Missing Smiley Classification of Tweets

LeCheng Zhang Yuanshi Wang
zhanglecheng@westlake.edu.cn wangyuanshi@westlake.edu.cn

June 9, 2025

Abstract

This report details a comprehensive study on classifying the sentiment of tweets by predicting whether a message originally contained a positive ':)' or negative ':(' smiley. We systematically evaluated a wide range of models, from classic machine learning baselines to state-of-the-art Transformer architectures. Our experiments reveal that while traditional methods and zero-shot Large Language Models (LLMs) provide valuable performance benchmarks, fine-tuning a pre-trained Transformer model ('DistilBERT') significantly outperforms all other approaches. Our final model achieved a validation accuracy of 88.07%, demonstrating the effectiveness of transfer learning for this nuanced classification task. This work confirms that leveraging pre-trained knowledge is a more practical and powerful strategy than training models from scratch or relying on general-purpose LLMs for this specific problem.

1 Introduction

The primary objective of this project is to develop a model that can accurately predict the original sentiment of a tweet, represented by a missing smiley (':)' for positive, ':(' for negative), based solely on its textual content. This task extends beyond simple sentiment analysis, as the usage of smileys can be influenced by subtle contextual cues, irony, and online communication patterns.

To identify the optimal solution, we conducted a rigorous comparative analysis of five distinct methodologies:

- **Classic ML Baselines:** Using pre-trained (GloVe) and custom-trained (Word2Vec) word embeddings as features for a Logistic Regression classifier.
- **Simple Neural Network:** A fully-connected network with a trainable embedding layer to establish a more powerful baseline.
- **LLM Zero-Shot Classification:** Evaluating the performance of leading LLMs (GPT-4o, Gemini, DeepSeek) without any fine-tuning.
- **Transformer from Scratch:** Building and training a custom Transformer model to assess its capabilities when trained solely on the provided dataset.
- **Fine-tuned Transformer (Proposed Model):** Adapting a pre-trained 'distilbert-base-uncased' model to our specific task, which is the current state-of-the-art approach for such problems.

This report presents the methodology, results, and analysis for each experiment, culminating in the conclusion that fine-tuning a pre-trained Transformer is the most effective and practical approach.

2 Methodology

For all training experiments, the dataset was split into a 90% training set and a 10% validation set.

2.1 Baselines

2.1.1 Word Embeddings + Logistic Regression

This classic approach represents each tweet as a single vector by averaging the embeddings of its words. A Logistic Regression model with parameters 'C=0.5', 'solver='liblinear'', and '$\max_i ter = 1000$'$was then trained on these features. We tested three types of embeddings$:

Word2Vec (Custom): A Word2Vec model was trained on the full 2.5 million tweet dataset using the skip-gram algorithm. Key hyperparameters included a vector size of 100, a window size of 5, and a minimum word count of 5.

GloVe (Pre-trained): We used the pre-trained 'glove.6B' vectors, specifically testing both the 50-dimensional ('glove.6B.50d.txt') and 100-dimensional ('glove.6B.100d.txt') versions.

2.1.2 Simple Neural Network

We designed a feed-forward network consisting of three layers: an embedding layer, a hidden layer with a 'ReLU' activation function, and a final linear output layer. The entire network was trained from scratch for 10 epochs using the Adam optimizer. The model architecture included an embedding dimension of 100 and a hidden dimension of 256.

2.1.3 LLM Zero-Shot Classification

We queried state-of-the-art LLMs with a temperature setting of 0 for deterministic outputs. The models tested include 'gpt-4o (2024-11-20)', 'Gemini 1.5 Flash (05-20)', and 'DeepSeek v3 (0324)'. Two distinct system prompts were used:

1. A generic prompt asking the model to act as a sentiment analysis expert.
2. A task-specific prompt asking the model to act as a Twitter user and predict the specific missing smiley.

