International Research Journal of Multidisciplinary Scope (IRJMS), 2026;7(3):1873-1887 IRJMS

Original Article | ISSN (O): 2582-631X DOI:10.47857/irjms.2026.v07i03.011806

A Hybrid Framework for Large-scale Tweet Sentiment Analysis Using Classical Machine Learning, Transformer Models and Uncertainty Estimation

Demmelash Abate[1]*, Nilay Mistry[2]

[1]National Forensic Sciences University, School of Doctoral Studies and Research, Ahmedabad, Gujarat, India, [2]National Forensic Sciences University, School of Cyber Security and Digital Forensics, Ahmedabad, Gujarat, India. *Corresponding Author's Email: demme.abe@gmail.com

Abstract

The article suggests a hybrid framework to analyse Twitter sentiment at scale by combining traditional machine learning methods with deep learning using transformers in a multi-layer stacking configuration. The system is evaluated using a dataset of 74,921 tweets, each labelled with one of three sentiment labels: negative, neutral, or positive. The architecture is 3-tier. Tier A is based on classical classifiers: Logistic Regression, Calibrated Linear SVM and Multinomial Naive Bayes, using TF-IDF features (unigrams to trigrams), with character-level n-grams added. Tier B is a response to a narrow-focused DistilBERT encoder to learn rich contextual embeddings. Tier C fuses out-of-fold predictions from tiers A and B with a stacked meta-learner that combines lexical accuracy and context depth. Removal of noise, normalisation to kenisation and stopword filtering enable complete text preprocessing, ensuring data quality and consistency throughout the pipeline. The experimental results show that the stacked Meta Logistic Regression achieves an accuracy and macro-F1 of 0.9705, which are much better than those of all the standalone classical and transformer baselines. The model has a near-perfect discriminative performance (macro-AUC = 0.994) and a very good calibration (ECE = 0.0043). Aspect and time analyses are also useful to comprehend the dynamics of sentiment and evolving attitudes towards airline services. These findings validate the hypothesis that hybrid stacking is an effective method for leveraging the complementary nature of lexical and contextual representations, thereby improving generalisation and achieving superior performance. The work provides a reproducible, explainable, operationally applicable model of sentiment analysis in operationally sensitive, high-stakes domains.

Keywords: Machine Learning, Natural Language Processing, Sentiment Analysis, TF-IDF, Twitter, Uncertainty Quantification.

Introduction

Sentiment analysis is a fundamental NLP task that extracts subjective information from large volumes of text. Social media, notably Twitter, is generating unprecedented amounts of user-generated content that expresses opinions, satisfaction and behavioural trends. Twitter text is noisy and context-dependent, so it is hard to classify. Lexical patterns can be classified using classical machine learning methods, but they are not very useful for sarcasm, negation and semantic dependencies. Transformer-based models are very good at understanding context, but are very costly and hard to explain. To overcome these weaknesses, the present study proposes a hybrid stacking approach that combines classical and deep learning models, leveraging both surface linguistic features and contextual embeddings for improved generalisation and performance. Stacking and ensemble learning, in general, have become effective techniques to improve the predictive accuracy of heterogeneous models. The method is tested on a large-scale Twitter dataset of 74,921 labelled examples across three sentiment categories: negative, neutral and positive. It uses a social media-specific preprocessing methodology, state-of-the-art feature engineering based on word- and character-level TF-IDF representations and transformer-based contextual encoding with DistilBERT. An outstanding feature of this work is the leakage-free meta-feature space generated from out-of-fold (OOF) predictions, which enables robust stacking with the meta-learner multinomial Logistic Regression (LR). The studies collectively support the use of hybrid/stacked ensembles to benefit from the complementary strengths in practical deployments (1, 2). In addition to classification performance, the research focuses on the model's reliability, including calibration

(Received 15th April 2026; Accepted 12th July 2026; Published 30th July 2026)

analysis and uncertainty estimation via Monte Carlo Dropout. A study shows that optimised MC Dropout variants that improve both accuracy and calibration, addressing known weaknesses of vanilla MC Dropout in probability calibration. The two canonical early families of sentiment analysis tools are lexicon-based and supervised tools, particularly Naive Bayes, SVMs and Logistic Regression, which were all consistently found to be, across surveys, the most effective supervised and unsupervised baselines that can be systematically adopted before the rise of CNNs, RNNs and transformer-based models like BERT and DistilBERT (3-5).

Hybrids (representation+sequence layers/attention for refinement) are favoured in modern work. The TRABSA is based on a RoBERTa encoder with attention and a BiLSTM, achieving ~94% accuracy, generalising to external datasets and providing interpretability via SHAP/LIME. Likewise, BERT–BiLSTM–Attention hybrids achieve F1>0.94 on multi-class Twitter sentiment tasks and ablation studies support the inclusion of the sequence layer and attention. Further validation of the hybrid ensembling is the improvement of robustness over datasets (BERT, GPT-2, RoBERTa, XLNet, DistilBERT) for multi-transformer ensembles (6, 7).

Deep learning approaches, including CNNs, LSTMs and attention-based models, improved representation learning for short texts. More recently, transformer-based models such as BERT, RoBERTa and XLM-R have achieved state-of-the-art results on Twitter sentiment benchmarks. Research indicates that rich contextual encoding: Bidirectional self-attention captures long-range and global dependencies better than sequential RNNs and local CNNs (8, 9). Hybrid designs such as RoBERTa-LSTM, BERT+CNN/BiGRU and multi-transformer ensembles further push the SOTA in Twitter sentiment and related tweet classification tasks (10-12).

Research on fake reviews focuses on linguistic cues, reviewer behaviour and network-based features. Few studies integrate sentiment evolution and time-aware deep learning, especially on Twitter. Text-only deep models: CNNs, RNNs, LSTMs, BiLSTMs and Transformers (BERT/RoBERTa) learn semantic and syntactic cues and outperform classic ML on benchmark review datasets (13, 14).

Sentiment intensity and emotion: The combination of lexicon-based indicators of emotion or sentiment scores and embeddings provides a significant enhancement as compared to text alone and the combination captures both emotion extremes of exaggeration and polarity (15, 16).

Aspect-based sentiment analysis and aspect replication (CNN+LSTM, GCN+SenticNet) detect spammers who selectively manipulate sentiments on specific product aspects. Behavioural and network characteristics: Reviewer activity dynamics, collusive groups and graph (reviewer-review-item graph, GNN/GCN) features enhance the identification of campaign organisation and fake reviewers (17, 18).

IEEE studies contrasting ML and DL on Twitter highlight the need for reproducibility in large-batch streaming setups. Neural networks don't provide certainty estimates or be over- or underconfident. To counter this, various forms and sources of uncertainty have been identified and methods for measuring and quantifying uncertainty in neural networks have been proposed. In high-risk applications, such as robot navigation or the navigation of high-speed drones, it is essential to incorporate uncertainty estimates to make safe decisions; in other sectors with little labelled data, it is even more critical (19, 20).

Calibration methods are classified into four categories: post-hoc calibration, regularisation methods, uncertainty estimation and composition methods. Recent advancements in calibrating large language models (LLMs) are also covered, alongside open issues and potential future directions.

