Emotion Detection from Short Text using Traditional and Deep Learning Models: A Comparative NLP Approach

1. Introduction

Emotion plays a critical role in human communication, shaping perception and decision-making (Zanwar et al., 2022; Plaza-del-Arco et al., 2024). In NLP, emotion detection aims to infer emotional states such as *joy*, *sadness*, *anger*, or *fear* from text. Its growing relevance spans applications in mental health, conversational AI, and social media analysis, where understanding emotional cues enhances personalization (Plaza-del-Arco et al., 2024; Gafar et al., 2023).

Unlike sentiment analysis, which simplifies text into polarity categories, emotion classification tackles fine-grained emotional states—posing greater difficulty due to ambiguity, overlap, and contextual subtlety (S and Geetha, 2024; Tulika Chutia and Nomi Baruah, 2024).

Traditional models like MNB and SVM perform well on sparse TF-IDF features (Gafar et al., 2023), but lack contextual depth (Zanwar et al., 2022). Deep models such as BiLSTM and BERT address this through sequence modelling and self-attention mechanisms, capturing richer semantic dependencies (Zanwar et al., 2022).

However, challenges persist—datasets often suffer from class imbalance (Plaza-del-Arco et al., 2024), and comparative studies integrating tailored preprocessing for both traditional and deep models remain scarce (Zanwar et al., 2022). This project addresses these gaps by developing a full NLP pipeline using the Emotions Dataset (Praveen, 2020), applying model-specific preprocessing, and evaluating performance on both the full 6-class and filtered 4-class subsets to assess generalizability and class balance effects.

2. Objective

This project aims to build a multi-class text classification system to detect one of six emotional states from short-form English text. The specific objectives are to:

- Achieve ≥89% test accuracy and a macro-averaged F1-score of ≥0.88
- Compare the performance of four models: Multinomial Naive Bayes, Support Vector Machines, BiLSTM, and fine-tuned BERT
- Assess the impact of model-specific preprocessing pipelines and class imbalance on performance

- Complete the full pipeline—from data exploration to model evaluation—within 6 weeks using Python, Scikit-learn, PyTorch, and Hugging Face Transformers

By evaluating traditional and deep learning approaches within a unified framework, this study contributes to the ongoing advancement of emotion detection in text—a rapidly growing area in applied NLP research.

3. Literature Review

Emotion detection has become a rapidly evolving subfield of natural language processing (NLP), as systems strive to interpret nuanced human emotions in domains such as social media, conversational agents, and mental health support. Classifying discrete emotional states from text presents persistent challenges due to semantic ambiguity, contextual sensitivity, and the inherently subjective nature of emotional expression. This review examines key studies published between 2019 and 2024 to position the current research within the broader discourse, identify dominant methodologies, evaluate key contributions, and expose gaps that inform this project's direction.

3.1 Evolving Methodologies in Emotion Detection

Tulika Chutia and Nomi Baruah (2024) review over 330 studies and highlight a methodological shift from traditional classifiers (e.g., Naive Bayes, SVM) to advanced neural architectures such as CNNs, BiLSTMs, and Transformers. Deep learning models consistently outperform traditional methods in contextual understanding and macro-F1 scores, though baseline models like Naive Bayes remain competitive on sparse, TF-IDF-based inputs due to their efficiency. The authors also critique the widespread reliance on Ekman's six basic emotions, arguing that such taxonomies oversimplify affective expression. They advocate for integrating lexical, semantic, and syntactic features—a strategy adopted in this project through architecture-specific preprocessing pipelines.

3.2 Traditional Models Still Matter: Insights from Targeted Classifiers

Gafar et al. (2023) assess traditional models for detecting the underexplored emotion *guilt*, using a binarized dataset compiled from multiple corpora. Their comparison of MNB, SVM, and Logistic Regression against CNN and BiLSTM baselines revealed that MNB achieved the highest F1 score (0.72), outperforming the more complex BiLSTM model. These findings highlight the potential of traditional models with engineered feature sets (e.g., TF-IDF, BoW), especially in low-resource or domain-specific contexts. The study also emphasizes the interpretability and generalizability of traditional classifiers, supporting their role as robust baselines in this research.

3.3 Domain Adaptation and Generalization in Transformer Models

Zanwar et al. (2022) explore domain adaptation in emotion detection by introducing hybrid architectures that combine transformers (BERT, RoBERTa) with BiLSTM layers and 435 psycholinguistic features. Their hybrid models significantly outperform standalone transformers on out-of-domain datasets, achieving improved macro-F1 scores. Through evaluations on eight datasets, the study reveals the susceptibility of pre-trained models to domain shift. These insights justify this project's use of both full and filtered datasets to assess model robustness. The authors exemplify methodological pluralism by blending contextual embeddings with feature engineering—an approach mirrored in this study's experimental design.

3.4 Impact of Preprocessing on Transformer-Based Emotion Detection

Rezapour (2024) compares transformer models—including DistilBERT, ELECTRA, and a Twitter-specific RoBERTa—on the GoEmotions dataset. The Twitter-RoBERTa model achieved the highest accuracy (92%). Crucially, the study finds that common preprocessing steps such as removing punctuation and stop words negatively affect performance, as these elements often carry critical emotional cues. This finding supports the use of preprocessing pipelines that retain such features for transformer-based models, reinforcing this project's decision to apply model-specific preprocessing strategies.

3.5 Theoretical and Practical Gaps in Emotion Analysis: A Meta-Review

Plaza-del-Arco et al. (2024) conduct a meta-review of 154 studies, identifying three systemic gaps in the field: (1) lack of demographic and cultural contextualization in datasets; (2) inconsistent emotion taxonomies and terminology; and (3) overreliance on Ekman's basic emotions, which may not account for culturally nuanced affective states. The authors call for interdisciplinary integration, particularly with psychology, to improve emotion modelling. These critiques inform this study's use of both 6-class and 4-class datasets, allowing for analysis of simplification effects on performance and biases, and supporting the broader aim of developing inclusive, generalizable emotion detection systems.

3.6 Synthesis and Research Justification

Four consistent themes emerge from the literature:

1. Deep learning models—especially transformers—excel at contextual understanding but require careful tuning and data alignment.
2. Traditional models remain effective in sparse, domain-specific, or small-scale settings.
3. Performance improves when feature inputs and preprocessing are tailored to model architecture.

4. Persistent gaps exist in cross-domain generalizability, cultural representation, and taxonomy standardization.

These findings directly motivate this project's comparative evaluation of MNB, SVM, BiLSTM, and BERT across both full and filtered datasets, using architecture-specific preprocessing. This design enables a structured investigation into the trade-offs between model complexity, feature sensitivity, and emotion taxonomy scope—offering a nuanced contribution to ongoing research in emotion classification.

4. Materials and Methods

This study implements a structured experimental pipeline to develop, train, and evaluate four models for emotion detection from text: two traditional classifiers (Multinomial Naive Bayes and Support Vector Machine) and two deep learning models (BiLSTM and BERT). Each model is applied to both the full and a filtered version of the Emotions Dataset for NLP.

4.1 Dataset Description

The dataset, curated by Praveen (2020) and sourced from Kaggle, contains over 20,000 short English sentences labelled with six emotions: *joy*, *sadness*, *anger*, *fear*, *love*, and *surprise*. It is pre-split into training, validation, and test sets. While well-suited for emotion classification tasks due to its clean and concise structure, class imbalance—especially in *love* and *surprise (see Figure 1)*—poses a known challenge (Plaza-del-Arco et al., 2024; Gafar et al., 2023).

To investigate the effects of this imbalance, two dataset configurations are used:

- **Full 6-class dataset:** Contains all emotion labels.
- **Filtered 4-class dataset:** Includes only the four most frequent emotions—*anger*, *fear*, *joy*, and *sadness*—to facilitate balanced evaluation.

Model design choices, preprocessing techniques, and evaluation strategies are informed by established NLP literature and best practices.

