Emotion Detection using BiLSTM on Social Media

Samia Rahman Misty
School of Computer Science and Electronic Engineering, University of Essex
Email: sm23788@essex.ac.uk

***Abstract*—Emotion recognition in digital communications offers valuable insights into social dynamics, and deep learning is a promising technique for enhancing this analysis. In this study, we propose a deep learning approach for emotion classification in Twitter messages using Twitter dataset. It labels six emotions: sadness, joy, love, anger, fear, and surprise. The data preprocessing includes cleaning the text, expanding abbreviations, removing stopwords, stemming, and eliminating non-alphabetic characters and URLs. The data is tokenized and padded before being processed by a Bidirectional Long Short-Term Memory (BiLSTM) network, achieving 92.06% accuracy with Stratified K-fold cross-validation. The study highlights BiLSTM's effectiveness for emotion analysis in social media.**

***Index Terms*—Emotion detection, Deep Learning, BiLSTM.**

I. Introduction

Emotion recognition has garnered significant research interest due to its applications in mental health and human-computer interaction. This study focuses on classifying emotions in Twitter messages using a Bidirectional Long Short-Term Memory (BiLSTM) model, trained on 416,809 labeled tweets. BiLSTM's ability to capture sequential text dependencies makes it ideal for emotion detection in social media [6].

Recent advancements highlight diverse approaches. Saadon et al. [9] proposed facial emotion detection using digital image speckle correlation, while Mellouka et al. [7] achieved over 90% accuracy in facial recognition with CNN-LSTM models. Huang et al. [3] improved facial emotion recognition accuracy from 77.37% to 83.37% using transfer learning. Binali et al. [1] explored hybrid systems for online emotion classification. Desmet et al. [2] applied sentiment mining for suicide prevention, achieving F-scores of 68.86%, with potential for improvement.

The paper is organized as follows: Section II explores the data characteristics, while Section III details the methodology. The evaluation metrics are presented in Section IV, and Section V concludes with future directions.

II. Dataset & Feature Analysis

Twitter dataset [4] contains 416,809 Twitter messages, labeled with six emotions, with the following distribution: joy (33.8%), sadness (29.1%), anger (13.8%), fear (11.5%), love (8.3%), and surprise (3.6%).

Table I summarizes key linguistic features, offering insights into various aspects of language analysis. Phonological features highlight the prevalence of shorter syllable counts, while the Morphological section reflects structural diversity with millions of root words and thousands of prefixes and suffixes. Syntactic patterns emphasize the dominance of nouns, adjectives, and verbs, showcasing descriptive and action-oriented expressions. The Semantic features reveal rich synonym networks and effective named entity recognition. Pragmatic markers such as "still" and "also" indicate nuanced transitions, while the Stylistic section highlights concise sentence structures and high lexical diversity. This overview captures essential patterns critical for linguistic studies.

TABLE I: Summary of Linguistic Features

Feature Type	Details
Phonological	**Syllable Count:** Common: 3–20 syllables; Less frequent: 20–40 syllables; Rare: $\geq$ 40 syllables.
Morphological	**Root Words Analysis:** Root Words: 386K; Lemmatized: 27K. **Word Formation:** Root Words: 379K; Prefixes: 30K; Suffixes: 44K; Pluralization: 27K.
Syntactic	**POS Tags:** NN: 1.89M, JJ: 715K, VB: 152K; RB, IN, VBD <199K. **SVO:** Subjects: 475K; Verbs: 991K; Objects: 381K.
Semantic	**Feature Distribution:** Synonyms: 13.2M, Antonyms: 2.6M, Hypernyms: 10.6M. **NER Distribution:** Person: 41K; Organization: 16.6K; Location: 9.3K.
Pragmatic	**Markers:** "Still" & "Also": >8K; Others: <4K. **Speech Acts:** Statements: 409K; Requests: 7.3K.
Lexical	**Common Words:** "Feel" (58.7%), "Like" (10.2%), "I'm" (8.2%); Others: <(3.5%). (Fig. 1) **2-grams:** "Feel like" (50K); "I'm feel" (20K+) ; Others: $\leq$ (10K). *Note: The calculations for common words are based on a comparison of the top 10 most frequent words.*
Stylistic	**Sentence Length :** <20 words dominate. (Fig. 2) **Voice:** 100% Active. **TTR:** 0.97 (High diversity).