Explicit predictive uncertainty is uncommon in Twitter sentiment research. However, some recent hybrid research already exists that supports explainability (SHAP/LIME) along with uncertainty, which helps give confidence in the feature attributions (21, 22).

In general, hybrid deep models (e.g., transformer + attention+BiLSTM) for tweets can be very accurate and provide local explanations via SHAP/LIME visualisations; they focus only on per-tweet sentiment classification, not on temporal evolution (23).

Several Twitter studies perform trend or time-series sentiment analysis (e.g., natural disaster tweets with polarity indices and time-series inspection, trend analysis with CNN–LSTM), but do

not provide per-time-step explanations, such as SHAP/LIME trajectories or evolving attention patterns (24).

Evidence supports the claim that SHAP, LIME and attention-based visualisations are now standard for per-tweet explainability on Twitter but are rarely integrated with explicit temporal sentiment models (time-series, longitudinal user trajectories, or event-level dynamics). This joint space of explainable temporal sentiment on Twitter remains a clear research opportunity (25). The purpose of this study is to:

(a) build a multi-tier stacked ensemble of classical ML and transformer models;
(b) test the framework on a large-scale airline Twitter corpus; and
(c) demonstrate the superiority of the hybrid stacking over stand-alone baselines by conducting extensive cross-validation.

Methodology

The initial stage of the research process involves loading a dataset of 74,921 tweets, each containing metadata such as follower count, followees, retweets and likes, as well as the content of the tweet and the sentiment of the users who followed, were followed, retweeted, or liked it. Integrity checks ensure that all required columns are present and label distribution checks for an imbalance among negative, neutral and positive attitudes.

Text Preprocessing and Label Normalisation: A comprehensive reprocessing pipeline for social media text is applied to tweets. This involves deleting URLs and retweet labels, deanonymizing user mentions to a generic token, extracting hashtag text, demojifying emojis and lowercasing and normalising whitespace. Sentiment labels are normalised with a mapping operator to achieve consistency across different encodings (e.g., numeric or abbreviated labels). Modelling is performed only with the three target classes (negative, neutral, positive).

Feature Engineering for Classical Models: Two complementary TF-IDF feature representa-tions were used to capture the linguistic properties of Twitter data. To begin with, word-level n-grams [1-3] were employed to capture semantic and contextual patterns in lexical sequences, which are particularly well-suited to modelling sentiment-bearing phrases and short expressions typical of tweets. Second, the character-level n-grams of three to five characters [3-5] were added to accommodate the informal nature of social media text, allowing the model to handle spelling variants, abbreviations, run-on words and typographical noise. A FeatureUnion was then used to combine these two feature spaces, creating a single, high-dimensional, sparse representation, a union of semantic and sub-lexical information. Such a hybrid representation makes a powerful contribution to linear classifiers, improving their generalisation across different linguistic patterns while maintaining computational efficiency.

The collection of n documents was represented by Equation [1]. In each document, the TF–IDF weighting scheme was used to transform it into a numerical feature representation, with the term importance calculated using Equation [2].

Let the corpus,

$$D = \{d_1, d_2, \dots, d_n\} \quad [1]$$

TF-IDF representation:

For terms t in a document d, TF-IDF is defined as:

$$Tfidf(t,d) = tf(t,d).\log(\frac{N}{df(t)}) \quad [2]$$

Tier-A Models: Classical Learners with Out-of-Fold (OOF) Predictions

In the initial model training stage, a list of known classical machine learning classifiers was used to give good baseline representations of sentiment classification; namely, a calibrated Linear Support Vector Machine (SVM) as shown in Equation [3] with sigmoid probability calibration, multinomial Logistic Regression and Multinomial Naive Bayes; these models were chosen since they have demonstrated good performance in text classification problems and since they complement each other in terms of learning behavior Stratified K-fold cross-validation K=3 was used to train to achieve a strong performance estimation, as well as to avoid information leakage during the later ensembling, thus maintaining the original class distribution in each fold. Models were trained on

the training partition and evaluated on the held-out validation partition, where out-of-fold (OOF) predicted class probabilities were produced for each iteration, as shown in Equation [4].

Linear SVM (calibrated):

$$f(x) = W^T + b \quad [3]$$

probability calibration:

$$P(y=c/x) = \frac{1}{1+\exp(Af(x)+B)} \quad [4]$$

$$\hat{p}_i^{(m)} \in [0,1]^C \text{ Such that } \sum_{C=1}^{C} \hat{P}_{i,c}^{(m)} = 1 \quad [5]$$

Tier-B Model: Transformer-based Learning (DistilBERT)

A transformer model (DistilBERT-base-uncased) is also trained for sentiment classification using a similar stratified K-Fold strategy in Equation [6]. Tweets are coded and tokenised with a sequence length of 128. The Trainer API from Hugging Face is used to perform version-safe training with training arguments. Validation-fold softmax probabilities are also used as OOF predictions, as shown in Equation [7], creating a second set of high-level semantic features.

Given tokenised input Z_i:

$$h_i = \text{DistilBERT}(z_i) \quad [6]$$

Out-of-fold probabilities:

$$\hat{p}_i(\text{BERT}) \quad [7]$$

Meta-feature Construction

The classical model (Tier-A) and the transformer model (Tier-B) OOF probabilities are combined to create a single large meta-feature matrix as defined in Equation [8]. A row represents a tweet and a column is a probability output of a base learner. The original sentiment labels are used as targets for the meta-learner.

Meta-feature matrix:

$$X_{meta} = [P_A P_B] \in R^{N*C(M+1)} \quad [8]$$

Tier-C Model: Stacked Meta-learner

The final sentiment prediction, in Equation [9], is done using a multinomial Logistic Regression model trained on the collection of the meta-features. This hierarchical approach uses shallow lexical structures and deep context-specific representations. Logistic Regression is tested as the main meta-learner for its stability and calibration quality, while Gradient Boosting is tested for comparison. Multinomial logistic regression.

$$P(y=c/x_{meta}) = \frac{\exp(B_C^T X_{mata})}{\sum_{K=1}^{C} \exp(\beta_K^T X_{meta})} \quad [9]$$

Model Evaluation and Comparison

A well-defined system of quantitative and visual performance measures was used to systematically assess model performance, ensuring predictive accuracy and robustness across sentiment classes. The metrics were overall accuracy and the macro-average F1 score, defined in Equation [10], a balanced performance measure for class imbalance. In the meantime, the misclassification and inter-class confusion patterns were compared with the visualisations of confusion matrices. One-vs.-Rest Receiver Operating Characteristic (ROC) curves were also used to assess discriminative ability and the macro-averaged Area Under the Curve (AUC) was summarised, as defined in

Equation [11]. A comprehensive comparison of classical, transformer-based and stacked ensemble models was made possible by creating tier-wise model comparison graphs.