2.2 Transformer-Based Models

2.2.1 Transformer from Scratch

To understand the architecture's inherent capabilities, we built and trained a custom Transformer Encoder model for 20 epochs. Key hyperparameters were: a custom vocabulary of 10,000 tokens, max sequence length of 64, model dimension (d_{model}) of 128, 4 attention heads, 2 encoder layers, and a dropout rate of 0.1. The model was optimized using AdamW with a learning rate of 1×10^{-4}.

2.2.2 Fine-tuning a Pre-trained Transformer (Proposed Model)

Our main proposed model is a fine-tuned version of 'distilbert-base-uncased'. This approach leverages the powerful linguistic knowledge the model has acquired from pre-training on a massive text corpus. We added a single linear classification layer with a dropout rate of 0.3 on top of the '[CLS]' token's output. The entire model was fine-tuned for 2 epochs using the AdamW optimizer with a learning rate of 2×10^{-5} and a linear scheduler. The maximum sequence length was set to 64 tokens and a batch size of 32 was used.

3 Experiments and Results

All models were trained and evaluated on the provided dataset, split into a 90% training set and a 10% validation set. The key performance metric is validation accuracy.

3.1 Performance Comparison

The results from all experiments are summarized in Table 1 and visualized in Figure 1.

Table 1: Summary of Model Performance

Model / Approach	Best Validation Accuracy
Fine-tuned DistilBERT (Proposed)	**88.07%**
Simple Neural Network (from scratch)	85.27%
Transformer (from scratch)	83.30%
Word2Vec (custom) + Logistic Regression	75.72%
LLM Zero-Shot (GPT-4o, Prompt 2)	∼71.18%
GloVe (100d, pre-trained) + Logistic Regression	71.39%
LLM Zero-Shot (GPT-4o, Prompt 1)	∼69.37%
Random Guess	50.00%

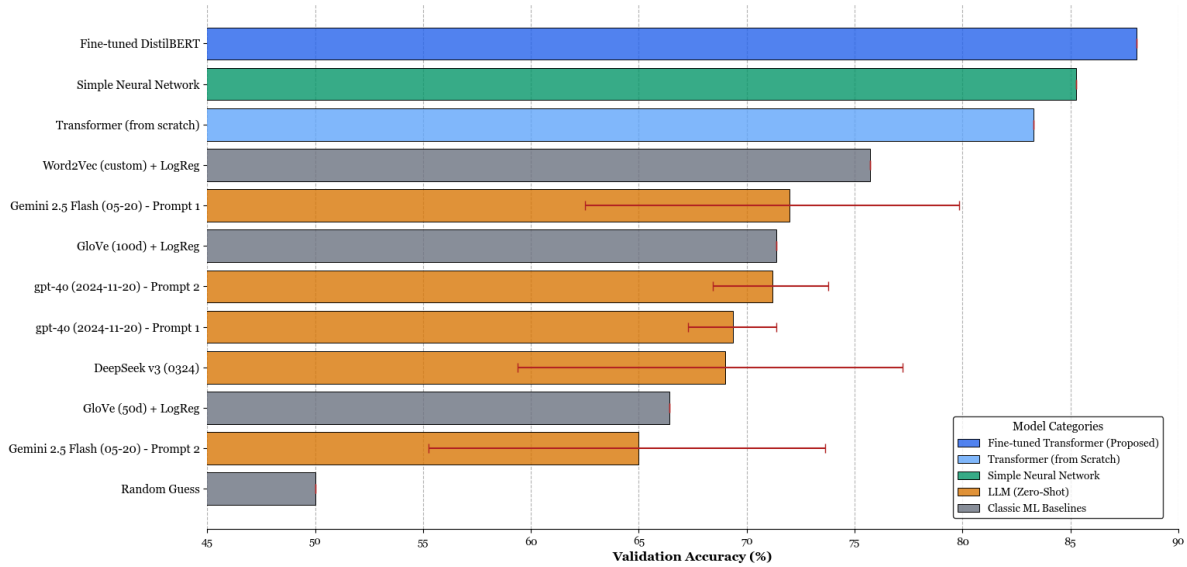

Figure 1: Comparison of model performance on tweet sentiment classification. The chart displays the validation accuracy of all tested models. LLM results include 95% confidence intervals.