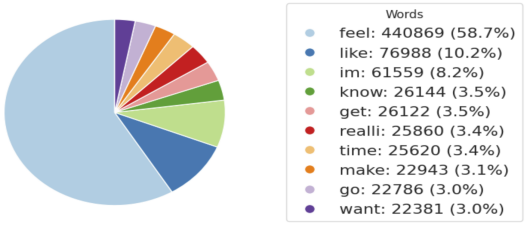

Fig. 1: Top 10 Most Common Words (Non-Lemmatized)

III. Methodology

A. Data Pre-processing

The following checks and transformations ensure the dataset's suitability for emotion classification -

- **Chat Abbreviations:** A dictionary was used to identify and expand common abbreviations, but no significant abbreviations are found that required modification.

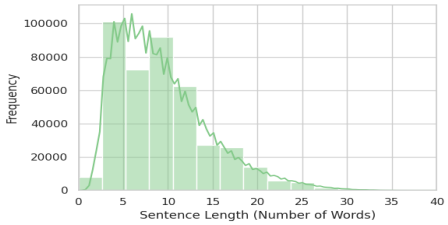

Fig. 2: Sentence Length Distribution

- **Unnecessary Columns:** An extraneous index column labeled 'Unnamed' is identified and removed.
- **Stopwords:** English stopwords are checked, and 414,302 rows containing stopwords are removed to reduce noise.
- **Text Standardization:** All text is converted to lowercase for consistency.
- **Text Cleaning:** Non-alphabetic characters, newline characters, numeric values, punctuation, special characters, duplicate rows, NaN values, and URLs are examined and removed.
- **Stemming:** Words are reduced to their root forms (e.g., "running" to "run") to maintain uniformity and minimize variability within the dataset.

The pre-processing steps ensure the dataset is clean, consistent, and ready for analysis by standardizing the data, removing irrelevant information, and reducing noise. After pre-processing, the dataset is split into 90% for training and 10% for testing, improving the accuracy and relevance of the emotion classification model.

B. Classification Method

In this study, we focus on classifying emotions in Twitter messages by leveraging the sequential nature of text data. Recognizing the importance of understanding both past and future word dependencies, we utilize the Bidirectional Long Short-Term Memory (BiLSTM) model. BiLSTM is particularly suited for emotion detection as it captures context from both preceding and succeeding words, enhancing the understanding of nuanced emotional cues [6].

Mathematically, BiLSTM operates at each time step t:

- The forward LSTM computes a hidden state $\overrightarrow{h_t}$:

$$\overrightarrow{h_t} = \text{LSTM}_{\text{forward}}(x_t, \overrightarrow{h_{t-1}}) \tag{1}$$

- The backward LSTM computes a hidden state $\overleftarrow{h_t}$:

$$\overleftarrow{h_t} = \text{LSTM}_{\text{backward}}(x_t, \overleftarrow{h_{t+1}}) \tag{2}$$

The combined hidden state is:

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \tag{3}$$

where $[\cdot;\cdot]$ denotes concatenation of the forward and backward states.

This structure allows BiLSTM to capture bidirectional dependencies, offering a richer context for emotion classification. Pre-processed Twitter messages are tokenized into sequences of integers and padded for a uniform input size, allowing effective training of the BiLSTM model.