Macro-F1:

$$\text{Macro-F1} = \frac{1}{C}\sum_{c=1}^{C}\frac{2P_cR_c}{P_c+R_c} \quad [10]$$

ROC-AUC:

$$\text{AUC-micro} = \frac{1}{C}\sum_{c=1}^{C}\int_0^1 \text{TPR}_c(\text{FPR}_c^{-1}(u))du \quad [11]$$

Temporal and Aspect-Level Analysis

Following sentiment classification, a series of post hoc analyses was conducted to enhance interpretability and reveal temporal and thematic trends in popular opinion. Temporal dynamics were also examined by plotting daily sentiment trends, complemented by a 7-day moving average to smooth short-term fluctuations and emphasise longer-term trends, as defined in Equation [12]. Additionally, aspect-based sentiment analysis was performed using a domain-specific lexicon that included the most important dimensions of service quality: price, time, comfort, food and Wi-Fi in Equation [13].

7-day moving average:

$$\overline{V}_{d,s}\frac{1}{7}\sum_{j=d-6}^{d} V_{j,s} \quad [12]$$

Daily Sentiment Volume:

$$V_{d,s} = \sum_{i=1}^{N}\Pi(t_i = d, y_i = s) \quad [13]$$

Calibration and Uncertainty Estimation

A systematic assessment of the model's calibration and uncertainty was conducted to test the reliability of probabilistic predictions in Equation [14]. Monte Carlo Dropout was also used, along with calibration, to investigate the predictive uncertainty of inferences from the transformer-based model. Stochastic-generated approximations of predictive entropy and probability variance, some forward models and quantifying confidence dispersion. The uncertainty analysis provides additional assurance of the model's soundness and identifies situations in which the predictions should be approached with caution.

Expected Calibration Error (ECE):

$$\text{ECE} = \sum_{b=1}^{B}\frac{|B_b|}{N}|\text{acc}(B_b) - \text{conf}(B_b)| \quad [14]$$

Model Persistence and Reproducibility

To make the model components reproducible and enable future experimentation, all trained components were systematically stored. This contains serialized artefacts of the stacked meta-learner, base model settings and related feature mappings. Middle results, especially out-of-fold (OOF) prediction matrices defined in Equation [15], were written in the efficient format, enabling easy reuse in full analyses or across diverse ensemble strategies. The workflow facilitates a transparent replication process, is auditable and can be deployed at scale by explicitly saving both intermediate and final artefacts in Equation [16], a practice consistent with best practices in reproducible machine learning research.

OOF prediction matrix:

$$\mathbf{P}\ \text{OOF} \in \mathbb{R}^{N*C(M+1)} \quad [15]$$

Stored artefacts:

$$\Theta = \{\theta_A, \theta_B, \theta_{meta}\} \quad [16]$$

The research experiment evaluated a three-level hybrid stacking approach for sentiment classification of large-scale Twitter data, as shown in Figure 1. The data were collected from 74,921 tweets labelled as negative, neutral, or positive. The first step in pre-processing tweets is to remove noise, normalise the text, tokenise it and filter out common stop words. Word-level TF-IDF features were extracted for all three classifiers (Logistic Regression, Calibrated Linear SVM, Multinomial Naive Bayes) for Tier A and comprised of unigrams, trigrams and character-level n-grams. To encode sentence representations contextually, DistilBERT was used in Tier B. In tier C, the classical models' predictions were stacked with DistilBERT to produce predictions using the stacked meta-learner, Meta Logistic Regression. The accuracy, macro-F1, macro-AUC and Expected Calibration Error were used to assess the models. Then, a hybrid stacking model was compared with standalone models to verify the effectiveness.

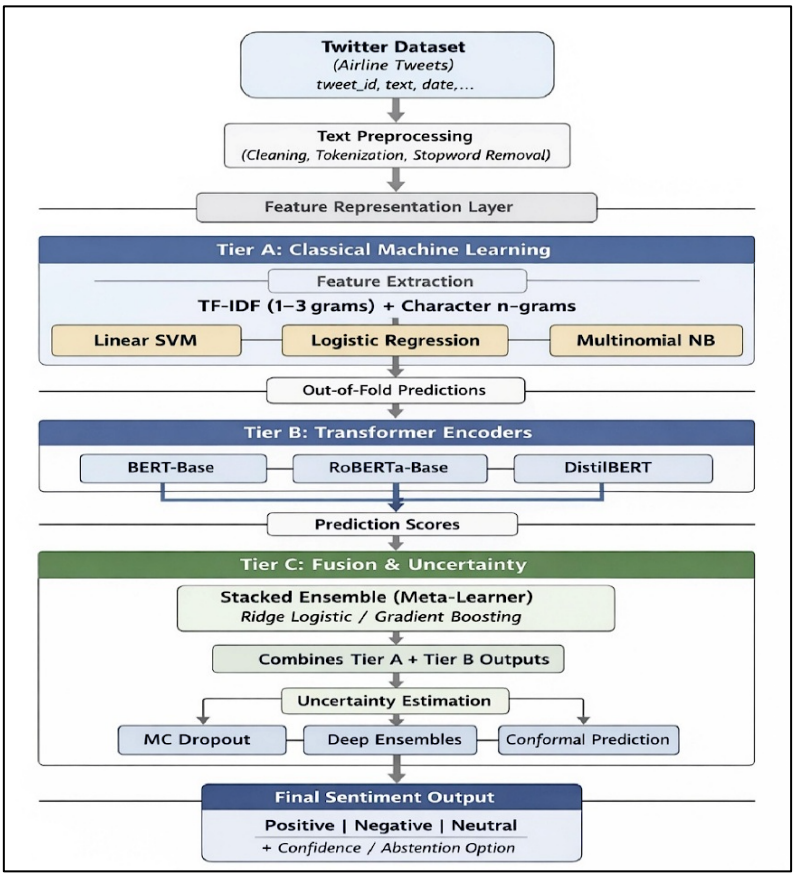

Figure 1: Hybrid Framework for Large-scale Tweet Sentiment Analysis Using Classical ML and Transformer Models

Result and Discussion

Dataset Characteristics

The data used in the experimental evaluation consisted of 74,921 annotated tweets across 3 sentiment classes: positive, neutral and negative. The distribution of classes suggests that neutral sentiment is the most prevalent, followed by positive and negative sentiment. It is important to balance the modelling techniques applied. Table 1 represents the sample of the dataset.

Table 1: Sample of Twitter Airline Sentiment Dataset with Metadata and Labelled Sentiment Classes

Tweet	Like Count	Retweet Count	Date	Time	Sentiment
Flying Line Year Tried Book Ticket Khartoum Ca...	0	0	1/4/2023	7:16:44	Neutral
Never Get Tired Lying, Let's Give U Bad News, Tplf...	0	0	1/4/2023	6:59:03	Negative
Great Business-Class Service, Always Comfy Seat...	0	0	1/4/2023	6:54:32	Positive
Visit An Amazing Place, Earth's Origin, Get Unfo...	1	1	1/4/2023	6:02:17	Positive
Administration Funding Aid Helping Gov Genocid...	0	1	1/4/2023	5:32:15	Positive

Table 2 shows that the sentiment distribution is highly skewed, with neutral sentiment the most frequent (31,669 instances), followed by positive (28,637 instances) and negative (14,615 instances). Some of this is due to the nature of social media in general–the amount of user-generated content is relatively high and most of it is not extremely emotional, but informative or a bit opinionated. The content is neutral, which may be interpreted as descriptive or event-based and high, which may indicate a positive user experience. The negative emotion is less common but very important, as it often stems from dissatisfaction and/or service-related issues.

