Emotion Recognition in Chat Conversations

Mid-Term Project Report

Group ID: 25PB15
Department of Computer Science and Engineering, IIT Patna
Course: CS3103 - Machine Learning

Team Members:
Ansh Prem (2302CS01) - *Literature Review & Planning*
Himanshu Kumar (2301AI39) - *Data Analysis (Preprocessing)*
Atul Raj Chaudhary (2301AI40) - *Model Development (RoBERTa)*
Abhay Pratap Singh (2301AI48) - *Model Development(DistilBERT / & BiLSTM)*
Sparsh Rastogi (2301A156) - *Stacking and Model Evaluation*

November 11, 2025

1 Introduction

Emotions play a vital role in human communication, influencing thoughts, behavior, and decision-making. In text-based conversations such as messaging platforms, customer support chats, and educational forums, understanding emotions helps enhance communication quality and user experience. However, detecting emotions from text is challenging due to the absence of tonal and facial cues.

This project, ”Emotion Recognition in Chat Conversations”, aims to design and implement machine learning models that can classify emotions such as joy, anger, sadness, fear, surprise, and neutral from conversational text. Multiple models are developed and compared, including baseline machine learning classifiers, an LSTM-based sequential model, and BERT-based contextual transformers. The ultimate objective is to identify the best-performing model or an ensemble of models for accurate, real-time emotion detection.

2 Brief Related Work

The field of emotion recognition in text is an active area of research, with significant advancements moving from traditional machine learning to complex deep learning architectures.

Several studies highlight the dominance of transformer models. A comparative study of BERT, RoBERTa, and DistilBERT found that transformer-based approaches consistently outperform traditional techniques, especially on contextually rich dialog datasets [3]. This is reinforced by work on BERT-based models for digital mental health, which achieved high accuracy in classifying emotions from chat logs to provide adaptive feedback [2].

Other research focuses on the complexity of conversational context. For instance, SemEval-2024 Task 3 emphasized the need for multimodal pipelines (text, audio, visual) to jointly detect both emotions and their causes in dialogues [1]. Similarly, Płaza et al. (2022) explored emotion recognition in contact center systems to create behavioral profiles for agents and clients, noting the challenges of working with automatic transcriptions [6].

Finally, alternative methodologies are also explored. Cardone et al. (2023) proposed a fuzzy-based classification method to analyze user reviews [7], while Yuming et al. (2024) developed an explanation framework based on psychological theories to make the ”black box” nature of AI emotion analysis more transparent [9]. Our project builds on these foundations by comparing several strong modern baselines on a common task.

3 Methodology

3.1 Data Source and Preprocessing

The data was sourced from text files, divided into `train.txt`, `val.txt`, and `test.txt`. The dataset contains text samples labeled with six emotions: joy, sadness, anger, fear, love, and surprise.

A rigorous preprocessing pipeline was applied to all datasets:

- **Lowercasing:** All text was converted to lowercase.
- **Punctuation and Number Removal:** All punctuation marks and digits were removed.
- **URL Removal:** Any web URLs were stripped from the text.
- **Stopword Removal:** Common English stopwords (e.g., ”is”, ”the”, ”a”) were removed using the NLTK library.
- **Lemmatization:** Words were reduced to their base or dictionary form (e.g., ”feeling” → ”feel”) using the NLTK `WordNetLemmatizer`.
- **Duplicate Removal:** Any fully duplicate rows or rows with identical text but different labels were removed to ensure data quality.

3.2 Modeling Approaches

Four distinct modeling strategies were implemented and compared.

3.2.1 Baseline Models with TF-IDF

As a baseline, traditional machine learning models were trained on TF-IDF (Term Frequency-Inverse Document Frequency) vector representations of the text. The models included:

- Logistic Regression
- Decision Tree Classifier
- Support Vector Machine (SVC)
- Random Forest Classifier

3.2.2 Bidirectional LSTM (BiLSTM)

A sequential deep learning model was built using TensorFlow/Keras. This model captures word order and temporal dependencies. The architecture is as follows:

1. **Embedding Layer:** Converts integer-tokenized text into dense 128-dimension vectors.

2. **Bidirectional LSTM Layer:** A 64-unit LSTM that processes the sequence in both forward and backward directions to capture context from both ends.
3. **Dropout Layer:** A 50% dropout for regularization to prevent overfitting.
4. **Dense Output Layer:** A 6-unit output layer with `softmax` activation to classify into the six emotion categories.

