

Investigation of fine-tuned BERT models for sentiment analysis in COVID-19 tweets using a fuzzy logic-based ensemble approach

Mustafa Sefa Evgin*, Sinan Toklu

Department of Computer Engineering, Faculty of Technology, Gazi University, Ankara, Turkiye

Cite this article as: Evgin, M. S., & Toklu, S. (2026). Investigation of fine-tuned BERT models for sentiment analysis in COVID-19 tweets using a fuzzy logic-based ensemble approach. *J Comp Electr Electron Eng Sci*, 4(1), 17-26.

Received: 19.01.2026

Accepted: 24.02.2026

Published: 25.04.2026

ABSTRACT

Aims: With the beginning of the COVID-19 pandemic, social media applications such as twitter used more than usual because people started to work at their homes rather than offices. Thus, data on this application has become more important to manage crisis of COVID-19. While conventional deep learning methods have shown success in sentiment analysis, they often encounter challenges in capturing the inherent semantic ambiguity and informal linguistic structures prevalent on social media platforms. To find ambiguity on these texts propose ensemble model enhanced by fuzzy logic designed to improve sensitivity and capability.

Methods: Architecture uses BERT model, fine-tuned for specific data to supply dynamic attributes for MLP, LSTM and BiLSTM elements. Their shared executive is regulated via Mamdani Fuzzy Interference System. Then dynamic weights results from defuzzification after mapping prediction of confidence and validation accuracy value on 7 level fine-grained rule set.

Results: Experiment performed on Kaggle Corona NLP dataset resulted in 91.28% accuracy, 91.23% F1 score and 91.38% precision. System's robust performance is demonstrated by Mean Square Error of 0.2301.

Conclusion: Relative analysis demonstrates dominance of this approach against traditional models. Fuzzy Ensemble model proposes more trustworthy solution for obstruse tweets with successfully straining noise and dealing semantic uncertainties which is naturally present in social media data.

Keywords: Sentiment analysis, BERT, fuzzy logic, ensemble learning, COVID-19, natural language processing

INTRODUCTION

Natural language processing has appreciated as need to interpret of unstructured textual information raised. During the COVID-19 pandemic, social media platforms provided a basis for expressing their concerns and psychological situations thus examination of this data become crucial (Singh et al., 2022). Accurate information on this data is essential for politician for strategic perspective. Though current literature confirms the effectiveness of BERT (Devlin et al., 2019) and LSTM (Hochreiter & Schmidhuber, 1997) frameworks, semantic ambiguity and noise are still among the biggest problems on social media data (Dhanalakshmi et al., 2024).

Yet, human language is intrinsically filled with ambiguity and uncertainty. Standard algorithms relying on 'crisp' logic often struggle here, as they try to force text into rigid 'positive' or 'negative' boxes. This binary approach inevitably ignores the nuanced gray areas that are fundamental to natural communication. To mathematically model this type of uncertainty, current literature advocates for the integration of fuzzy theories into natural language processing (Howell & Ertugan, 2017). This approach proves particularly superior in high-stakes fields like medical diagnosis. In such critical contexts, text-based systems enhanced with fuzzy logic have

been shown to deliver significantly more precise and reliable outcomes than standard NLP methods (Omogrebe et al., 2020). In this context, it is widely accepted in the literature that hybrid and fuzzy-based approaches offer a more stable architecture compared to individual deep learning models, especially in datasets where sentiment transitions are not sharp (Sherin et al., 2025). Furthermore, the creation of optimized and compressed representations of texts based on information theory has entered the literature as a new approach to enhancing the performance of machine learning models (Kale et al., 2024).

When the literature is examined, it is observed that a large portion of research in the field of COVID-19 sentiment analysis relies on standard deep learning architectures such as LSTM or CNN, as seen in the study (Karaca & Aslan, 2021). Although these models produce successful results, hybrid approaches in which modern transformer architectures like BERT are used within a Fuzzy Ensemble structure are limited. Even though hybrid studies on COVID-19 tweet datasets have increased recently, gaps still exist. For instance, in a study where sampling methods were employed to address imbalance in the dataset (Kumar et al., 2024), an accuracy rate of 89.00%

Corresponding Author: Mustafa Sefa Evgin, 24833301023@gazi.edu.tr



This work is licensed under a Creative Commons Attribution 4.0 International License.

was achieved by combining BERT and CNN models. Another study conducted on same dataset remained at the level of 86% while using hybrid model. Conversely, research employing conventional deep learning layers like BiLSTM (Schuster & Paliwal, 1997) and GRU saw performance plateau at an 85% accuracy rate (Shahriar & Sarker, 2023). These findings serve as implicit evidence that shifting towards Transformer-based architectures is essential for surpassing this threshold.

Nevertheless, a common limitation in prior research is the reliance on static techniques to handle model uncertainty. In contrast, fuzzy logic has emerged as a superior alternative, particularly for complex tasks like detecting rare textual events (Arslan et al., 2021) where classical methods struggle. Consequently, our work seeks to bridge this gap by integrating fine-tuned BERT architectures with dynamic fuzzy logic weighting.

To handle with challenges, we propose ensemble framework that integrates a fine-tuned BERT with 7-level Mamdani-type fuzzy inference system. Our model starts with fine-tuning on BERT architecture on COVID-19 specific dataset for capturing context dependent shift in vocabulary. Then contextual representation is processed by three parallel individual base learner MLP (Rumelhart et al., 1986), LSTM and BiLSTM for extract diverse architectural features. Sentiment decisions are conducted via dynamic fuzzy weighting mechanism which evaluates both model confidence and validation accuracy. Our work have three contribution: Firstly fined-grained, 7 level fuzzy logic module provides better sensitivity over traditional 3-level systems; secondly dynamic 'reliability filter' mitigates noise with penalizing high confidence but formerly inaccurate prediction; lastly model achieves better performance leap which surpasses soft, hard voting and recent hybrid benchmarks in COVID-19 sentiment analysis domain.

RELATED WORK

This section provides inclusive examinations of theoretical keystone to support our search: deep learning-based sentiment analysis, transformer driven hybrid architectures and the useful application of fuzzy logic.

Deep Learning and Hybrid Approaches

Over the years recurrent neural networks (RNNs) and machine learning method have served as headstone for sentiment analysis, operating as typical tools for mapping text into vector space and classifying. Comprehensive literature reviews in this area emphasize that deep learning models' performance performs quietly better than conventional models for text classification (Miaee et al, 2021). One of the examples for it, analyzing of children's stories where feature engineering integrated with Random Forest algorithms produced notably more steady result than traditional lexicon-based approaches (Bilal et al., 2023). Similarly, several studies confirmed the benefit of weight updated classifiers for text-based sentiment analysis across variety of datasets (Bilal et al., 2024). Furthermore, research emphasize that bringing Deep Belief Network with feature selection considerably improves classification accuracy, mainly by alleviating noise essential in high-dimensional data (Ruangkanokmas et al., 2016).

Transformer-Based (BERT, RoBERTa) Models

As the BERT model became important prominent, researchers focused on merging BERT with other deep learning models

more than before. A prominent purpose includes combining BERT with CNNs for long text classification. Thus, hybrid attitude supports model to instantaneously catch local features by the CNN while keeping global context provided by BERT (Chen et al., 2022).

Similar hybrid approach used to evaluate investor sentiment in the energy sector. While channeling BERT outputs directly to BiLSTM model, this architecture performed significant proficiency in finding the progressive dependencies which essential in financial literature (Cai et al., 2020). Recent study finds a triple hybrid model by implementing RoBERTa (Liu et al, 2019) with BiLSTM and CNN layers to examine ceramic product comments. Results shows that this combined structure achieved considerably higher accuracy when compared individual models operating alone (Yang et al., 2025). Furthermore, research on COVID-19 tweets revealed that combining BERT with LSTM model resulted in better outcome compared with conventional methods (Dhanalakshmi et al., 2024). This hybrid integration showed considerably more efficient than conventional models in catching nuances in the dataset. Similarly, comprehensive research shows that architectures combining BERT with CNN and BiLSTM layers notably exceed conventional Word2Vec-based techniques, achieving better results both accuracy and F1 scores (Bello et al., 2023).