Table 2: Distribution of Sentiment Classes in the Twitter Dataset

Sentiment	Count
Positive	28,637
Negative	14,615
Neutral	31,669
Total	74,921

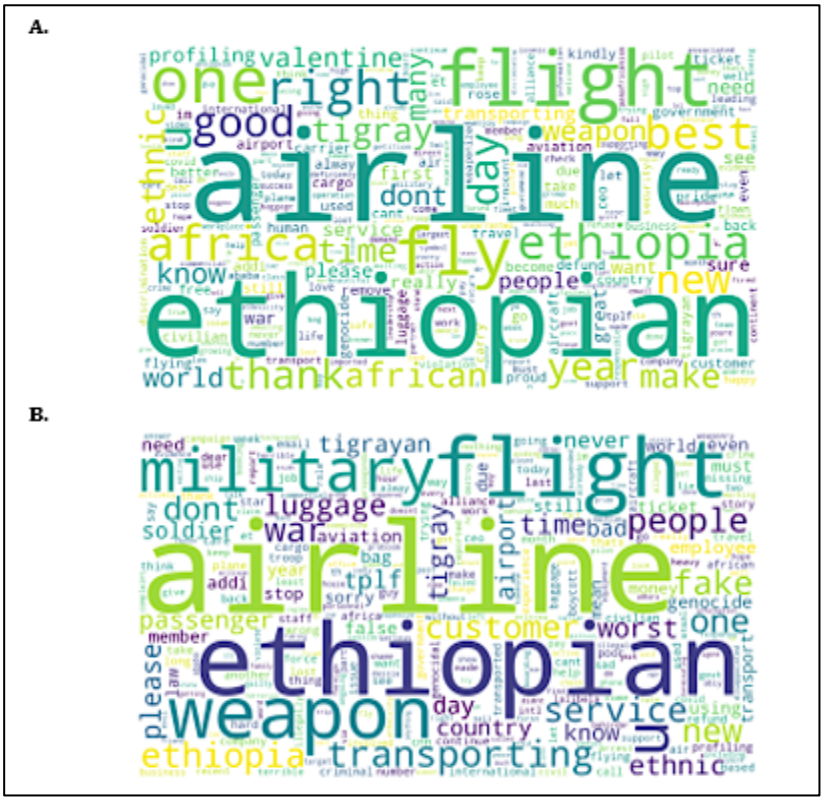

Figure 2: Word Cloud Visualisation. A) Positive Word Cloud Visualisation of Positive Sentiment Tweets, B) Negative Word Cloud Visualisation of High-frequency Terms in Tweet Corpus

Positive Tweets Word Cloud

Figure 2A has the word cloud that shows the most meaningful words that are most likely to appear in the positive-sentiment tweets and the size of the words is relatively similar to the corpus frequency and significance. The prominence of words such as Ethiopian, flight and airline, as well as the word "one," indicates that good user feedback is mostly focused on the overall airline experience and brand name. The positive, appreciative language: great, best, proud, support, love, suggests that the customers are highly satisfied with the services provided by the airline and have a high emotional investment in the airline. In addition, the terms service, time, passenger and ticket describe vital areas of activity in which good experiences are created, such as quality of service, punctuality and customer care. Contextual words such as African, international and aviation provide a broader picture of the airline's image and global presence. On the whole, the word distribution analysis reveals that emotional appreciation and functional service quality are the two factors contributing to positive sentiment, indicating that the dataset can be used for aspect-based sentiment analysis

(ABSA) and to understand the factors driving customer satisfaction in the airline industry.

Negative Tweets Word Cloud

The word cloud in Figure 2B shows that the words in the corpus of tweets are more frequent; the bigger the word, the more frequent it is. These words are leading words, i.e., airline, flight, time, work, indicating that the user conversation is primarily focused on the operational side of airline services (scheduling, service efficiency, customer experience). This discrepancy can cause classifiers to be overly sensitive to the majority classes in a modelling context, resulting in poor performance on minority classes, especially negative sentiment. To ensure fair and reliable sentiment classification, class weighting, resampling techniques and macro-averaging evaluation metrics should be employed.

Temporal Patterns in Twitter Sentiment Dynamics

The temporal dynamics of Twitter sentiment are clearly observed in Figures 3A and 3B (daily counts and trend over 7 days), indicating that Twitter activity is low, with steep spikes due to short-lived, powerful events that elicit multi-high activity across all negative, neutral and positive sentiment groups. Figure 3A presents daily sentiment counts together with overall sentiment trends. The pre-eminence of neutral feeling is always apparent in these peaks, suggesting that many high-volume events are informational rather than emotionally polarised. Positive sentiment, conversely, tends to follow, while negative sentiment tends to occur in short bursts, suggesting a rapid response to scandals or upheaval.

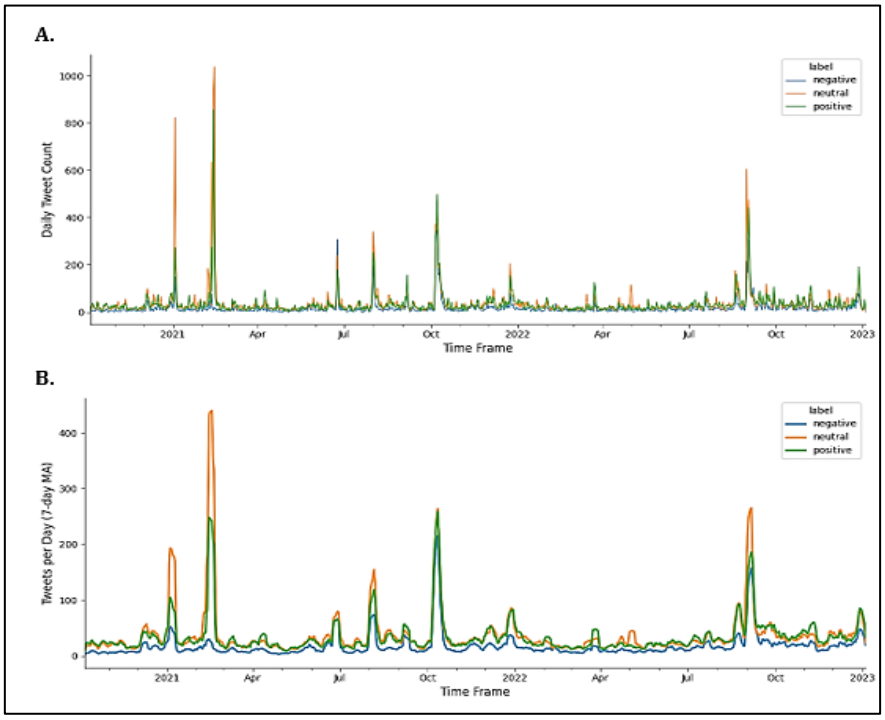

Figure 3: Tweet Sentiment Trend. A) Daily Tweet Sentiment Trend, B) Seven-Day Moving Average Tweet Sentiment Trends

The seven-day moving-average panel as shown in Figure 3B reveals that engagement dynamics in Twitter remained stable over time, particularly at the start and end of 2021, indicating that the sentiment of the platform is driven mainly by events and that engagement cycles occur for informational, supportive and critical responses within the platform's dynamic social environment.

