Capstone Two Final Report

TF-IDF vs Transformer-Based Sentiment Analysis (TweetEval)

Author: Albert Vempala | **Program:** Data Science Career Track
Artifacts: Production-ready notebook, model metrics, and executive slide deck

Executive Summary

This project evaluates whether a modern transformer-based sentiment model materially outperforms a strong classical baseline (TF-IDF + Logistic Regression) on short, noisy social text (TweetEval). The dataset contains **45,615** training samples, **2,000** validation samples, and **12,284** test samples with three sentiment classes. Data quality checks found **0** missing values across all splits and **29** duplicate texts in the training split.

On the held-out test set, the transformer achieved **Accuracy 0.7123** and **Macro F1 0.7140**, compared to TF-IDF+LogReg at **Accuracy 0.5947** and **Macro F1 0.5911**. This is an absolute improvement of **+0.1176** in accuracy and **+0.1229** in macro F1. Based on these results and observed error patterns, the transformer is recommended as the production model, with TF-IDF retained as a low-cost fallback and regression baseline.

1. Business Problem and Objective

Organizations use sentiment signals to inform customer experience, brand monitoring, and product decisions. On social media text, bag-of-words approaches often fail on context-dependent sentiment (negation, sarcasm, and compositional meaning). The objective is to quantify the value of transformer-based modeling relative to a TF-IDF baseline and recommend a deployable approach based on measurable lift and interpretability.

2. Data Description and Quality Checks

The analysis uses TweetEval sentiment data with three classes (negative, neutral, positive). A consistent train/validation/test split is used throughout. Data quality checks were performed for missing values and duplicate texts.

Table 1. Split sizes, data quality checks, and label distribution.

Split	Rows	Missing	Duplicate texts	Negative	Neutral	Positive
Train	45615	0	29	7093	20673	17849
Validation	2000	0	0	312	869	819
Test	12284	0	0	3972	5937	2375

3. Exploratory Data Analysis (EDA)

EDA focused on class balance and typical tweet length. The figures below are exported directly from the production-ready notebook.

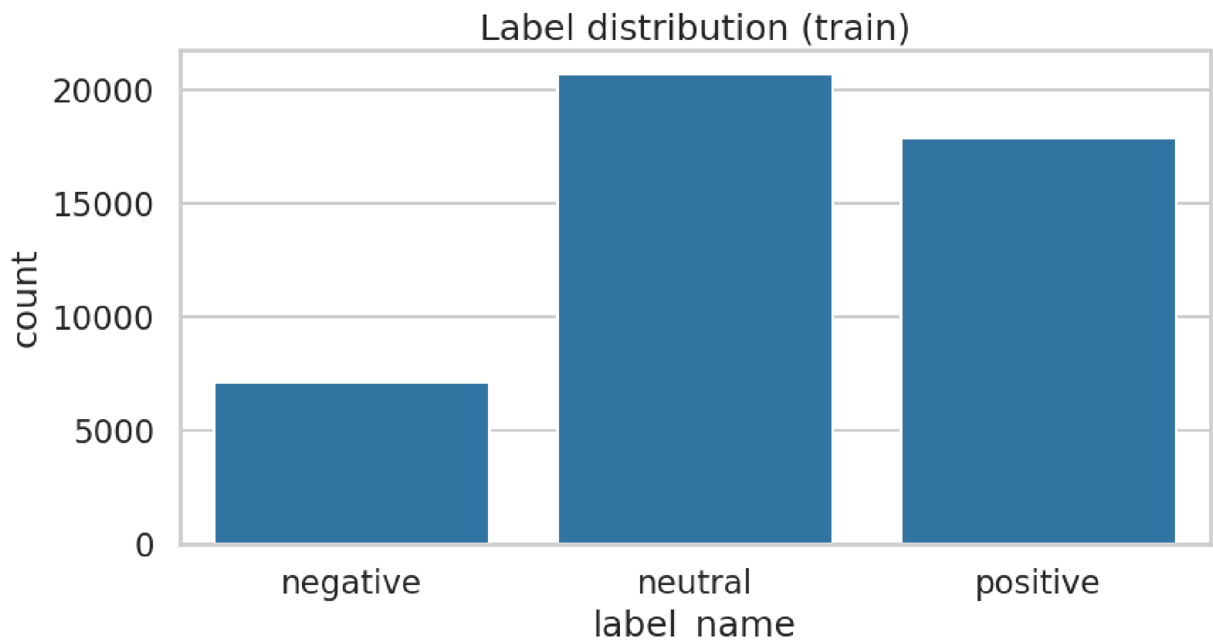

Figure 1. Training set label distribution (negative/neutral/positive).