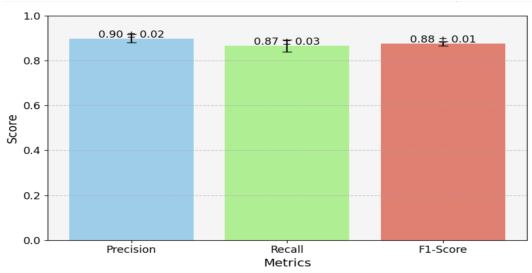

Fig. 3: Mean PRF with Standard Deviation

1) Model Architecture: The model architecture is as follows:

- **Input Layer:** Receives tokenized and padded sequences of preprocessed Twitter messages.
- **Embedding Layer:** Maps input tokens to 100-dimensional dense vectors to capture semantic similarities.
- **Bidirectional LSTM Layer:** 128 units that learn dependencies from both past and future sequences for enhanced context.
- **Batch Normalization:** Stabilizes training by normalizing hidden layer activations, improving generalization.
- **Dropout Regularization:** A rate of 0.5 to prevent overfitting by randomly disabling neurons during training.
- **Fully Connected (Dense) Layer:** 64 units with ReLU activation to learn complex patterns.
- **Output Layer:** 6 units, one for each emotion, with softmax activation to provide probabilities.
- **Compilation:** Compiled with the Adam optimizer and sparse categorical cross-entropy loss, monitoring accuracy.

IV. Experiments & Evaluation

A. Training Evaluation Metrics

To evaluate the performance and robustness of the emotion classification model, we use stratified 5-fold cross-validation. This approach ensures that each fold maintains the same class distribution as the entire dataset, reducing bias and providing a reliable estimate of the model's generalization.

1) Overall Metrics (Mean ± Standard Deviation): The mean and standard deviation of key evaluation metrics across the five folds are as follows:

- **Precision:** 0.90 ± 0.02 — 90% of positive predictions are correct, with low variability across folds.
- **Recall:** 0.87 ± 0.03 — 87% of actual positives are identified, with slight variability across folds.
- **F1-Score:** 0.88 ± 0.01 — A balanced metric combining precision and recall, reflecting consistent performance.

Fig. 3 summarizes the overall metrics of model's consistency and effectiveness across the folds.

2) Performance Metrics Analysis: Table II displays consistent performance across all folds, with minimal variability in precision, recall, and F1-Score. The Matthews Correlation Coefficient (MCC) is highest in Fold 2, reflecting strong

Fold	Precision	Recall	F1-Score	MCC	Perplexity
Fold 1	0.879	0.900	0.883	0.883	1.161
Fold 2	0.896	0.860	0.875	0.975	1.160
Fold 3	0.898	0.884	0.884	0.883	1.162
Fold 4	0.870	0.919	0.890	0.891	1.1645
Fold 5	0.879	0.903	0.887	0.886	1.166

TABLE II: Metrics Per Fold for Precision, Recall, F1-Score, Matthews Correlation Coefficient (MCC), and Perplexity.

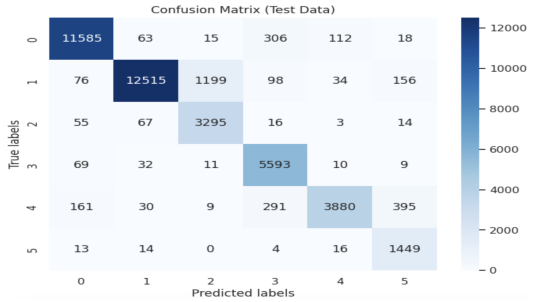

Fig. 4: Confusion Matrix of Test Dataset

classification quality. Stable and low perplexity scores further indicate the model's efficiency and reliability for emotion classification tasks.

B. Testing Evaluation Metrics

1) Accuracy and Confusion Matrix: The model achieves 92.06% accuracy on the test set, demonstrating strong generalization. Fig. 4 shows the confusion matrix and confirms minimal misclassifications, reflecting reliable performance across emotional categories.

2) ROC-AUC Curve: Fig. 5 shows an AUC of 1.00 for all classes, except class 2 (Love), which has 0.99, indicating excellent discriminative power with negligible performance loss for class 2.