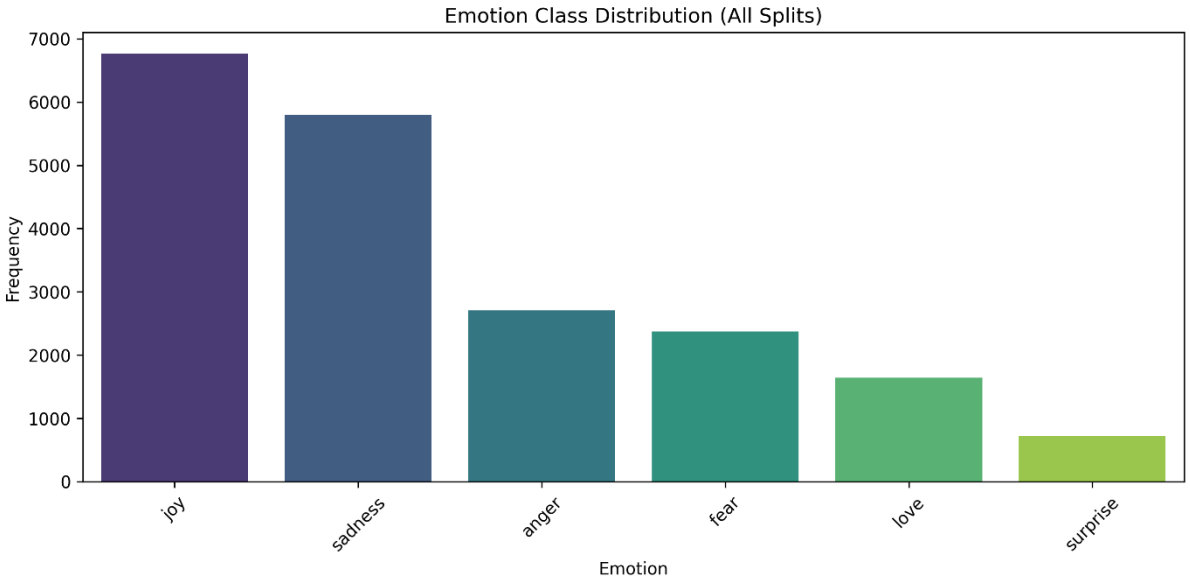

Figure 1- Emotion class frequency distribution in the full 6-class dataset.

This plot shows the imbalance in emotion labels, with 'joy' and 'sadness' dominating the dataset. This motivated both the use of class-weighted metrics and the creation of a 4-class filtered subset to enhance model robustness on underrepresented classes.

4.2 Model Comparison and Selection

This study compares four models—two traditional and two deep learning—selected from a broader set of candidates based on empirical performance and literature guidance.

- Traditional Models:
  From five evaluated classifiers (MNB, SVM, Logistic Regression, Random Forest, KNN), Multinomial Naive Bayes (MNB) and Support Vector Machine (SVM) were selected. MNB was chosen for its efficiency and effectiveness on sparse, TF-IDF features, while SVM was selected for its strong generalization and ability to handle non-linear class boundaries (Tulika Chutia & Nomi Baruah, 2024; Gafar et al., 2023).

- Deep Learning Models:
  Among RNN, LSTM, GRU, BiLSTM, and Transformer-based models, BiLSTM and BERT were selected. BiLSTM captures bidirectional context in sequential data, improving over unidirectional RNNs, while BERT provides state-of-the-art results in sentence-level tasks through deep contextual embeddings and self-attention (Zanwar et al., 2022).

These selections ensure a comprehensive comparison across levels of model complexity, interpretability, and computational cost.

4.3. Preprocessing Pipeline

4.3.1 For Traditional Models (MNB and SVM)

For traditional models, a structured preprocessing pipeline was applied to convert raw text into sparse TF-IDF feature vectors:

- Text was lowercased and cleaned of URLs, numbers, and punctuation.
- Stopwords were filtered using the NLTK library.
- Lemmatization was applied to normalize word forms.
- **Negation marking** was used to preserve emotional polarity (e.g., *not happy → not_happy*), a technique shown to enhance feature discrimination in high-dimensional space (Mukherjee et al., 2021).
- As shown in **Figure 2 (left panel)**, this process reduced text length while preserving semantic content.

TF-IDF vectorization was performed using unigrams and bigrams, capped at 5,000 features with a minimum document frequency (min_df=5) to exclude rare terms.

4.3.2 Preprocessing for Deep Learning Models (BiLSTM and BERT)

Deep learning models were provided with minimally altered, natural sentence structures to preserve contextual integrity:

- Negation marking was omitted, as deep models infer semantic inversion from context.
- Tokenization was model-specific:
    - BiLSTM: Text was tokenized and padded to 64 tokens; vocabulary was built from the training set.
    - BERT: Used the bert-base-uncased tokenizer from Hugging Face's Transformers library, with padding and truncation set to max_length=64.

BERT did not require manual feature engineering due to its pre-trained architecture and capacity to learn semantic, syntactic, and contextual features via self-attention. Input data was formatted into the Hugging Face 'datasets.Dataset' format and converted to PyTorch tensors (input_ids, attention_mask, labels).

As shown in **Figure 2 (right panel),** most text sequences remained under 64 tokens after cleaning, validating the selected padding and truncation strategy.

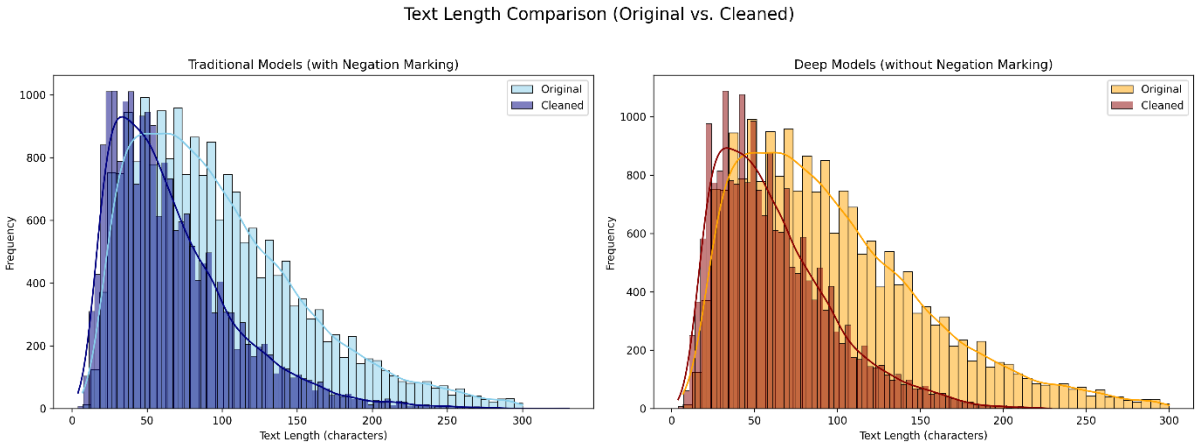

Figure 2 - Text Length Distribution: Traditional Models vs. Deep Models.

This comparison confirms the suitability of a 64-token input length for deep models while demonstrating the relative sparsity of traditional representations.

4.4 Model Implementation

4.4.1 Traditional Models

Traditional classifiers were implemented using Scikit-learn, with TF-IDF feature matrices as input. Two models were used:

- MultinomialNB()
- LinearSVC()

A majority class classifier was established as a baseline for both the full and filtered datasets.

4.4.2 BiLSTM

The BiLSTM model was built using PyTorch, with input sequences embedded using pre-trained GloVe vectors (6B, 100d). The embedding matrix was initialized for the full vocabulary and populated with GloVe vectors where available.
The model architecture included:

- Embedding Layer → BiLSTM (units ∈ {32, 64, 128, 256}) → Dropout → Dense SoftMax Output

4.4.3 BERT

The bert-base-uncased model from Hugging Face Transformers was fine-tuned for emotion classification. A classification head was added and trained using the Trainer API with the AdamW optimizer.

- Inputs were formatted as input_ids, attention_mask, and labels using the Hugging Face datasets.Dataset and converted into PyTorch tensors, consistent with earlier preprocessing steps.

4.5 Hyperparameter Tuning and Optimization

4.5.1 Traditional Models

Hyperparameter tuning for Multinomial Naive Bayes and Support Vector Machine was performed using GridSearchCV with 5-fold stratified cross-validation.

- MNB: alpha ∈ [0.1, 0.5, 1.0]
- LinearSVC: C ∈ [0.01, 0.1, 1, 10]

Models were evaluated using Accuracy and Macro F1-score, ensuring sensitivity to class imbalance.

4.5.2 BiLSTM

- BiLSTM optimization involved grid search over **learning rate**, **dropout**, and **hidden unit size**.
  Early stopping was applied based on validation loss to prevent overfitting and accelerate convergence.
- **6-class parameter grid**:

Table 1 - Hyperparameter Grid for BiLSTM – Full 6-Class Emotion Classification Task

Configuration	Hidden Units	Dropout Rate	Learning Rate	Description
1	64	0.3	0.001	Balanced baseline
2	64	0.5	0.001	Increased regularization
3	128	0.3	0.001	Larger hidden size
4	128	0.5	0.001	Larger + higher dropout
5	256	0.3	0.001	High capacity
6	128	0.3	0.0005	Lower learning rate
7	128	0.5	0.0005	Lower LR + high dropout
8	64	0.3	0.0005	Small model + low LR

Interpretation:

- **Higher hidden unit sizes** (128, 256) tested whether deeper representations improved learning.
- **Dropout values** (0.3 vs. 0.5) assessed the model's sensitivity to overfitting.
- **Learning rates** (0.001 and 0.0005) controlled convergence stability, particularly in configurations with larger hidden layers.