Performance of Tier-A Classical Models

Table 3 shows that Tier-A classical machine learning models provide a strong baseline for sentiment classification. The best model (with an accuracy of about 0.95) is Logistic Regression, which achieves a high macro-F1 score, indicating its strength when using TF-IDF features and multinomial loss optimisation. Calibration is also competitive (with ~0.94 accuracy), has good margins and strong separation, provides probabilistic outputs with useful margins, can be used in combination with ensemble learning downstream and is well-calibrated. Despite being lighter, Multinomial Naive Bayes is less accurate, with an average Macro-F1 score and thus is not as

effective at capturing more complex feature dependencies. The overall results indicate that more complex discriminative models, such as LogReg and SVMs, are more useful for text data with a high number of features (high-dimensional, sparse) than simpler probabilistic models. Besides, their strong performance at the individual level makes them a strong base learner in the suggested stacking design and their advantages complement the higher-level meta-learners, thereby improving classification accuracy and generalisation.

Table 3: Performance Evaluation of Tier-A Classical Machine Learning Models for Sentiment Classification

Model	Accuracy	Macro-F1	Observations
Calibrated Linear SVM	0.94	High	Strong baseline, excellent margin separation
Logistic Regression	0.95	High	Effective with TF-IDF + multinomial loss
Multinomial NB	Lower	Moderate	Underperforms compared to discriminative methods

Tier-B (DistilBERT) Transformer

The strong performance of the Tier-B transformer model trained on DistilBERT in Table 4 suggests that it can capture contextual and semantic nuances in text. According to Table 4, the model achieves an accuracy of around 96.5% and a macro-F1 of 0.9649, indicating significant, well-balanced classification performance across the sentiment classes. Transformer architecture is an enhancement of more traditional n-gram-based models that capitalise on deep contextual embeddings, allowing the model to encode more fine-grained linguistic evidence, including sarcasm, negation and long-range dependencies, in tweets. This greatly increases its ability to capture the slightest differences in sentiment that classical models often overlook. In addition, the out-of-fold (OOF) prediction strategy will be employed to ensure that the generated probabilities do not overfit the training data and are unbiased and generalizable. With such good OO predictions, they can also serve as strong meta-features in the stacking framework, further enhancing performance in the meta-learning steps (26).

Table 4: Classification Performance of the Tier-B DistilBERT Transformer Model

Model	Accuracy	Macro-F1
DistilBERT	0.965	0.9649

Tier-C Meta-Classifier

The results of the Tier-C meta-classifier show the efficacy of the proposed stacked ensemble framework for sentiment classification. Table 5 shows that the Meta Logistic Regression model performs better than both Tier-A classical models and the Tier-B DistilBERT transformer model in terms of accuracy and macro-F1. This illustrates the advantages of combining different learning paradigms on a common basis, such as TF-IDF. Tier-A models show robust semantic relationships grounded in lexical and frequency patterns.

Table 5: Performance of Tier-C Meta-classifier

Model	Accuracy	Macro-F1
Meta Logistic Regression	0.9705	0.9705

Performance Comparison of Meta-learner Models for Sentiment Classification

As shown in Table 6, both meta-learning models perform well. Still, the Meta Gradient Boosting (Meta GB) model ranks higher across most evaluation measures than the Meta Logistic Regression (Meta LR) model. In particular, the overall accuracy of Meta GB [97.05] is higher than that of Meta LR [96.50] and the macro-averaged precision, recall and F1-score are also higher, which means that Meta GB is better balanced in performance across all sentiment classes. Meta GB shows some improvements in detecting negative sentiment and has a higher class-level recall [0.9444 vs. 0.9256], suggesting that it is more sensitive to minority-class sentiment. Furthermore, for both classes (neutral and positive), Meta GB outperforms other models in F1 score, indicating that it is highly capable of capturing both majorities and minorities. The gap between the macro and weighted averages are relatively small across both models, indicating that it is a stable performer, even in the face of class imbalance. Still, the consistent gains of Meta GB indicate its greater capacity to learn intricate relationships in the joint

feature space. These findings affirm that gradient boosting is a more robust meta-learner in the proposed stacking model, providing better generalisation and stronger performance on large-scale sentiment classification tasks. Research indicates that a meta-learner gradient boosting ensemble outperformed several standalone machine learning models, achieving accuracies above 90%. The study confirms that combining diverse learners can improve prediction accuracy and robustness, supporting the effectiveness of ensemble learning strategies (27).

Table 6: Performance Comparison of Meta-Learner Models for Sentiment Classification

Model	Class	Precision	Recall	F1-Score	Support
	Negative	0.9424	0.9256	0.9339	14,615
	Neutral	0.9727	0.9813	0.9770	31,669
Meta LR	Positive	0.9677	0.9671	0.9674	28,637
	Accuracy			0.9650	74,921
	Macro Avg	0.9609	0.9580	0.9594	74,921
	Weighted Avg	0.9649	0.9650	0.9649	74,921
	Negative	0.9433	0.9444	0.9439	14,615
	Neutral	0.9825	0.9803	0.9814	31,669
Meta GB	Positive	0.9711	0.9730	0.9720	28,637
	Accuracy			0.9705	74,921
	Macro Avg	0.9656	0.9659	0.9658	74,921
	Weighted Avg	0.9705	0.9705	0.9705	74,921

Figure 4 indicates the confusion matrix for the Meta Logistic Regression (Meta-LR) model and provides a fine-grained measure of the classify-cation performance across the three sentiment classes: negative, neutral and positive. The results indicate a well-generalised model, capable of forecasting all classes.

Interestingly, the true-positive rates for the negative and positive classes are high. The neutral category encompasses the largest number of correctly classified instances [31,076], with only a few misclassified into either the negative [284] or positive [309] categories. This implies that the model is appropriate in capturing the delicacy of the language employed in neutral sentiment. This is a hard-to-categorise class because it is ambigu-ous and borders on other sentiment classes.

