FIFA WC TWEETS SENTIMENT ANALYSIS

2022

Enrique Ulises Báez Gómez Tagle

Master’s in Data Science – NLP Final Project

~25k tweets from Day 1 (Qatar 2022)

Analyze public **sentiment** on **Twitter** during **FIFA World Cup 2022** using **classical ML, deep learning, and Transformers**.

Motivation & Problem

Why this matters?

- Twitter is a live **global "sensor"** of fan **emotions**.
- Sentiment **informs** media, sponsors, and organizers **about engagement**.
- **NLP challenges**: short, noisy, multilingual, sarcastic text.

Task: classify tweets into positive / neutral / negative.

Research Background

2016-2018

TF-IDF + Logistic Regression / Linear SVM → baselines

2018-2020

CNN / BiLSTM capture local context and sequence.

2020-2025

Transformers (BERT/RoBERTa) dominate Twitter tasks.

Our contribution

Head-2-Head comparison of all three families on the same dataset & split.

Dataset Overview

8,489

Positive (37.7%)

8,251

Neutral (36.6%)

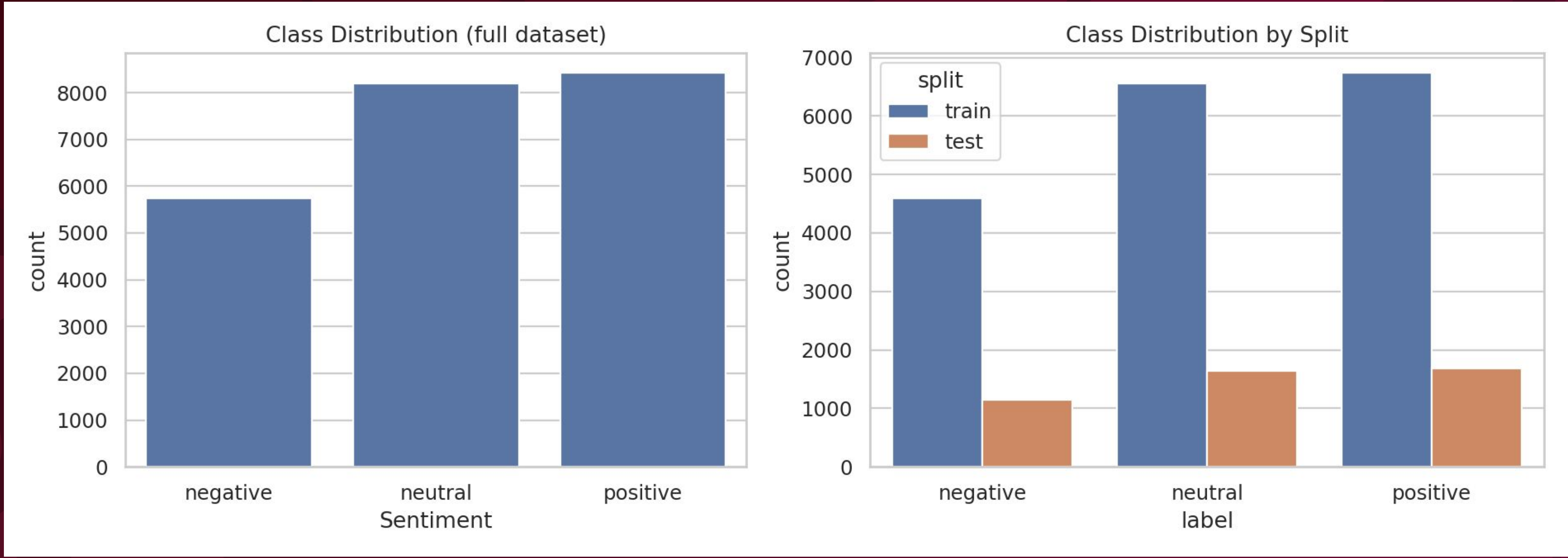
5,784

Negative (25.7%)

22,524

tweets after cleaning

Imbalance Ratio = 1.47 < 1.5); no resampling



Class Distribution (full dataset)
count
8000
7000
6000
5000
4000
3000
2000
1000
0
negative
neutral
positive
Sentiment
Class Distribution by Split
7000
6000
5000
4000
3000
2000
1000
0
split
train
test
negative
neutral
positive
label

Preprocessing Pipeline

Lowercasing → **Tags** → **Hashtags**

Tags

Mentions → USR
URLs → URL

Hashtags

preserved as tokens
(hash_topic)