- **4-class parameter grid**:

Table 2 - Hyperparameter Grid for BiLSTM – 4-Class Emotion Subset

Configuration	Hidden Units	Dropout Rate	Learning Rate	Description
1	32	0.3	0.001	Lightweight baseline
2	32	0.3	0.0005	Same with smaller LR
3	64	0.3	0.001	Slightly larger model
4	64	0.3	0.0005	Lower LR
5	64	0.6	0.0005	Aggressive regularization
6	32	0.6	0.0005	Tiny model + strong dropout

Interpretation:

- **Higher dropout values** (e.g., 0.6) were used to mitigate overfitting, which is more pronounced in smaller datasets.
- **Lower learning rates** (0.0005) were tested to enhance training stability in compact network architectures.

Metrics included **Accuracy**, **Macro F1**, and **ROC-AUC**, with results visualized using confusion matrices and ROC curves.

4.5.3 BERT

BERT was fine-tuned on the full 6-class dataset using the following hyperparameter combinations:

Table 3 - Hyperparameter Grid for BERT Fine-Tuning

Configuration	Learning Rate (lr)	Epochs
1	2×10^{-5}	3
2	3×10^{-5}	3
3	5×10^{-5}	4

Interpretation:

- **Learning rates** between *2e-5* and *5e-5*, and **epochs** ranging from *3 to 4*, were chosen to balance convergence speed with generalization.
- The grid aimed to test whether slightly longer training or higher learning rates could yield marginal gains without introducing overfitting.

5. Results

This section reports the performance of four emotion classification models—Multinomial Naive Bayes (MNB), Support Vector Machine (SVM), BiLSTM, and BERT—on both the full six-class and filtered four-class versions of the dataset.

Models were evaluated using accuracy, macro-averaged F1-score, and confusion matrices.
For deep learning models, additional analyses—such as ROC curves and training dynamics—are presented in the following sections to provide deeper insight into model behaviour and generalization.

5.1. Traditional Machine Learning Model Performance

5.1.1 Overview of Accuracy and Macro F1-Score

Table 4 - Performance Metrics of Traditional Models

Model	Dataset	Accuracy	Macro F1-score
Multinomial NB	All 6 Emotions	0.829	0.750
Multinomial NB	Major 4 Emotions	0.892	0.871
SVM	All 6 Emotions	0.872	0.822
SVM	Major 4 Emotions	0.922	0.910

Table 4 summarizes the performance metrics for MNB and SVM on both the full and filtered datasets.

5.1.2 Comparative Visualizations

As shown in **Figure 3** and **Figure 4**, both classifiers demonstrate improved performance on the 4-class subset, highlighting the challenges of modelling imbalanced classes in the full dataset.

Accuracy Comparison Across MNB and SVM on Both Datasets

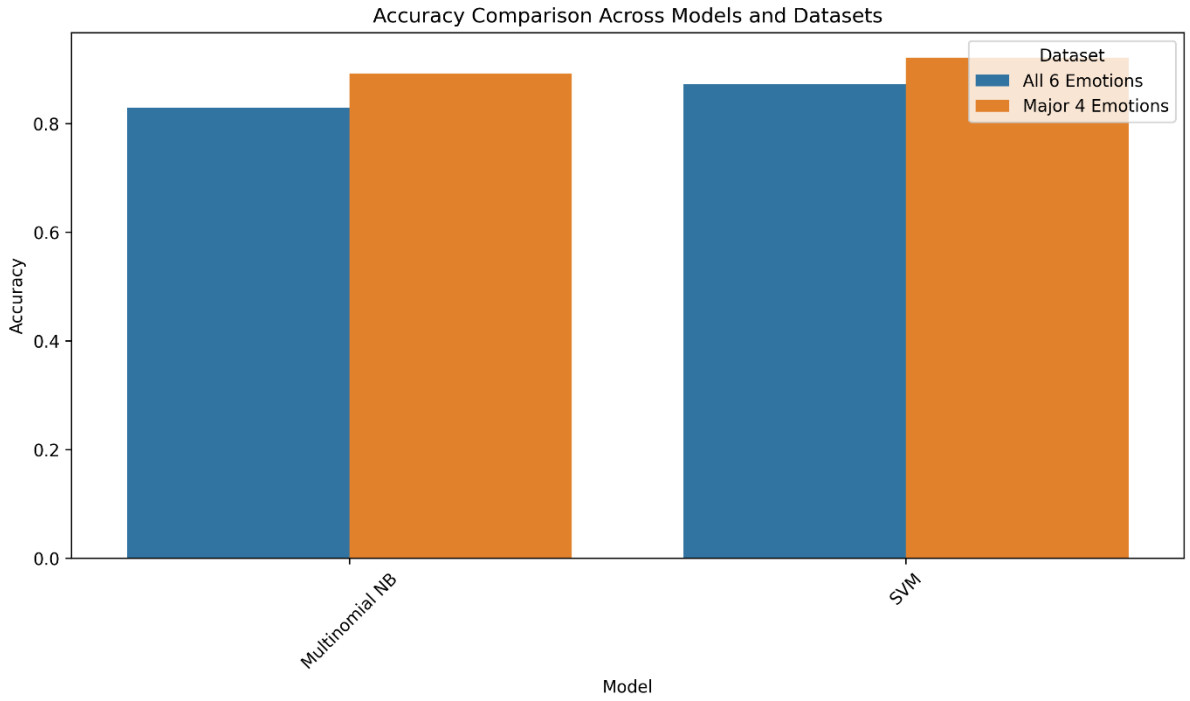

Figure 3- Accuracy comparison for MNB and SVM across the full 6-class and filtered 4-class datasets. SVM consistently outperformed MNB, especially in the filtered setting.

Macro F1-Score Comparison Across MNB and SVM

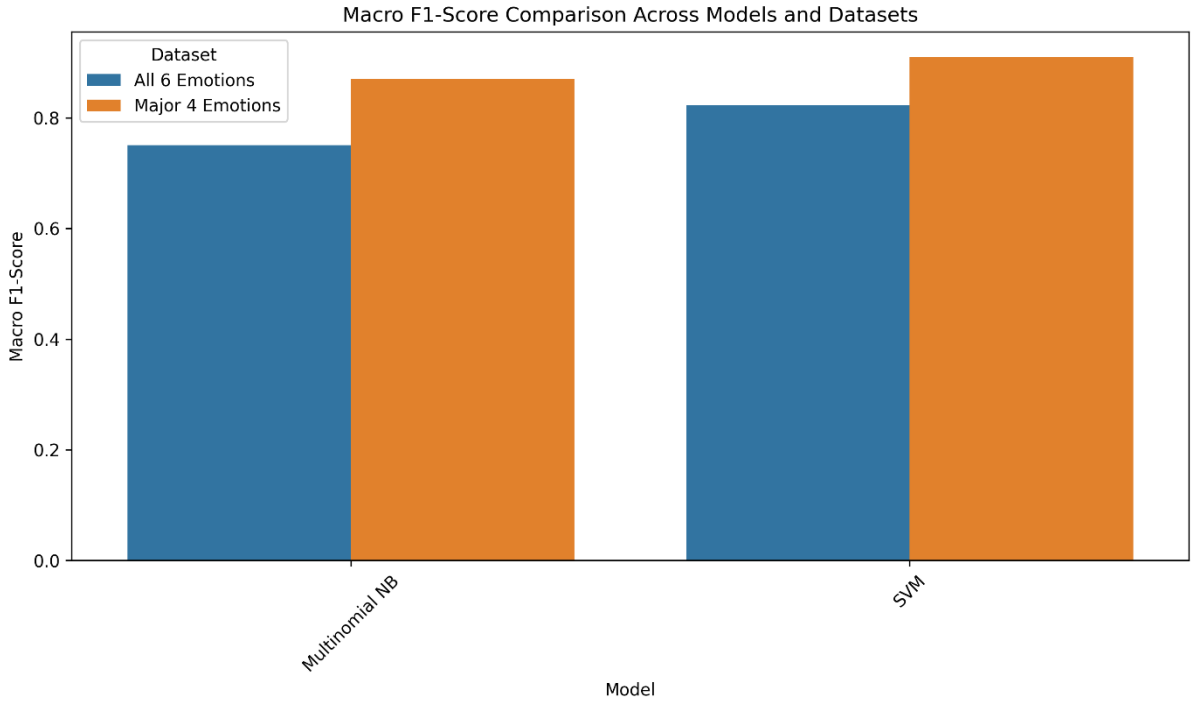

Figure 4- Macro F1-score comparison showing performance improvements for both models on the more balanced 4-class dataset.

5.1.3 Confusion Matrices

Confusion matrices offer insight into the distribution of correct and incorrect predictions for each emotion class. Both models show strong performance in identifying *joy* and *sadness*, with more misclassifications observed in *love* and *surprise*.

MNB

- **Figure 5** shows the confusion matrix for MNB on the full dataset.
- **Figure 6** displays the results on the filtered 4-class subset.