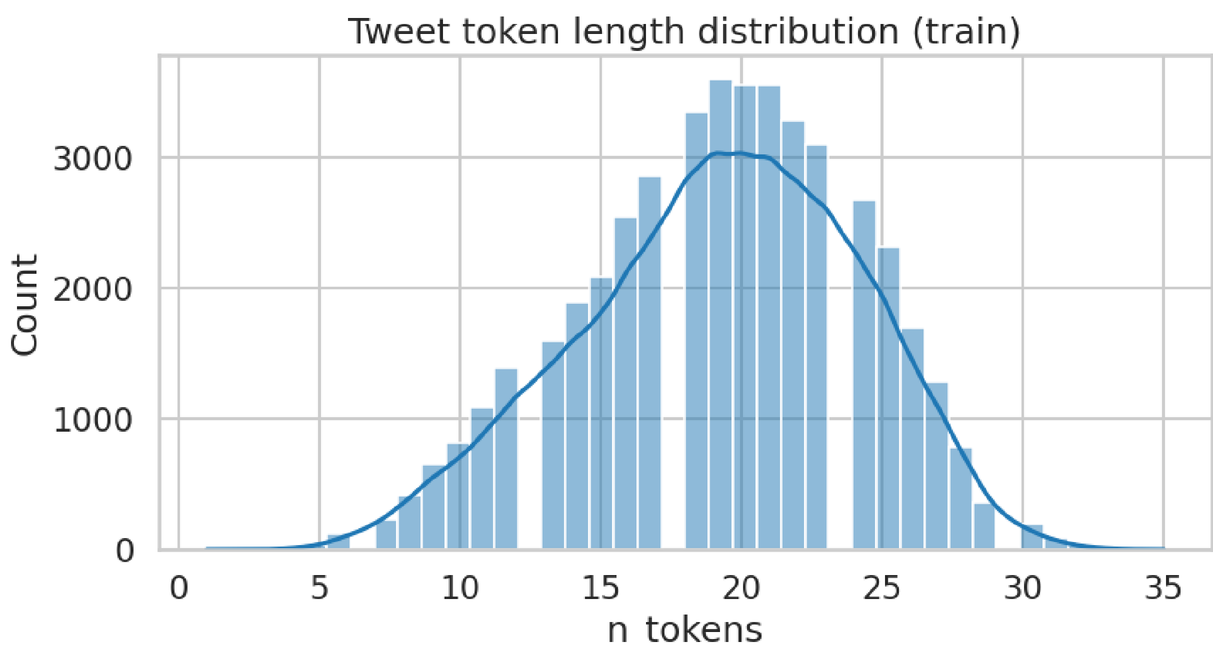

Figure 2. Tweet token length distribution (train).

4. Modeling Approach

Two model families were compared under a consistent evaluation protocol: (A) TF-IDF + Logistic Regression baseline with tuning on the validation set, and (B) a fine-tuned transformer model evaluated on the same test set. Because class imbalance exists, macro-averaged metrics (macro precision/recall/F1) are emphasized alongside accuracy.

4.1 Baseline: TF-IDF + Logistic Regression

The baseline uses TF-IDF word n-grams (1,2) with sublinear term frequency scaling and document-frequency filtering (min_df=2, max_df=0.95). Logistic Regression is trained with class balancing and tuned via grid search on regularization strength (C). The notebook-selected best parameters were C=2.0.

4.2 Transformer: Fine-Tuned Model

The transformer model is fine-tuned for sentiment classification and evaluated on the held-out test set. Transformers represent word order and context, enabling more reliable handling of negation and compositional meaning.

5. Results and Evaluation

Performance is reported using Accuracy and macro-averaged Precision/Recall/F1. Macro metrics are emphasized because they weight each class equally.

Table 2. Model performance (validation and test).

Model / Split	Accuracy	Macro Precision	Macro Recall	Macro F1
TF-IDF+LogReg (Val)	0.6675	0.6376	0.6615	0.6455
TF-IDF+LogReg (Test)	0.5947	0.5858	0.5996	0.5911
Transformer (Test)	0.7123	0.7040	0.7303	0.7140

