Hope Speech Detection: A Comparative Study of Machine Learning, Deep Learning, and Transformer-Based Models

GodsGift Uzor[*,†], Mohammadsaleh Pajuhanfard[*,†], Michael Beebe[*,†], Maaz Amjad[*,†] and Victor Sheng[*,†]

Texas Tech University, Lubbock, Texas, USA

Abstract

Hope speech detection is a crucial task in Natural Language Processing (NLP) that involves identifying and classifying text that expresses optimism, encouragement, or positivity. This study evaluates and compares classical machine learning (SVM), deep learning (LSTM, BiLSTM, CNN), and transformer-based models (BERT, SBERT, GPT-2) for binary and multiclass hope speech classification. The binary classification task distinguishes between "Hope" and "Not Hope," while the multiclass classification further categorizes hope speech into Generalized Hope, Realistic Hope, Unrealistic Hope, Sarcasm, and Not Hope. The results indicate that SBERT outperforms all models, achieving the highest binary and multiclass accuracies of 92.81% and 84.29%, respectively. For binary classification, BERT follows with 85.12%, GPT-2 with 84.49%, BiLSTM with 79.34%, SVM with 78.50%, CNN with 75.13%, and LSTM with 72.00%. In the multiclass task, GPT-2 achieves 77.76%, closely trailing SBERT, while BERT achieved 77.13%, followed by SVM at 66.98%, BiLSTM at 63.35%, and both CNN and LSTM at 59.20% and 59.00%, respectively. Transformer-based models, particularly SBERT and BERT, demonstrate superior precision, recall, and F1-scores, excelling in distinguishing subtle categories like sarcasm and unrealistic hope, whereas LSTM, BiLSTM, and CNN struggle with nuanced hope categories due to limitations in capturing long-range dependencies or contextual depth. Despite these advancements, challenges persist in detecting underrepresented hope categories, such as unrealistic hope and sarcasm, especially in low-resource settings. This study highlights the effectiveness of transformer-based models for hope speech detection while identifying avenues for improving deep learning approaches.

Keywords

Bidirectional Encoder Representations from Transformers (BERT), Sentence-Bidirectional Encoder Representations from Transformers (SBERT), Long Short-Term Memory (LSTM), Support Vector Machine (SVM), Convolutional Neural Network (CNN), Generative Pre-trained Transformer 2 (GPT-2), Bi-directional Long Short-Term Memory (Bi-LSTM), Natural Language Processing (NLP)

1. Introduction

Hope speech detection, an emerging field in NLP, seeks to identify positive and encouraging content within unstructured text, such as social media posts, where emotions and sentiments are often expressed with subtlety and context-dependent nuances. Recognizing emotions in text is essential for understanding human behavior and intent, complementing research in voice and facial expression analysis [1]. The challenge lies in capturing underlying meanings, such as distinguishing genuine hope from sarcasm or non-hopeful sentiments, amidst indirect phrasing, irony, and overlapping linguistic features that traditional methods like keyword matching fail to address [2]. Machine learning models, ranging from classical to deep learning and transformer-based architectures, offer robust solutions by modeling complex textual relationships and extracting emotionally significant patterns.

This study was conducted within the IberLEF 2025 framework [3] and evaluates seven machine

IberLEF 2025, September 2025, Zaragoza, Spain

[*]Corresponding author.

[†]These authors contributed equally.

godsgift.uzor@ttu.edu (G. Uzor); ryan.pajuhanfard@ttu.edu (M. Pajuhanfard); michael.beebe@ttu.edu (M. Beebe); maaz.amjad@ttu.edu (M. Amjad); victor.sheng@ttu.edu (V. Sheng)

https://github.com/GodsGift-Uzor (G. Uzor); https:https://github.com/Ryan-Pajuhan (M. Pajuhanfard); https://github.com/michael-beebe (M. Beebe)

0000-0003-3069-98553 (G. Uzor); 0009-0000-8524-2313 (M. Pajuhanfard); 0009-0004-9151-6607 (M. Beebe); 0000-0002-5969-9085 (M. Amjad); 0000-0003-4960-174X (V. Sheng)

learning models for hope speech detection in English social media posts: Support Vector Machine (SVM) [4], Long Short-Term Memory (LSTM) [5], Bidirectional Long Short-Term Memory (BiLSTM) [6], Convolutional Neural Network (CNN) [7], Bidirectional Encoder Representations from Transformers (BERT) [8], Sentence-BERT (SBERT) [9], and Generative Pre-trained Transformer 2 (GPT-2) [10]. We address two classification tasks: binary classification (Hope vs. Not Hope) and multiclass classification (Generalized Hope, Realistic Hope, Unrealistic Hope, Sarcasm, Not Hope). Our main contributions are: (1) a comprehensive comparison of classical, machine learning, deep learning, and transformer-based models to delineate their strengths and limitations in hope speech detection; (2) an in-depth error analysis identifying linguistic and contextual factors affecting performance, particularly for nuanced categories like sarcasm and unrealistic hope; and (3) insights into the generalizability of these models for other text classification tasks, such as sentiment analysis or emotion detection.

The paper is structured as follows: Section 2 reviews related work on hope speech detection datasets, models, and benchmarks. Section 3 details the dataset, preprocessing, feature extraction, model training, and evaluation metrics. Section 4 presents the results and discussions for each model, analyzing performance trends. Section 5 provides an error analysis, highlighting misclassification patterns. Section 6 summarizes findings, discusses limitations, and proposes future directions.

2. Related Works

Hope speech detection has emerged as a vital area in NLP, with the aim of identifying and amplifying positive supportive content on social media networks to counteract negativity. Recent efforts have advanced this field through the creation of annotated datasets, development of sophisticated models, and the establishment of collaborative benchmarks, particularly for multilingual and nuanced contexts. This section synthesizes key contributions and organizes them into three interconnected themes: dataset development, modeling innovations, and community-driven benchmarking. By critically evaluating these studies, we elucidate their contributions, limitations, and direct implications for our comparative analysis of classical machine learning (SVM), deep learning (LSTM, BiLSTM, CNN), and transformer-based models (BERT, SBERT, GPT-2) for binary and multiclass hope speech classification in English tweets, highlighting how previous work informs our methodology and corroborates our findings.

The creation of annotated datasets has laid the foundation for hope speech detection, enabling robust model training and evaluation. Chakravarthi [11] introduced the HopeEDI dataset, a pioneering multilingual collection of YouTube comments in English, Tamil, and Malayalam, manually labeled for hope speech. This resource established a critical benchmark for studying positivity in social media networks, but its focus on YouTube and its limited language scope restrict its applicability to the concise and contextually diverse texts present in Twitter tweets. Our study leverages English tweets, aligning with the PolyHope dataset proposed by Balouchzahi et al. [2], which classifies tweets into binary (Hope vs. Not Hope) and multiclass categories (Generalized Hope, Realistic Hope, Unrealistic Hope). PolyHope's hierarchical structure mirrors our classification objectives, improving granularity, but its scope in English only limits generalizability across various languages. In contrast, García-Baena et al. [12] developed the SpanishHopeEDI dataset, focusing on Spanish Twitter posts related to LGBT issues. Although this dataset underscores the social significance of hope, its domain-specific nature reduces its applicability to the diverse emotional expressions in our study. These datasets highlight a trade-off between specificity and versatility, guiding our selection of a broadly applicable English Twitter dataset to balance nuanced classification with practical relevance.

Innovations in modeling have propelled hope speech detection by harnessing transformer-based architectures to capture intricate emotional and contextual nuances. Sidorov et al. [13] demonstrated that transformers excel over traditional methods in detecting hope and regret, using contextual embeddings to model complex emotional patterns. This aligns with the findings of this study, where the SBERT and BERT models achieved superior performance compared to the other models used, thus reaffirming the efficacy of transformer-based models for nuanced tasks. Balouchzahi et al. [14] extended transformer-based approaches to Urdu texts, addressing hope and hopelessness detection in low-resource languages.

Their work reveals challenges in data scarcity and linguistic diversity, providing insight for future multilingual extensions of our English-focused study. Similarly, Sidorov et al. [15] introduced MIND-HOPE, a framework for the detection of fine-grained multilingual hope, emphasizing nuanced emotional dimensions. This parallels our multiclass approach, although its broader linguistic scope contrasts with our focus on a single language. The persistent challenge of distinguishing subtle hope subtypes, as noted in our error analysis (e.g., Realistic vs. Unrealistic Hope), suggests that integrating domain-specific features with transformer embeddings, as implied by MIND-HOPE, could enhance performance, a direction our SBERT results support.