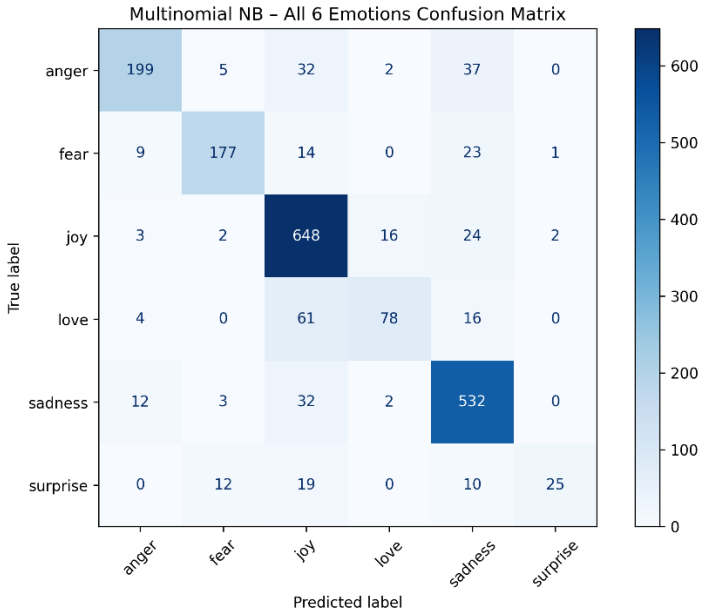

Figure 5-Confusion matrix for Multinomial NB on all six emotions.

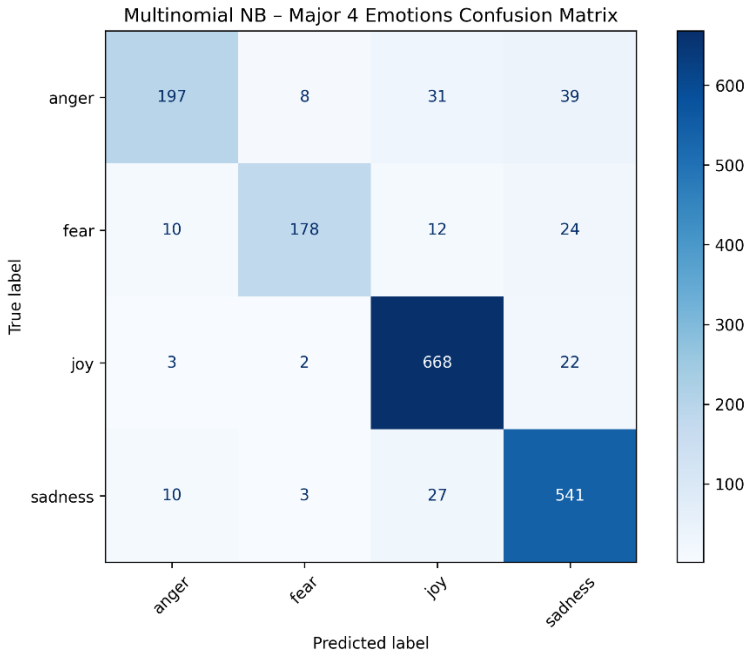

Figure 6-Confusion matrix for Multinomial NB on major four emotions.

SVM

- **Figure 7** shows the confusion matrix for SVM on the full dataset.
- **Figure 8** presents result for the 4-class subset.

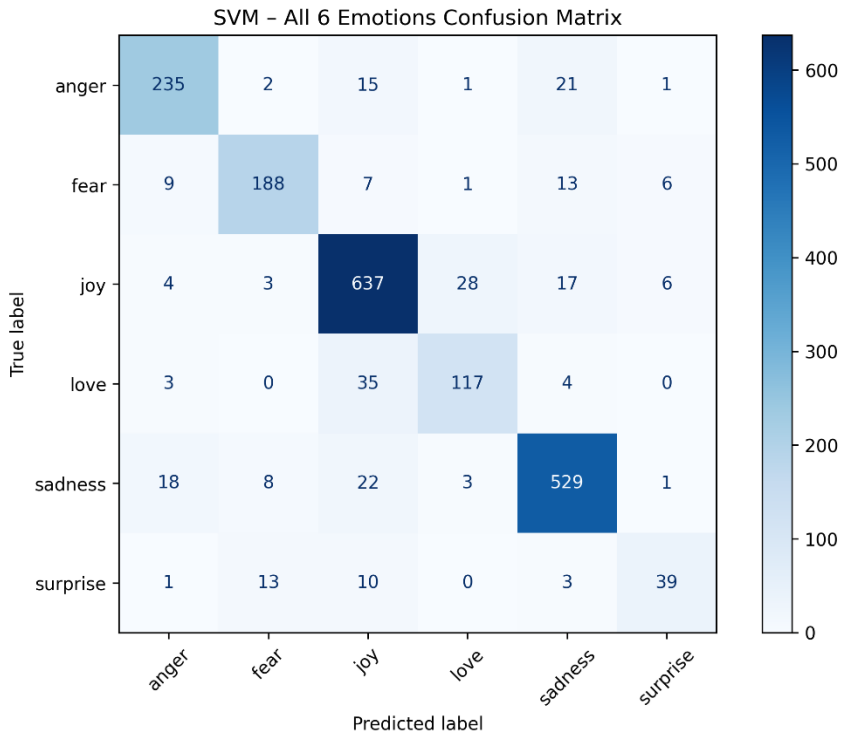

Figure 7- Confusion matrix for SVM on all six emotions.

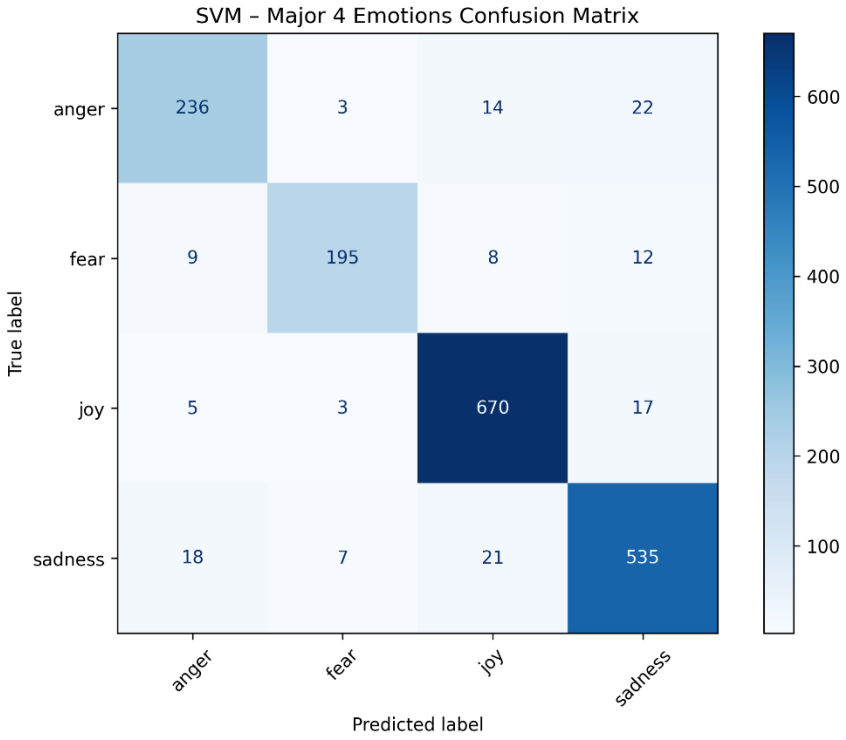

Figure 8- Confusion matrix for SVM on major four emotions.

5.1.4 Key Observations from Traditional Models

- SVM consistently outperformed MNB on both accuracy and F1 across datasets.
- All models performed better on the **4-class dataset**, likely due to improved class balance.
- Both classifiers frequently confused *love* and *joy* and struggled with the *surprise* class.
- SVM showed the highest macro F1 of **0.910** on the 4-class task, demonstrating robustness to noise and inter-class similarity.

5.2. Deep Learning Model Performance — BiLSTM

5.2.1 Overview of BiLSTM Performance

Table 5 - BiLSTM Performance Summary on Both Datasets

Dataset	Hidden Units	Dropout	Learning Rate	Accuracy	Macro F1	Loss	ROC-AUC
All 6 Emotions	64	0.3	0.001	0.909	0.8642	0.2400	0.9942
Major 4 Emotions	32	0.6	0.0005	0.9555	0.9447	0.2043	0.9937

The BiLSTM model, optimized through structured hyperparameter tuning, demonstrated strong performance on both the 6-class and 4-class datasets. As shown in **Table 2**, results were particularly high on the 4-class task, with the best configurations yielding robust accuracy and macro F1-scores across settings.

5.2.2 Confusion Matrices

The confusion matrices in **Figures 9 and 10** detail the class-wise prediction results. The model exhibited strong discriminatory ability between *joy*, *sadness*, and *anger*, with lower confusion compared to traditional models.