The LSTM architecture is defined by its internal gates:

$$\begin{aligned}
f_t &= \sigma(W_f[h_{t-1}, x_t] + b_f) && \text{(Forget Gate)} \\
i_t &= \sigma(W_i[h_{t-1}, x_t] + b_i) && \text{(Input Gate)} \\
\tilde{C}_t &= \text{tanh}(W_C[h_{t-1}, x_t] + b_C) && \text{(Candidate Cell State)} \\
C_t &= f_t * C_{t-1} + i_t * \tilde{C}_t && \text{(Cell State Update)} \\
o_t &= \sigma(W_o[h_{t-1}, x_t] + b_o) && \text{(Output Gate)} \\
h_t &= o_t * \text{tanh}(C_t) && \text{(Hidden State)}
\end{aligned}$$

3.2.3 Transformer Models (Bagged)

Fine-tuned transformer models from the Hugging Face library were used for their powerful contextual understanding. To improve stability and performance, an ensemble bagging approach was used:

- **DistilBERT:** The `distilbert-base-uncased` model was fine-tuned 3 times with different random seeds (42, 1, 27). The final prediction is the average of the probabilities from these three models.
- **RoBERTa:** The `roberta-base` model was also fine-tuned 3 times with the same seeds, and its predictions were averaged.

All models were trained for 3 epochs using the `Trainer` API.

3.2.4 Ensemble Stacking

The most advanced approach combined the predictions of the deep learning models. A meta-classifier (Logistic Regression) was trained using the probability outputs (logits) of the BiLSTM, bagged DistilBERT, and bagged RoBERTa models as its input features. This stacking model learns to weigh the predictions from each base model to make a final, more accurate classification.

4 Key Results

All models were evaluated on the held-out test set. The Random Forest model was the strongest baseline, but all deep learning models significantly outperformed it. The ensemble stacking model achieved the highest accuracy.

The final stacking model not only had the highest overall accuracy but also showed strong and balanced performance across all emotion classes, as shown in its classification report (Table 2).

Table 1: Model Accuracy Comparison on Test Set

Model	Accuracy
Baseline Models (TF-IDF):	
Logistic Regression	86.85%
Decision Tree	86.20%
Support Vector Machine (SVC)	86.75%
Random Forest	88.35%
Deep Learning Models:	
Bidirectional LSTM (BiLSTM)	91.30%
DistilBERT (3 seeds)	91.30%
RoBERTa (3 seeds)	92.65%
Ensemble:	
Stacking (LR Meta-Model)	**93.40%**

Table 2: Classification Report for Stacking Meta-Classifier (Test Set)

Emotion	Precision	Recall	F1-Score	Support
Anger	0.93	0.94	0.94	275
Fear	0.91	0.91	0.91	224
Joy	0.94	0.97	0.95	695
Love	0.87	0.80	0.83	159
Sadness	0.97	0.97	0.97	581
Surprise	0.78	0.70	0.74	66
Accuracy			**0.93**	**2000**
Macro Avg	0.90	0.88	0.89	2000
Weighted Avg	0.93	0.93	0.93	2000

5 Conclusion

This project successfully demonstrated the effectiveness of various machine learning models for emotion recognition in chat conversations. While traditional TF-IDF models provided a robust baseline (up to 88.35% accuracy with Random Forest), deep learning architectures offered superior performance.

The BiLSTM model, capturing sequential information, achieved 91.30% accuracy. Bagged transformer models, leveraging pre-trained contextual knowledge, performed even better, with DistilBERT reaching 93.20%.

The highest accuracy of 93.40% was achieved by a stacking ensemble model. This model used the outputs from the BiLSTM, DistilBERT, and RoBERTa models as features for a final Logistic Regression classifier. This result highlights that combining the predictive power of diverse, strong models (sequential, transformer, etc.) through ensembling can yield state-of-the-art results for this task.

6 References

References

[1] SemEval-2024 Task 3: Multimodal Emotion Cause Analysis in Conversations. https://arxiv.org/abs/2405.13049

[2] BERT-Based Emotion Detection and Content Recommendation for Digital Mental Health. https://ijsrem.com/download/bert-based-emotion-detection-and-real-time-content-recommendation-system-for-

[3] Emotion Detection with Transformers: A Comparative Study. https://arxiv.org/pdf/2403.15454.pdf

[4] Multimodal Sentiment Analysis Integrating Text, Audio, and Video. https://ieeexplore.ieee.org/document/10863949/

[5] Karunya, S. G., & Sathish, A. (2025). Intelligent emotion sensing using BERT BİLSTM and generative AI for proactive customer care. *Scientific Reports, 15*, 34192.

[6] Płaza, M. et al. (2022). Emotion recognition method for call/contact centre systems. *Applied Sciences, 12*(21), 10951.

[7] Cardone, B., Di Martino, F., & Miraglia, V. (2023). A fuzzy-based emotion detection method to classify the relevance of pleasant/unpleasant emotions posted by users in reviews of service facilities. *Applied Sciences, 13*(10), 5893.

[8] Machova, K., Szaboova, M., Paralic, J., & Micko, J. (2023). Detection of emotion by text analysis using machine learning. *Frontiers in Psychology, 14*, 1190326.

[9] Li, Y., Chan, J., Peko, G., & Sundaram, D. (2024). An explanation framework and method for AI-based text emotion analysis and visualisation. *Decision Support Systems, 178*, 114121.