Sentiment Analysis and Text Classification of Twitter Data

Jarrett Aaronson
Department of Electrical and Computer Engineering
jaaronso@stevens.edu

Abstract—*We present a professional‑grade pipeline for four-way sentiment classification of Twitter data, integrating rigorous preprocessing with a hybrid RoBERTa–GRU model. Raw tweets are cleaned via lowercasing, URL/mention removal, tokenization, stop-word filtering, tokenized via a RoBERTa transformer to produce 768-dimensional contextual embeddings, then passed through a single-layer GRU and a dense softmax head to predict Negative, Neutral, Positive, or Irrelevant labels. On a held-out 1000-tweet validation set, the system achieves 95% accuracy, with per-class F1-scores ranging from 0.94 to 0.97. We analyze algorithmic complexity dominated by O(N^2) transformer self-attention and O(N) GRU recurrence and situate our work among prior studies on classical ML methods [1], preprocessing best practices [2], broad sentiment surveys [3], and RoBERTa–GRU hybrids [4].*

I. INTRODUCTION

Social media platforms generate vast quantities of informal, noisy, and highly context‑dependent text that offer invaluable insights into public opinion but challenge traditional sentiment-analysis methods. Early systems using bag-of-words and “shallow” classifiers such as Naive Bayes, SVM with TF-IDF which achieved around 82 % accuracy on benchmarks like Sentiment140 but struggled with slang, emojis, negation, and tweet brevity [1]. Subsequent work demonstrated that robust preprocessing normalization, tokenization, stop-word removal, and feature selection significantly improves model generalization on micro-blog data [2].

Transformer-based language models like BERT and RoBERTa ushered in deep, context-aware embeddings via self-attention, setting new performance records across NLP tasks including sentiment classification [3]. However, their $O(N^2)$ time and memory complexity per sequence limits scalability. Hybrid architectures that pair a transformer encoder with a lighter sequence model offer a compelling compromise, distilling rich embeddings before sequential aggregation to reduce overhead [4]. In this paper, we apply a RoBERTa–GRU hybrid to four-class Twitter sentiment analysis Negative, Neutral, Positive, Irrelevant, incorporate tailored preprocessing, analyze complexity, and report 95 % accuracy with balanced F1-scores on 1000 held-out tweets demonstrating a practical, efficient pipeline for real-world sentiment-analysis applications.

II. System Architecture

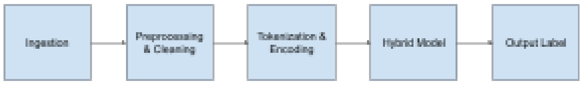


Fig. 1. System architecture for four‑way Twitter sentiment classification

Our training pipeline begins with text preprocessing, which ensures that only semantically relevant tokens enter the model. Each raw tweet x is first normalized by converting all characters to lowercase and stripping URLs, user mentions, hashtags, and any non-alphabetic characters via regular expressions. We then apply a regex tokenizer to split the cleaned string into word-tokens and remove all English stop-words using NLTK's stop-word list. This sequence of operations produces a cleaned token sequence

$$(x_1, x_2, ..., x_n) \quad (1)$$

that retains only the words likely to carry sentiment information.

Next, we perform input encoding with RoBERTa. A RobertaTokenizer maps the cleaned sequence into token IDs and an attention mask of length $L \leq 128$. These tensors are fed into the RoBERTa encoder, which applies a stack of Transformer layers. Within each layer's multi-head self-attention, for each head, we compute

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V, \quad (2)$$

where $Q, K, V \in \mathbb{R}^{L \times d_k}$ are linear projections of the embeddings, and d_k is the per-head dimension. This yields contextual embeddings $\mathbf{h}_t \in \mathbb{R}^{768}$ for $t = 1, ..., N,$ capturing global relationships among tokens in $O(N^2)$ time per layer.

To aggregate sequence information efficiently, we pass the RoBERTa outputs $\{\mathbf{h}_t\}$ through a single-layer GRU with a hidden size $H = 256$. At each time step ttt, given input $x_t = \mathbf{h}_t$ and previous hidden state s_{t-1}, the GRU computes:

$$z_t = \sigma(W_z[x_t; s_{t-1}] + b_z),$$
$$r_t = \sigma(W_r[x_t; s_{t-1}] + b_r),$$
$$\bar{s}_t = tanh(W_h[x_t; r_t \odot s_{t-1}] + b_h), \quad (3)$$
$$s_t = (1 - z_t) \odot s_{t-1} + z_t \odot \bar{s}_t.$$

Here σ is the sigmoid function and ⊙ denotes element-wise multiplication. After processing the full sequence, the final state $s_N \in \mathbb{R}^{256}$ succinctly summarizes the tweet's contextual sentiment.