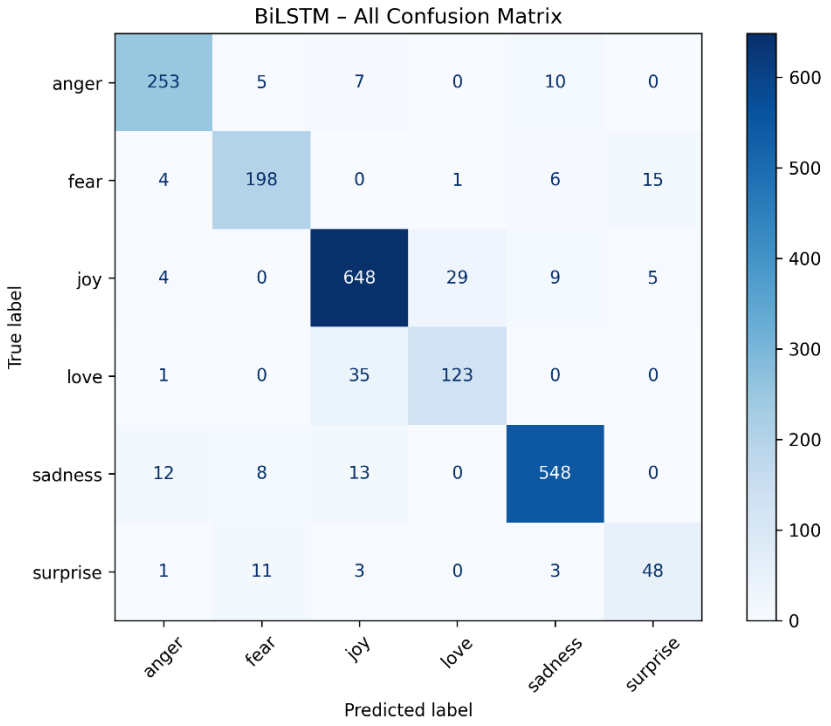

Figure 9-Confusion matrix for BiLSTM model on the 6-class dataset.

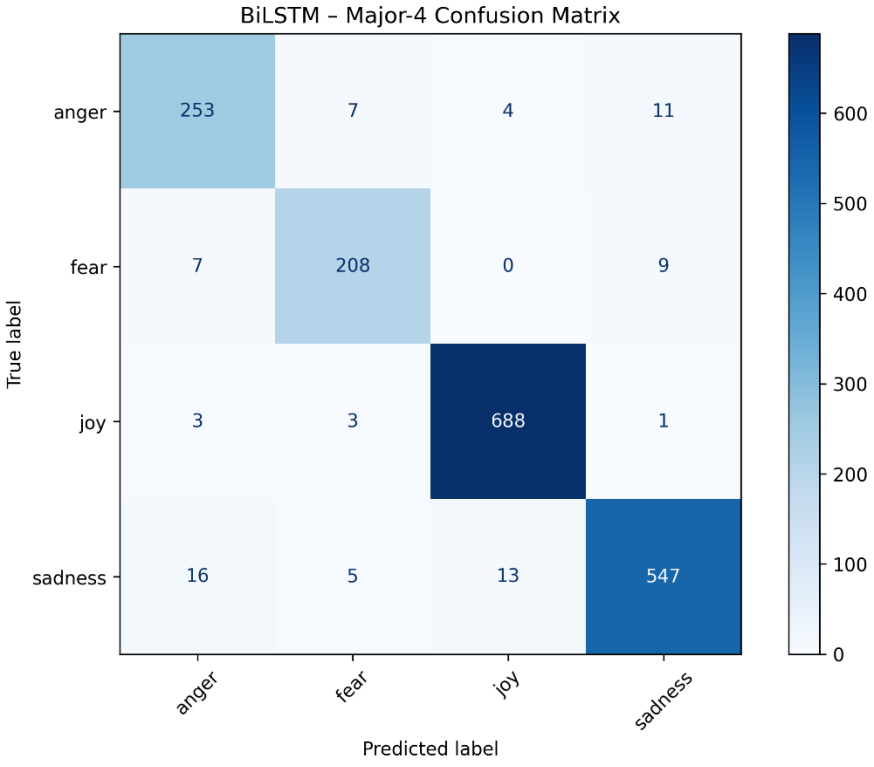

Figure 10- Confusion matrix for BiLSTM model on the major 4-class dataset.

5.2.3 ROC Curve Analysis

ROC curves for each emotion class are plotted in Figures 11 and 12. For both datasets, the Area Under the Curve (AUC) values were exceptionally high across all classes (≥ 0.99).

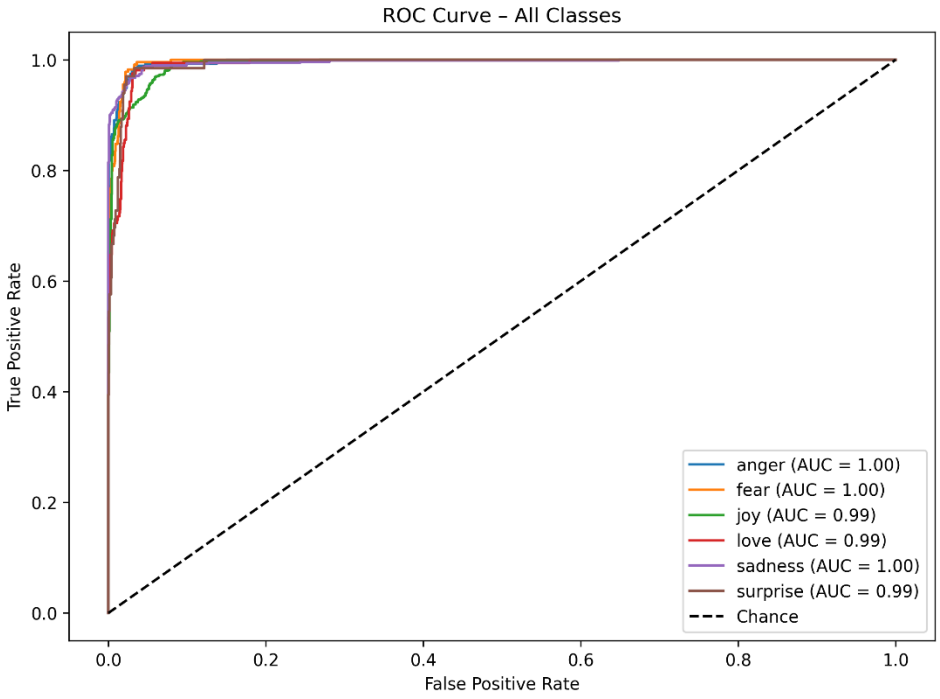

Figure 11- ROC curve for BiLSTM model on the 6-class dataset.

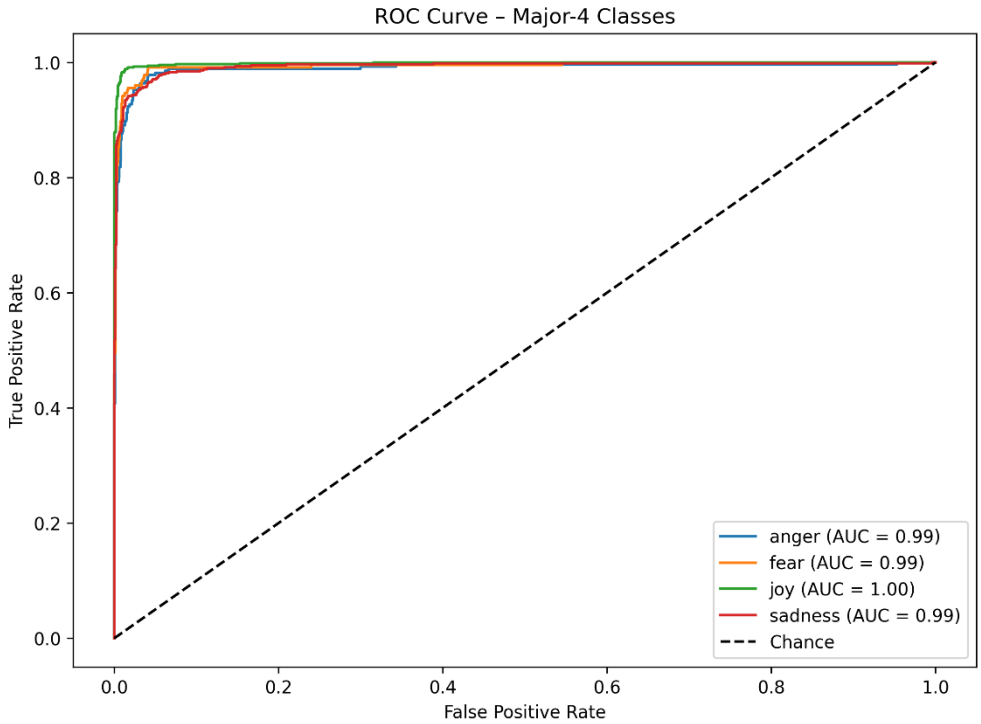

Figure 12- ROC curve for BiLSTM model on the 4-class dataset.

5.2.4 Training Curves

Figures 13–16 show the training and validation accuracy and loss trends across epochs. The model converged smoothly, with no evidence of overfitting.