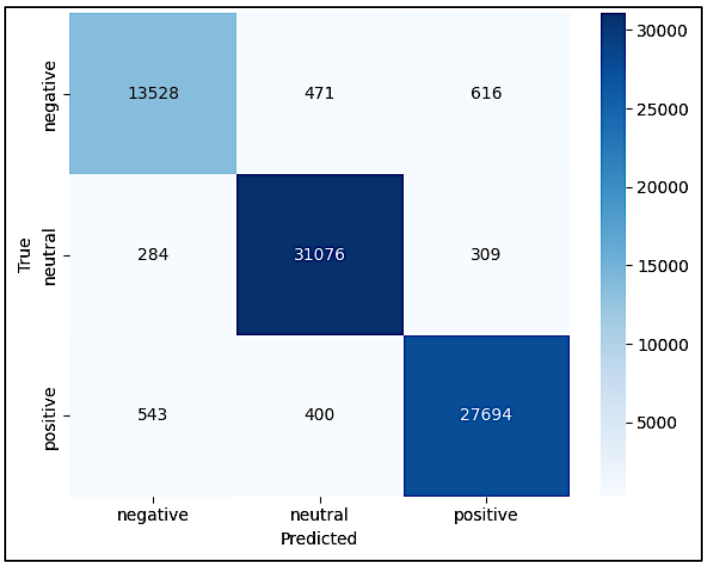

Figure 4: Confusion Matrix of the Meta Logistic Regression (Meta-LR) Classifier for Sentiment Classification

For the positives, their performance is good, with 27.694 correctly classified and there are slightly fewer positives classified as negatives [543] and neutrals [400]. The numbers indicate the model's sensitivity to polarity when a strong positive signal is present. The negative category is more difficult to classify. However, the outcomes are great. The precision and recall rates are greater than 94%. The most common source of errors is a mismatch with a predicted neutral [471] or positive [616]

word. This increased misclassification is due to the difficulty of the negative statement, with some linguistic features resembling expected neutral or positive words. The confusion matrix indicates low misclassification rates across all categories. The stability and robustness of the Meta-LR model are evidenced by the uniformly low misclassification rates across all categories. These findings demonstrate the effectiveness of the stacking-based meta-learning approach, which combines lexical representations (Tier A) with semantic representations (Tier B) to generate more effective decision boundaries. As a result, the model is better generalised and better classified than an individual baseline.

ROC AUC Analysis

Figure 5 plots the Receiver Operating Characteristic curves of the Meta Logistic Regression model as one-vs.-rest on sentiment in three different classes: negative, neutral and positive. The Meta Logistic Regression model differentiates these two classes very well: 0.996 for neutral, 0.995 for positive and 0.992 for negative. The model's AUC score is 0.994, indicating strong performance across all classes.

The curves are also close on the left side, indicating that the model correctly recognises the sentiment class and makes no mistakes. It can differentiate the classes, but it errs in several ways.

The AUC score for the negative class is lower than that for the neutral and positive classes. This difference isn't that significant in real life. The results show that the Meta Logistic Regression model is good at distinguishing between classes, even when negative sentiment is complex or hard to understand. The Meta Logistic Regression model is robust to negative sentiment. Simmonds and Higgins showed that random-effects logistic regression offers a flexible and statistically robust framework for analysing binary outcomes. The model directly maximises the binomial likelihood and can be extended to various advanced analytical settings, improving the reliability of statistical inference (28).

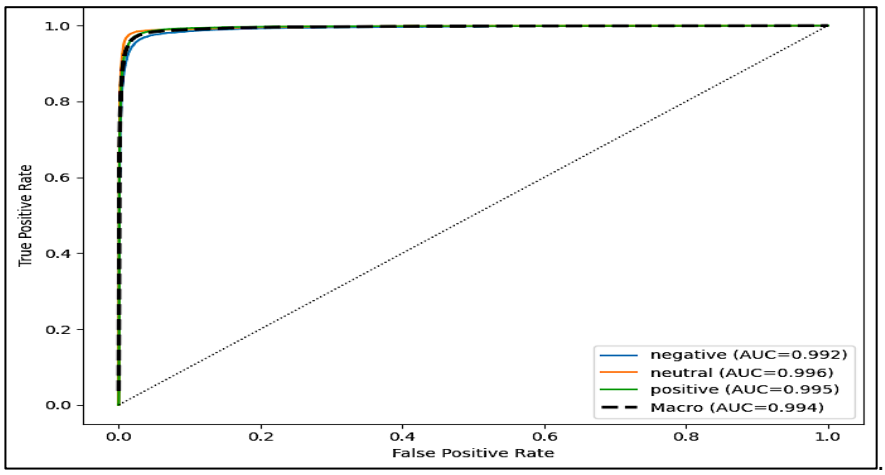

Figure 5: One-vs.-Rest ROC Curve Analysis of the Meta Logistic Regression (Meta-LR) Model

Calibration and Reliability

In Figure 6 of the reliability diagram, the model's calibration performance is shown: the predicted probabilities lie along the diagonal reference line, which represents perfect calibration. The expected calibration error (ECE) is 0.0043, indicating that the model's probability predictions are reasonable and that there is little gap between the estimated and actual accuracy. The model is neither overfitting nor underfitting the class probabilities, since the points are not far from the ideal calibration line within each confidence interval. level. In the healthcare analytics field, for instance, where predictions can be as important as confidence scores and the believability of the prediction is as much a factor in the score as the accuracy. We show that the Meta-LR model can serve as both a discriminator and a calibrator in a pragmatic decision-support system that requires interpretable and reliable probability estimates. Research shows that language models are commonly miscalibrated, particularly in low-shot learning scenarios. They concluded that uncertainty estimation and prediction reliability can be greatly enhanced by applying recalibration techniques, which highlights the significance of calibrated confidence scores in NLP applications (29).

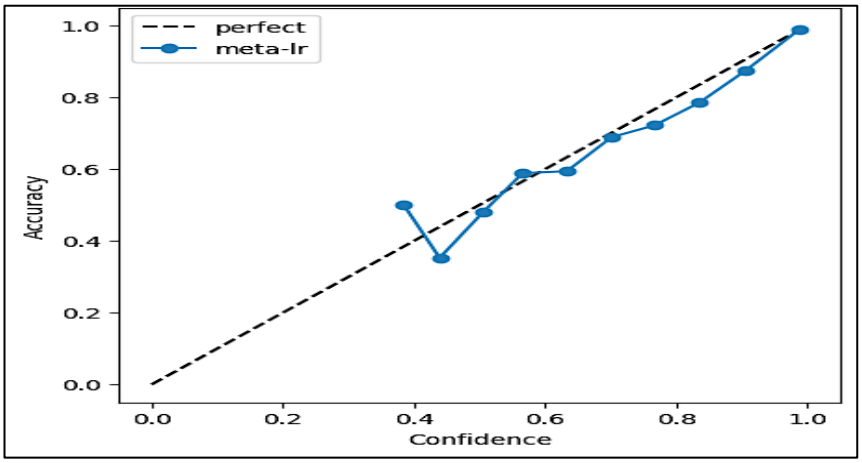

Figure 6: Reliability Diagram and Calibration Performance of the Meta Logistic Regression (Meta-LR) Model

Figure 7: Aspect-based Sentiment Heatmap Using Meta-learner Probabilities

Aspect-Sentiment Heatmap Using Meta-learner Probabilities

From the aspect-sentiment heatmap in Figure 7, the proposed hybrid learning model effectively captures positive sentiment across all service dimensions, indicating positive customer perceptions. Remarkably, Wi-Fi has the highest positive sentiment (+0.571), followed by comfort (+0.478), price (+0.443), food (+0.438) and service (+0.420), suggesting that these are significant factors of customer satisfaction. That connectivity is an important element of service analysis in today's services. Contrarily, sentiment distribution over time indicates that the proportion of neutrality (0.352) is higher. In contrast, the proportion of positive sentiment (0.371) is lower, suggesting that service efficiency and waiting time should be optimised. Unlike in other aspects, the likelihood of negative sentiment is minimal, but the price has the highest negative sentiment (-0.322), suggesting that cost is an issue for users. In addition, moderate neutrality in some terms offers the opportunity to optimise service differentiation, shifting perceptions toward more positive experiences. Overall, the meta-learning methodology supports stability, discriminative ability and generalisation by providing a smooth, continuous probability distribution across all fronts, thereby justifying the efficiency of combining lexical and contextual representations in fine-grained aspect-based sentiment analysis in a real-world setting. Recent research demonstrated that logistic regression can serve as an effective meta-learner with fast convergence and strong classification performance in the few-

shot learning setting. The results confirm the logistic regression's potential for improving the accuracy and efficiency of the meta-learning frameworks (30).

