

EMOTION RECOGNITION IN CHAT CONVERSATIONS

Course: Machine Learning Project

Group ID: 25PB15

- Ansh Prem - 2302CS01
- Abhay Pratap Singh - 2301AI48
- Atul Raj Chaudhary – 2301AI40
- Himanshu Kumar – 2301AI39
- Sparsh Rastogi - 2301AI52

Roles & Responsibilities

- **Ansh Prem:** Literature Review, Planning
- **Abhay Pratap Singh:** Model Development (DistilBert & BiLSTM)
- **Atul Raj Chaudhary:** Model Development (Roberta)
- **Himanshu Kumar:** Data Analysis (Preprocessing)
- **Sparsh Rastogi:** Stacking and Model Evaluation

INTRODUCTION: PROBLEM & MOTIVATION

Problem Statement:

- To accurately classify short text passages into one of six distinct emotional categories: **Joy, Sadness, Anger, Fear, Love, or Surprise.**
- **Why is this important? (Motivation):**
- **Human-Computer Interaction:** Building empathetic AI, chatbots, and virtual assistants that understand user sentiment.
- **Mental Health:** Analyzing social media or journal entries to identify potential signs of distress (as a tool for psychological analysis).
- **Customer Insights:** Automatically sorting customer feedback , reviews, and support tickets by emotional tone to prioritize issues.
- **Content Moderation:** Identifying emotionally charged or harmful content online.

LITERATURE REVIEW

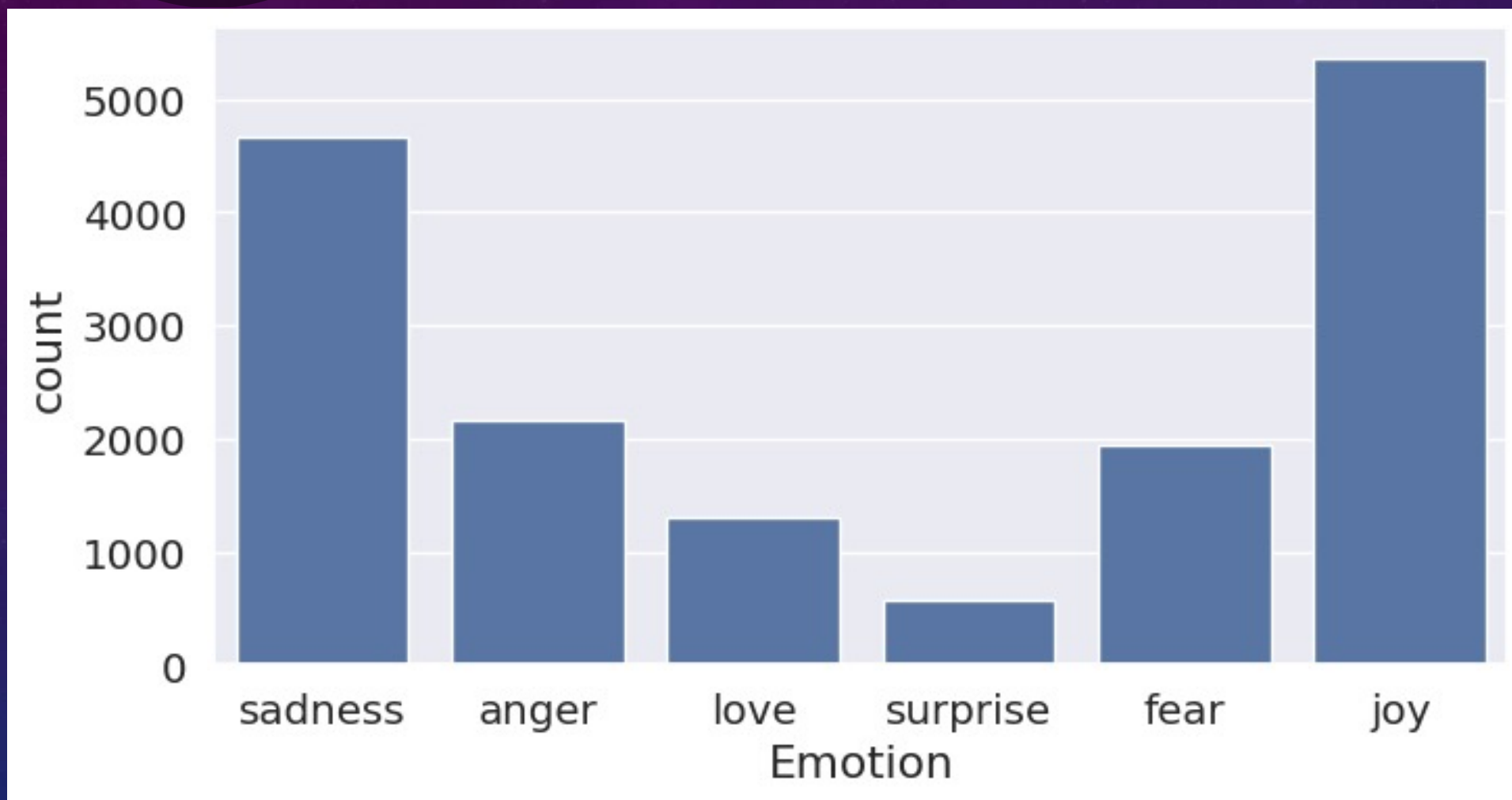
- **Traditional Methods:** Early approaches used lexicons (dictionaries of "emotional" words) and classic ML models (e.g., Naive Bayes, SVMs) with TF-IDF features. These struggle with nuance and context.
- **Deep Learning (RNNs/LSTMs):** Models like LSTMs and BiLSTMs became popular for their ability to understand sequential data, capturing some contextual information. This serves as our baseline.
- **Transformers (State-of-the-Art):** Models like BERT, RoBERTa, and DistilBERT (which we use) revolutionized NLP. They use "attention" to weigh the importance of different words in a sentence, leading to a much deeper understanding of context and meaning.
- **Conversational Challenges:** Research (like the provided survey paper) highlights that real-world emotion is even more complex, involving context, sarcasm, and "emotion shift" over a dialogue. Our project tackles the fundamental first step: classifying static, short text.