Community-driven benchmarking efforts have standardized hope speech detection, fostering collaborative progress across languages. Chakravarthi et al. [16] expanded the HopeEDI dataset to include English, Spanish, Tamil, Malayalam, and Kannada, introducing a three-class classification scheme (Hope, Not Hope, Not Intended). Their baselines provide a valuable reference, but their focus on coarser categories contrasts with our fine-grained multiclass approach, which tackles nuanced subtypes like sarcasm. García-Baena et al. [17] advanced multilingual benchmarking at IberLEF 2024, emphasizing hope detection as expressions of expectations and diversity in English and Spanish social media. Their findings highlight multilingual challenges, which our English-focused study partially addresses, with transformers outperforming SVM, consistent with their results. Butt et al. [18] further refined the benchmarking at IberLEF 2025, focusing on detecting optimism, expectation, and sarcasm in Spanish and English. Their emphasis on sarcasm detection sustains our findings, where SBERT excels (F1=0.98), indicating the strength of transformer models in capturing ironic nuances. Complementing this, Butt et al. [19] proposed a multiclass framework for optimism, expectation, and sarcasm detection, aligning with the objective of our study nuanced classification goals and providing a comparative lens for our error analysis, which reveals confusion among hope subtypes.

These contributions collectively chart the trajectory of hope speech detection, from foundational datasets to advanced models and standardized benchmarks. This study integrates these advances by systematically evaluating a diverse model set, confirming that transformer-based approaches (SBERT, BERT, GPT-2) surpass traditional machine learning (SVM) and deep learning (LSTM, BiLSTM, CNN) methods, particularly for nuanced categories like sarcasm and Generalized Hope. However, our error analysis underscores challenges in distinguishing Realistic and Unrealistic Hope, echoing limitations in previous works [15, 14]. This highlights the need for richer, multilingual datasets and hybrid modeling strategies, positioning our work as a stepping stone toward addressing these gaps and advancing hope speech detection.

3. Methodology

3.1. Dataset Description

In this study, two datasets were used: the training dataset and the development dataset of English tweets from the first half of 2022 [20]. The dataset had no missing values, ensuring its reliability for analysis. Table 1 presents a statistical description of the datasets.

3.2. Data Collection and Preprocessing

The dataset consists of two text-based datasets: `en_train.csv` and `en_dev.csv`, which serve as the training and development (test) sets, respectively. These datasets are loaded into Pandas DataFrames for further processing.

3.2.1. Text Pre-processing

Several Natural Language Processing (NLP) techniques are applied to clean and standardize the text data:

- Emoji Conversion: Emojis are converted into textual descriptions to provide more context.

	Training Dataset	Development Dataset
Size	5,233 samples	1,902 samples
Avg. Text Length (SD)	187.80 (93.83) chars	186.94 (91.49) chars
Text Length Range	24 - 886 chars	29 - 766 chars
Binary Distribution		
Hope	2,426 (46.36%)	899 (47.27%)
Not Hope	2,807 (53.64%)	1,003 (52.73%)
Text Length Distribution		
25th Percentile	115 words	114 words
50th Percentile (Median)	172 words	172 words
75th Percentile	252 words	253.75 words
Multiclass Distribution		
Not Hope	2,245 (42.90%)	816 (42.90%)
Generalized Hope	1,284 (24.54%)	467 (24.55%)
Sarcasm	692 (13.22%)	252 (13.25%)
Realistic Hope	540 (10.32%)	196 (10.30%)
Unrealistic Hope	472 (9.02%)	171 (8.99%)

Table 1
Dataset Statistical Description for Training and Development Sets

- Text Normalization: All text is converted to lowercase.
- URL removal: URLs and web links are removed using regular expressions.
- Mention and Hashtag Handling: User mentions (`@username`) are replaced with the placeholder `USER_MENTION`, while hashtags (`#hashtag`) are retained but stripped of the # symbol.
- Punctuation removal: All punctuation marks are removed.
- Whitespace Standardization: Extra spaces are removed and the text is trimmed.

3.3. Feature Extraction

To convert text into suitable numerical representations for machine learning models, several feature extraction techniques are used. For transformer-based models such as BERT, SBERT, and GPT-2, pre-trained tokenizers from the Hugging Face library are utilized. Specifically, BERT uses the `BertTokenizer` from the `bert-base-uncased` model, while SBERT uses the `AutoTokenizer` from the `sentence-transformers/all-MiniLM-L6-v2` model. GPT-2 employs the `GPT2Tokenizer`, with a special padding token manually added to accommodate batch processing. These tokenizers apply special tokens, truncation, and padding to a fixed maximum length, and the output is converted into PyTorch `Dataset` objects to facilitate efficient batch processing. For traditional neural architectures such as LSTM, Bi-LSTM, and CNN, word-level tokenization is performed using Keras `Tokenizer`, followed by sequence conversion and zero-padding using `pad_sequences` to standardize input lengths for consistent training. Additionally, TF-IDF vectorization is applied to capture term importance, using unigram and bigram features with a vocabulary size limited to 10,000. These vectorized features are used particularly for classical machine learning baselines. To address class imbalance in the TF-IDF feature space, the Synthetic Minority Over-sampling Technique (SMOTE) is applied, generating synthetic examples of the minority class to ensure balanced training. All resulting datasets, tokenized or vectorized are then formatted into structures compatible with PyTorch or TensorFlow to enable seamless training and evaluation across different model architectures.

3.4. Model Training and Evaluation

Several models were trained and evaluated for text classification, with their results presented in this report.

3.4.1. Support Vector Machine (SVM) with Hyperparameter Tuning

Support Vector Machine (SVM) models are supervised learning algorithms used for classification and regression tasks, which works by finding the optimal hyperplane that maximally separates data points of different classes in a high-dimensional space [4]. This model was selected for its effectiveness on high-dimensional data and its robustness to overfitting. A hyperparameter search is performed using `GridSearchCV` with cross-validation (CV=5), and the best combination of hyperparameters (`C`, `kernel`, `gamma`) is selected for training the selected dataset.

3.4.2. Bidirectional Encoder Representations from Transformers (BERT)

BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based model designed to pre-train deep bidirectional representations from unlabeled text [8], reading entire sequences simultaneously in both directions using the transformer encoder architecture, unlike traditional language models that process text sequentially. It is pre-trained on two unsupervised tasks: Masked Language Modeling (MLM), where random tokens in the input are masked and the model learns to predict them based on surrounding context, and Next Sentence Prediction (NSP), where the model predicts whether one sentence logically follows another. This pre-training enables BERT to capture complex linguistic features and semantics beyond traditional models, making it a foundational model in modern NLP due to its bidirectional context awareness and ability to generalize across tasks; thus, it was selected for this study. For training, BERT is fine-tuned using the Hugging Face Trainer framework with a learning rate of 2e-5, a batch size of 4, 10 epochs, and a weight decay of 0.01.

3.4.3. Sentence-BERT (SBERT)

Sentence-BERT (SBERT) is a modification of the BERT architecture designed to generate semantically meaningful sentence embeddings [9], addressing the computational inefficiency of the original BERT in sentence-pair tasks by using a siamese or triplet network structure to compute fixed-size vector representations for entire sentences. In SBERT, two or more sentences are passed independently through a shared BERT-based encoder, and the resulting embeddings are compared using similarity measures such as cosine similarity to facilitate tasks like semantic textual similarity, clustering, and sentence classification. SBERT is fine-tuned on Natural Language Inference (NLI) and semantic similarity datasets using objectives such as contrastive loss or regression on similarity scores, improving efficiency for pairwise comparisons while retaining the contextual richness of transformer-based encodings, thus enabling the use of vector-based retrieval and classification techniques in downstream NLP applications. For this study, SBERT was trained with a learning rate of 2e-5, a batch size of 4, 10 epochs for both binary and multiclass tasks, using CrossEntropyLoss, with its transformer-based nature supported by epoch evaluation and model saving for the binary task, and warmup plus logging without evaluation during training for the multiclass task.

3.4.4. Generative Pre-trained Transformer 2 (GPT-2)

Generative Pre-trained Transformer 2 (GPT-2), developed by OpenAI [10], is a large-scale, autoregressive language model based on the Transformer architecture, specifically utilizing the decoder component to predict the next word in a sequence given the preceding context. It consists of stacked Transformer decoder blocks that employ masked (causal) self-attention mechanisms and feedforward layers to model long-range dependencies in text, with positional embeddings added to token embeddings to preserve word order. GPT-2 is pre-trained on a large corpus of text using unsupervised learning and fine-tuned for specific NLP tasks such as text classification, question answering, summarization, and language generation, leveraging its ability to generate coherent and contextually relevant text by capturing complex language patterns during pre-training, making it effective for a wide range of applications without task-specific architecture modifications. For this study, GPT-2 was trained with a learning rate

of 2e-5, a batch size of 1, 3 epochs for both binary and multiclass tasks, using CrossEntropyLoss, with a heavy model and small batch size noted for its computational demands.