C. Analysis

The emotion classification model shows strong performance with high precision, recall, and F1-score metrics during training. Precision (0.90) and recall (0.87) reflect the model's ability to accurately identify positive instances, with minimal variation across folds. The F1-score of 0.88 ensures a balanced performance, maintaining a good trade-off between precision and recall. The Matthews Correlation Coefficient (MCC) values, especially the highest MCC of 0.975 in Fold 2, indicate strong classification quality. Stable perplexity scores further suggest model efficiency.

On the test set, the model achieves 92.06% accuracy, confirming its generalization ability. These scores are derived using built-in methods from Keras for deep learning models [5] and from Scikit-learn for machine learning models [10]. The confusion matrix shows balanced performance with few misclassifications, while the ROC-AUC curve shows excellent class separation (AUC of 1.00 for most classes, 0.99 for class 2), indicating the model's ability to distinguish between emotions effectively.

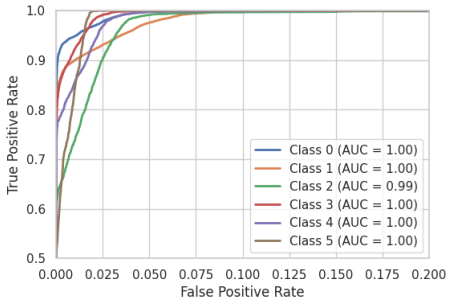

Fig. 5: ROC-AUC Curve for All Classes

The model demonstrates consistent and reliable emotion classification, with strong performance on both training and test data, making it suitable for applications like social media analysis and mental health monitoring.

V. Conclusion & Future Works

The proposed model demonstrates strong performance in emotion classification, with high precision, recall, and F1-scores across multiple folds, indicating its effectiveness in predicting emotions such as joy, sadness, and surprise. The stable results and solid generalization to unseen data reflect its reliability.

To further enhance its capabilities, incorporating pre-trained word embeddings like GloVe [8] can improve the model's understanding of word semantics and contextual meaning. Additionally, exploring advanced architectures like Transformer-based models, such as BERT or GPT, could enable the model to capture more nuanced emotions and better handle the complexities of social media data, leading to even more accurate and context-aware predictions.

References

[1] H. Binali, C. Wu, and V. Potdar. Computational approaches for emotion detection in text. In *4th IEEE international conference on digital ecosystems and technologies*, pages 172–177. IEEE, 2010.

[2] B. Desmet and V. Hoste. Emotion detection in suicide notes. *Expert Systems with Applications*, 40(16):6351–6358, 2013.

[3] Z.Y. Huang, C.C. Chiang, J.H. Chen, Y.C. Chen, H.L. Chung, Y.P. Cai, and H.C. Hsu. A study on computer vision for facial emotion recognition. *Scientific Reports*, 13(1):8425, 2023.

[4] Kaggle. Emotions dataset. https://www.kaggle.com/datasets/nelgiriyewithana/emotions, 2024. [Accessed: Jan. 1, 2025].

[5] Keras. Keras: The python deep learning api. https://keras.io/, 2024. [Accessed: Jan. 30, 2024].

[6] G. Liu and J. Guo. Bidirectional lstm with attention mechanism and convolutional layer for text classification. *Neurocomputing*, 337, 2019.

[7] W. Mellouk and W. Handouzi. Facial emotion recognition using deep learning: review and insights. *Procedia Computer Science*, 175:689–694, 2020.

[8] J. Pennington, R. Socher, and C.D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[9] J.R. Saadon, F. Yang, R. Burgert, S. Mohammad, T. Gammel, M. Sepe, M. Rafailovich, C.B. Mikell, P. Polak, and S. Mofakham. Real-time emotion detection by quantitative facial motion analysis. *Plos one*, 18(3):e0282730, 2023.

[10] Scikit-learn. Scikit-learn: Machine learning in python. https://scikit-learn.org/stable/index.html, 2024. [Accessed: Jan. 30, 2024].