While BERT models show extraordinary potential in contextual learning, current studies emphasize a critical point; they mostly need additional optimization layer. This additional optimization layer is critical for effectively managing high noise and semantic discrepancies which occurs commonly faced in raw data. (Elgabry & Hamdi, 2025).

METHODS

Ethical approval was not required for this study as it does not involve human participants, animal subjects, or identifiable personal data. All procedures were carried out in accordance with the ethical rules and principles.

In this section, we present technical infrastructure and methodological details of proposed hybrid model which aimed performing sentiment analysis of COVID-19 tweets with higher accuracy, F1 score and minimized mean square error. Methodology begins with dataset preparation and preprocessing then fine-tuned BERT model used for contextual text representation and parallel deep learning models MLP, LSTM and BiLSTM processing respectively. Finally, probabilities which generated from respectively models are combining via Mamdani-type fuzzy logic system and it offers decision mechanism closer to human thinking and reasoning system than conventional system. Fuzzy system then elaborates principles of proposed ensemble model.

All experimental processes and model training phases performed on Google Colab Pro cloud computing platform. In this platform we used NVIDIA A100 Tensor Core GPU with 40GB of VRAM and High-RAM, approximately 25GB. Software environment configured with Python 3.10 and TensorFlow 2.x and KerasNLP library. Individual models (BERT+MLP, BERT+LSTM, BERT+BiLSTM) have comparable model complexities, ranging from 109.5 million to 109.9 million trainable parameters, with BiLSTM variants displaying highest computational cost during training (869.28 second). In terms of interference latency, deep learning

methods shows high efficiency with an average processing time of approximately 2.00 ms per sample. Rule-based fuzzy logic module performs computational overhead of 22.67 ms per sample due to CPI-bound defuzzification but aggregate systems stay approximately 22.47 ms per second. That means our model processing product roughly 40 tweets per second and it proves that proposed ensemble model is computationally viable for real time social media monitoring application despite increased architectural complexity.

Dataset

In this study, the “Corona NLP” dataset obtained from the Kaggle platform was utilized. Due to its structure harboring varying sentiment intensities regarding COVID-19, this dataset is also preferred as a primary data source in recent studies where language models are tested within the framework of information theory (e.g., in the TexShape architecture) (Kale et al., 2024). The dataset contains tweets posted about COVID-19 and their sentiment labels (positive, negative, neutral). A total of 41,157 training and 3,798 test data points were used.

Training dataset includes 15398 negative, 7713 neutral and 18046 positive tweets. Test dataset includes 1633 negative, 619 neutral and 1546 positive tweets.

Data preprocessing: The following steps were applied to clean the noise from the raw tweet data:

- **Cleaning:** URLs, HTML tags, and “@user” expressions were removed using Regex.
- **Normalization:** All texts were converted to lowercase, and non-ASCII characters were filtered out.
- **Tokenization:** Texts were converted into numerical vectors (Input IDs and Attention Masks) using the bert-base-uncased tokenizer, in accordance with the BERT model. The maximum sequence length was determined as 60.

Proposed Model Architecture: Fuzzy-BERT Ensemble

In this study, a hierarchical ensemble architecture integrating BERT-based contextual feature extraction with a fuzzy logic-based decision fusion mechanism has been developed to detect the sentiment status of COVID-19 tweets. This architecture, the detailed flowchart of which is given in **Figure 1**, fundamentally consists of three main modules: (1) Contextual feature extraction (BERT), (2) Parallel base learners, and (3) Fuzzy decision fusion module. The architecture begins with the retraining of all layers of the pre-trained BERT-base-uncased model as trainable=True. Contrary to the “frozen weights” approach common in the literature, the model was subjected to a fine-tuning process to more accurately represent words such as “quarantine,” “mask,” and “symptom,” whose meanings shifted within the COVID-19 context. Contextual features obtained from the BERT layer were transferred to three parallel sub-models (BERT+MLP, BERT+LSTM, and BERT+BiLSTM) to approach the data from different perspectives. Thus, dense, sequential, and bidirectional information processing capabilities were combined within the same architecture.

In combining the prediction probabilities produced by these sub-models, a Mamdani-type fuzzy inference system (FIS) was used instead of classical weighted average methods. In this system, the effect of each model on the final decision depends not on fixed coefficients, but on dynamic variables

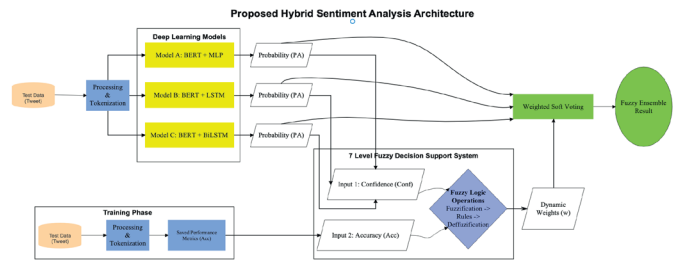


Figure 1. Proposed 7-level (fine-grained) Mamdani fuzzy logic module contextual feature extraction (fine-tuned BERT)
BERT: Bidirectional encoder representations from transformers

such as “model confidence” and “model accuracy.” Triangular Membership Functions (Trimf), shown in Equation (1), were preferred for the fuzzification of these variables due to computational efficiency (Jang et al, 1997):

$$\mu_A(x) = \max\left(\min\left(\frac{x-a}{b-a}, \frac{c-x}{c-b}\right), 0\right) \quad (1)$$

In Equation (1) *a*, *b*, and *c* represent the corner points of the triangular membership function. Numerical values are converted into linguistic variables such as “low,” “medium,” and “high” through these functions, and IF-THEN rules based on expert opinion are activated.

In the final stage of the inference mechanism, the fuzzy set formed by the triggered rules needs to be converted into a crisp numerical weight value (*W_i*). For this operation, the centroid (center of gravity) method given in Equation (2) was used in the Defuzzification stage (Mendel, 1995):

$$W_i = \frac{\int \mu_C(z) \cdot z \, dz}{\int \mu_C(z) \, dz} \quad (2)$$

Here, $\mu_C(z)$ denotes the membership value of the aggregated fuzzy set. These calculated weights are combined with the prediction *P_i* of each sub-model to obtain the final output of the ensemble architecture. Thus, both the strong representation capacity of contextual deep learning models and the flexible decision mechanism of fuzzy logic are integrated within a single architecture.

In contrast to traditional Word2Vec or GloVe methods, the BERT (bidirectional encoder representations from transformers) model was employed on the preprocessed tweet texts at the input layer of the architecture to capture the context-dependent meanings of words based on their position within the sentence.

BERT was specifically selected as the primary feature extractor due to its superior bidirectional context-capture capabilities and its proven robustness as a benchmark in recent COVID-19 literature, which allows for a more focused evaluation of the proposed fuzzy ensemble’s impact on handling sentiment uncertainty.

At this stage, all layers of the bert-base-uncased model were set to trainable, and the model was subjected to a fine-tuning process on the COVID-19 dataset; this ensured that the model could distinguish between medical terms with negative connotations and their usage in daily language. In this process, where high-density contextual vectors of dimension (N, 768) were generated for each tweet, the training of the model was carried out using the Adam optimizer. To prevent overfitting, a Dropout rate of 0.3 was applied; furthermore, the training parameters were set as a learning rate of 2e-5, a batch size of 16, and 4 epochs.