Emojis & emoticons → **Punctuation** → **Output**

Emojis & emoticons

kept (😊 🔥 😭)

Punctuation

! ? retained

Output

clean tokens ready for
vectorization/embedding

Model Families

Traditional ML (TF-IDF features)

1. Logistic Regression
2. Linear SVM
3. Random Forest

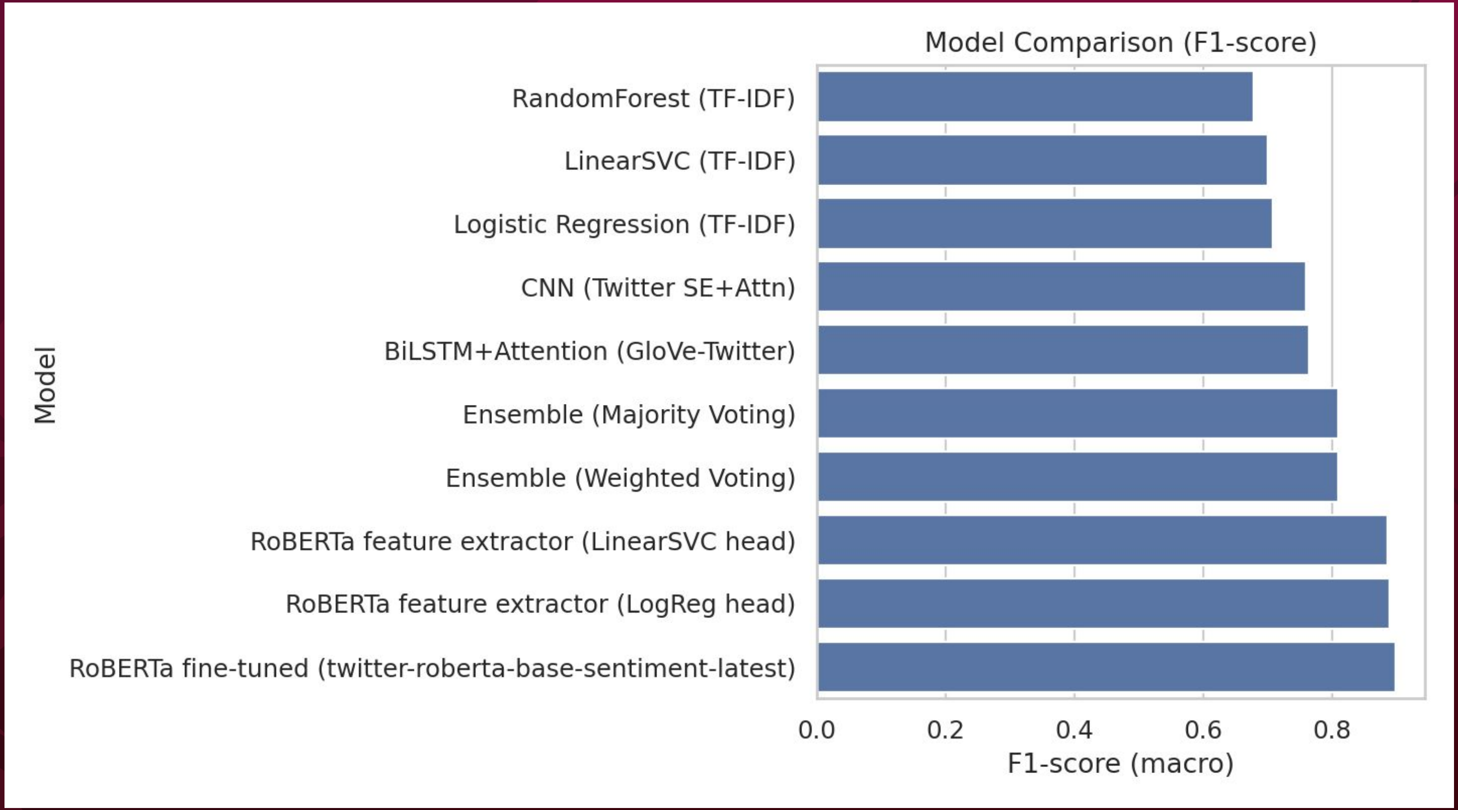
Deep Learning (GloVe embeddings)

1. CNN (SE + Attention)
2. BiLSTM + Attention

Transformers (RoBERTa)

1. Frozen feature extractor → LR / SVM heads
2. Fine-tuned end-to-end classifier
3. Ensembles: weighted and majority (fixed tie-breaker)

