BERT Sentiment Analysis

Fine-tuning a DistilBERT model for review sentiment analysis

Sentiment Analysis

- Sentiment Analysis is a branch of Natural Language Processing (NLP) that categorizes text input according to the emotional tone of the writer.
- Use cases include determining consumer attitude toward a business or product
- Essential for automating customer feedback and for market research

Purpose

BERT

(Bidrectional Encoder Representations from Transformers)

A type of Large Language Model, contrasted to GPT models

Optimized for specific use cases in Natural Language Processing.

HuggingFace

An open-source community for AI data science tools and collaboration

Provides a wide selection of pre-trained BERT models for various applications.

This project

Fine-tune and train a BERT model

Train this model for sentiment analysis based on customer reviews.

DistilBERT

DistilBERT is a distilled version of the BERT model

Compared to base BERT model:

40% less parameters	60% faster	Preserves 95% of performance

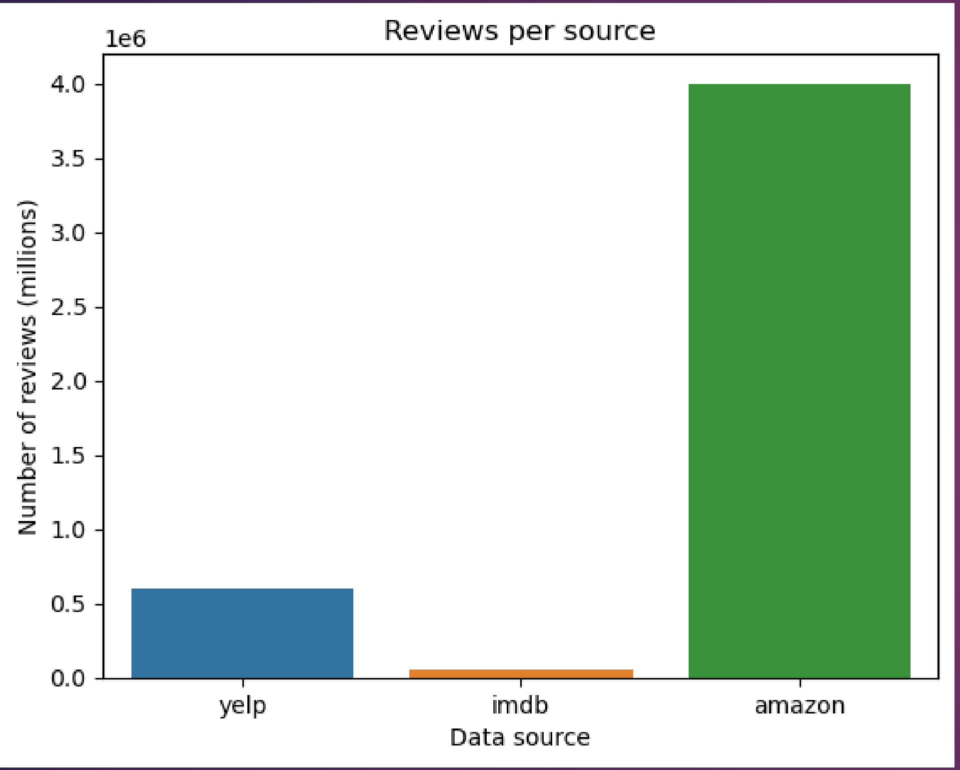
Objectives

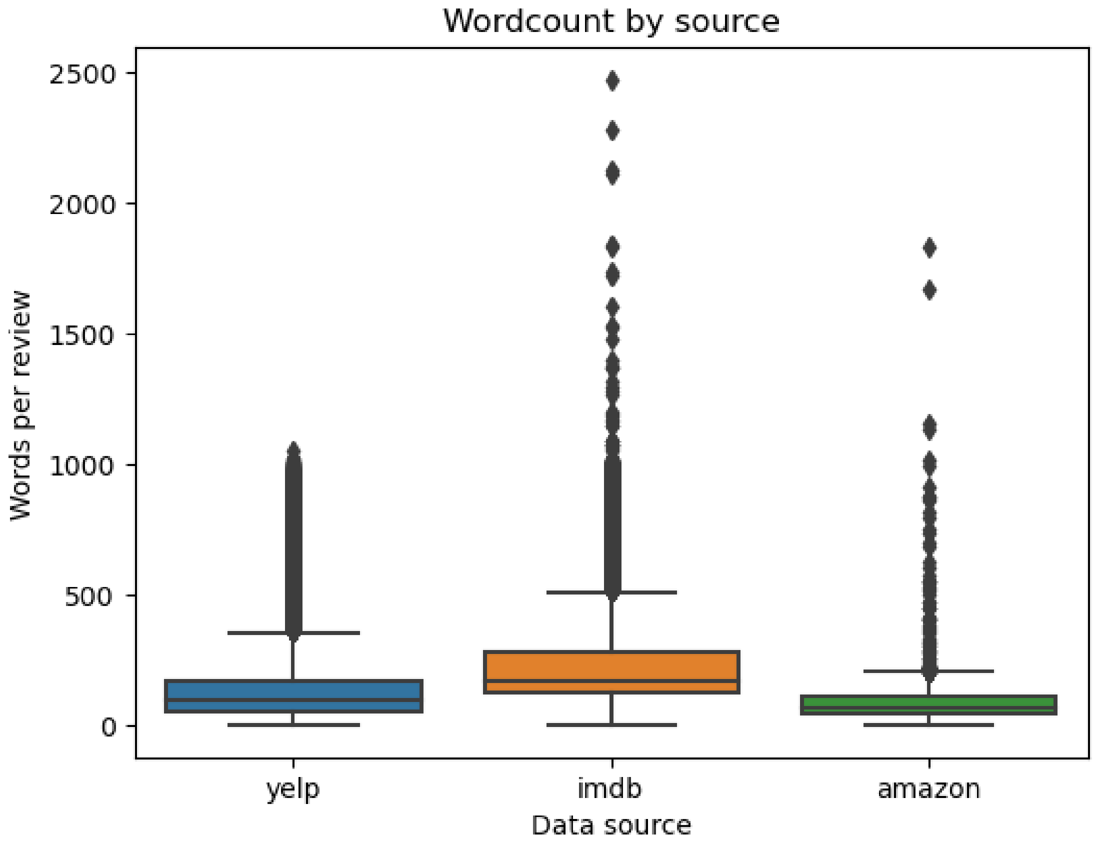
- Create a supervised Machine Learning model for binary classification of text input
- Given text input, predict whether it belonged to a review with a high or low star rating.

Data Sources

- Three Kaggle Datasets with labeled data:
  - “Yelp Review Sentiment Dataset”: ~600 thousand business reviews (Ilham Firdausi Putra)
  - “IMDB Dataset of 50K Movie Reviews”: 50 thousand movie reviews (Lakshmipathi N)
  - “Amazon Reviews for Sentiment Analysis”: 4 million product reviews (Adam Bittlingmayer)
- Data sets were cleaned and unified for consistent model training

Data distribution





Methodology

- Preprocess data into the form DistilBERT expects
- Use DistilBERT's Tokenizer to turn text into numbers