Parallel Base Learners

Dynamic vectors which obtained from BERT instantaneously feeds three distinct parallel subordinate model (MLP, LSTM and BiLSTM) to capture semantic dimension of data and utilize the principle of ‘architectural diversity’. While the MLP component of this hybrid design models features through dense and non-linear transformations, it has been observed in the deep learning literature that such hybrid utilization of different architectural paradigms (such as Transformer and MLP) yields significant improvement compared to single-type architectures (Bashar, 2025). The BERT+MLP model was trained for 4 epochs using the Adam optimizer and a learning rate of 2×10^{-5} , in accordance with the parameters specified in **Table 1**.

Parameter	Value
Learning rate	2×10^{-5}
Epoch	4
Batch size	16
Optimizer	Adam
Hidden layer units	128
Hidden activation	ReLU
Dropout rate	0.3
Output activation	Softmax

BERT: Bidirectional encoder representations from transformers, MLP: Multi-layer perceptron, ReLU: Rectified linear units

Parameter	Value
Learning rate	2×10^{-5}
Epoch	4
Batch size	16
Optimizer	Adam
LSTM units	64
Dropout rate	0.3
Output activation	Softmax
Loss function	Sparse categorical cross entropy

BERT: Bidirectional encoder representations from transformers, LSTM: Long short-term memories

Parameter	Value
Learning rate	2×10^{-5}
Epoch	4
Batch size	16
Optimizer	Adam
BiLSTM units	64
Dropout rate	0.3
Output activation	Softmax
Loss function	Sparse categorical crossentropy

BERT: Bidirectional encoder representations from transformers, BiLSTM: Bidirectional long short-term memories

Capturing context in text classification by protecting progressive relationship and forwarding flow of words within the text through memory cells, contrasting conventional models (Airlangga, 2024). BERT+LSTM model was trained 4 epochs with one recurrent layer, 64 neurons, using the Adam optimizer with a learning rate of 2×10^{-5} , based on the parameters presented in **Table 2**.

BiLSTM last branch of parallel base learners integrated in the system but has different features when compared with LSTM. BiLSTM scans text in both forward and backwards directions, so learning the future context of the word; this reduces ‘vanishing gradient’ problem and gives more stable results for maintaining semantic integrity, especially for long sentences (Rahman, 2025). Correspondingly BERT+BiLSTM model is trained for 4 epoch, one recurrent layer and total 128 neuron 64 for the forward pass and 64 for the backward pass with the configuration setting in the **Table 3**, while employing Adam optimizer and with learning rate of 2×10^{-5} .

Each sub-model generates its own probability distribution (P_{model}) regarding the class of the tweet (positive, negative, neutral).

Fuzzy Decision Fusion Module

In this study, a Mamdani-type fuzzy inference system (FIS) has been designed to optimally combine the predictions of the individual models (BERT+MLP, BERT+LSTM, BERT+BiLSTM). Proposed system essentially consists of three sub-layer and its decision mechanism close to human thinking and reasoning system by transforming input values into linguistic variables.

Fuzzification

Normalized triangular membership functions (Trimf) were used between the range of [0,1] for input and output variables of the system. The reason for choosing normalized triangular membership function, it able to reduce quantities cost due to their linear features also offers to quick responses for real time application. The variables expressed in the system with three main component which shown in **Figure 2**.

The first input which is labeled ‘model confidence’ and model confidence is equal to peak SoftMax probability score originated from model’s specific prediction. Divided from conventional three-layer classification, we charted this variable

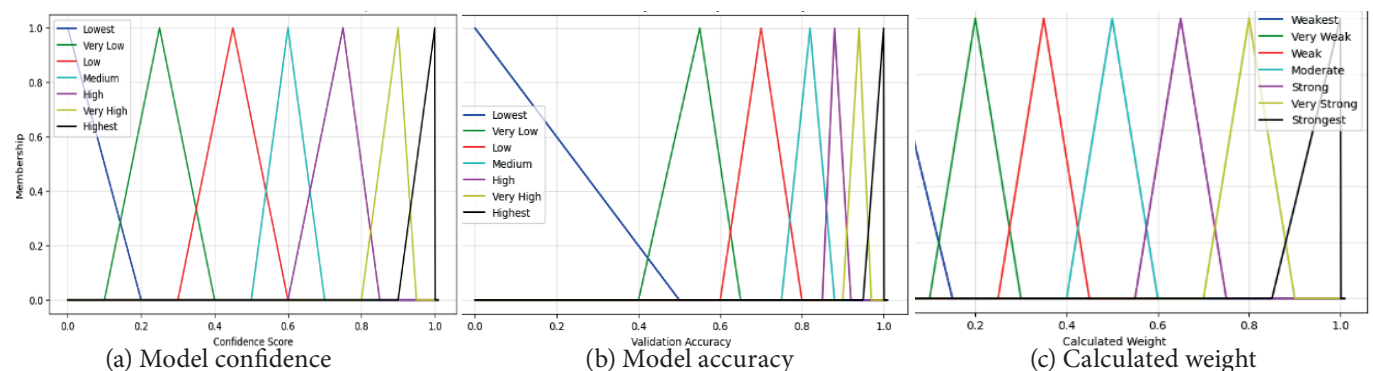


Figure 2. Membership functions of the proposed fuzzy logic system

onto more elaborate seven level scale varying from ‘lowest’ to ‘highest’ for drastically raise the system’s performance.

Model accuracy is a second input, and it reflect model’s collective performance for validation dataset. Reflecting approach taken with confidence scores, we also charted accuracy onto seven level granular scale. Detailed segmentation is designated for handle the model’s uncertainty with higher precision, especially addressing the crucial ambiguity often found in between 50% and 60%.

‘Weight’ decides how much effect a specific model holds over ensemble’s final decision and acts as a final output of our inference system. Dependable with our input variables, this output is categorized into seven-layer spectrum that ranges from ‘weak’ to ‘strong’.

After fuzzification system assesses linguistic terms while using ‘IF-THEN’ conditional statement by domain experts. In harmony with improved 7 level granularity, we used a set of 12 precise rules to rule decision making process. This inclusive rule base provides for dual purpose first it suppresses underperforming models and simultaneously refining weight allocation for nuanced scenarios, such as instances of ‘high accuracy’ and ‘medium confidence’.

Model’s final decision weight is stemmed from the collaboration between calculated input variables and fuzzy logic interference mechanism. Inclusive analysis of rule base and main decision logic chart exhibited in **Table 4**. In this regard, linguistic terms equal to seven tier fuzzy spectrum, expressed from L1 (lowest) throughout L7 (highest).

Table 4. Rule base of the Mamdani fuzzy inference system

Rule no	Input 1: Confidence	Operator	Input 2: Accuracy	Output: Weight
1	l7 (highest)	AND	l7 (perfect)	l7 (strongest)
2	l6 (very high)	AND	l7 (perfect)	l7 (strongest)
3	l5 (high)	AND	l7 (perfect)	l6 (very strong)
4	l7 (highest)	AND	l6 (very good)	l7 (strongest)
5	l5 (high)	AND	l6 (very good)	l6 (very strong)
6	l4 (medium)	AND	l6 (very good)	l5 (strong)
7	l7 (highest)	AND	l5 (good)	l6 (very strong)
8	l4 (medium)	AND	l5 good)	l4 (moderate)
9	l7 (highest)	AND	l4 (average)	l5 (strong)
10	l3 (low)	OR	l3 (low)	l3 (weak)
11	-	-	l2 (very low)	l2 (very weak)
12	l1 (lowest)	OR	l1 (lowest)	l1 (weakest)

Standard protocols (rules 1-9) evaluate both inputs together by the ‘AND’ operator. However, to reinforce system reliability, we applied a distinct strategy for low-performance scenarios.