3.4.5. Long Short-Term Memory (LSTM) Network

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) designed to handle sequential data and long-range dependencies more effectively than traditional RNNs by introducing gates (input, forget and output) to regulate information flow, allowing it to retain relevant information over long sequences while minimizing issues such as vanishing gradients, and was selected for its effectiveness in NLP tasks [5]. The LSTM network in this study is trained with an architecture that includes an embedding layer with a 5000-word vocabulary and an embedding size of 128, followed by two LSTM layers where the first (128 units) returns sequences and the second (64 units) produces final features, a dropout layer with a rate of 0.5 to mitigate overfitting, a fully connected dense layer with 64 neurons and ReLU activation, and an output layer with softmax activation for classification.

3.4.6. Bi-directional Long Short-Term Memory (BiLSTM) Network

Bidirectional Long Short-Term Memory (BiLSTM) is a recurrent neural network (RNN) architecture that extends the traditional LSTM by processing input sequences in both forward and backward directions [6], incorporating past and future context for each time step, making it particularly effective for tasks where meaning depends on the surrounding context, such as NLP applications. A BiLSTM consists of two LSTM layers: one processes the sequence from the first to the last token, and the other from the last to the first token, with their outputs typically concatenated at each time step to enable the network to learn richer, context-aware representations, which is advantageous in sequence labeling and classification tasks where bidirectional dependencies are important. For this study, the BiLSTM model was trained with a learning rate of 1e-3, a custom batch size, 10 epochs for the binary task, and 15 epochs for the multiclass task. The binary task utilized BCELoss, while the multiclass task employed CrossEntropyLoss, with a lighter model and standard optimizer noted for efficiency.

3.4.7. Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNNs), originally developed for image processing [7], have been adapted for text classification by treating word embeddings as input features, applying one-dimensional convolutional filters over token embedding sequences to capture local patterns like n-grams critical for identifying sentiment or emotional cues, and using pooling layers such as max-pooling to focus on salient signals, offering computational efficiency and adeptness at modeling short-range dependencies, making them a strong baseline for hope speech detection. The CNN implementation begins with an embedding layer transforming input token IDs into dense vectors of dimension 300, initialized with pretrained word embeddings (e.g. GloVe) to leverage general linguistic knowledge, reshaped into a matrix for one-dimensional convolution. Multiple convolutional layers are applied with filter sizes of 3, 4, and 5 (capturing trigrams to pentagrams), followed by ReLU activation for non-linearity, max-pooling over time to reduce the output of each filter to a single value, and concatenation of the resulting features into a unified vector. To mitigate overfitting, a dropout layer with a rate of 0.5 is applied before a fully connected layer maps the feature vector to class logits, and the model is optimized using the Adam optimizer with a learning rate of 0.001 and cross-entropy loss, trained for 10 epochs with a batch size of 32, effectively capturing local textual patterns relevant to distinguishing hopeful, sarcastic, and non-hopeful expressions in the PolyHope dataset.

3.5. Evaluation Metrics

Each model is evaluated using a set of metrics to comprehensively assess its performance. Accuracy measures the proportion of correct predictions, providing an overall indicator of the effectiveness of the model. A detailed classification report includes precision, recall and F1-score for each class, where

precision is calculated as the ratio of true positives to the sum of true positives and false positives (TP / (TP + FP)), recall as the ratio of true positives to the sum of true positives and false negatives (TP / (TP + FN)), and F1-score as the harmonic mean of precision and recall (2 * (precision * recall) / (precision + recall)). These metrics offer insights into the model's performance across individual classes, particularly useful for imbalanced datasets. In addition, a confusion matrix displays misclassifications, revealing patterns of errors between classes.

3.6. Training and Evaluation Pipeline

For each model, the following process is executed: the model is trained on the training dataset, predictions are generated on the test dataset, performance metrics are computed and stored, and the results are compared across models. The final results allow a comparative analysis of different models to determine the best performing approach for the given text classification task. The strengths and shortcomings of the best performing and least performing approaches are also discussed in the following section.

4. Results and Discussions

4.1. SVM Model Analysis

This report provides an evaluation of the Support Vector Machine (SVM) model trained for binary and multiclassification. Performance metrics, including accuracy, precision, recall, F1-score, and confusion matrix, are analyzed to assess the effectiveness of the model.

4.1.1. Binary Model Performance

The SVM model achieved an accuracy of 78.50% for binary classification. The detailed classification report is shown in Table 2.

Class	Precision	Recall	F1-score	Support
Hope	0.79	0.74	0.76	899
Not Hope	0.78	0.83	0.80	1003
Accuracy		0.7850		
Macro Avg	0.79	0.78	0.78	1902
Weighted Avg	0.79	0.78	0.78	1902

Table 2
SVM Binary Classification Report

The confusion matrix for the SVM binary model is presented in Table 3.

	Hope	Not Hope
Hope	664	235
Not Hope	174	829

Table 3
SVM Binary Confusion Matrix

The model exhibits relatively balanced precision and recall for both classes. The recall for "Not Hope" is slightly higher (0.83) compared to "Hope" (0.74), suggesting that the model is better at correctly identifying Not Hope instances. However, there is a trade-off, as precision for "Not Hope" (0.78) is slightly lower than for "Hope" (0.79). From the confusion matrix presented in Table 3, the model correctly classified 664 instances of "Hope" but misclassified 235 instances as "Not Hope." The model correctly classified 829 instances of "Not Hope" but misclassified 174 instances as "Hope". The overall performance is acceptable with an accuracy of 78.50%.

4.1.2. Multi-Class Model Performance

The Support Vector Machine (SVM) model was evaluated on a multiclass classification task, achieving an overall accuracy of 66.98%. Table 4 presents the detailed performance analysis, including precision, recall, and f1-score for each class.

Class	Precision	Recall	F1-Score	Support
Generalized Hope	0.55	0.59	0.57	467
Not Hope	0.69	0.86	0.76	816
Realistic Hope	0.57	0.25	0.35	196
Sarcasm	0.94	0.75	0.83	252
Unrealistic Hope	0.66	0.35	0.45	171
Accuracy		0.6698		
Macro Avg	0.68	0.56	0.59	1902
Weighted Avg	0.67	0.67	0.65	1902

Table 4
SVM Multiclass Classification Report

Table 5 presents the confusion matrix for the SVM multiclass model performance insights on different classes.

	Generalized Hope	Not Hope	Realistic Hope	Sarcasm	Unrealistic Hope
Generalized Hope	276	155	23	2	11
Not Hope	91	702	6	7	10
Realistic Hope	82	58	49	1	6
Sarcasm	14	46	1	188	3
Unrealistic Hope	41	63	7	1	59

Table 5
SVM Multiclass Confusion Matrix

The model performed best in the "Sarcasm" class, with the highest precision (0.94) and f1-score (0.83). However, the "Realistic Hope" class had the lowest recall (0.25), indicating that many instances of this class were misclassified. The model also struggled with the "Unrealistic Hope" class, showing a relatively low f1-score of 0.45.

Summary

The Support Vector Machine (SVM) model demonstrated moderate effectiveness in classifying emotional expressions from English tweets, achieving 78.50% accuracy in binary classification and 66.98% in multiclass classification. The model showed relatively balanced precision and recall in distinguishing between *Hope* and *Not Hope* classes, and performed particularly well in identifying the *Sarcasm* class, which likely contains more distinctive linguistic patterns. However, notable weaknesses were observed in classifying nuanced hope-related categories such as *Generalized Hope*, *Realistic Hope*, and *Unrealistic Hope*, which were frequently misclassified. These errors are likely due to the limitations of TF-IDF-based feature representations, which lack contextual understanding and semantic depth. Furthermore, the lexical overlap and subtle differences between hope-related classes, combined with class imbalance in the dataset, contributed to reduced performance. Overall, while the SVM model is effective in detecting more distinct emotional categories, it struggles with nuanced sentiment differentiation, motivating the adoption of deep learning and transformer-based methods that better capture contextual and semantic nuances in social media texts.

4.2. BERT Model Analysis

4.2.1. BERT Binary Model Performance

The performance of the BERT model was evaluated on the test dataset, achieving an overall accuracy of 85.12%, indicating strong performance in distinguishing between the "Hope" and "Not Hope" classes. Classification metrics, including precision, recall, and F1-score for each class, are presented in Table 6.

Class	Precision	Recall	F1-score	Support
Hope	0.84	0.84	0.84	899
Not Hope	0.86	0.86	0.86	1003
Accuracy		0.8512		
Macro Avg	0.85	0.85	0.85	1902
Weighted Avg	0.85	0.85	0.85	1902

Table 6
BERT Binary Classification Report

The model demonstrates balanced performance across both classes, with precision, recall, and F1-scores around 0.84 for "Hope" and 0.86 for "Not Hope." The macro average (0.85) and weighted average (0.85) confirm that the model maintains consistent performance without significant bias toward one class.

The confusion matrix in Table 7 illustrates the classification results. The model correctly classified 756 out of 899 "Hope" samples (84.1%) and 863 out of 1003 "Not Hope" samples (86.1%), 143 "Hope" instances were misclassified as "Not Hope" (false negatives) and 140 "Not Hope" instances were misclassified as "Hope" (false positives).