- Hyperparameter tuning with Optuna library
- Train on subset of review data
- Evaluate on unseen data

Model Training

Accelerator library allowed for distributed processing

Google Colab provided necessary power

HuggingFace Transformers library facilitated the training process

Best hyperparameters

- Using the Optuna library on a subset of training data, these hyperparameters were found:
- {'learning_rate': 4.122342215733177e-05, 'num_train_epochs': 1, 'gradient_accumulation_steps': 2, 'per_device_train_batch_size': 12, 'evaluation_strategy': 'epoch', 'per_device_eval_batch_size': 5, 'warmup_steps': 391, 'weight_decay': 0.08139944860406301}

Results

Primary metric: F1-score

- Neither false positives nor false negatives were more important to minimize

F1-score for both categories was ~0.96

Precision and recall were approximately equal at 0.96

Exceptionally strong predictive power

Future Work

- Some interesting extensions to this project could include:
  - A case-sensitive model could capture nuances indicated by capitalization
  - A model that detects and eliminates non-English reviews from the dataset
  - A full BERT model, rather than DistilBERT could slightly improve the predictive power
  - A model trained on a larger subset of the data may decrease variance

References

- HuggingFace Transformers documentation
  https://huggingface.co/docs/transformers/index
- Kaggle datasets
  Yelp: https://www.kaggle.com/datasets/ilhamfp31/yelp-review-dataset
  IMDB: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews/
  Amazon: https://www.kaggle.com/datasets/bittlingmayer/amazonreviews

Thanks to

The HuggingFace community, for the models, documentation, and support forums

My Springboard mentor, Chris Esposo, for introducing me to BERT models and non-stop support along the way