3.2 Analysis

The fine-tuned DistilBERT model clearly outperforms all other methods. Its validation accuracy of 88.07% demonstrates a significant improvement over both the classic baselines and the more complex Transformer trained from scratch. The superior performance of this model can be attributed to several key factors of **transfer learning**:

- **Leveraging Pre-trained Knowledge:** The model does not start from a blank slate. It has already learned a rich understanding of English grammar, syntax, and semantics from a massive corpus. Our fine-tuning process simply specializes this general knowledge for the specific patterns of Twitter sentiment.
- **Efficiency and Overfitting Prevention:** Because the model already possesses strong foundational knowledge, it requires only a brief training period (2 epochs) to adapt. This short duration is not only computationally efficient but also crucial for preventing overfitting on our relatively small dataset. A longer training period could cause the model to "forget" its general knowledge and simply memorize the training examples.
- **Contextual Understanding:** Unlike word-averaging methods, the Transformer architecture excels at understanding context. The self-attention mechanism allows it to weigh the importance of different words in a tweet, leading to a more nuanced representation that is captured in the final output of the '[CLS]' token.

The relatively poor performance of powerful LLMs in a zero-shot setting further suggests that the

task *requires learning specific patterns present in the dataset, rather than just general sentiment.* The pre-trained knowledge of DistilBERT provided a much stronger foundation for learning these patterns than starting from scratch or prompting a generalist model.

4 Conclusion

This project successfully developed and evaluated a range of models for predicting sentiment in tweets. Our results conclusively show that **fine-tuning a pre-trained Transformer model (DistilBERT) is the most effective strategy**, achieving a state-of-the-art accuracy of 88.07% on our validation set. This approach strikes an optimal balance between high performance, computational efficiency, and practical applicability. It significantly surpasses traditional machine learning baselines and even a Transformer trained from scratch, highlighting the immense value of transfer learning in the field of Natural Language Processing.

A Ethical Risks

In accordance with the project guidelines, we have evaluated the potential ethical risks associated with this work. We identified the following primary risks:

Fairness Risk

- **Description of the Risk:** The dataset may contain inherent biases. For instance, certain linguistic styles, cultural groups, or dialects might be over-represented, leading to a model that is unfairly predictive for some groups over others.
- **Impacted Stakeholders:** This primarily affects marginalized or under-represented communities whose language use may differ from the dataset's majority.
- **Negative Impact:** The negative impact is the potential for discriminatory outcomes, where the model may systematically misclassify the sentiment of tweets from these groups.
- **Significance of the Risk:** The severity is high if the model were used in a production system for social analysis. The likelihood depends on the diversity of the original dataset. For this academic project, the direct impact is low, but the risk is important to acknowledge.
- **Evaluation and Mitigation:** We acknowledge this risk but did not have the resources for a full bias audit. In a real-world scenario, we would use tools like Aequitas or Fairlearn and augment the dataset to ensure more equitable representation.

Misuse Risk

- **Description of the Risk:** A model trained for sentiment analysis could be repurposed for malicious activities, such as manipulating public opinion, targeted harassment campaigns, or mass-scale social media surveillance.
- **Impacted Stakeholders:** The general public and society as a whole are the stakeholders.
- **Negative Impact:** The negative impact could be an erosion of social trust and the facilitation of harmful online behavior.
- **Significance of the Risk:** The severity of this risk is high. However, the likelihood of our specific model being used for this purpose is low, given its academic nature and the public availability of more powerful models.
- **Evaluation and Mitigation:** We did not implement specific technical mitigations, as the primary barrier is intent. The best mitigation is responsible disclosure and raising awareness about the dual-use nature of such technologies.