	Hope	Not Hope
Hope	756	143
Not Hope	140	863

Table 7
Confusion Matrix for BERT Binary Model

These results demonstrate that the BERT model is effective in distinguishing between the binary classes, achieving a balanced performance across both classes.

4.3. BERT Multiclass Model Performance

The BERT multiclass model achieved an overall accuracy of 77.13%, demonstrating a strong ability to classify hope-related expressions while highlighting areas for improvement. Tables 4.3 and 4.3 present the classification report and confusion matrix for the BERT multiclass model.

Class	Precision	Recall	F1-Score	Support
Generalized Hope	0.71	0.70	0.70	467
Not Hope	0.82	0.85	0.83	816
Realistic Hope	0.63	0.58	0.60	196
Sarcasm	0.99	0.89	0.94	252
Unrealistic Hope	0.59	0.61	0.60	171
Accuracy		0.7713		
Macro Avg	0.75	0.73	0.74	1902
Weighted Avg	0.77	0.77	0.77	1902

Table 8
BERT Multiclass Classification Report

"Not Hope" has the highest performance with an F1-score of 0.83, reflecting the model's strong ability to identify non-hopeful content. "Sarcasm" achieves the best precision (0.99) and a high F1-score (0.94),

indicating that the model is highly confident in correctly identifying sarcastic expressions. "Generalized Hope" (F1 = 0.70), "Realistic Hope" (F1 = 0.60), and "Unrealistic Hope" (F1 = 0.60) show moderate performance, suggesting these categories may have more ambiguous or overlapping characteristics. The macro average F1-score of 0.74 indicates that the model performs fairly well across all classes, though some imbalances remain.

True / Pred	Generalized Hope	Not Hope	Realistic Hope	Sarcasm	Unrealistic Hope
Generalized Hope	328	81	40	0	18
Not Hope	75	695	14	2	30
Realistic Hope	40	22	114	0	20
Sarcasm	5	14	2	225	6
Unrealistic Hope	16	38	12	0	105

Table 9
BERT MultiClass Confusion Matrix

Most classes have a high true positive count, but misclassifications occur primarily between "Generalized Hope", "Realistic Hope", and "Unrealistic Hope", indicating difficulty in differentiating nuanced forms of hope. "Not Hope" is classified with high accuracy, with 695 correct predictions out of 816 and relatively few misclassifications. "Generalized Hope" is often misclassified as "Not Hope" (81 instances) and "Realistic Hope" (40 instances). "Unrealistic Hope" is sometimes confused with "Not Hope" (38 instances) and "Realistic Hope" (12 instances). "Sarcasm" is well classified, with only minor confusion with "Not Hope" (14 instances).

The BERT multiclass model performs well in distinguishing Not Hope and Sarcasm, while struggles remain in differentiating nuanced hope categories, particularly Generalized, Realistic, and Unrealistic Hope. Fine-tuning with more representative training data and incorporating context-aware embeddings may help address these challenges.

Summary

The BERT model demonstrated strong overall performance in both binary and multiclass classification tasks involving emotional content from English tweets. In binary classification, it achieved an accuracy of 85.12%, with balanced precision and recall across the *Hope* and *Not Hope* classes. For the multiclass task, BERT attained an accuracy of 77.13%, with particularly high performance in the *Not Hope* and *Sarcasm* categories. However, the model exhibited weaknesses in distinguishing between nuanced hope-related categories—*Generalized Hope*, *Realistic Hope*, and *Unrealistic Hope*—which were frequently confused due to overlapping linguistic features and limited semantic separation. These misclassifications likely stem from subtle contextual differences in expression and the class imbalance in the training data. Despite these challenges, the contextual embedding capabilities of the BERT model mark a significant improvement over traditional models, and its performance can be further enhanced through targeted data augmentation, better class representation, and continued fine-tuning.

4.4. SBERT Model Analysis

4.4.1. SBERT Binary Model Performance

The performance of the SBERT model was evaluated on the training dataset, achieving a high overall accuracy of 92.81%, reflecting strong capability in differentiating between the "Hope" and "Not Hope" classes. Table 10 presents detailed classification metrics, including precision, recall, and F1-score.

The model demonstrates strong and balanced performance across both classes, with F1-scores of 0.92 for "Hope" and 0.93 for "Not Hope". The macro and weighted averages, both at 0.93, indicate that the SBERT model generalizes well across the class distribution without significant bias.

Table 11 shows the confusion matrix. Out of 2,426 "Hope" instances, the model correctly predicted 2,287 (94.3%), while misclassifying 139 as "Not Hope." For the 2,807 "Not Hope" instances, 2,570 were correctly predicted (91.6%), with 237 misclassified as "Hope."

Class	Precision	Recall	F1-score	Support
Hope	0.91	0.94	0.92	2426
Not Hope	0.95	0.92	0.93	2807
Accuracy			0.9281	
Macro Avg	0.93	0.93	0.93	5233
Weighted Avg	0.93	0.93	0.93	5233

Table 10
SBERT Binary Classification Report

	Hope	Not Hope
Hope	2287	139
Not Hope	237	2570

Table 11
Confusion Matrix for SBERT Binary Model

These results indicate that SBERT is highly effective in extracting semantic features for binary classification, achieving reliable and consistent results across both classes.

4.4.2. SBERT Multiclass Model Performance

The SBERT multiclass model achieved an overall accuracy of 84.29%, showing robust performance in identifying various expressions of hope and non-hope. Tables 4.4.2 and 4.4.2 present the classification report and confusion matrix for the SBERT multiclass model.

Class	Precision	Recall	F1-Score	Support
Generalized Hope	0.81	0.84	0.83	1284
Not Hope	0.93	0.90	0.91	2245
Realistic Hope	0.56	0.71	0.63	540
Sarcasm	0.99	0.97	0.98	692
Unrealistic Hope	0.72	0.58	0.64	472
Accuracy			0.8429	
Macro Avg	0.80	0.80	0.80	5233
Weighted Avg	0.85	0.84	0.84	5233

Table 12
SBERT Multiclass Classification Report

"Not Hope" achieved the highest F1-score (0.91), reflecting the model's reliability in detecting non-hopeful content. "Sarcasm" continues to show outstanding classification performance with a precision of 0.99 and an F1-score of 0.98. Among the hope-related categories, "Generalized Hope" achieved strong results with an F1-score of 0.83, while "Realistic Hope" and "Unrealistic Hope" had lower scores of 0.63 and 0.64, respectively. These values suggest ongoing challenges in distinguishing the nuanced subtypes of hope. The macro average F1-score of 0.80 and a weighted average of 0.84 indicate that the model maintains a relatively balanced performance across different classes, despite class imbalance.

True / Pred	Generalized Hope	Not Hope	Realistic Hope	Sarcasm	Unrealistic Hope
Generalized Hope	1078	71	115	2	18
Not Hope	111	2010	63	4	57
Realistic Hope	109	23	383	0	25
Sarcasm	9	5	3	668	7
Unrealistic Hope	22	60	117	1	272

Table 13
SBERT MultiClass Confusion Matrix

The confusion matrix reveals that most predictions for "Generalized Hope", "Not Hope", and "Sarcasm"

are highly accurate. "Generalized Hope" is most frequently misclassified as "Realistic Hope" (115 instances), while "Unrealistic Hope" is often confused with "Realistic Hope" (117 instances) and "Not Hope" (60 instances). These confusions suggest overlapping linguistic features between realistic and unrealistic hope. Nevertheless, the SBERT model performs consistently across most categories and outperforms BERT in several hope-related subcategories.

Overall, SBERT proves to be a robust model for multiclass classification tasks involving nuanced emotional and contextual categories. Further fine-tuning and data augmentation could help address remaining challenges in accurately separating closely related hope classes.

Summary

The SBERT model exhibited strong performance in both binary and multiclass emotional classification tasks on English tweets, achieving accuracies of 92.81% and 84.29% respectively. Its primary strengths lie in its ability to capture deep semantic relationships through contextualized sentence embeddings, enabling high precision and recall across distinct categories such as *Not Hope* and *Sarcasm*. In particular, the model achieved F1-scores above 0.90 for these classes, reflecting its robustness in identifying linguistically distinguishable content. Furthermore, the model demonstrated substantial improvement over traditional and transformer-based baselines in differentiating hope-related expressions, particularly *Generalized Hope*, which achieved an F1-score of 0.83. However, performance declined for more nuanced categories such as *Realistic Hope* and *Unrealistic Hope*, which were frequently confused with each other and with *Not Hope*. These misclassifications are likely attributable to overlapping lexical and syntactic structures, as well as subtle semantic differences that require a deeper understanding of context and intent. The challenges underscore the difficulty of distinguishing between nuanced emotional states, particularly in informal social media language. Despite these limitations, SBERT's contextual embeddings provide a solid foundation for emotion classification, with further gains possible through domain-specific fine-tuning and enrichment of underrepresented categories in the training data.

