali, F., Uddin, M., Kehar, A., Manickam, S. (2024). Exploring aspect-based sentiment analysis: An in-depth review of current methods and prospects for advancement. Knowledge and Information Systems, 66: 3639-3669. https://doi.org/10.1007/s10115-024-02104-8
[33] Malik, M.S.I., Rehan, M., Nawaz, A. (2024). Analyzing cross-lingual approaches: A case study for detecting multilingual hope expressions in YouTube comments. Pattern Recognition and Image Analysis, 34: 831-843. https://doi.org/10.1134/S105466182470072X
[34] Latrech, J., Kodia, Z., Ben Azzouna, N.B. (2024). Twit-CoFiD: A hybrid recommender system based on tweet sentiment analysis. Social Network Analysis and Mining, 14: 123. https://doi.org/10.1007/s13278-024-01288-9
[35] Rahman, E., Carruthers, J.D.A., Rao, P., Rahman, Z., Esfahlani, S.S., Webb, W.R. (2024). From posts to perceptions: Sentiment and psychological analysis of aesthetic enhancements on social media. Aesthetic Plastic Surgery, 49: 1478-1494. https://doi.org/10.1007/s00266-024-04455-7
[36] Mohawesh, R., Maqsood, S., Althebyan, Q. (2023). Multilingual deep learning framework for fake news detection using capsule neural network. Journal of Intelligent Information Systems, 60: 655-671. https://doi.org/10.1007/s10844-023-00788-y
[37] Liyih, A., Anagaw, S., Yibeyin, M., Tehone, Y. (2024). Sentiment analysis of the Hamas-Israel war on YouTube comments using deep learning. Scientific Reports, 14: 13647. https://doi.org/10.1038/s41598-024-63367-3
[38] Ba Alawi, A., Bozkurt, F. (2024). Performance analysis of embedding methods for deep learning-based Turkish sentiment analysis models. Arabian Journal for Science and Engineering, 50: 7299-732. https://doi.org/10.1007/s13369-024-09360-4
[39] Levallois, C. (2024). Umigon-lexicon: Rule-based model for interpretable sentiment analysis and factuality categorization. Language Resources and Evaluation, 59: 913-930. https://doi.org/10.1007/s10579-024-09742-y
[40] Punetha, N., Jain, G. (2024). Optimizing sentiment analysis: A cognitive approach with negation handling via mathematical modelling. Cognitive Computation, 16: 624-640. https://doi.org/10.1007/s12559-023-10227-3

NOMENCLATURE

NLP	Natural Language Processing
BERT	Bidirectional Encoder Representations from Transformers
CNN	Convolutional Neural Network
Bi-LSTM	Bidirectional Long Short-Term Memory
DNN	Deep Neural Network
GPT	Generative Pre-trained Transformer
LLM	Large Language Model
MAE	Mean Absolute Error
RL	Reinforcement Learning
KNetwork	Knowledge Network
ML	Machine Learning
AI	Artificial Intelligence
SAR	Sarcasm Detection
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
ARIMA	AutoRegressive Integrated Moving Average
F1-Score	Harmonic Mean of Precision and Recall
Umigon	User Model Integration for Generalized Online Needs
SN	Social Networks
SA	Sentiment Analysis