Especially, rules 10 and 12 uses ‘OR’ operator as fool proofed. This ensures that if accuracy or instantaneous confidence drops below a critical level, model’s effects is automatically limited. With doing this system effectively filters out noise and avoiding unstable interpreters from crooking the final ensemble. Instead of 49 rules, we generated 12 rules for the reason decrease calculation cost and focus on only critic points.

Defuzzification

Defuzzification is a process that converting fuzzy output into clear numerical value with using rules. In this study we used centroid (center of gravity) which is recognized as one of the most common and reliable method in fuzzy interference system. This method calculates geometric center of the area which is formed by all activated members functions and its enabling system to generate precise weight value. Centroid method considers contribution of all active rules, so allowing for smoother and continuous transition in output value. Obtained clear value is assigned as a final constant of the relevant deep learning model in voting process.

RESULTS

In this section experimental results of 7 level fuzzy logic ensemble architecture which is developed for sentiment analysis for COVID-19 tweets with ‘Corona NLP’ dataset are presented in detail. To measure proposed method’s efficiency and resistance against noise we used Accuracy, F1-score, precision, recall and mean square error (MSE) as principal performance sings. Resulting data then subjected to a three-stage analysis for ensuring a multidimensional evaluation of model’s reliability. First performance differences among proposed model, individual models and conventional ensemble models were examined throughout comparative analysis (ablation study). Then classification behaviors of sub-models (BERT+MLP, BERT+LSTM, BERT+ BiLSTM) and ensemble structures were visualized by confusion matrices and (reciever operating characteristic) ROC curves. Finally, case study was performed to test context sensitivity of model beyond arithmetical data.

Ablation Study and Comparative Analysis

To show the efficiency of proposed Fuzzy Logic-based ensemble architecture (fuzzy ensemble), comparative analysis such as ablation study were carried out not only for fine-tuned BERT models but also unweight soft voting and hard voting methods because they are widely used in the literature and **Table 5** provides detailed information of performance metrics for all evaluated methods.

A review of **Table 5** identifies BERT+LSTM as the extraordinary performer among the standalone architectures, obtaining an accuracy of 90.20%. However, the data shows a considerable

Table 5. Performance benchmarking of the proposed framework against standard ensemble and individual models

Model architecture	Model type	Accuracy	F1-score	Precision	Recall	MSE
BERT+MLP	Individual model	89.86%	89.80%	89.86%	89.86%	0.2569
BERT+LSTM	Individual model	90.20%	90.13%	90.25%	90.20%	0.2527
BERT+BiLSTM	Individual model (Best)	89.81%	89.74%	89.90%	89.81%	0.2598
Hard voting	Standard ensemble	90.96%	90.91%	91.05%	90.96%	0.2340
Soft voting	Standard ensemble	91.12%	91.07%	91.22%	91.12%	0.2348
Fuzzy ensemble	Proposed method	91.28%	91.23%	91.38%	91.28%	0.2301

MSE: Mean squared error, BERT: Bidirectional encoder representations from transformers, MLP: Multi-layer perceptron, LSTM: Long short-term memories, BiLSTM: Bidirectional long short-term memories

leap in predictive capability upon the implementation of ensemble strategies. The proposed 7-level fuzzy ensemble framework exceeded both standalone models and the conventional soft voting approach, obtaining an excellent accuracy of 91.28%. McNemar testing at this stage emphasize a specific improvement; the proposed framework successfully corrected 6 complex instances that the soft voting approach had previously misclassified. While the numerical variance might seem slight at first glance, the fundamental metrics reveal a substantial advantage. Proposed methods prove it reliability with boosting precision to 91.38% and reduced mean squared error (MSE) to 0.2301. These figures confirm that 7 level structure overtakes at filtering out false positives, indicating a system that is both accurate and inherently robust against data noise.

Table 6 shows our approach has considerable performance gain and outperformed the benchmark baseline by approximately 15.73% in terms of F1 score. Our approach

also performed better accuracy, F1 score, Precision and Recall against Benchmark research.

Visual Performance Analysis of Sub-Models (CM, ROC, and MSE)

Supplementing the aggregate metrics in Table 5, we scrutinized the confusion matrices (CM) and ROC curves for each sub-architecture (BERT+MLP, BERT+LSTM, and BERT+BiLSTM). This granular analysis is necessary for expose the specific class-based discrimination capabilities of the models and to understand the synergy that allows them to complement one another.

A relative review of the matrices in Figure 3-5 shows that each architecture develops a unique proximity for specific sentiment classes. Especially, the BERT+MLP model (Figure 3a) proved most effective for the ‘positive’ category, securing 1428 correct predictions, while correctly identifying 1490 instances in the ‘negative’ class. In contrast, the BERT+BiLSTM model

Method	Accuracy	F1 score	Precision	Recall	MSE
LSTM + Word2Vec (Karaca & Aslan, 2021)	-	75.50%	88.65%	-	-
BERT + CNN (Kumar et al., 2024)	89%	90%	91%	90%	-
Hybrid model (Shahriar & Sarker, 2025)	86%	86%	86%	86%	-
Fuzzy BERT ensemble	91.28%	91.23%	91.38%	91.28%	0.2301

MSE: Mean squared error, LSTM: Long short-term memories, BERT: Bidirectional encoder representations from transformers

