Transformer-Based Sentiment Classification Across Multiple Datasets

CMPE 346 Final Project

Devrim Kıvrak - 25/05/2025

Project Summary

I explored the task of sentiment classification using transformer based models.

This presentation summarizes the pipeline and results.

Motivation

Sentiment analysis is vital for understanding user feedback in many domains.

I aimed to evaluate BERT based model performances across different datasets.

Main question is how accurately can we classify sentiment from various text datasets?

Chosen Datasets

I used IMDB, Yelp, and Amazon Reviews datasets. They differ in domain and linguistic style, but based on similar text-label system.

Data Preparation

I applied tokenization, truncation at 256 tokens, and label normalization. Class distributions were balanced for BERT based models.

Model List

Five models were used: BERT, DistilBERT, RoBERTa, XLM-RoBERTa, and ALBERT. All were fine-tuned using HuggingFace Transformers.

Training Details

Models were trained for 1 epoch with a batch size of 16. I used AdamW optimizer with default HuggingFace settings.

Due to the length of the training, training and testing datasets' length reduced significantly.

Evaluation Metrics

We evaluated models using Accuracy and F1 Score. These metrics reflect both precision and recall.

Results – BERT

BERT performed best on the Yelp dataset (F1: 0.8082). It showed moderate performance elsewhere.

Results – DistilBERT

DistilBERT achieved the highest F1 on Yelp (0.9038). It maintained strong results across all sets.

Results – RoBERTa

RoBERTa achieved the best F1 on IMDB (0.8898). It also performed well on Amazon reviews.

Results – XLM-RoBERTa

XLM-RoBERTa struggled on IMDB (F1: 0.3554) but performed better on Yelp. Amazon performance was unstable.

Results – ALBERT

ALBERT achieved the highest F1 on Amazon (0.9533). It consistently ranked near the top across all datasets.

Comparison Table

Model	IMDB (F1)	Yelp (F1)	Amazon (F1)
BERT	0.7085	0.8082	0.9331
DistilBERT	0.8398	0.9038	0.9205
RoBERTa	0.8898	0.8491	0.9488
XLM-RoBERTa	0.3554	0.8430	0.705
ALBERT	0.8839	0.8655	0.9533

Review

Larger models provided better results but demanded longer training. There was a correlation between training time and accuracy.

Limitations

I limited training to one epoch due to compute limitation.

No hyperparameter tuning was performed, parameters remained same.

Future Work

I plan to increase the number of training epochs and use larger datasets to improve accuracy and F1-scores.

Thank you for listening.
I would like to answer if you have any questions.