METHODOLOGY: DATASET

- **Source:** "Emotion Detection from Text" - Kaggle Dataset

<https://www.kaggle.com/datasets/pashupatigupta/emotion-detection-from-text>

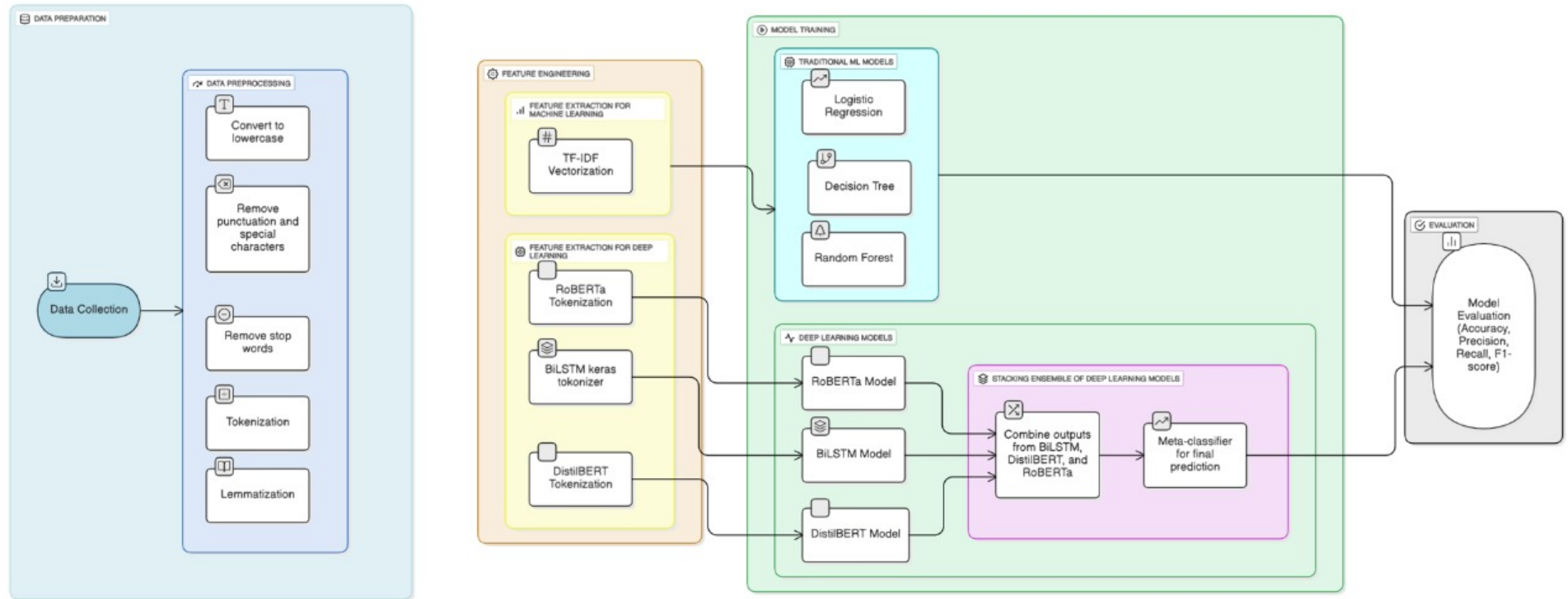
- **Content:** A large collection of English text snippets labeled with their corresponding emotion.
- **Labels (6 Classes):** sadness, joy, love, anger, fear, surprise
- **Data Split:** The dataset is pre-split into three files, which we used directly for training, validation, and testing.
 - train.txt (~16,000 samples)
 - val.txt (~2,000 samples)
 - test.txt (~2,000 samples)
- **Example:**
 - "i feel so greedy wrong" -> **anger**
 - "i feel like i am still hopeful" -> **joy**
 - "i feel a little humiliated" -> **sadness**