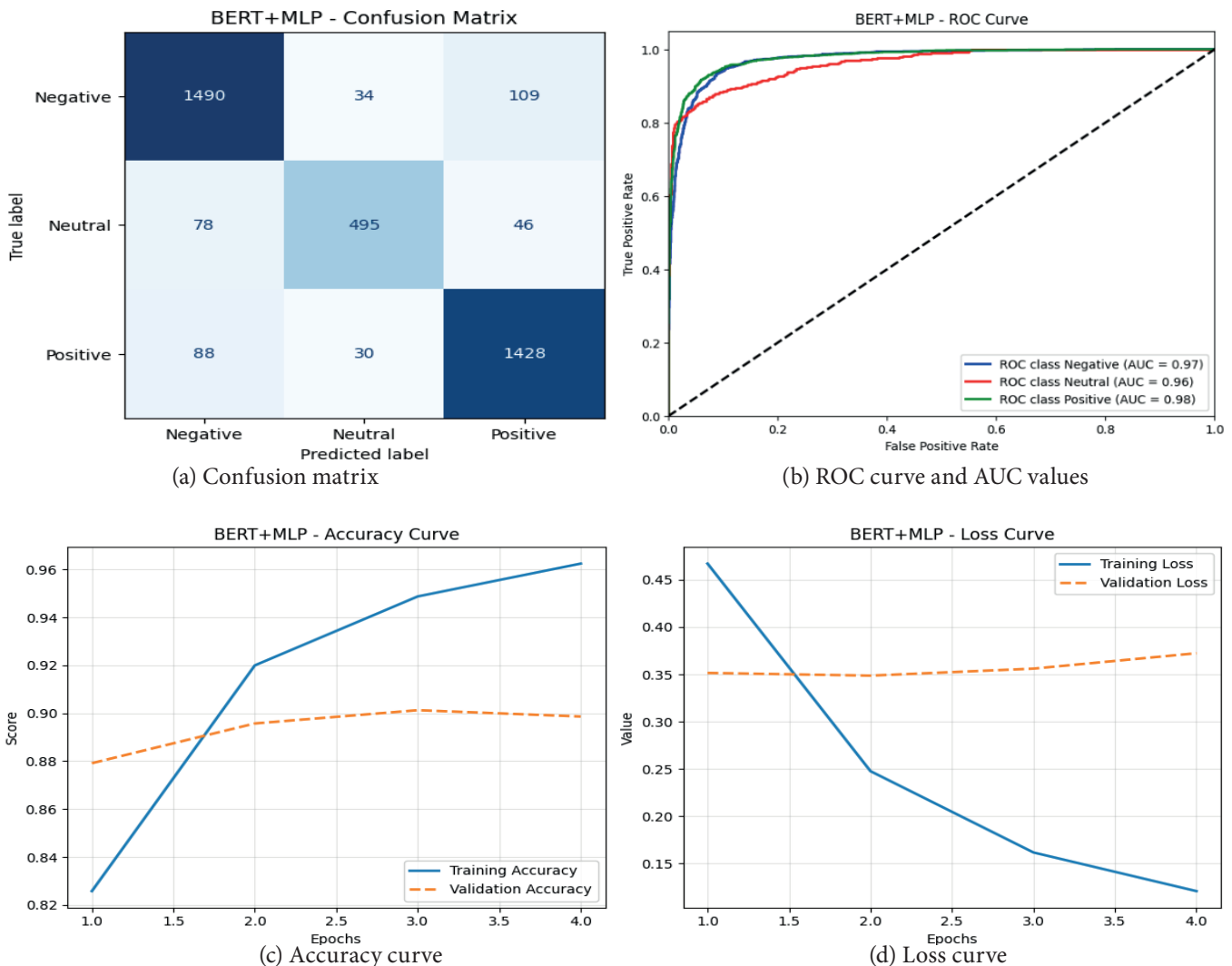


Figure 3. BERT+MLP performance analysis

BERT: Bidirectional encoder representations from transformers, MLP: Multi-layer perceptron, ROC: Receiver operating characteristic, AUC: Area under curve

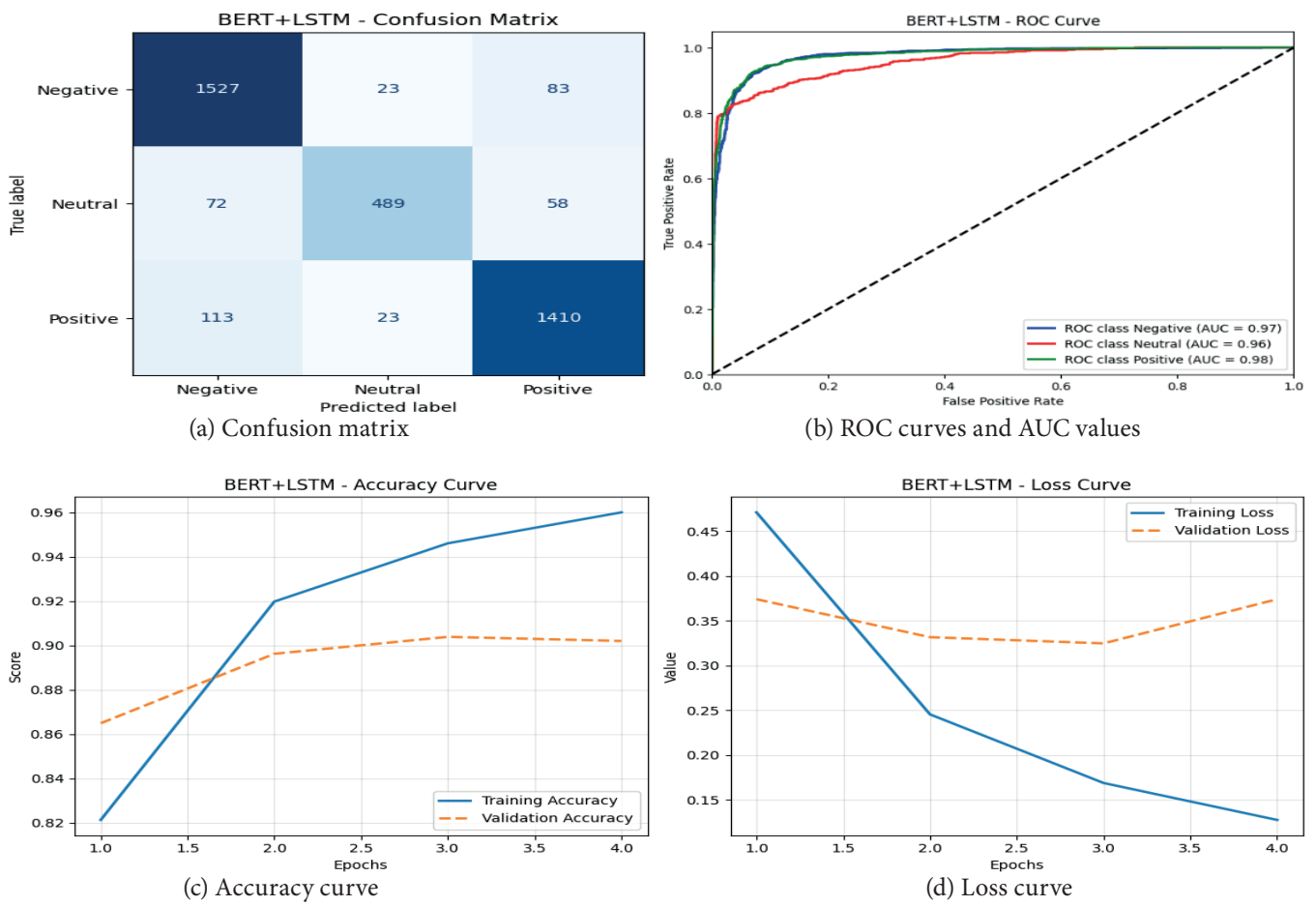


Figure 4. BERT+LSTM performance analysis

BERT: Bidirectional encoder representations from transformers, LSTM: Long short-term memories, AUC: Area under curve, ROC: Receiver operating characteristic