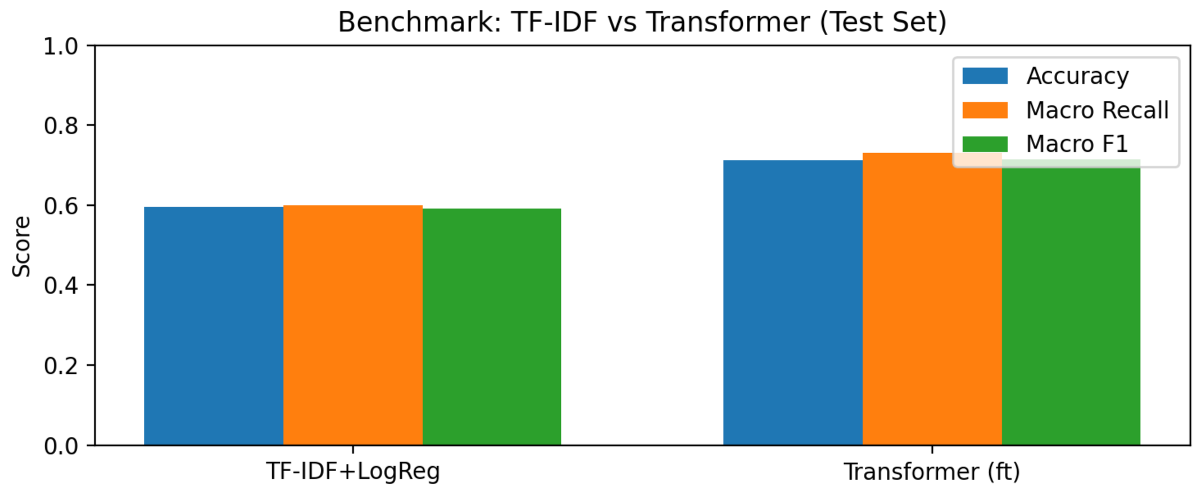

Figure 3. Benchmark comparison (test): TF-IDF+LogReg vs Transformer (accuracy, macro recall, macro F1).

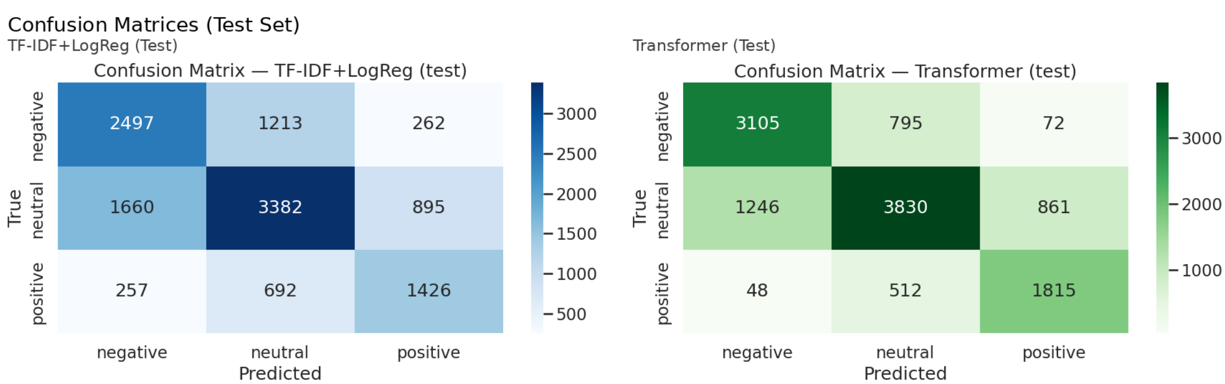

Figure 4. Confusion matrices (test): TF-IDF+LogReg vs Transformer.

5.1 Interpretation and Error Modes

The transformer increases macro recall from 0.5996 to 0.7303 (Δ +0.1307), reducing misses across classes. Macro precision increases from 0.5858 to 0.7040 (Δ +0.1182), and macro F1 increases by +0.1229. The confusion matrices show fewer boundary confusions between neutral and the other classes under the transformer, consistent with improved context modeling.

6. Recommendations and Deployment Plan

Adopt the transformer as the production model (Accuracy 0.7123, Macro F1 0.7140). Retain TF-IDF+LogReg as a low-cost fallback and regression baseline. Operationalize with batch inference, monitoring of macro metrics, and periodic refresh to address drift.

7. Limitations and Future Work

Limitations include the domain specificity of TweetEval, residual ambiguity in very short tweets, and higher compute cost. Future work: domain adaptation to target business data, uncertainty thresholds for ambiguous text, and inference cost optimization (batching/quantization/distillation).

8. Submission Checklist (GitHub)

Upload the model metrics CSV, final report PDF, and slide deck to your GitHub repository and submit the repo link for mentor review. Ensure notebooks run end-to-end without path issues and include a README.md that summarizes the problem, approach, results, and recommendation.