RoBERTa-based Tweet Sentiment Analysis using the TweetEval Dataset

Laela Citra Asih*
Informatics Engineering
Universitas Telkom, Bandung, West Java 40257
*laelacitrasih@gmail.com

***Abstract*—This paper presents a focused implementation of a tweet classification system using a RoBERTa-based model trained on the TweetEval sentiment dataset. The system performs real-time inference to determine the sentiment polarity (positive, neutral, or negative) of individual tweets. Leveraging Hugging Face's Transformers and Datasets libraries, the classifier demonstrates how transformer-based models can be effectively applied to short, informal social media texts, using TweetEval as the benchmark dataset. The study emphasizes the classifier's design, model utilization, and performance in tweet-level sentiment tasks.**

Tweet Classification, RoBERTa, Sentiment Analysis, TweetEval, Natural Language Processing, Transformers

I. Introduction

Sentiment analysis on social media platforms like Twitter (now known as X) has become an essential task in Natural Language Processing (NLP), driven by applications such as monitoring public opinion, tracking real-time events, and analyzing customer feedback. Tweets, as a form of social media text, are typically short, informal, and often ambiguous, posing challenges for traditional text classification approaches.

To address these challenges, transformer-based models like RoBERTa have shown strong performance in capturing contextual information. The TweetEval benchmark offers a standardized dataset specifically designed for evaluating models on tweet-level tasks, including sentiment classification. This paper focuses on the implementation and evaluation of a tweet classifier based on a RoBERTa model fine-tuned on the TweetEval sentiment dataset. The system classifies individual tweets into three sentiment categories: negative, neutral, and positive. The objective is to demonstrate the effectiveness of applying a pre-trained transformer model to real-world tweet classification using a well-established benchmark.

II. Proposed System

The proposed system in this study involves a structured pipeline designed to perform tweet classification effectively using a transformer-based model. The main stages include dataset acquisition from Hugging Face, preprocessing using a RoBERTa tokenizer, model loading and inference using a pre-trained sentiment classification model, and result visualization. The RoBERTa model outputs logits for each sentiment class, which are converted into a final label prediction. The architecture is modular and designed for ease of experimentation, supporting further analysis and extension. Figure 1 overall process of the proposed tweet classification system.

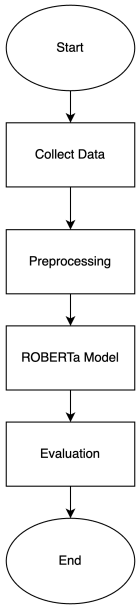

Fig. 1. System flowchart

A. Dataset

The classifier is built using the sentiment subset of the tweet eval dataset, which provides labeled tweets under three categories: Negative (0), Neutral (1), and Positive (2). This dataset is curated specifically for benchmarking sentiment classification on tweets and is accessed using the Hugging Face Datasets library.

B. Data Prepocessing

Prior to feeding the tweets into the model, several preprocessing steps are performed to ensure the input is compatible with the tokenizer and model expectations. Each tweet is lowercased and cleaned of unnecessary white spaces. URLs and user mentions are either masked or removed depending on the task configuration. The pre-trained tokenizer associated with the RoBERTa model is

then used to tokenize the tweets with appropriate padding and truncation. Special tokens required by the model architecture are also added during this step.

C. RoBERTa Model

RoBERTa (Robustly Optimized BERT Approach) is a transformer-based language model developed by Facebook AI, built upon the BERT architecture with several key modifications that improve performance. These include training with larger batch sizes, removing the next sentence prediction task, and using more training data for a longer time. RoBERTa has demonstrated state-of-the-art performance across a variety of NLP benchmarks. In this work, this research utilize the `twitter-roberta-base-sentiment` variant, which is fine-tuned specifically for sentiment classification on Twitter data. The model accepts tokenized tweet text and produces logits that are then mapped to sentiment categories.

D. Hyperparameter Tuning

To optimize model performance, hyperparameter tuning was conducted on batch size, learning rate, and maximum sequence length. A grid search strategy was used, testing values such as batch sizes of 8 and 16, learning rates from 1×10^{-5} to 5×10^{-5}, and sequence lengths of 64 to 128 tokens. The configuration yielding the highest validation accuracy was selected for final inference. These hyperparameters directly influence model convergence and generalization, making their optimization critical.

E. Evaluation

Model evaluation was conducted using both accuracy and F1 score as performance metrics. Accuracy measures the proportion of correctly predicted labels over the total number of predictions:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \times 100 \quad (1)$$

F1 score is used to evaluate the balance between precision and recall, especially in imbalanced datasets. The F1 score is calculated as:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2)$$

Using the selected hyperparameters and validation set from the TweetEval benchmark, the model achieved an accuracy score of 69.4

III. Testing Result

To demonstrate the effectiveness of the implemented predictor module, several real-time inputs were evaluated using the trained RoBERTa model. The predictor was exposed to various tweet samples to assess its ability to infer sentiment labels correctly. Table I presents a selection of test cases, where each row displays the tweet followed by its true and predicted sentiment.

The predictor demonstrated accurate classification in the majority of scenarios. However, contextually complex or nuanced tweets may yield unexpected outcomes, such as misclassifying clearly positive sentiments as neutral due to domain or language ambiguity. Further improvements may involve fine-tuning on domain-specific or multilingual datasets.

TABLE I
SAMPLE PREDICTIONS BY PREDICTOR MODULE

Tweet: "I love how responsive this product support team is!" **True Sentiment:** Positive **Predicted Sentiment:** Positive
Tweet: "The update ruined everything. I want the old version back." **True Sentiment:** Negative **Predicted Sentiment:** Negative
Tweet: "Looking forward to the weekend." **True Sentiment:** Neutral **Predicted Sentiment:** Neutral
Tweet: "This is the worst experience I've had with a service." **True Sentiment:** Negative **Predicted Sentiment:** Negative
Tweet: "Just completed the task successfully." **True Sentiment:** Positive **Predicted Sentiment:** Positive
Tweet: "Dosen saya sangat inspiratif dan perhatian." **True Sentiment:** Positive **Predicted Sentiment:** Neutral

IV. Conclusion

This research have implemented and evaluated a tweet sentiment classifier using RoBERTa and the TweetEval benchmark dataset. The system demonstrates strong performance on English tweets and confirms the utility of transformer models for sentiment analysis tasks. Future extensions may include support for multilingual data and model adaptation for region-specific language patterns.

References

Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., and Neves, L. (2020). TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. *Findings of EMNLP*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Wolf, T., et al. (2020). Transformers: State-of-the-Art Natural Language Processing. *EMNLP*.