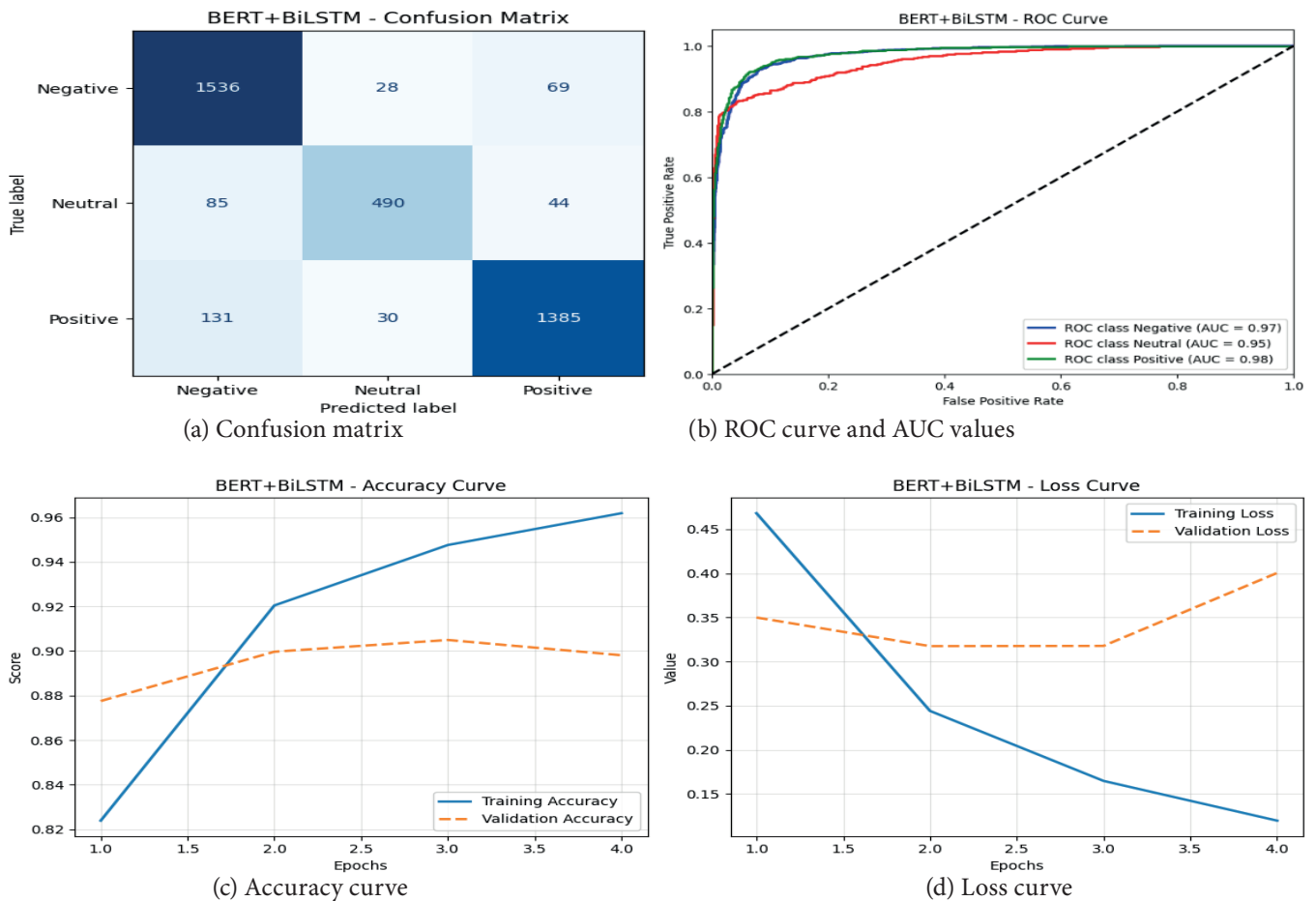


Figure 5. BERT+BiLSTM performance analysis

BERT: Bidirectional encoder representations from transformers, BiLSTM: Bidirectional long short-term memories, ROC: Receiver operating characteristic, AUC: Area under curve

(Figure 5a) exhibits a distinct advantage in identifying ‘negative’ sentiments. With 1536 correct predictions, this architecture demonstrated significantly superior to all competing models in this specific category. However, this same model adopted a more conservative attitude toward the ‘positive’ class, recording 1385 correct predictions covering the MLP variant in this specific regard. In contrast, the BERT+LSTM model (Figure 4a) bridged the gap between these two extremes, exhibiting a balanced performance profile with 1527 correct identifications for negative cases and 1410 for positive ones.

These findings confirm the study’s major hypothesis; though solitary models tend to exhibit blind spot toward specific classes, the ensemble framework succeeds in exploiting this heterogeneity. Moreover, the examination of ROC curves of all sub-model (Figures 3b, 4b and 5b) demonstrated that the area under curve (AUC) continuously decreases in range between 0.95 and 0.98 yet it does not surpass them indicative of state-of-the-art learning ability of fundamental architectures.

Comparative Analysis of Ensemble Methods

To show performance improvement of our model over soft voting method we used McNemar’s statical test. Test shows p-value is 0.1094. Result is slight above standard significance threshold of $\alpha=0.05$.

Figures 6-8 help direct relative analysis between the proposed fuzzy logic framework and traditional ensemble techniques like hard and soft voting. These visuals clearly explain the performance advantages of our approach over standard methods.

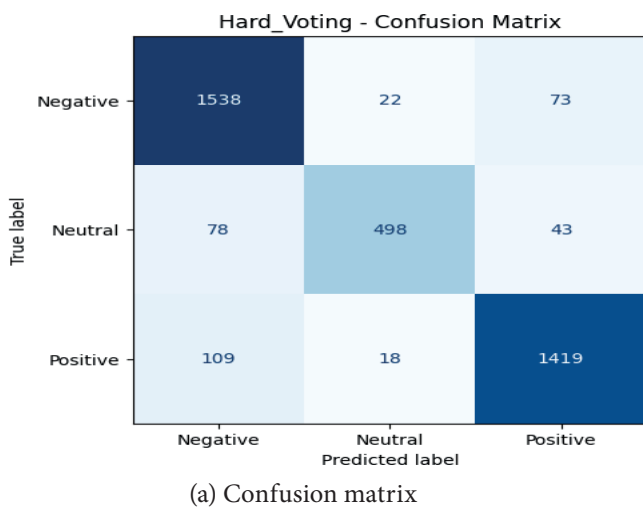


Figure 6. Hard voting

Figure 6 shows that hard voting has 1538 correct prediction for ‘negative’ class and 1419 for ‘positive’. Significantly, because hard voting disregards confidence scores, it introduces hardness that fails to accommodate borderline samples situated along critical decision boundaries.

Soft voting performed slightly better than hard voting while raising positive prediction count to 1418. Nevertheless, our proposed fuzzy ensemble model (Figure 8a) performed better than others with 1420 positive, 1543 negative and 504 neutral correct predictions. It is look like one digit increase but +1 gain serves as a quantitative proof. This represents irony tweet previously misclassified by Soft Voting that our Fuzzy Logic model rules successfully classified it on their correct category.