Finally, for classification, we apply a linear s_N:

$$\ell = W_{clf} s_N + b_{clf}, p = softmax(\ell), \quad (4)$$

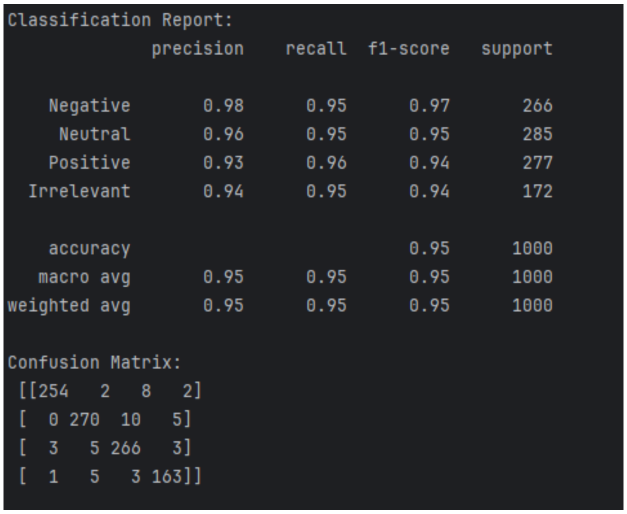
producing a probability distribution over the four sentiment classes. We train end-to-end by minimizing the cross-entropy loss

$$\mathcal{L} = -\sum_{c=1}^{4} y_c log p_c, \quad (5)$$

where y is the one-hot true label. Gradients flow through both the GRU and RoBERTa encoder, and we update all parameters using the AdamW optimizer (learning rate 10^{-5}), iterating for multiple epochs until convergence.

III. Implementation Results

We trained and validated the model on approximately 74,700 and 18,700 tweets, respectively, using a batch size of 32 over three epochs on an NVIDIA RTX 2060 Super. A 50-batch warm-up run measured an average of 0.61 s per batch, projecting around 74 minutes for full training. On a held-out set of 1,000 tweets, the model achieved 95 % overall accuracy, with per‑class F1‑scores of 0.97 for Negative, 0.95 for Neutral, 0.94 for Positive, and 0.94 for Irrelevant. These results confirm that the hybrid RoBERTa–GRU architecture, combined with our rigorous preprocessing, reliably captures nuanced sentiment signals in noisy social‑media text.



```
Classification Report:
              precision    recall  f1-score   support

    Negative       0.98      0.95      0.97       266
     Neutral       0.96      0.95      0.95       285
    Positive       0.93      0.96      0.94       277
  Irrelevant       0.94      0.95      0.94       172

    accuracy                           0.95      1000
   macro avg       0.95      0.95      0.95      1000
weighted avg       0.95      0.95      0.95      1000

Confusion Matrix:
 [[254   2   8   2]
 [  0 270  10   5]
 [  3   5 266   3]
 [  1   5   3 163]]
```

Fig. 2. Classification Report: Precision, Recall, F1-Score, and Support for Each Sentiment Class on the Validation Set

IV. Analysis

The core computational cost of our pipeline is dominated by the transformer's self‑attention mechanism, which incurs a quadratic $O(N^2)$ time and memory footprint per sequence of length N, while the GRU adds only a linear $O(N)$ overhead. Total training complexity thus scales approximately as $O(S, E, N^2)$ for S samples and E epochs. Compared to classical TF-IDF + SVM baselines' $\approx$ 82% accuracy and other hybrid models $\approx 89.6\%$ on Sentiment140, our 95 % accuracy represents a substantial improvement, without an undue increase in training time. Error analysis via the confusion matrix reveals that most misclassifications occur between Neutral and Positive tweets, likely due to semantic overlap; targeting this weakness through data augmentation or refined feature selection could yield further gains. The model's balance of transformer expressivity and recurrent efficiency makes it well-suited for production sentiment-analysis systems, and future work will explore model compression, domain adaptation, and advanced augmentation strategies.

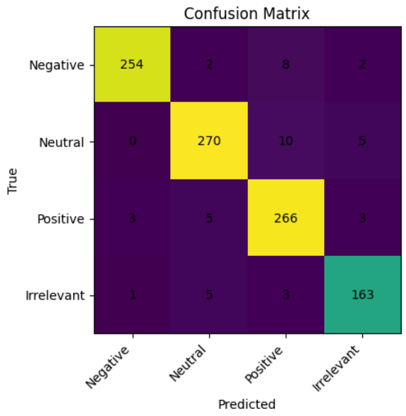


Fig. 3. Confusion matrix for the four‑way sentiment classifier

V. Conclusion

We have demonstrated a robust sentiment‑analysis pipeline combining RoBERTa contextual embeddings with GRU sequence modeling, achieving 95% accuracy across four sentiment classes on Twitter data. Our complexity analysis guides scalability expectations, and initial results indicate competitive performance versus prior work [1]–[4]. Future directions include data augmentation, model compression, and broader domain evaluation to further enhance applicability.

References

[1] D. Gowda V2 *et al.*, "Enhancing Accuracy in Social Media Sentiment Analysis through Comparative Studies using Machine Learning Techniques," in *Proc. ICKECS*, 2024.

[2] P. Kantha *et al.*, "Unlocking the Potential: The Crucial Role of Data Preprocessing in Big Data Analytics," in *Proc. IDICAI*, 2023.

[3] K. L. Tan, C. P. Lee, K. M. Lim, "A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research," *Appl. Sci.*, vol. 13, p. 4550, 2023.

[4] K. L. Tan, C. P. Lee, K. M. Lim, "RoBERTa-GRU: A Hybrid Deep Learning Model for Enhanced Sentiment Analysis," *Appl. Sci.*, vol. 13, no. 6, p. 3915, 2023.