Training Curves – All 6 Emotions

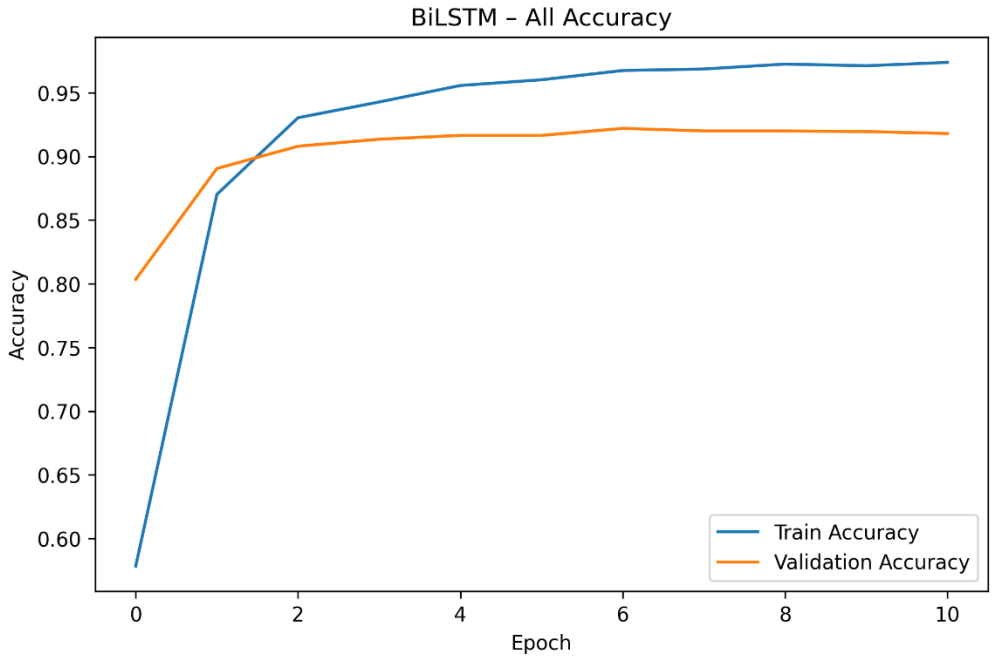

Figure 13- BiLSTM training vs validation accuracy on the 6-class dataset.

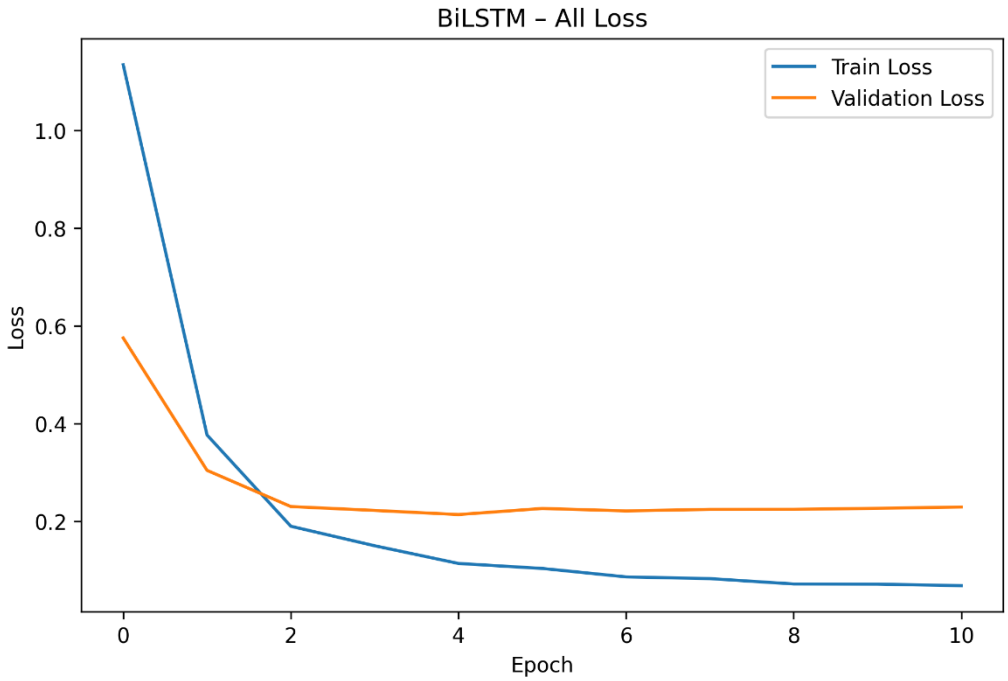

Figure 14- BiLSTM training vs validation loss on the 6-class dataset.

Training Curves – Major 4 Emotions

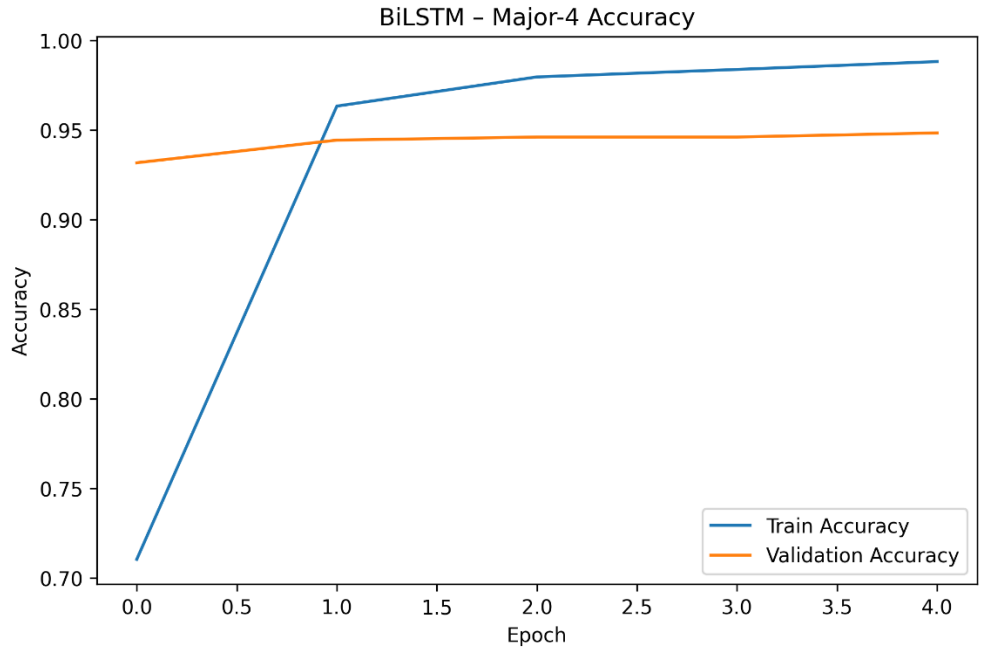

Figure 15- BiLSTM training vs validation accuracy on the 4-class dataset.

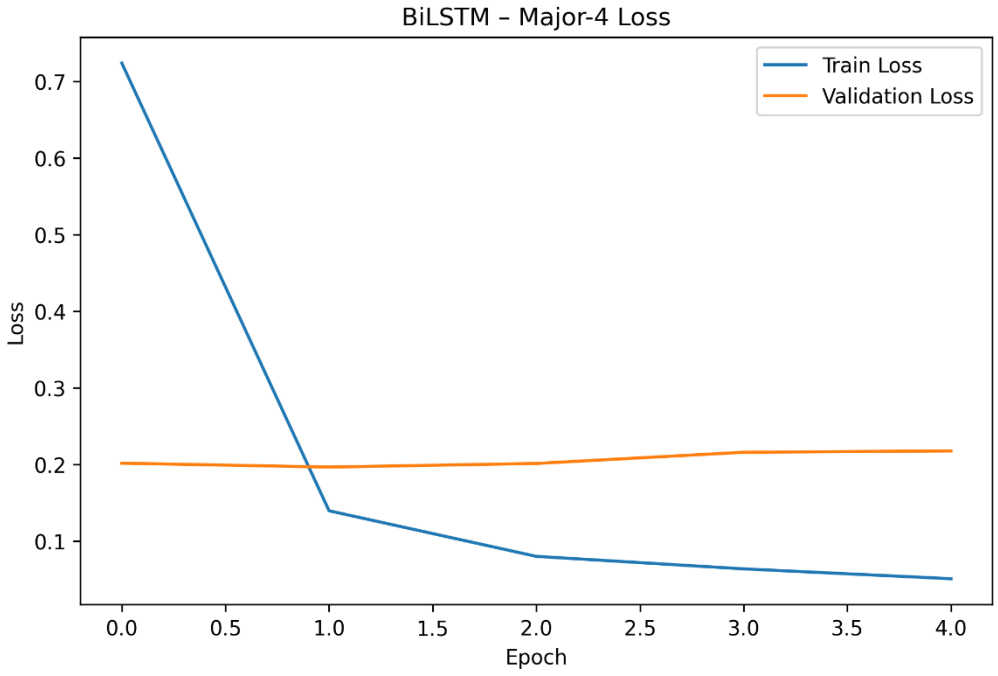

Figure 16- BiLSTM training vs validation loss on the 4-class dataset.