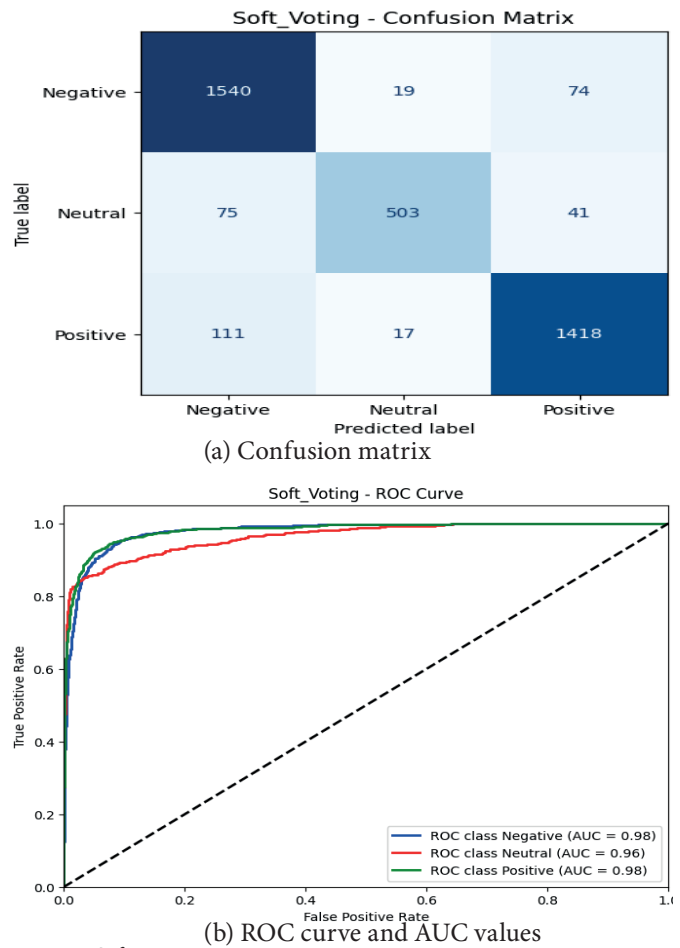


Figure 7. Soft voting

ROC: Reciever operating characteristic, AUC: Area under curve

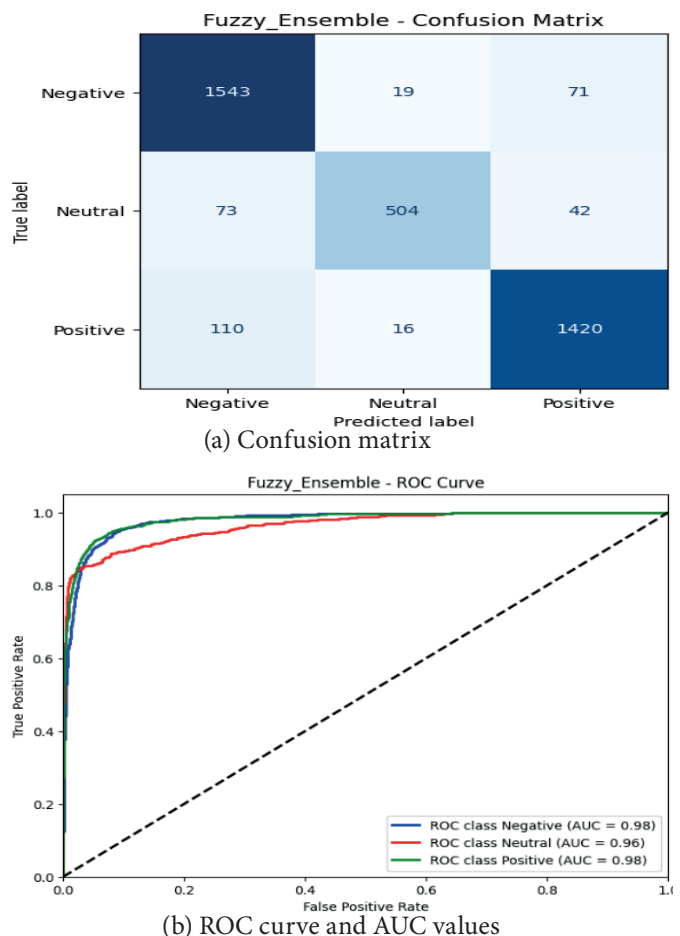


Figure 8. Proposed fuzzy ensemble

ROC: Reciever operating characteristic, AUC: Area under curve

In the visual analysis, the ROC curve for the fuzzy ensemble (Figure 8b) confirms the system's superiority. It delivers peak accuracy together with unmatched classification robustness, a claim validated by AUC values that consistently reach 0.98 across all categories.

DISCUSSION

Comparing with current literature our model shows better advantage. Especially our model surpassed 0.7550 F1 score reported in the reference study for their LSTM model (Karaca & Aslan, 2021) with reaching 0.9123. This performance leap is largely attributed to our dynamic weighting mechanism which effectively lessens 'contextual uncertainty' issue mostly cited in previous search (Alam, 2025). Additionally, although high performance was reported in a study proposing a hybrid RoBERTa-BiLSTM-CNN structure (Yang et al., 2025), the model in question does not include a weighting mechanism based on error rate (MSE) as in our research.

The proposed ensemble architecture aligns with the necessity of "managing uncertainty and ambiguity in natural language," which is emphasized in previous studies (Ming et al., 2024), (Alam, 2025). When more detailed comparison conducted, proposed Fuzzy logic-based model shows better performance from their contemporary rivals for COVID-19 tweets. For instance, while an accuracy rate of 89.00% was achieved with the proposed BERT-CNN structure in a study using a limited time range (March 16-31) of the same data source (Kumar et al., 2024), another study conducted on same dataset remained at the level of 86% while using hybrid model (Shahriar & Sarker, 2025). This indicates that, instead of static weighting, fuzzy logic rules that dynamically process the confidence and past performance values of each sample are more successful in managing the diversity and uncertainty in the dataset. While approaches in the literature such as TexShape rely on weighted linear combinations to balance data processing objectives (Kale et al., 2024), the model proposed in this study has unwearied this balance through linguistic rules based on expert opinion. To cope with the noise inherent and ambiguity available in social media data (Nandi et al., 2025), this ensemble model proposed in the study conducted in 2025 offers a more versatile decision mechanism in ambiguous situations, as opposed to traditional methods.

Traditional soft voting methods are vulnerable to indiscriminate reliance when it comes to model confidence. To solve this problem, the Mamdani method is used as a reliability filter since its configuration provides historical accuracy. Within the protocol, if the model's past performance is not sufficient, even a high-confidence prediction is amerced via the Rule Base. This mechanism enhances noise resistance. Consequently, the observed rise in precision (to 0.9138) stems from the fuzzy logic functioning as a 'safety valve,' which systematically removes false positives.

Conventional type-1 fuzzy logic mostly relies on basic 3-level member structure, but our research enlarges this to 7-level fine-grained frameworks. This increased granularity serves to highly improve the correctional power of the model. To validate this advantage over a plain Soft Voting and to bring out context dependency that raw numbers might miss, we focused on disagreements between respective outputs in cases where the two-voting mechanism were a odd. In this case, our study reveals that the Fuzzy Logic attempted to solve failures when text was particularly highly complex, for example, irony

or slang. Within this framework, 'ambiguity' points out to the semantic 'gray areas' where sentiment transitions are not clear, so it is not mathematically easy to find a 'crisp' logic to create a definitive label. Furthermore, tweets that are 'hard-to-understand' are defined as examples where high-dimensional noise such as slang, irony, humor or context-dependent keywords create conflicts with the literal meaning of the text. For find a solution, our model operates as a sensitive 'sentiment filter' that figures out these ambiguities by assessing sample-specific uncertainty. As an example, the system was tested against tweets that are usually misinterpreted by the traditional architectures. One of the examples is the test set entry: *'Im at a pioneer supermarket in brooklyn and its nice to see thats some places havent completely lost their mind and bought everything and everything covid.'* This text is prone to misinterpretation by shallow models because keywords like 'covid' and 'lost' often trigger negative associations regarding hardship, obscuring the true supportive sentiment.

Standard soft voting faltered here, mislabeling the text as 'neutral' due to an over-reliance on the pessimistic connotations of specific keywords. The proposed fuzzy ensemble, however, successfully decoded the nuance; it recognized the statement as an empathetic observation rather than a complaint, correctly assigning it to the 'positive' class. This instance underscores the utility of our 7-level granular architecture. It functions as a highly sensitive 'sentiment filter,' specifically excelling in ambiguous, gray-area categories like 'neutral' typically the hardest to pinpoint.

Limitations

Even though this model achieves significant performance, several limitations must be acknowledged. First, model success is mostly dependent on fine-tuning process especially adjusted for COVID-19 vocabulary such as "mask", "quarantine" and "symptom", so this can limit its usage for other sentiment analysis. Also, preprocessing process and bert-base-uncased tokenizer restrict evaluation to English language texts. Our 7 level fuzzy logic module is additional computational burden (22.67 ms per sample) during the defuzzification process. It can be usage restriction with wider dataset and real time streaming. We used 12 precise rules instead of 49 rules. It balanced performance but may not catch all possible nuanced scenario.

CONCLUSION

The obtained accuracy rate of 91.28% is at a competitive level with the results of recent studies. All results when compared to benchmark studies above with 91.28% accuracy and 91.38% precision rate achieved in this study. Thus, the improvement is not merely statistical; it represents a tangible qualitative leap over standard methodologies. Our future work will prioritize collection of far more diverse, real word-datasets to unsure the model can adapt effectively to wide range of different context. Generalizability is the primary objective here. At the same time, we intend to benchmark our current architecture against other powerful transformer variants such as DistilBERT and RoBERTa. Future work also can contain different deep learning models such as CNN, RNN and Gans. Finally moving from theory to practice we plan to deploy this model on live tweet streams while integrating Explainable Artificial Intelligence to ensure every decision remains transparent to the user.

ETHICAL DECLARATIONS

Ethics Committee Approval

Ethical approval was not required for this study as it does not involve human participants, animal subjects, or identifiable personal data.

Peer Review Process

This manuscript was subject to external peer review.

Conflict of Interest

The authors declare no conflicts of interest related to this study.

Financial Disclosure

The authors received no financial support for the conduct or publication of this research.