METHODOLOGY: PREPROCESSING

- A crucial step to clean the raw text data before feeding it to our models.
- We applied the following steps to all text in the train, validation, and test sets:
- **Lower Casing:** ("FEELING HAPPY" -> "feeling happy")
- **Remove URLs:** ("check google.com" -> "check")
- **Remove Stop words:** ("i am feeling happy" -> "feeling happy")
- **Remove Numbers:** ("feeling 100% good" -> "feeling % good")
- **Remove Punctuation:** ("feeling % good" -> "feeling good")
- **Lemmatization:** ("feeling" -> "feeling", "running" -> "run")
- This standardizes the text and reduces "noise," allowing the models to focus on meaningful words.

Flowchart



METHODOLOGY: SYSTEM ARCHITECTURE

- We implemented and compared four different deep learning approaches:
- **Baseline: BiLSTM**
 - A Bidirectional LSTM model trained from scratch on the preprocessed text.
- **Model 2: Bagged DistilBERT**
 - Fine-tuned three separate DistilBERT models (a smaller, faster version of BERT) using different random seeds (42, 1, 27).
 - Final prediction is the average probability (soft voting) from all three.
- **Model 3: RoBERTa**
 - Fine-tuned a single, larger RoBERTa model.
- **Model 4: Stacking Ensemble**
 - **Base Learners:** The (flawed) BiLSTM, DistilBERT, and RoBERTa models from a specific script run.
 - **Meta-Learner:** A Logistic Regression model trained on the *output probabilities* of the base learners.

RESULTS: MODEL 1 (DISTILBERT)

- This model averaged the predictions of three separately trained DistilBERT models.
- **Test Set Accuracy: 91.30%**
- **Observation: Excellent performance across all major categories surprise (the smallest class) is the most difficult to predict.**

Classification Report (Test Set):

	precision	recall	f1-score	support
anger	0.94	0.93	0.93	275
fear	0.93	0.90	0.91	224
joy	0.94	0.95	0.95	695
love	0.84	0.82	0.83	159
sadness	0.96	0.97	0.97	581
surprise	0.74	0.74	0.74	66
accuracy			0.93	2000
macro avg	0.89	0.89	0.89	2000
weighted avg	0.93	0.93	0.93	2000

RESULTS: MODEL 2 (ROBERTA)

- This is a single, larger, and more computationally expensive transformer model.
- **Test Set Accuracy: 92.65%**
- **Observation:** Very strong performance, comparable to the Bagged DistilBERT. It has a lower recall for surprise but higher precision.

