CS 583 Research Project Report
Sentiment Classification of Obama and Romney Tweets Using RoBERTa

Yukta Salvi
University of Illinois Chicago
ysalv@uic.edu

Please Lukau
University of Illinois Chicago
pluka@uic.edu

Abstract

This project investigates sentiment classification of political tweets from the 2012 U.S. presidential election. We evaluate classical machine learning baselines using TF–IDF with Logistic Regression and Linear SVM, and compare them against a fine-tuned transformer model, `cardiffnlp/twitter-roberta-base-sentiment`. We preprocessed the data set to normalise noise, collapse label 2 (mixed sentiment) into the neutral class, and train a single classifier for the Obama and Romney models. RoBERTa substantially outperforms all baselines, achieving validation accuracies and achieves 75.5% for tweets on Obama and 71.3% for tweets on Romney. Finally, we developed a comprehensive prediction pipeline that can read the instructor's unlabeled test files and generate predictions in the required Lisp-style format.

1 Introduction

Political sentiment analysis is challenging due to the short, noisy, and context-dependent nature of tweets. The dataset for this project consists of two Excel sheets—Obama and Romney, each containing tweets labelled with sentiment toward the respective candidate. The tweets' label was classified into negative (-1), neutral (0), positive (1), or mixed (2). Following project guidelines, our task focuses exclusively on the three primary labels {-1, 0, 1}.

We preprocess the dataset to normalise noise, collapse label 2 (mixed sentiment) into the neutral class, and train separate Obama and Romney models. RoBERTa substantially outperforms all baselines, achieving validation accuracies of 75.5% (Obama) and 71.3% (Romney). Finally, we developed a comprehensive prediction pipeline that can read the instructor's unlabeled test files and generate predictions in the required Lisp-style format.

In this project, we worked on sentiment classification of political tweets from the 2012 U.S. presidential election, focusing on tweets about **Barack Obama** and **Mitt Romney**. The data is provided as an Excel file with two sheets:

- **Obama** – tweets annotated with sentiment labels toward Obama,
- **Romney** – tweets annotated with sentiment labels toward Romney.

Each tweet is assigned a discrete sentiment label (e.g., –1, 0, 1, 2). For this project, we focused on the three main classes: +1 (positive), –1 (negative), and 0 (neutral).

The objective of this work is to clean and preprocess the tweets, develop multiple classification models, and evaluate them on a held-out test set. Our goal is to determine which model best classifies the sentiment of political tweets.

2 Techniques

2.1 Data Preprocessing

Tweets were cleaned using a custom preprocessing function that:

- lowercases text,
- removes @mentions,
- removes URLs,
- removes HTML-like fragments,
- removes punctuation and collapses whitespace,
- preserves hashtag words but drops the # symbol.

Handling Label 2 (Mixed Sentiment)

The dataset contains a fourth label, 2, representing mixed or ambiguous sentiment. Because the project evaluates only the {-1, 0, 1} sentiment classes, we considered two strategies:

1. removing label 2 tweets entirely,
2. mapping label 2 to the neutral class (0).

Removing label 2 significantly reduced dataset size and increased class imbalance. Mapping label 2 to the neutral class yielded a more stable dataset and better performance across all models. Therefore, we applied the transformation:

$$\texttt{label 2} \rightarrow \texttt{0 (neutral)}$$

globally during preprocessing, before training any model.

2.2 Baseline Models

We implemented two classical baselines using scikit-learn:

1. **TF–IDF + Logistic Regression** using unigrams and bigrams with balanced class weights. Although the model performed reasonably well, accuracy remained limited because political tweets often contain sarcasm, implicit sentiment, and topic references that are not captured by surface-level n-grams.
2. **TF–IDF + Linear SVM** using the same TF–IDF representation with a max-margin classifier. While SVMs typically outperform Logistic Regression on sparse text data, the improvement here was marginal. The accuracy plateaued due to high variability in tweet length and vocabulary, leading to feature sparsity even with bi-grams.

Tables 1 and 2 summarize performance for Obama and Romney.

Table 1: Baseline performance on Obama tweets (validation split).

Model	Accuracy	Macro F1
TF–IDF + Logistic Regression	0.672	0.66
TF–IDF + Linear SVM	0.659	0.64

Table 2: Baseline performance on Romney tweets (validation split).

Model	Accuracy	Macro F1
TF–IDF + Logistic Regression	0.629	0.60
TF–IDF + Linear SVM	0.612	0.58

2.3 Fine-Tuning RoBERTa

Before turning our attention to RoBERTa, we also experimented with SBERT and DistilBERT. While both models provided meaningful sentence-level embeddings, they were not able to adequately capture the nuanced and context-dependent nature of political tweets. This motivated us to adopt a model specifically designed for social media text.

We therefore fine-tuned `twitter-roberta-base-sentiment` separately for the Obama and Romney datasets using an 80/20 stratified train–validation split. The model, available on Hugging Face, is pretrained on large-scale Twitter data, making it well-suited for sentiment analysis in this domain.

Training details:

- labels {-1, 0, 1} mapped to {0, 1, 2},
- tokenization with a 96-token maximum length,
- batch size 16 (train) / 32 (eval),
- learning rate 5×10^{-5},
- weight decay 0.01,
- early stopping with patience 1,
- model selection based on macro-F1.

Each fine-tuned model was saved under `./models/roberta_obama_cardiff` and `./models/roberta_romney_cardiff`.

2.4 Demo-Day Prediction Pipeline

The instructor's final test files contain two columns and no header: tweet number and raw tweet. Our prediction pipeline:

1. reads the test file using `header=None`,
2. applies the same cleaning used during training,
3. loads the correct fine-tuned RoBERTa model,
4. generates sentiment predictions in {-1, 0, 1},

5. outputs results in the required format:

```
(setf x *(
(1 0)
(2 -1)
...
) )
```

This ensures compatibility with the instructor's evaluation scripts.

3 Experiment Results

Table 3 summarizes the validation results for Obama and Romney.

Table 3: RoBERTa validation results (20% stratified split).

Candidate	Accuracy	Macro Precision	Macro Recall	Macro F1
Obama	0.755	0.755	0.744	0.748
Romney	0.713	0.709	0.673	0.687

Class-wise performance for Obama is balanced across all sentiment labels. For Romney, the positive class is more challenging (F1 = 0.60), reflecting class imbalance and more ambiguous language in supportive Romney tweets.

4 Conclusion and Lessons Learned

We built a full sentiment classification pipeline for political tweets, from preprocessing to baseline models, transformer fine-tuning, and final demo-day predictions. Classical TF–IDF models achieved moderate performance, while fine-tuned RoBERTa models performed substantially better, especially for Obama tweets.

Key lessons include:

- consistent preprocessing is critical for reliable performance,
- mixed-sentiment labels should be normalized to reduce noise,
- transformer models excel at capturing the nuanced language of political discourse,
- class imbalance remains a major challenge for low-resource sentiment classes.

Future improvements may include sarcasm detection, domain adaptation, or multi-task modeling of both candidates simultaneously.

References

[1] Hugging Face Transformers. `https://huggingface.co/transformers/`

[2] Cardiff NLP Twitter RoBERTa.
`https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment`

[3] Pedregosa et al., Scikit-learn: Machine Learning in Python, JMLR, 2011.