4.5. GPT-2 Model Analysis

4.5.1. GPT-2 Binary Model Performance

The GPT-2 model achieved an overall accuracy of 84.49% on the evaluation dataset, demonstrating a solid performance in distinguishing between "Hope" and "Not Hope" expressions. Table 14 presents detailed metrics, including precision, recall, and F1-score.

Class	Precision	Recall	F1-score	Support
Hope	0.81	0.87	0.84	899
Not Hope	0.88	0.82	0.85	1003
Accuracy		0.8449		
Macro Avg	0.85	0.85	0.84	1902
Weighted Avg	0.85	0.84	0.85	1902

Table 14
GPT-2 Binary Classification Report

The GPT-2 model shows fairly balanced performance across the two classes, with an F1-score of 0.84 for "Hope" and 0.85 for "Not Hope". The high recall for "Hope" (0.87) reflects the model's strength in identifying hopeful content, while the slightly lower recall for "Not Hope" (0.82) suggests the model occasionally misclassifies non-hopeful instances as hopeful.

Table 15 presents the confusion matrix. The model correctly predicted 783 of 899 "Hope" instances (87.1%) and 824 of 1003 "Not Hope" instances (82.1%), indicating generally reliable classification with room for improvement in reducing false positives and negatives.

These results suggest that GPT-2 is capable of capturing contextual signals relevant to hope classification, though improvements may be gained through fine-tuning with additional labeled data or context-enhanced training strategies.

	Hope	Not Hope
Hope	783	116
Not Hope	179	824

Table 15
Confusion Matrix for GPT-2 Binary Model

4.5.2. GPT-2 Multiclass Model Performance

The GPT-2 multiclass model achieved an overall accuracy of 77.76%, indicating reliable but slightly lower performance compared to SBERT in identifying various expressions of hope and non-hope. Tables 4.5.2 and 4.5.2 present the classification report and confusion matrix for the GPT-2 multiclass model.

Class	Precision	Recall	F1-Score	Support
Generalized Hope	0.70	0.71	0.71	467
Not Hope	0.86	0.84	0.85	816
Realistic Hope	0.52	0.66	0.58	196
Sarcasm	0.98	0.91	0.94	252
Unrealistic Hope	0.68	0.57	0.62	171
Accuracy		0.7776		
Macro Avg	0.75	0.74	0.74	1902
Weighted Avg	0.79	0.78	0.78	1902

Table 16
GPT-2 Multiclass Classification Report

"Not Hope" and "Sarcasm" achieved the highest F1-scores of 0.85 and 0.94, respectively, reflecting the model's effectiveness in classifying clearly defined emotional tones. "Generalized Hope" showed solid performance with an F1-score of 0.71, but "Realistic Hope" (0.58) and "Unrealistic Hope" (0.62) performed less reliably, highlighting challenges in detecting more nuanced or abstract expressions of hope. The macro and weighted averages (0.74 and 0.78, respectively) suggest balanced but somewhat conservative performance, likely influenced by the limited size and complexity of the dataset.

True / Pred	Generalized Hope	Not Hope	Realistic Hope	Sarcasm	Unrealistic Hope
Generalized Hope	332	64	60	1	10
Not Hope	73	689	33	4	17
Realistic Hope	37	17	130	0	12
Sarcasm	4	5	6	230	7
Unrealistic Hope	25	26	22	0	98

Table 17
GPT-2 Multiclass Confusion Matrix

The confusion matrix shows strong predictions for "Generalized Hope", "Not Hope", and "Sarcasm", while the majority of misclassifications are concentrated around the ambiguous "Realistic Hope" and "Unrealistic Hope" classes. For example, "Unrealistic Hope" was confused 26 times with "Not Hope" and 22 times with "Realistic Hope", suggesting semantic and contextual overlaps. Similarly, "Generalized Hope" was misclassified 60 times as "Realistic Hope".

In conclusion, while GPT-2 effectively distinguishes more explicit emotional tones like "Not Hope" and "Sarcasm", its performance on finer-grained hope categories indicates potential for improvement through contextual enhancement or targeted data augmentation.

Summary

The GPT-2 model demonstrated solid performance in both binary and multiclass emotion classification tasks, achieving accuracies of 84.49% and 77.76% respectively. It has a high capacity to model contextual dependencies and capture syntactic nuance, particularly evident in the strong classification outcomes

for the *Not Hope* (F1 = 0.85) and *Sarcasm* (F1 = 0.94) classes. Additionally, the model performed well on the *Generalized Hope* category, with an F1-score of 0.71, reflecting its general effectiveness in detecting overt hopeful sentiment.

However, GPT-2 exhibited limitations in accurately distinguishing finer-grained emotional categories such as *Realistic Hope* and *Unrealistic Hope*, which achieved relatively lower F1-scores of 0.58 and 0.62, respectively. These misclassifications likely stem from overlapping lexical features, ambiguous sentiment expressions, and a limited representation of these classes in the training data. The confusion matrix also revealed frequent mislabeling between these subtypes and with the *Not Hope* class, suggesting that the model struggled to unravel subtle affective cues embedded in colloquial or metaphorical tweets.

The results indicate that while GPT-2 benefits from pre-trained language modeling and exhibits strong generalization to clearly delineated emotional categories, its performance on nuanced emotional distinctions remains constrained by semantic ambiguity and class imbalance. Improvements such as domain-specific fine-tuning can improve the classification precision.

4.6. LSTM Model Analysis

4.6.1. LSTM Binary Model Performance

The LSTM model achieved an accuracy of 72% for binary classification. Tables 18 and 19 present the classification report and confusion matrix of the binary model classification, respectively.

Class	Precision	Recall	F1-score	Support
Hope	0.72	0.66	0.69	899
Not Hope	0.72	0.77	0.74	1003
Accuracy		0.7200		
Macro Avg	0.72	0.72	0.72	1902
Weighted Avg	0.72	0.72	0.72	1902

Table 18
LSTM Binary Classification Report

	Hope	Not Hope
Hope	597	302
Not Hope	234	769

Table 19
LSTM Binary Classification Confusion Matrix

The LSTM model demonstrates a moderate classification performance with an accuracy of 72%. The classification report seen in Table 18 reveals that the "Not Hope" class has a slightly better recall (0.77) compared to the "Hope" class (0.66), indicating that the model is more effective at identifying instances of the "Not Hope" class. The F1-score is balanced between the two classes, with the "Not Hope" class scoring 0.74 and the "Hope" class scoring 0.69.

From the confusion matrix seen in Table 19, we observe that the model correctly classifies 597 instances of the "Hope" class and 769 instances of the "Not Hope" class. However, 302 instances of the "Hope" class were misclassified as the "Not Hope" class, and 234 instances of the "Not Hope" class were misclassified as the "Hope" class. This suggests that while the model performs reasonably well, there is still room for improvement, particularly in reducing misclassification rates.

Future enhancements may include hyperparameter tuning, adding more training data, or using a more complex model architecture to improve performance.

4.6.2. LSTM Multi-class Model Performance

The LSTM model's performance across multiple classes demonstrates an overall accuracy of 59%. The classification report and confusion matrix are presented in Table 20 and Table 21. The classification

Class	Precision	Recall	F1-score	Support
Generalized Hope	0.51	0.55	0.53	467
Not Hope	0.69	0.68	0.68	816
Realistic Hope	0.29	0.35	0.32	196
Sarcasm	0.91	0.78	0.84	252
Unrealistic Hope	0.28	0.25	0.26	171
Accuracy		0.5900		
Macro Avg	0.54	0.52	0.53	1902
Weighted Avg	0.60	0.59	0.59	1902

Table 20
LSTM Classification Report for Multiclass Classification

	Generalized Hope	Not Hope	Realistic Hope	Sarcasm	Unrealistic Hope
Generalized Hope	255	110	66	4	32
Not Hope	132	553	57	13	61
Realistic Hope	68	46	68	2	12
Sarcasm	11	33	8	197	3
Unrealistic Hope	35	60	33	1	42

Table 21
LSTM Multiclass Confusion Matrix

report indicates that "Sarcasm" achieved the highest precision (0.91) and f1-score (0.84), suggesting it is the most effectively classified class. Conversely, "Unrealistic Hope" has the lowest recall (0.25) and f1-score (0.26), implying that the model struggles to correctly classify instances of this class.

The confusion matrix seen in Table 21 reveals that the "Not Hope", the largest category, is reasonably well classified, with 553 correct predictions out of 816 cases. However, there are significant misclassifications, particularly for the "Realistic Hope" and the "Unrealistic Hope" classes, where many instances are misclassified as other categories.

To improve performance, future work may explore increasing the dataset size, adjusting class weighting, fine-tuning hyperparameters, or employing a more advanced model architecture to better capture class-specific patterns.

Summary

The LSTM model displayed moderate performance in both binary and multiclass classification tasks, with accuracies of 72% and 59%, respectively. Its primary strength lies in its ability to capture temporal dependencies in sequential data, which contributed to relatively balanced F1-scores for the binary classification of *Hope* (0.69) and *Not Hope* (0.74). The model also showed reliable identification of explicitly defined emotions, such as *Sarcasm*, which achieved the highest multiclass F1-score of 0.84 and precision of 0.91. These results suggest that LSTM effectively captures clear sentiment signals embedded in tweet structures.