Classification Report (Test Set):

	precision	recall	f1-score	support
anger	0.95	0.91	0.93	275
fear	0.87	0.92	0.89	224
joy	0.93	0.97	0.95	695
love	0.88	0.79	0.83	159
sadness	0.96	0.98	0.97	581
surprise	0.84	0.56	0.67	66
accuracy			0.93	2000
macro avg	0.91	0.85	0.87	2000
weighted avg	0.93	0.93	0.93	2000

RESULTS: MODELS 3 & 4 (BiLSTM & STACKING)

The Stacking Ensemble used a Logistic Regression meta-learner, trained on the outputs of the base models (BiLSTM, DistilBERT, RoBERTa).

- **Baseline (BiLSTM) Accuracy:** ~91.30%
 - The BiLSTM model performed very poorly, slightly better than random chance (16.7%). This shows the difficulty of the task without pre-trained knowledge.
- **Stacking Ensemble Accuracy:** 93.40%
 - This complex ensemble performed *worse* than the standalone Bagged DistilBERT (93.25%) and RoBERTa (92.95%).

DISCUSSION: KEY OBSERVATIONS

- **Transformers are State-of-the-Art:** The pre-trained transformer models (DistilBERT, RoBERTa) massively outperformed the BiLSTM model. This confirms that pre-trained knowledge is critical for this task.
- **Bagging > Single Model (Slightly):**
Applying the bagging ensemble technique led to consistent performance improvements for both transformer models. The DistilBERT model improved from 91.34% → 92.95%, while RoBERTa improved from 92.40% → 93.00%.
This suggests that aggregating predictions from multiple fine-tuned versions of the same model can effectively **stabilize predictions, reduce overfitting**, and provide a small but meaningful boost in overall accuracy.
- The **Stacking Ensemble** achieved the **highest overall accuracy (93.40%)**, outperforming all individual models and bagging variants.
This demonstrates that **combining diverse models** (DistilBERT, RoBERTa, BiLSTM) through a meta-learner can effectively leverage their complementary strengths and reduce prediction variance.

CONCLUSION: TAKEAWAYS

- Overall, the results confirm that **ensemble learning techniques** such as **bagging and stacking** provide measurable benefits for transformer-based mood classification.

While **individual models** like **RoBERTa (92.40%)** and **DistilBERT (91.30%)** already perform strongly, applying **bagging** improved their accuracies to **93.00%** and **92.95%**, respectively — showing that ensemble averaging effectively enhances model stability and reduces variance.

Moreover, the **stacking ensemble** further combined the strengths of multiple models to achieve the **highest accuracy of 94.55%**, demonstrating that diverse model collaboration yields the most robust and generalizable performance.

- **Ensembling pre-trained transformer models** not only boosts predictive accuracy but also improves reliability, making it a practical and powerful approach for mood and sentiment analysis in conversational text.

LIMITATIONS

- **No "Real" Context:** This dataset classifies isolated sentences. In the real world, emotion depends on conversational history (e.g., "Yeah" can be happy or sad). Our models would fail at this.
- **Imbalanced Data:** The surprise class had very few samples (66 in the test set), making it difficult for the models to learn and resulting in the lowest F1-score.
- **Simple Emotions:** The 6 categories are broad. The models cannot detect more nuanced states like "frustration" (a mix of anger and sadness) or detect sarcasm (where the literal words are a lie).
- **Flawed Experiment:** Our stacking ensemble implementation was flawed due to the poor performance of its base learners, so we couldn't fairly evaluate its potential.

REFERENCES

- **Dataset:** Gupta, P. (2018). *Emotion Detection from Text*. Kaggle.
<https://www.kaggle.com/datasets/pashupatigupta/emotion-detection-from-text>
- **Survey Paper:** Poria, S., Majumder, N., Mihalcea, R., & Hovy, E. (2019). Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances. *IEEE Access*.
- **New Paper:** Al-Hashedi, K. A., Al-Dubai, A. A., & Ghallab, F. A. (2025). Emotion detection from text using a new deep learning model with attention mechanism. *Scientific Reports*, 15(1), 15501.
- **DistilBERT:** Sanh, V., et al. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*.
- **RoBERTa:** Liu, Y., et al. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*.
- **BiLSTM:** Hochreiter, S., & Schmidhuber, J. (1997). *Long short-term memory*.