Tier Model Comparison

A comparison of tier-by-tier models estimated using the out-of-fold (OOF) procedure, as shown in Figure 8, indicates that the single-learner and the proposed stacked meta-learning model achieve high classification performance. Logistic Regression and Calibrated Linear SVM were the best-performing methods, achieving accuracies and Macro-F1 scores exceeding 0.90. This is another sign of their ability to learn lexical patterns from the dataset they have been assigned. Multinomial Naive Bayes, on the other hand, was even worse, with an F1 score of 0.72 (macro). These results show once again that Naive Bayes is less effective at capturing contextual dependencies in sentiment data. It is also true that contextualised embeddings can better capture semantic nuances, with DistilBERT (a deep learning model) scoring around 0.96 on both metrics, suggesting that they can be more effective than traditional embeddings. It should also be noted that the proposed Stacked Meta-LR model achieves the best performance, with accuracy and Macro-F1 nearly equal to 0.97, which implies that heterogeneous models leverage complementary advantages in both lexical and contextual representations. The slight differences in accuracy and Macro-F1 across all models also suggest similar class predictions and robustness to class imbalance. These findings demonstrate that the hybrid stacking approach achieves a high degree of generalization and improves performance over single machine learning and deep learning models and is thus highly applicable to fine-grained sentiment analysis.

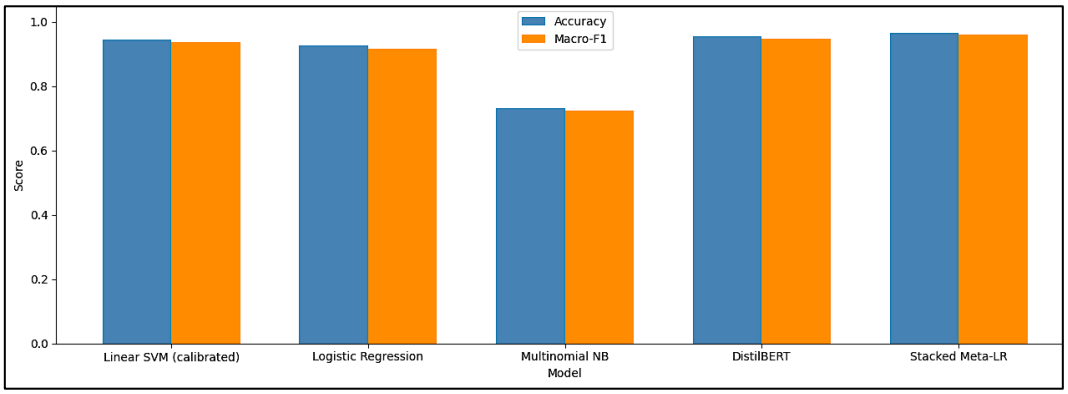

Figure 8: Tier-wise Model Performance Comparison Using Out-of-Fold (OOF) Evaluation

Conclusion

The article presents a fully scalable hybrid sentiment analysis system that effectively integrates both classical machine learning and transformer-based deep learning. It leverages them in a stacked ensemble. The suggested approach, based on 74,921 annotated Twitter (non-filtered) examples, demonstrates its ability to distinguish between superficial lexical rules and underlying semantics, thereby addressing the key problems posed by noisy, informal social media text.

The experiment's outcomes are clear: the stacked meta-learning approach is significantly better than base models alone. Despite the high baseline accuracy of classical algorithms, such as Logistic Regression and Calibrated Linear SVM, which can effectively utilise high-dimensional TF-IDF features, the DistilBERT transformer model can also provide additional contextual knowledge. The meta-learner, leveraging these complementary strengths, significantly improves performance and the Meta Logistic regression model achieves 0.9705 accuracy and macro-F1. In addition, it is highly discriminative, with nearly perfect ROC-AUC and highly reliable, with good calibration and low Expected Calibration Error. The paper also addresses the interpretation and practical use of sentiment analysis, as well as classification performance, through temporal and aspect-based approaches. The results of these analyses lead us to conclude that user behaviour patterns are significant, that sentiment changes are event-driven and that the most significant service dimensions influencing public perception have been identified. Another is uncertainty estimation, in which evidence supporting it is incorporated to

enhance predictions. Thus, the framework is appropriate for the use of decision-support systems in sensitive fields. The findings establish that hybrid stacking is a practical method for sentiment analysis, improving generalisation, strength and reliability compared to individual models. The approach is particularly practical when estimating predictive accuracy and confidence is necessary. Future studies can discuss methods for integrating more advanced transformer architectures, domain pretraining and combining multimodal data sources to make further progress.

Abbreviations

ABSA: Aspect-based Sentiment Analysis, AUC: Area Under the Curve, CEC: Expected Calibration Error, LIME: Local Interpretable Model-agnostic Explanations, OOF: Out of Fold, ROC: Receiver Operating Characteristic, SHARP: SHapley Additive exPlanations, SVM: Support Vector Machine, TF-IDF: Term Frequency–Inverse Document Frequency

Acknowledgment

I want to express my sincere thanks to the School of Doctoral Studies and Research at National Forensic Sciences University for providing research facilities and leadership that enabled me to complete this work. Their help and supportive atmosphere have greatly assisted me in this work.

Author Contributions

Demmelash Abate: Conceptualization, Data Curation, Methodology, Model Training, Validation, Formal Analysis, Writing-Original Draft, Nilay Mistry: Supervision, Validating Methods, Evaluation, Review, Editing, Visualization.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability

Data presented in this study are available upon request from the corresponding author.

Declaration of Artificial Intelligence (AI) Assistance

The author declares no use of Artificial Intelligence (AI) for the write-up of this manuscript. The authors take full responsibility for the content's originality, interpretation and accuracy.

Ethical Approval

Not Applicable.

Funding

The author did not receive any funding for this research.