Author Contributions

Concept: MSE, ST; Design: MSE, ST; Control: MSE, ST; Resources: MSE, ST; Materials: MSE, ST; Data Collection and/or Processing: MSE, ST; Analysis and/or Interpretation: MSE, ST; Literature Review: MSE, ST; Writing the Article: MSE, ST; Critical Review: MSE, ST.

REFERENCES

- Airlangga, G. (2024). Spam detection in YouTube comments using deep learning models: a comparative study of MLP, CNN, LSTM, BiLSTM, GRU, and attention mechanisms. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, 4(4), 1533-1538.
- Alam, M. S., Mrida, M. S. H., & Rahman, M. A. (2025). Sentiment analysis in social media: how data science impacts public opinion knowledge integrates natural language processing (NLP) with artificial intelligence (AI). *American Journal of Scholarly Research and Innovation*, 4(1), 63-100.
- Anwar, Z., Afzal, H., Altaf, N., Kadry, S., & Kim, J. (2024). Fuzzy ensemble of fine-tuned BERT models for domain-specific sentiment analysis of software engineering dataset. *PLOS ONE*, 19(5), e0300279. <https://doi.org/10.1371/journal.pone.0300279>
- Arslan, S., Orman, Z., & Akan, A. (2021). A novel fuzzy logic-based text classification method for tracking rare events on Twitter. *IEEE Access*, 9, 36915-36929. <https://doi.org/10.1109/ACCESS.2021.3062345>
- Bashar, M. K., Monjur, O., Islam, S., Shams, M. G., & Quader, N. (2025). Exploring synergistic ensemble learning: uniting CNNs, MLP-mixers, and vision transformers to enhance image classification. *arXiv*. <https://arxiv.org/abs/2504.09076>
- Bello, A., Ng, S. C., & Leung, M. F. (2023). A BERT framework for sentiment analysis of tweets. *Sensors*, 23(1), 506. <https://doi.org/10.3390/s23010506>
- Bilal, A. A., Erdem, O. A., & Toklu, S. (2023). Applying sentiment analysis on children's stories. *Gazi University Journal of Science Part A: Engineering and Innovation*, 10(1), 71-84.
- Bilal, A. A., Erdem, O. A., & Toklu, S. (2024). Children's sentiment analysis from texts by using weight updated tuned with random forest classification. *IEEE Access*, 12, 70103-70116. <https://doi.org/10.1109/ACCESS.2024.3400992>
- Cai, R., Qin, B., Chen, Y., Zhang, L., Yang, R., Chen, S., & Wang, W. (2020). Sentiment analysis about investors and consumers in energy market based on BERT-BiLSTM. *IEEE Access*, 8, 171408-171416. <https://doi.org/10.1109/ACCESS.2020.3024417>
- Chen, X., Cong, P., & Lv, S. (2022). A long-text classification method of Chinese news based on BERT and CNN. *IEEE Access*, 10, 34094-34105. <https://doi.org/10.1109/ACCESS.2022.3161545>
- Dhanalakshmi, P., Reddy, U. J., Ravikanth, G., Samathoti, P., & Ramu, G. (2024). COVID-19 Twitter data analysis using LSTM and BERT techniques. *International Journal of Engineering Trends and Technology*, 72(1), 219-228.
- Elgabry, M., & Hamdi, A. (2025). Confidence-credibility aware weighted ensembles of small LLMs outperform large LLMs in emotion detection. *arXiv*. <https://arxiv.org/abs/2512.17630>
- Garcia-Plaza, A. P., Fresno, V., & Martinez, R. (2008). Web page clustering using a fuzzy logic based representation and self-organizing maps. In *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*.
- Haroon, N. H., Abdulsada, Z. R., Sharif, H., Ahmed, W. S., Saleem, M., & Jawad, I. A. (2023). Social media analysis using fuzzy natural language processing with an extension of semantic queries. In *Proceedings of AICERA/ICIS*. <https://doi.org/10.1109/AICERA59538.2023.10420086>
- Howells, K., & Ertugan, A. (2017). Applying fuzzy logic for sentiment analysis of social media network data in marketing. *Procedia Computer Science*, 120, 664-670. <https://doi.org/10.1016/j.procs.2017.11.293>
- Jang, J. S. R., Sun, C. T., & Mizutani, E. (1997). *Neuro-fuzzy and soft computing: a computational approach to learning and machine intelligence*. Prentice Hall.
- Kale, K., Esfahanizadeh, H., Elias, N., Baser, O., Médard, M., & Vishwanath, S. (2024). TexShape: information theoretic sentence embedding for language models. *arXiv*. <https://arxiv.org/abs/2402.05132>
- Karaca, Y. E., & Aslan, S. (2021). Sentiment analysis of COVID-19 tweets using LSTM. *Journal of Computer Science (IDAP Special Issue)*, 366-374.
- Kumar, G., Agrawal, R., Sharma, K., Gundalwar, P. R., Kazi, A., Agrawal, P., ..., & Salagrama, S. (2024). Combining BERT and CNN for sentiment analysis: a case study on COVID-19. *IJACSA*, 15(10), 676-686.
- Liu, M., Zhang, H., Xu, Z., & Ding, K. (2024). The fusion of fuzzy theories and natural language processing: a state-of-the-art survey. *Applied Soft Computing*, 162, 111818. <https://doi.org/10.1016/j.asoc.2024.111818>
- Mendel, J. M. (1995). Fuzzy logic systems for engineering: a tutorial. *Proceedings of the IEEE*, 83(3), 345-377. <https://doi.org/10.1109/5.364486>
- Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning-based text classification: a comprehensive review. *ACM Computing Surveys*, 54(3), 1-40. <https://doi.org/10.1145/3439726>
- Nandi, S., Subrahmanyam, P., & Singh, S. (2025). Fuzzy-based ensemble learning for sentiment analysis in social media data. *Global Journal of Engineering Innovations & Interdisciplinary Research*, 5(1), 1-8.
- Rahman, M. M., Shiplu, A. I., Watanobe, Y., & Alam, M. A. (2025). RoBERTa-BiLSTM: a context-aware hybrid model for sentiment analysis. *arXiv*. <https://arxiv.org/abs/2406.00367>
- Singh, C., Imam, T., Wibowo, S., & Grandhi, S. (2022). A deep learning approach for sentiment analysis of COVID-19 reviews. *Applied Sciences*, 12(8), 3709. <https://doi.org/10.3390/app12083709>
- Seth, T., & Muhuri, P. K. (2024). Enriching word embeddings with fuzzy systems for NLP tasks. In *IEEE FUZZ Conference*. <https://doi.org/10.1109/FUZZ-IEEE60900.2024.10611949>
- Sherin, A., Lokesh, S., Deepa, S. N., & Jeya, I. J. S. (2025). Fusion of deep recurrent neural network models and fuzzy decision support system for tweet sentiment analysis. *Automatika*, 66(4), 28-50. <https://doi.org/10.1080/00051144.2025.XXXXXX>
- Xu, C., & Kechadi, M.-T. (2024). An enhanced fake news detection system with fuzzy deep learning. *IEEE Access*, 12, 88009-88022. <https://doi.org/10.1109/ACCESS.2024.3418340>
- Yang, L., Wang, J., & Qiu, W. (2025). RoBERTa-based multi-feature integrated BiLSTM and CNN model for ceramic review analysis. *IEEE Access*, 13, 103681-103692.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv*. <https://arxiv.org/abs/1810.04805>
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ..., & Stoyanov, V. (2019). RoBERTa: a robustly optimized BERT pretraining approach. *arXiv*. <https://arxiv.org/abs/1907.11692>
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11), 2673-2681. <https://doi.org/10.1109/78.650093>
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533-536. <https://doi.org/10.1038/323533a0>