However, several limitations were observed. The multiclass model struggled with the classification of nuanced categories such as *Realistic Hope* and *Unrealistic Hope*, which received F1-scores of 0.32 and 0.26, respectively. The confusion matrices indicate a high rate of misclassification between these two categories and with the broader *Not Hope* class. This suggests that the model has difficulty learning fine-grained emotional distinctions due to overlapping vocabulary, implicit sentiment expressions, and underrepresentation of certain classes in the training set.

Moreover, the limited context retention in standard LSTM architectures can hinder the ability of the model to disambiguate subtle affective cues that are context-dependent or dispersed across longer tweet structures. The low recall for *Unrealistic Hope* (0.25) and *Realistic Hope* (0.35) indicates that the model tends to incorrectly generalize these classes and mistaking their ambiguous tone for more dominant or better-defined categories.

Overall, while the LSTM model exhibits adequate generalization for overt sentiment classification, its performance degrades significantly for semantically complex or sparsely represented emotional states. Future improvements may involve increasing training data diversity, implementing class rebalancing strategies, or adopting advanced architectures such as attention-based models that offer improved contextual representation and discriminative power.

4.7. Bi-LSTM Model Analysis

4.7.1. Bi-LSTM Binary Model Performance

The Bi-LSTM model achieved an overall accuracy of 79.34% on the evaluation dataset, indicating a solid performance in distinguishing between "Hope" and "Not Hope" expressions. Table 22 presents precision, recall, and F1-scores for each class.

Class	Precision	Recall	F1-score	Support
Hope	0.80	0.75	0.78	899
Not Hope	0.79	0.83	0.81	1003
Accuracy		0.7934		
Macro Avg	0.79	0.79	0.79	1902
Weighted Avg	0.79	0.79	0.79	1902

Table 22
Bi-LSTM Binary Classification Report

The Bi-LSTM model demonstrated relatively balanced performance across the two classes. The "Hope" class had a precision of 0.80 and a recall of 0.75, indicating the model's ability to detect hopeful expressions with reasonable accuracy, though some instances were misclassified. "Not Hope" showed slightly higher recall (0.83), suggesting a stronger ability to correctly identify non-hopeful content. Overall, the macro average F1-score of 0.79 reflects steady but improvable performance.

Table 23 presents the confusion matrix. The model correctly predicted 677 of 899 "Hope" instances (75.3%) and 832 of 1003 "Not Hope" instances (82.9%), showing that it was more likely to misclassify "Hope" as "Not Hope."

	Not Hope	Hope
Not Hope	832	171
Hope	222	677

Table 23
Confusion Matrix for Bi-LSTM Binary Model

These results suggest that while the Bi-LSTM model is competent at identifying expressions of hope and non-hope, it tends to produce more false negatives for hopeful content. Enhancing the training data with more representative hopeful examples or adding contextual embeddings could further improve its performance.

4.7.2. Bi-LSTM Multiclass Model Performance

The Bi-LSTM multiclass model achieved an overall accuracy of 63.35%, demonstrating moderate effectiveness in distinguishing between expressions of hope and non-hope. Tables 4.7.2 and 4.7.2 provide the classification report and confusion matrix for this model.

"Not Hope" and "Sarcasm" were classified most accurately by the Bi-LSTM model, with F1-scores of 0.73 and 0.85, respectively. In contrast, hope-related subcategories showed weaker performance. "Generalized Hope" achieved an F1-score of 0.54, while "Realistic Hope" and "Unrealistic Hope" lagged behind at 0.38 and 0.39, respectively. The relatively low precision and recall for these subtypes indicate that the model struggles to distinguish between nuanced hope expressions. The macro and weighted F1-scores of 0.58 and 0.64 highlight the imbalance in model performance across classes.

Class	Precision	Recall	F1-Score	Support
Generalized Hope	0.52	0.57	0.54	467
Not Hope	0.75	0.72	0.73	816
Realistic Hope	0.46	0.33	0.38	196
Sarcasm	0.88	0.83	0.85	252
Unrealistic Hope	0.34	0.47	0.39	171
Accuracy		0.6335		
Macro Avg	0.59	0.58	0.58	1902
Weighted Avg	0.65	0.63	0.64	1902

Table 24
Bi-LSTM Multiclass Classification Report

True / Pred	Generalized Hope	Not Hope	Realistic Hope	Sarcasm	Unrealistic Hope
Generalized Hope	266	106	42	8	45
Not Hope	120	586	23	13	74
Realistic Hope	72	27	64	5	28
Sarcasm	8	27	0	209	8
Unrealistic Hope	45	34	10	2	80

Table 25
Bi-LSTM Multiclass Confusion Matrix

From the confusion matrix, we observe that "Generalized Hope" is frequently misclassified as "Not Hope" (106 instances), while "Realistic Hope" is often confused with "Generalized Hope" and "Unrealistic Hope." Similarly, a large number of "Unrealistic Hope" predictions were incorrectly assigned to other categories. These misclassifications underscore the model's difficulty in capturing subtle semantic differences among hope-related classes. Overall, while the Bi-LSTM model performs reasonably on more distinct classes, improvements through advanced architectures or enriched context may be needed for finer-grained hope classification.

Summary

The Bi-LSTM model demonstrates competent performance in classifying emotional expressions in English tweets, particularly in the binary setting. With an accuracy of 79.34% in the binary classification task, the model shows balanced performance across the *Hope* and *Not Hope* classes. The model achieved F1-scores of 0.78 and 0.81 for *Hope* and *Not Hope*, respectively, reflecting its strength in modeling contextual dependencies in sequential data through bidirectional processing. The improved recall for *Not Hope* (0.83) compared to *Hope* (0.75) suggests a tendency to more accurately identify negative sentiment while occasionally misclassifying positive emotional content.

In the multiclass setting, the Bi-LSTM model achieved a moderate accuracy of 63.35%. While performance remained strong for more distinguishable categories such as *Not Hope* (F1-score = 0.73) and *Sarcasm* (F1-score = 0.85), the model exhibited noticeable weaknesses in recognizing nuanced hope-related subtypes. *Generalized Hope*, *Realistic Hope*, and *Unrealistic Hope* all achieved low F1-scores (0.54, 0.38, and 0.39, respectively). These discrepancies are supported by confusion matrix analysis, which revealed frequent misclassifications between semantically similar hope-related categories, such as confusion between *Generalized Hope* and *Not Hope*, and between *Realistic Hope* and *Unrealistic Hope*.

The primary weakness of the model appears to be its difficulty in distinguishing subtle semantic variations and overlapping sentiment expressions that often characterize hope-related subcategories. This limitation likely stems from insufficient class-specific features and the inherent ambiguity present in real-world tweets. The Bi-LSTM's reliance on sequential patterns may not fully capture the nuanced pragmatic or contextual cues necessary for such fine-grained distinctions.

Augmenting the dataset to balance underrepresented classes or incorporating syntactic and semantic enrichment could improve the model's generalization across complex emotional categories.

4.8. CNN Model Analysis

This report provides an evaluation of the CNN model trained for binary and multiclass classification tasks. The performance metrics, including accuracy, precision, recall, F1-score, and confusion matrix, are analyzed to assess the effectiveness of the model.

4.8.1. CNN Binary Model Performance

The performance of the CNN model was evaluated on the test dataset for binary classification, achieving an overall accuracy of 75.13%. This indicates a moderate ability to distinguish between "Hope" and "Not Hope" classes. The classification metrics, including precision, recall, and F1-score for both classes, are presented in Table 26.

Class	Precision	Recall	F1-score	Support
Hope	0.75	0.72	0.73	899
Not Hope	0.76	0.78	0.77	1003
Accuracy		0.7513		
Macro Avg	0.75	0.75	0.75	1902
Weighted Avg	0.75	0.75	0.75	1902

Table 26
CNN Binary Classification Report

The confusion matrix for the CNN binary model is presented in Table 27, showing the distribution of correct and incorrect classifications.

	Hope	Not Hope
Hope	647	252
Not Hope	221	782

Table 27
CNN Binary Confusion Matrix

The model demonstrates a reasonable balance between precision and recall for the "Hope" class, with a precision of 0.7454 and recall of 0.7197, resulting in an F1-score of 0.7323. For the "Not Hope" class, the precision is 0.7563, the recall is 0.7797, and F1-score is 0.7678. The macro average F1-score of 0.7500 indicates consistent performance across both classes. However, the accuracy of 75.13% is lower than that of BERT (85.12%) and SVM (78.50%). The confusion matrix reveals 252 "Hope" instances misclassified as "Not Hope" and 221 "Not Hope" instances misclassified as "Hope," indicating that the model struggles with some overlapping features between the classes. Future enhancements could involve adjusting convolutional filter sizes or incorporating additional contextual features to reduce these misclassification rates.