References

1. Sathya A, Mythili M. Evaluating Sentiment Classification to Specify Polarity by Lexicon-Based and Machine Learning Approaches for COVID-19 Twitter Data Sets. Journal of advanced applied scientific research. 2023. doi: 10.46947/joaasr542023678
2. Mahmood A, Kamaruddin S, Naser R, Nadzir MM. A Combination of Lexicon and Machine Learning Approaches for Sentiment Analysis on Facebook. Journal of System and Management Sciences. 2020. doi: 10.33168/jsms.2020.0310
3. Sham NM, Mohamed A. Climate Change Sentiment Analysis Using Lexicon, Machine Learning and Hybrid Approaches. Sustainability. 2022. doi:10.3390/su14084723
4. Srivastava R, Bharti P, Verma P. Comparative Analysis of Lexicon and Machine Learning Approach for Sentiment Analysis. International Journal of Advanced Computer Science and Applications. 2022. doi:10.14569/ijacsa.2022.0130312
5. Yadav A, Vishwakarma DK. Sentiment analysis using deep learning architectures: a review. Artificial intelligence review. 2020;53(6):4335-85. doi.10.1007/s10462-019-09794-5
6. Zhang X, Liu Y, Zhang T, *et al*. A BERT–LSTM–Attention Framework for Robust Multi-Class Sentiment Analysis on Twitter Data. Systems. 2025;13(11):964. doi.10.3390/systems13110964
7. Albladi A, Uddin MK, Islam M, Seals C. Twssenti: A novel hybrid framework for topic-wise sentiment analysis on social media using transformer models. arXiv preprint arXiv:250409896. 2025. doi:10.48550/arxiv.2504.09896
8. Suleiman D, Awajan A. Deep Learning Based Abstractive Text Summarization: Approaches, Datasets, Evaluation Measures and Challenges. Mathematical Problems in Engineering. 2020. doi:10.1155/2020/9365340
9. Rakshit P, Sarkar P, Roy S. Hybrid Deep Learning Approach for Sentiment Analysis on Twitter Data. Multimedia Tools and Applications. 2025;84(15):15271-92. doi:10.1007/s11042-024-19555-4
10. Tan KL, Lee C, Anbananthen K, Lim K. RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network. IEEE Access. 2022;10:21517-25. doi: 10.1109/access.2022.3152828
11. Xiao H, Luo L. An Automatic Sentiment Analysis Method for Short Texts Based on Transformer-BERT Hybrid Model. IEEE Access. 2024;12:93305-17. doi:10.1109/access. 2024.3422268
12. Chen T, Su LY-F, Ng YMM, Stasiewicz MJ, Wang Y-C. Transforming social media sentiment analysis with generative large language models: A case study from

the cultured meat debate. Food Quality and Preference. 2026;145:105968. doi: 10.1016/j.foodqual.2026.105968
13. Bathla G, Singh P, Singh R, Cambria E, Tiwari R. Intelligent fake reviews detection based on aspect extraction and analysis using deep learning. Neural Computing and Applications. 2022;34:20213-29. doi:10.1007/s00521-022-07531-8
14. Hájek P, Barushka A, Munk M. Fake consumer review detection using deep neural networks integrating word embeddings and emotion mining. Neural Computing and Applications. 2020;32:17259-74. doi:10.1007/s00521-020-04757-2
15. Zhang S, Zhu A, Zhu G, Wei Z, Li KC. Building Fake Review Detection Model Based on Sentiment Intensity and PU Learning. IEEE Transactions on Neural Networks and Learning Systems. 2023;34:6926-39. doi:10.1109/tnnls.2023.3234427
16. Manaskasemsak B, Tantisuwankul J, Rungsawang A. Fake review and reviewer detection through behavioral graph partitioning integrating deep neural network. Neural Computing and Applications. 2021;35:1169-82. doi:10.1007/s00521-021-05948-1
17. Phukon P, Potikas P, Potika K. Detecting Fake Reviews Using Aspect-Based Sentiment Analysis and Graph Convolutional Networks. Applied Sciences. 2025. doi:10.3390/app15073771
18. Kotha S. Distributed Fake Review Detection and Real-Time Anomaly Detection: A Technical Framework. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. 2025. doi:10.32628/cseit25112824
19. Gawlikowski J, Tassi CRN, Ali M, *et al*. A survey of uncertainty in deep neural networks. Artificial intelligence review. 2023;56(Suppl 1):1513-89. doi:10.1007/s10462-023-10562-9
20. Agarwal B, Nayak R, Mittal N, Patnaik S. Deep learning-based approaches for sentiment analysis. 2020. doi.10.1007/978-981-15-1216-2
21. Jahin MA, Shovon MSH, Mridha M, Islam MR, Watanobe Y. A hybrid transformer and attention based recurrent neural network for robust and interpretable sentiment analysis of tweets. Scientific reports. 2024;14(1):24882. doi:10.1038/s41598-024-76079-5
22. Jain R, Kumar A, Nayyar A, *et al*. Explaining sentiment analysis results on social media texts through visualization. Multimedia Tools and Applications. 2023;82:22613-29. doi:10.1007/s11042-023-14432-y
23. Sweidan AH, El-Bendary N, Taie S, Idrees A, Elhariri E. Explainable Deep Learning for COVID-19 Vaccine Sentiment in Arabic Tweets Using Multi-Self-Attention BiLSTM with XLNet. Big Data Cogn Comput. 2025;9:37. doi:10.3390/bdcc9020037
24. Mendon S, Dutta P, Behl A, Lessmann S. A Hybrid Approach of Machine Learning and Lexicons to Sentiment Analysis: Enhanced Insights from Twitter Data of Natural Disasters. Information Systems Frontiers. 2021;23:1145-68. doi:10.1007/s10796-021-10107-x
25. Thakur RS, Singh J. AI and NLP for Mental Health Prediction from Social Media: A Decade of Progress, Challenges and Explainability (2015–2025). International Journal for Research in Applied Science and Engineering Technology. 2025. doi:10.22214/ijraset.2025.73857
26. Ding M-L, Ma Y-L, You F-Q. Domain-Constrained Stacking Framework for Credit Default Prediction. Mathematics. 2025;13(21):3451. https://doi.org/10.3390/math13213451
27. Neshat M, Sergiienko NY, Rafiee A, Mirjalili S, Gandomi AH, Boland J. meta-learner deep gradient boosting model: A case study from Australian coasts. Energy. 2024;304:132122. doi:10.1016/j.energy.2024.132122
28. Zabor EC, Reddy CA, Tendulkar RD, Patil S. Logistic regression in clinical studies. Int J Radiat Oncol Biol Phys. 2022;112(2):271-7. doi.org/10.1016/j.ijrobp.2021.08.007
29. Haoyuan S, Yizhong M, Chenglong L, Jian Z, Lijun L. Hierarchical Bayesian support vector regression with model parameter calibration for reliability modeling and prediction. Reliability Engineering & System Safety. 2023;229:108842. doi:10.1016/j.ress.2022.108842
30. Abbas MA, Munir K, Raza A, *et al*. A novel meta learning based stacked approach for diagnosis of thyroid syndrome. Plos one. 2024;19(11):e0312313. doi:10.1371/journal.pone.0312313

How to Cite: Abate D, Mistry N. A Hybrid Framework for Large-scale Tweet Sentiment Analysis Using Classical Machine Learning, Transformer Models and Uncertainty Estimation. Int Res J Multidiscip Scope. 2026;7(3):1873-1887. DOI: 10.47857/irjms.2026.v07i03.011806