Overall Performance



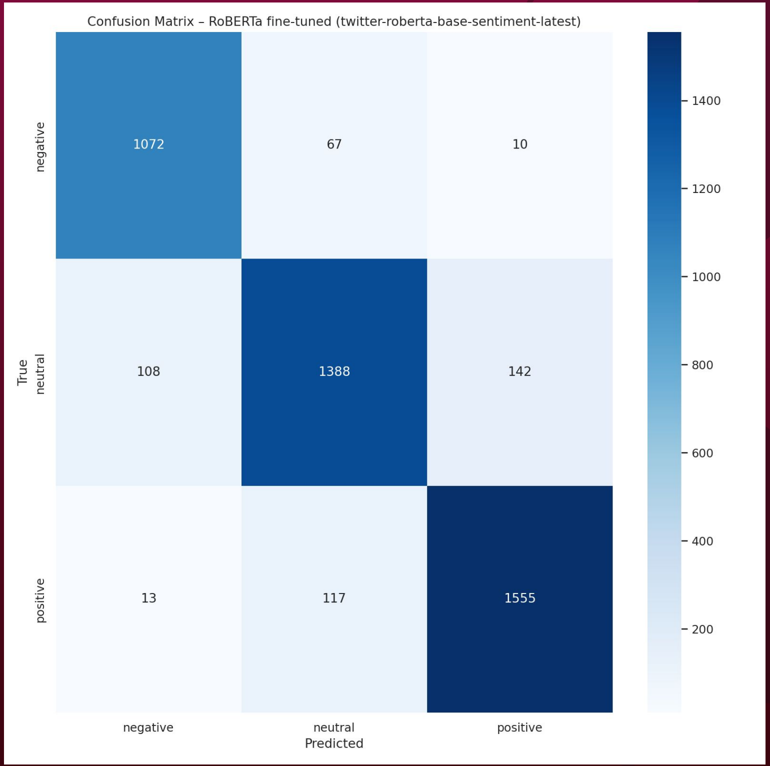
Overall Performance

Model	Accuracy	Precision	Recall	F1
RoBERTa fine-tuned (twitter-roberta-base-sentiment-latest)	0.897800	0.897500	0.901100	0.899000
RoBERTa feature extractor (LogReg head)	0.886900	0.889200	0.889000	0.889100
RoBERTa feature extractor (LinearSVC head)	0.884200	0.886700	0.886100	0.886300
Ensemble (Weighted Voting)	0.814200	0.811800	0.818900	0.814400
Ensemble (Majority Voting)	0.814200	0.811800	0.818900	0.814400
CNN (Twitter SE+Attn)	0.767900	0.766500	0.772300	0.768400
BiLSTM+Attention (GloVe-Twitter)	0.767000	0.764900	0.772500	0.767300
Logistic Regression (TF-IDF)	0.706800	0.705000	0.711900	0.707800
LinearSVC (TF-IDF)	0.698600	0.700100	0.699000	0.699400
RandomForest (TF-IDF)	0.677300	0.686900	0.671900	0.677400