5.2.5 Key Observations from BiLSTM Results

The BiLSTM model achieved **peak performance on the 4-class dataset**, with **95.55% accuracy**, **94.47% macro F1**, and an exceptional **ROC-AUC of 0.9937**. Confusion matrices and ROC curves indicated **strong class separation**, while

training curves showed **rapid convergence and no overfitting**, confirming the model's robustness.

5.3. Deep Learning Model Performance — BERT

5.3.1 Overview of BERT Performance

Table 6 - BERT Performance Summary (Best Hyperparameter Configuration)

Run Label	Learning Rate	Epochs	Accuracy	Macro F1	Loss	ROC-AUC
bert_lr2e-05_ep3	**2e-5**	**3**	**0.9350**	**0.9102**	**0.1688**	**≥ 0.99**

BERT achieved **state-of-the-art performance** among all models tested, with a best configuration (2e-5 learning rate, 3 epochs) yielding **93.5% accuracy**, **0.9102 macro F1-score**, and **ROC-AUC > 0.99** across all classes (see **Table 6**). These results confirm the effectiveness of **transformer-based architectures** for nuanced emotion classification.

5.3.2 Confusion Matrix

As shown in **Figure 17**, BERT achieved **high precision and recall across all emotion classes**, with **minimal misclassification**. Notably, it demonstrated improved distinction between *love* and *joy* compared to both BiLSTM and traditional models, underscoring its superior contextual understanding.

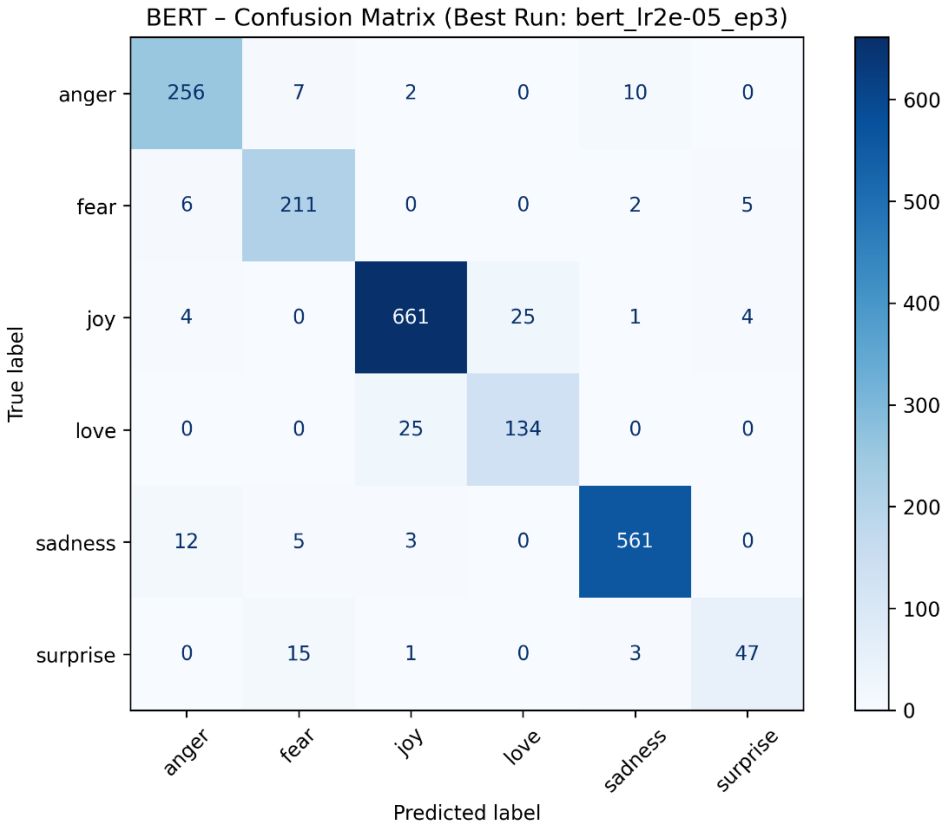

Figure 17- Confusion matrix for BERT on the 6-class emotion dataset (best run: lr=2e-5, epochs=3).

5.3.3 ROC Curve

BERT's per-class ROC curves are displayed in **Figure 18**, revealing nearly perfect discrimination across all classes. AUC values for five of the six emotions reached **1.00**, with *surprise* slightly lower at **0.99**.

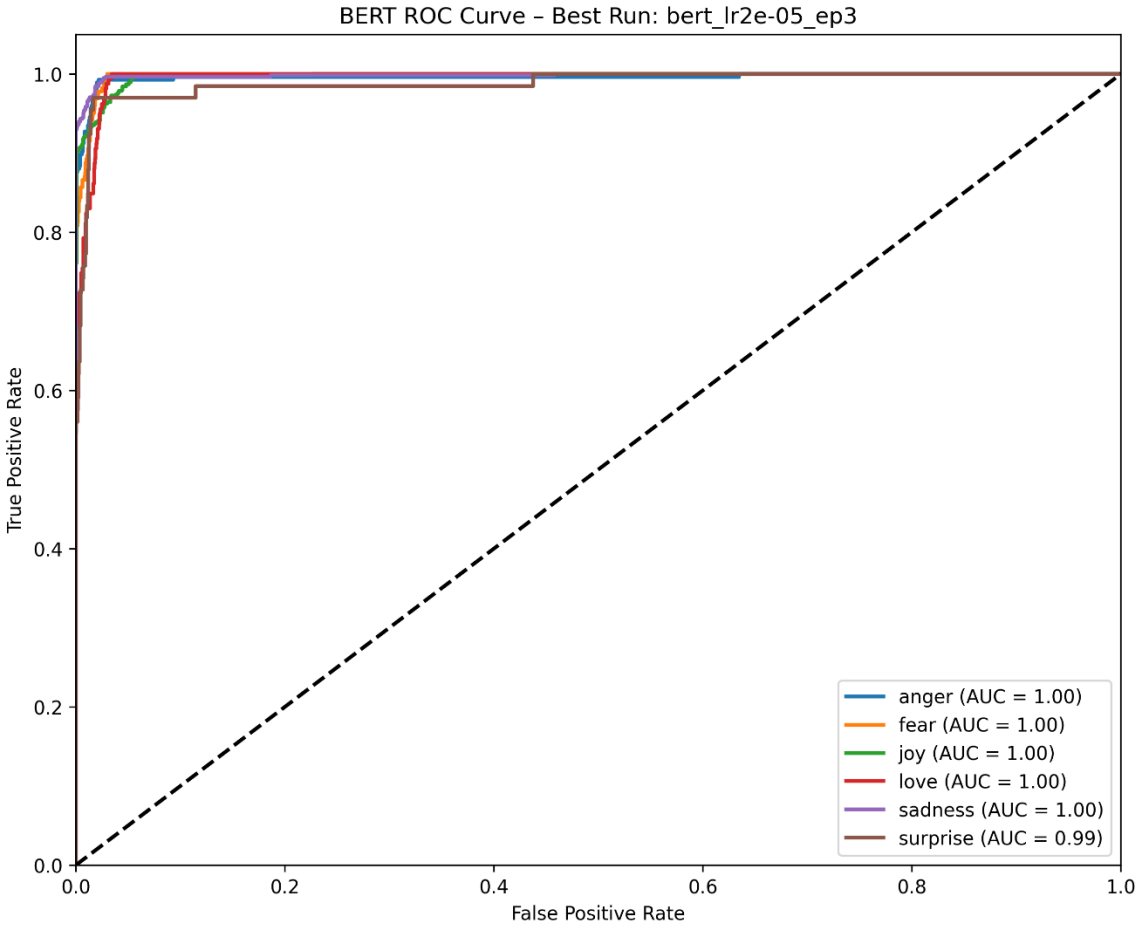

Figure 18- ROC curve for BERT model showing per-class AUC values on the full dataset.

5.3.4 Training Curves

Figures 19 and 20 plot the training and validation curves for BERT. The training accuracy plateaued early, while validation accuracy steadily improved over epochs, peaking at 93.5%. Loss values consistently declined on both training and validation sets.