4.8.2. CNN Multi-Class Model Performance

The CNN model was evaluated on a multiclass classification task, achieving an overall accuracy of 59.20%. The classification metrics, including precision, recall, and F1-score for each class, are presented in Table 28, consistent with the format used for the binary classification report.

The confusion matrix for the CNN multiclass model is presented in Table 29, with rows representing predicted labels and columns representing true labels due to the matrix orientation matching the support values (e.g., true "Not Hope" sums to 816).

The model shows varying performance across classes. "Sarcasm" performs best with an F1-score of 0.73, achieving a high recall (0.96), though some "Not Hope" instances (119) are mistaken for "Sarcasm." "Not Hope" has an F1-score of 0.66, correctly classifying 492 out of 816 instances but misclassifying 100 as "Generalized Hope." "Generalized Hope" and "Realistic Hope" both have an F1-score of 0.48, with notable confusion between each other (92 instances of "Generalized Hope" predicted as "Realistic Hope").

Class	Precision	Recall	F1-Score	Support
Not Hope	0.72	0.60	0.66	816
Sarcasm	0.59	0.96	0.73	252
Generalized Hope	0.56	0.42	0.48	467
Realistic Hope	0.44	0.53	0.48	196
Unrealistic Hope	0.49	0.54	0.51	171
Accuracy		0.5920		
Macro Avg	0.6273	0.5920	0.5925	1902
Weighted Avg	0.62	0.59	0.59	1902

Table 28
CNN Multiclass Classification Report

	Not Hope	Sarcasm	Generalized Hope	Realistic Hope	Unrealistic Hope
Not Hope	492	119	100	51	54
Sarcasm	3	243	0	0	6
Generalized Hope	87	52	196	92	40
Realistic Hope	26	12	38	103	17
Unrealistic Hope	17	11	12	39	92

Table 29
CNN Multiclass Confusion Matrix

"Unrealistic Hope" has an F1-score of 0.51, with lower correct predictions (92/171). The overall accuracy of 59.20% and macro F1-score of 0.5925, compared to BERT (77.13%) and SVM (66.98%), suggest that the CNN struggles with nuanced distinctions, particularly for underrepresented classes. This may stem from its reliance on local patterns rather than broader contextual understanding. Future improvements could include increasing filter diversity, adding training data, or integrating transformer-based embeddings to enhance performance.

Summary

The CNN model exhibits moderate success in classifying emotional expressions within English tweets, with varied performance across binary and multiclass classification tasks. In the binary task, the model achieved an overall accuracy of 75.13%, with balanced precision and recall across both classes. The *Hope* class attained an F1-score of 0.73, while the *Not Hope* class achieved a slightly higher F1-score of 0.77. The confusion matrix reveals that a substantial number of instances (252 *Hope* and 221 *Not Hope*) were misclassified, indicating a challenge in resolving semantic overlap between the two classes. Despite outperforming simpler models the CNN model lags behind BERT and SBERT in binary classification, suggesting limitations in capturing long-range dependencies and deeper contextual cues from textual inputs.

In the multiclass classification task, the CNN model attained a modest accuracy of 59.20%. Performance varied significantly across classes, with *Sarcasm* demonstrating the highest classification effectiveness (F1-score = 0.73, recall = 0.96), likely due to its distinctive lexical and syntactic patterns. Conversely, *Generalized Hope*, *Realistic Hope*, and *Unrealistic Hope* exhibited lower F1-scores (0.48, 0.48, and 0.51, respectively), reflecting the model's difficulty in differentiating nuanced subtypes of hope. The confusion matrix reveals notable misclassifications between hope-related categories, such as 92 instances of *Generalized Hope* misclassified as *Realistic Hope*, and frequent confusion between *Not Hope* and *Sarcasm*.

These patterns suggest that the CNN's reliance on local n-gram features via convolutional filters is insufficient for modeling the complex, often contextually-dependent distinctions required for emotion detection in natural language. The architecture's inability to model long-distance dependencies and pragmatic cues likely contributes to its lower performance on nuanced categories, particularly where subtle semantic shifts differentiate emotional subtypes.

To enhance classification performance, increasing the volume and diversity of training data, especially for underrepresented hope-related subcategories, may further mitigate the observed misclassification

patterns.

5. Error Analysis

To gain deeper understanding of the behavior of the models, a comprehensive error analysis was conducted across binary and multiclass classification tasks. This section highlights the prevalent error patterns and offers hypotheses regarding their underlying causes.

5.0.1. Binary Classification

Across all binary classifiers, most misclassifications involved confusion between *Hope* and *Not Hope* instances with subtle or ambiguous cues. Notably, the LSTM model exhibited the weakest performance (accuracy = 72.00%), with a marked tendency to misclassify *Hope* instances as *Not Hope* (302 false negatives), likely due to its limited contextual encoding capacity. In contrast, SBERT achieved the highest binary accuracy (92.81%) and demonstrated balanced precision and recall for both classes. Its performance suggests that deeper semantic embeddings are more effective at distinguishing nuanced expressions of hope.

GPT-2 and BERT models also performed competitively, with accuracies of 84.49% and 85.12%, respectively. However, both showed a mild recall asymmetry: GPT-2 was slightly more prone to misclassify *Not Hope* as *Hope*, while BERT's errors were more evenly distributed. The Bi-LSTM and CNN models achieved moderate accuracy (79.34% and 75.13%), but both exhibited a slight bias toward *Not Hope* classification. Table 30 presents the comparison of the binary classification performance.

Model	Accuracy	Precision (Macro)	Recall (Macro)	F1-Score (Macro)
SBERT	**92.81%**	0.93	0.93	0.93
BERT	85.12%	0.85	0.85	0.85
GPT-2	84.49%	0.85	0.85	0.84
BiLSTM	79.34%	0.79	0.79	0.79
SVM	78.50%	0.79	0.78	0.78
CNN	75.13%	0.75	0.75	0.75
LSTM	72.00%	0.72	0.72	0.72

Table 30
Binary Classification Performance Comparison

5.0.2. Multiclass Classification

Multiclass classification revealed consistent challenges in distinguishing fine-grained subtypes of hope, particularly *Realistic Hope* and *Unrealistic Hope*, which were frequently confused or misclassified as *Not Hope*. The LSTM and CNN models struggled the most (accuracies = 59.00% and 59.20%), with low F1-scores for the hope-related categories. These models frequently confused *Generalized Hope* with *Not Hope* and *Realistic Hope*, suggesting difficulty in capturing subtle semantic distinctions without richer contextual understanding.

SBERT again achieved the best overall multiclass performance (accuracy = 84.29%), effectively separating all classes, especially *Sarcasm* (F1 = 0.98) and *Generalized Hope* (F1 = 0.83). GPT-2 and BERT followed with strong accuracies of 77.76% and 77.13%, respectively, but showed confusion between *Unrealistic Hope* and *Realistic Hope*. The Bi-LSTM model attained an accuracy of 63.35%, with relatively better performance for *Sarcasm* and *Not Hope* but poor discrimination between hope subcategories. Table 31 presents the comparison of the multiclass classification performance.

5.0.3. Common Error Patterns and Hypotheses

A recurring pattern across models was the misclassification of *Realistic Hope* as *Not Hope* or *Generalized Hope*, indicating ambiguity in linguistic markers of grounded optimism. Similarly, *Unrealistic Hope* was

Model	Accuracy	Precision (Macro)	Recall (Macro)	F1-Score (Macro)
SBERT	**84.29%**	0.80	0.80	0.80
GPT-2	77.76%	0.75	0.74	0.74
BERT	77.13%	0.75	0.73	0.74
SVM	66.98%	0.68	0.56	0.59
BiLSTM	63.35%	0.59	0.58	0.58
CNN	59.20%	0.63	0.59	0.59
LSTM	59.00%	0.54	0.52	0.53

Table 31
Multiclass Classification Performance Comparison

often confused with both *Not Hope* and *Generalized Hope*, possibly due to overlapping emotional tone and lack of concrete outcomes in the text.

Sarcastic expressions were accurately identified by models with stronger contextual representation (e.g., SBERT, BERT), but often misclassified by simpler models (LSTM, CNN), suggesting that sarcasm detection benefits from deeper semantic modeling.

Overall, error trends suggest that model capacity for semantic understanding, rather than surface-level token features alone, is critical in capturing the nuance required for this classification task. Future improvements may benefit from integrating external knowledge bases or fine-tuning on context-rich paraphrased data.

6. Conclusion

This study conducted a comprehensive evaluation of seven machine learning and deep learning models—SVM, BERT, SBERT, GPT-2, LSTM, Bi-LSTM, and CNN—for both binary and multiclass emotion classification tasks centered on detecting expressions of hope in text. The models were assessed based on classification accuracy, precision, recall, F1-score, and confusion matrices, with additional error analysis to elucidate common failure patterns and linguistic challenges.