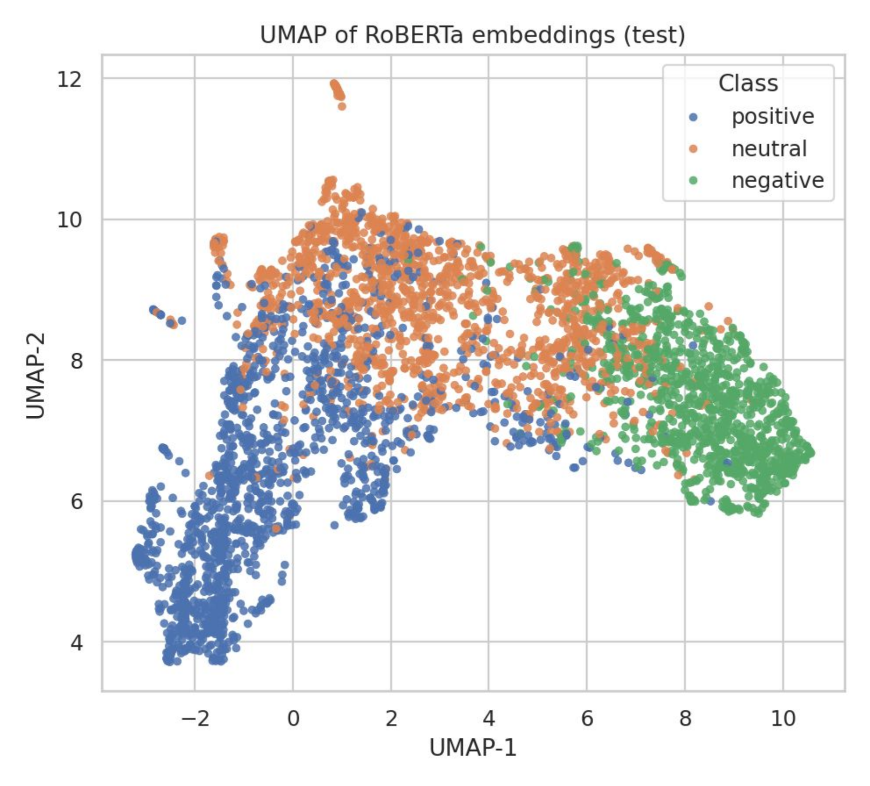
Per-Class Performance

Where errors happen

- Most **confusions: neutral ↔ positive.**
- **Negative** is well **separated** (high precision & recall).
- Neutral remains the **hardest boundary**.



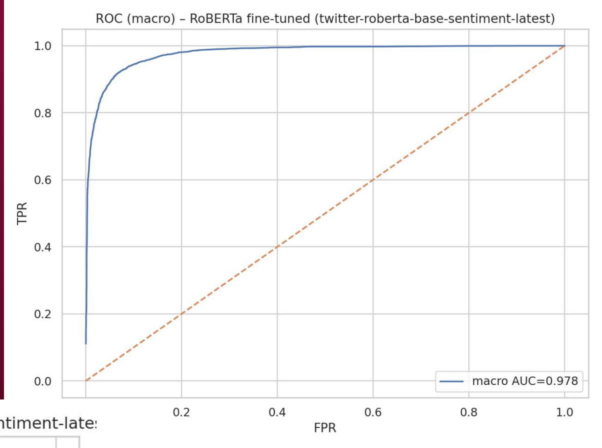
Model Behavior & Representation

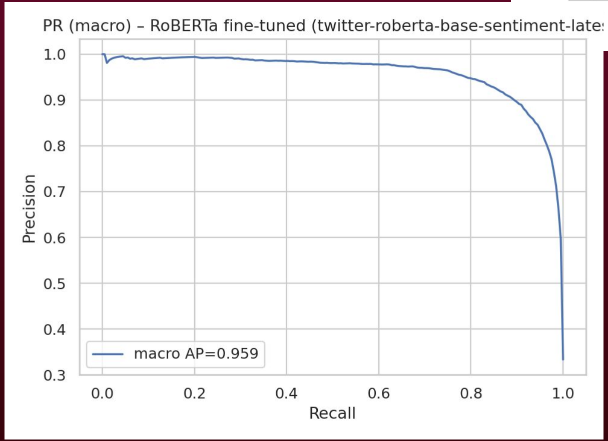


Model Behavior & Representation

Discrimination quality

- RoBERTa shows **strong separability** (high AUC).
- PR curves **confirm** robust **precision/recall** trade-offs.
- Indicates effective contextual encoding for sentiment polarity.





Discussion: Insights

What I learned

- RoBERTa **outperformed classical** ML models by **+19 pp** and **deep learning** models by **+13 pp** in **macro-F1**.
- **Contextual pretraining** handles **slang/sarcasm better** than GloVe.
- **TF-IDF** remains a **fast**, strong baseline for **low compute**.
- **Neutral** is the main **ambiguity**; data **near** the **boundary** drives **errors**.

Limitations & Future Work

Limitations

- **Temporal scope:** only Day 1 of World Cup.
- **English-only** subset (no multilingual coverage).
- **Limited GPU:** single-GPU training prevented multi-seed runs, longer sequences, or large backbones

Future Work

- **Cross-domain evaluation**: other days, clubs, or sports.
- **Cross-lingual robustness**: XLM-T, mDeBERTa, multilingual RoBERTa.
- **Larger or instruction-tuned transformers**: zero/few-shot setups.
- **Neutral-class refinement**.

Conclusions & Takeaways

- **Transformers dominate** on **context-rich** tweet sentiment.
- RoBERTa fine-tuned: Macro-F1 = 0.899 (best).
- GloVe + LSTM/CNN ≈ 0.76 → solid mid-tier trade-off.
- TF-IDF ≈ 0.70 → fast baselines for constrained compute.

Takeaway: Context is king for Twitter sentiment.

GitHub

https://github.com/enriquegomeztagle/MCD-NLP-SentimentAnalysisOfFIFATweets-FinalProject.git

"We Are 26"

Thanks!