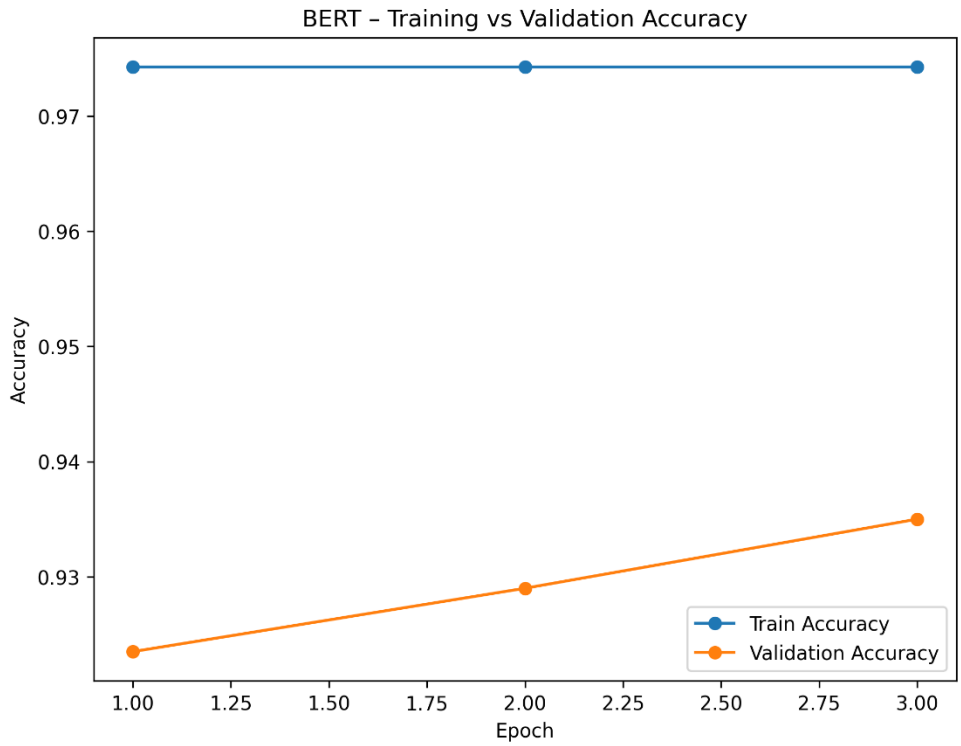

Figure 19 - Training vs validation accuracy for BERT (best run: lr=2e-5, epochs=3).

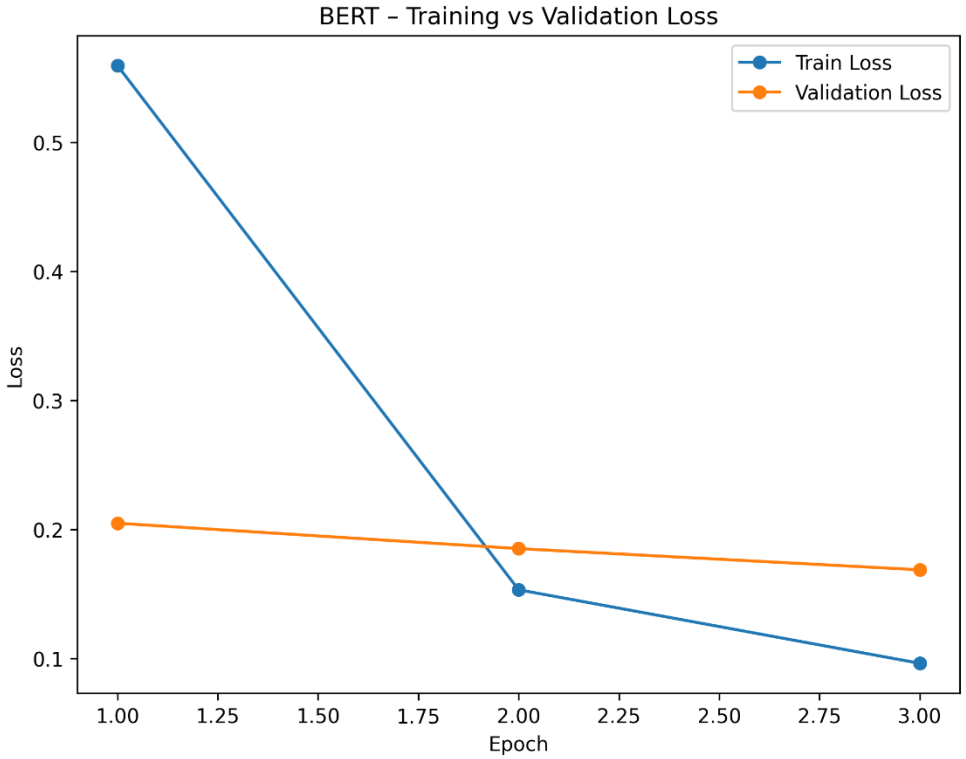

Figure 20- Training vs validation loss for BERT.

5.3.5 Key Observations from BERT Results

BERT delivered the best overall performance on the full 6-class dataset, with a macro F1-score of 0.9102 and 93.5% test accuracy.

- It outperformed all other models with just 3 epochs of fine-tuning.
- Confusion matrix and ROC analyses confirmed strong class-wise discrimination and robust generalization.
- Training curves showed stable convergence without overfitting, validating the effectiveness of the selected hyperparameters.

5.4 Summary of Key Findings

- **BERT** outperformed all models with **93.5% accuracy** and **0.9102 macro F1**, followed by **BiLSTM** on the 4-class dataset with **95.55% accuracy** and **0.9447 macro F1**.
- Traditional models (SVM, MNB) performed competitively, particularly on the balanced 4-class version, validating their usefulness as strong baselines.
- ROC curves and training dynamics indicate all models were well-optimized and free from overfitting.
- Confusion matrices showed that deep models significantly reduced inter-class confusion, especially for ambiguous emotions like *love*, *fear*, and *surprise*.

6. Discussion

This section interprets key findings, evaluates objectives, and outlines study strengths, limitations, and future directions.

6.1 Model Comparison and Performance

BERT and BiLSTM outperformed traditional models across all metrics, confirming their superior contextual understanding—consistent with findings by Zanwar et al. (2022). BERT achieved the highest performance on the 6-class task, while BiLSTM excelled on the 4-class dataset. Among traditional models, SVM was the strongest, especially on the balanced subset, aligning with Gafar et al. (2023), though its performance declined on the imbalanced full dataset.

6.2 Dataset Design and Preprocessing Impact

All models improved on the 4-class dataset, highlighting the negative impact of rare or overlapping emotion classes like *love* and *surprise*—an issue noted by Plaza-del-Arco et al. (2024). Traditional models benefited from negation marking, while deep models—

especially BERT—performed best with minimal preprocessing, supporting Rezapour's (2024) emphasis on preserving linguistic cues.

6.3 SMART Objectives Review

All SMART objectives were successfully achieved:

- **Specific:** Two datasets and four models implemented
- **Measurable:** Performance goals exceeded
- **Achievable:** Benchmarks met
- **Relevant:** Supports applications in conversational AI and sentiment-aware systems
- **Time-bound:** Completed within the project schedule

6.4 Strengths and Limitations

Strengths:

- Comprehensive comparison of traditional and deep models
- Model-specific preprocessing
- Evaluation across balanced and imbalanced datasets
- Use of multiple evaluation metrics and visualizations

Limitations:

- Difficulty classifying *love* and *surprise*
- Focus limited to sentence-level input
- No domain-specific BERT fine-tuning conducted

6.5 Future Work

Future improvements include:

- Data augmentation for underrepresented emotions
- Domain-adapted BERT fine-tuning
- Integration of context-aware models (e.g., BART, DialogXL)
- Exploration of multilingual and code-mixed datasets

Reference list

Praveen, G. (2020). *Emotions dataset for NLP*. [online] www.kaggle.com. Available at: https://www.kaggle.com/datasets/praveengovi/emotions-dataset-for-nlp.

Gafar, A., Hussain, N., Sidorov, G. and Gelbukh, A. (2023). *Guilt Detection in Text: A Step Towards Understanding Complex Emotions*. [online] arXiv.org. Available at: https://arxiv.org/abs/2303.03510 [Accessed 15 May 2025].

Mukherjee, P., Badr, Y., Doppalapudi, S., Srinivasan, S.M., Sangwan, R.S. and Sharma, R. (2021). Effect of Negation in Sentences on Sentiment Analysis and Polarity Detection. *Procedia Computer Science*, 185, pp.370–379. doi:https://doi.org/10.1016/j.procs.2021.05.038.

Plaza-del-Arco, F.M., Curry, A., Curry, A.C. and Hovy, D. (2024). *Emotion Analysis in NLP: Trends, Gaps and Roadmap for Future Directions*. [online] arXiv.org. Available at: https://arxiv.org/abs/2403.01222.

Rezapour, M. (2024). *Emotion Detection with Transformers: A Comparative Study*. [online] Available at: https://arxiv.org/pdf/2403.15454.

S, S.A.K. and Geetha, A. (2024). Emotion Detection from Text using Natural Language Processing and Neural Networks. *International Journal of Intelligent Systems and Applications in Engineering*, [online] 12(14s), pp.609–615. Available at: https://ijisae.org/index.php/IJISAE/article/view/4707.

Tulika Chutia and Nomi Baruah (2024). A review on emotion detection by using deep learning techniques. *Artificial intelligence review*, 57(8). doi:https://doi.org/10.1007/s10462-024-10831-1.

Zanwar, S., Wiechmann, D., Qiao, Y. and Kerz, E. (2022). *Improving the Generalizability of Text-Based Emotion Detection by Leveraging Transformers with Psycholinguistic Features*. [online] arXiv.org. Available at: https://arxiv.org/abs/2212.09465 [Accessed 15 May 2025].