Across all models, the SBERT model consistently outperformed others in both binary (92.81%) and multiclass (84.29%) settings. Its superior performance is attributed to its capacity to capture fine-grained sentence-level semantics, allowing for robust discrimination even among closely related categories such as *Realistic Hope*, *Unrealistic Hope*, and *Sarcasm*. BERT and GPT-2 also demonstrated strong performance, in binary classification, with accuracies of 85.12% and 84.49%, respectively. However, both showed limited ability to differentiate subtle hope-related nuances in the multiclass task.

Models such as SVM and CNN performed moderately well, with binary accuracies of 78.50% and 75.13%, respectively. Their performance in multiclass classification (66.98% for SVM and 59.20% for CNN) was hampered by difficulty in modeling complex contextual relationships, particularly in under-represented or overlapping classes. The LSTM and Bi-LSTM models achieved moderate results, with Bi-LSTM outperforming its unidirectional counterpart, suggesting that bidirectional context improves sensitivity to semantic variation, though not sufficiently for fine-grained classification.

Error analysis revealed that most models struggled with distinguishing between *Realistic Hope* and *Unrealistic Hope*, often misclassifying them as *Not Hope* or *Generalized Hope*. This suggests that subtle linguistic cues and implicit sentiment present significant challenges, especially for models lacking contextual depth. Moreover, sarcastic expressions were more reliably detected by transformer-based models, underscoring the importance of deep contextual understanding for nuanced sentiment analysis.

Overall, our findings emphasize the effectiveness of transformer-based architectures—especially SBERT—for tasks requiring semantic sensitivity and subtle affect detection.

6.1. Future Work

Overall, each model has its own strengths, and the choice of model depends on the specific classification task at hand. To achieve higher accuracy and robustness in both binary and multiclass classification

tasks, future improvements should focus on several strategic directions. For SVM, exploring non-linear kernel methods, such as Gaussian or polynomial kernels, and integrating contextual embeddings like BERT or SBERT as input features could better capture complex linguistic patterns, particularly for underrepresented classes. For transformer-based models like BERT, SBERT, and GPT-2, fine-tuning on larger, domain-specific datasets with a focus on underrepresented hope categories (e.g., "Realistic Hope" and "Unrealistic Hope") could mitigate current limitations, while experimenting with advanced techniques like few-shot learning or prompt engineering may enhance adaptability to low-resource settings. Additionally, leveraging multimodal data—such as combining text with user metadata, emojis, or temporal features—could provide richer context for all models, especially for CNN, which struggles with nuanced distinctions due to its reliance on local patterns; integrating attention mechanisms or hybrid CNN-transformer architectures could further improve its contextual understanding. For LSTM and Bi-LSTM, addressing their challenges in capturing long-range dependencies could involve incorporating attention mechanisms or exploring hybrid models that combine LSTM/Bi-LSTM with transformer layers to better model sequential and contextual relationships. Data augmentation strategies, such as back-translation, paraphrasing, or synthetic data generation using generative models like GPT-2, should be prioritized to balance the dataset and improve performance on minority classes across all models. Finally, developing ensemble approaches that combine the strengths of classical models (e.g., SVM), deep learning models (e.g., LSTM, CNN), and transformer-based models (e.g., BERT, SBERT) could yield more robust predictions, potentially through techniques like stacking or weighted voting, while also exploring the integration of external knowledge bases (e.g., sentiment lexicons or emotion ontologies) to enhance semantic understanding and classification accuracy.

Declaration on Generative AI

This manuscript benefited from generative AI tools for linguistic enhancement. The research design, data, and conclusions are entirely the work of the authors. The authors take full responsibility for the accuracy and integrity of the content.

References

[1] Uzor, Gods Gift G and Vadapalli, Hima B, Smartening E-therapy using Facial Expressions and Deep Learning, 2nd International Multidisciplinary Information Technology and Engineering Conference, IMITEC 2020 (2020) 9334115.

[2] Balouchzahi, Fazlourrahman and Sidorov, Grigori and Gelbukh, Alexander, PolyHope: Two-level hope speech detection from tweets, Expert Systems with Applications 225 (2023) 120078. doi:10.1016/j.eswa.2023.120078.

[3] González-Barba, José Ángel and Chiruzzo, Luis and Jiménez-Zafra, Salud María, Overview of IberLEF 2025: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS.org, 2025.

[4] Meyer, David and Wien, FT, Support vector machines, R News 1 (2001) 23–26.

[5] Hochreiter, Sepp and Schmidhuber, Jürgen, Long short-term memory, Neural Computation 9 (1997) 1735–1780.

[6] Schuster, Mike and Paliwal, Kuldip K., Bidirectional Recurrent Neural Networks, IEEE Transactions on Signal Processing 1 (1997) 2673–2681. doi:10.1109/78.650093.

[7] LeCun, Yann and Boser, Bernhard and Denker, John S and Henderson, Donnie and Howard, Richard E and Hubbard, Wayne and Jackel, Lawrence D, Backpropagation applied to handwritten zip code recognition, Neural Computation 1 (1989) 541–551.

[8] Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[9] Reimers, Nils and Gurevych, Iryna, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, arXiv preprint arXiv:1908.10084, 2019. URL: https://arxiv.org/abs/1908.10084.

[10] Radford, Alec and Wu, Jeffrey and Child, Rewon and Luan, David and Amodei, Dario and Sutskever, Ilya, Language Models are Unsupervised Multitask Learners, OpenAI Blog, 2019. URL: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

[11] Chakravarthi, Bharathi Raja, HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion, in: Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media, Association for Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 41–53. URL: https://aclanthology.org/2020.peoples-1.5.

[12] García-Baena, David and García-Cumbreras, Miguel Ángel and Jiménez-Zafra, Salud María and García-Díaz, José Antonio and Valencia-García, Rafael, Hope Speech Detection in Spanish: The LGTB Case, Language Resources and Evaluation (2023) 1–31. doi:10.1007/s10579-023-09688-7.

[13] Sidorov, Grigori and Balouchzahi, Fazlourrahman and Butt, Sabur and Gelbukh, Alexander, Regret and Hope on Transformers: An Analysis of Transformers on Regret and Hope Speech Detection Datasets, Applied Sciences 13 (2023) 3983. doi:10.3390/app13063983.

[14] Balouchzahi, Fazlourrahman and Butt, Sabur and Amjad, Maryam and Sidorov, Grigori and Gelbukh, Alexander, UrduHope: Analysis of hope and hopelessness in Urdu texts, Knowledge-Based Systems 308 (2025) 112746. doi:10.1016/j.knosys.2025.112746.

[15] Sidorov, Grigori and Balouchzahi, Fazlourrahman and Ramos, Luis and Gómez-Adorno, Helena and Gelbukh, Alexander, MIND-HOPE: Multilingual Identification of Nuanced Dimensions of HOPE, 2024. Unpublished manuscript.

[16] Chakravarthi, Bharathi Raja and Muralidaran, Vigneshwaran and Priyadharshini, Ruba and Cn, Subalalitha and McCrae, John P. and García, Miguel Ángel and Jiménez-Zafra, Salud María and Valencia-García, Rafael and Kumaresan, Prasanna and Ponnusamy, Rahul and García-Baena, David and García-Díaz, José, Overview of the shared task on hope speech detection for equality, diversity, and inclusion, in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, Association for Computational Linguistics, 2022, pp. 378–388. doi:10.18653/v1/2022.ltedi-1.58.

[17] García-Baena, David and Balouchzahi, Fazlourrahman and Butt, Sabur and García-Cumbreras, Miguel Ángel and Tonja, Atnafu Lambebo and García-Díaz, José Antonio and Jiménez-Zafra, Salud María and Others, Overview of HOPE at IberLEF 2024: Approaching hope speech detection in social media from two perspectives, for equality, diversity and inclusion and as expectations, Procesamiento del Lenguaje Natural 73 (2024) 407–419. doi:10.26342/2024-73-26.

[18] Butt, Sabur and Balouchzahi, Fazlourrahman and Amjad, Maryam and Jiménez-Zafra, Salud María and Ceballos, Hugo Jair Escalante and Sidorov, Grigori, Overview of PolyHope at IberLEF 2025: Optimism, Expectation or Sarcasm?, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), CEUR-WS.org, Valladolid, Spain, 2025.

[19] Butt, Sabur and Balouchzahi, Fazlourrahman and Amjad, Ali Imran and Amjad, Maryam and Ceballos, Hugo Jair Escalante and Jiménez-Zafra, Salud María, Optimism, Expectation, or Sarcasm? Multi-Class Hope Speech Detection in Spanish and English, ResearchGate, 2025. URL: https://www.researchgate.net/publication/379058018. doi:10.13140/RG.2.2.19761.90724.

[20] Balouchzahi, F and Sidorov, G and Gelbukh, A, Polyhope: Two-level hope speech detection from tweets, DOI: 10.48550, arXiv preprint ARXIV.2210.14136 (2022).