SetFit for Automated Essay Scoring

Leon Krug, Jannik Bundeli, Jannine Meier, Elena Nazarenko

Lucerne University of Applied Sciences and Arts
HOCHSCHULE LUZERN
Informatik

The Objective

Problem Definition

Accurately predict essay scores, capturing both the quantitative and qualitative aspects of writing

Dataset Overview

Real student essays that cover different topics and writing styles

Objective of the Challenge

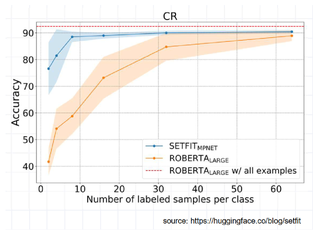
Score essays on a scale from 1 to 6

Multiclass Classification

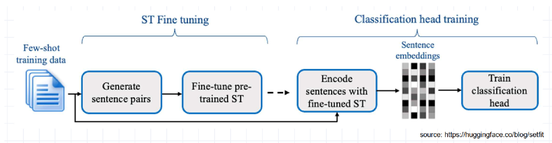
What is Setfit?

SetFit (Set-Factory for Few-shot Text Classification) is a powerful text classification framework designed to work well even with very few labeled examples (few-shot learning). It combines Sentence Transformers with efficient fine-tuning techniques to deliver strong performance using minimal training data.

Key Features:
Few-shot capable: Trains models with as few as 8–64 labeled samples.
No need for large GPUs: Very efficient, often runs on CPU.
Fast fine-tuning: Trains in seconds to minutes.
High accuracy: Competes with full fine-tuning methods on many benchmarks.



SetFit's two-stage Training



- positive and negative pairs
- generates dense vectors
- unseen example passes through the fine-tuned Sentence Transformer

The Dataset

Overview of Dataset Composition

Public Dataset (Train/Eval Dataset)
- **Persuade 2.0**: Around 13'000 essays.
- **Kaggle-Only**: Around 4'500 essays.

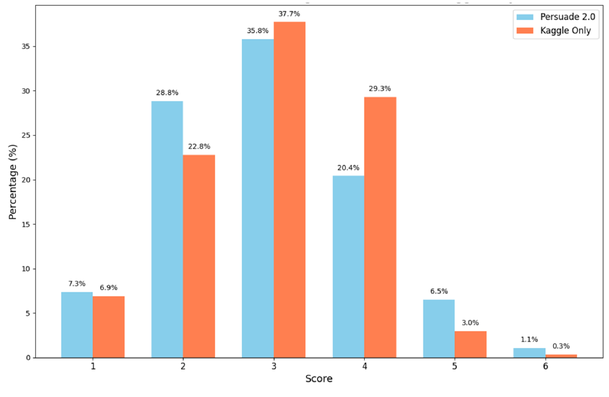
Private Dataset (Test Dataset)
- **Kaggle-Only**: Around 6'500 essays.

Dataset Segmentation

Cross-referencing
- Labeling essays originating from the Persuade 2.0 corpus within our training dataset

Creating two subsets
- Original (Mixed) Dataset
- Kaggle-Only Dataset

The Score Distribution



The Score Description

6	Clear mastery with few errors, outstanding critical thinking, appropriate evidence, well-organized, skilled language use.
5	Reasonable mastery with occasional errors, strong critical thinking, generally appropriate evidence, well-organized, good language use.
4	Adequate mastery with some lapses, competent critical thinking, adequate evidence, generally organized, fair language use.
3	Developing mastery with weaknesses, limited critical thinking, inconsistent evidence, limited organization, fair language use with weaknesses.
2	Little mastery with serious flaws, weak critical thinking, insufficient evidence, poor organization, limited language use with frequent errors.
1	Very little or no mastery, severely flawed, no viable point of view, disorganized, fundamental language flaws, pervasive grammar/mechanics errors.

Our Innovation

SetFit + Longformer Sentence Transformer

Challenge
Token limitations with normal base models (512 Tokens)

Old base model:
- Transformer model allenai/longformer-base-4096
- Not a SentenceTransformer

anchor	positive	negative
Children smiling and waving at camera	There are children present	The kids are frowning

Our base model
- SentenceTransformer based on allenai/longformer-base-4096
- Trained on:
  - all-nli-pair
  - all-nli-pair-class
  - all-nli-pair-score
  - all-nli triplet
  - Stsb
  - Quora
  - natural-questions

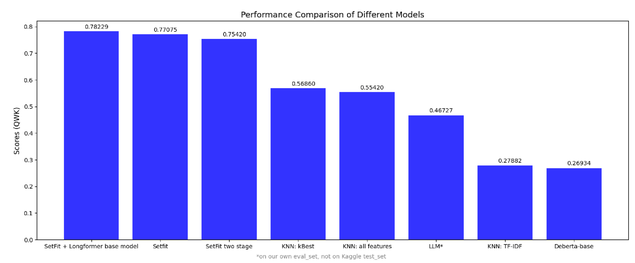
Maps sentences & paragraphs to a 768-dimensional dense vector space

Longformer as a Sentence Transformer
Old context size: 512 Tokens
New context size: 4096 Tokens
8x longer context

1 essay page (512 tokens)

8 essay pages (4096 tokens)

The Results



Quadratic Weighted Kappa Metrics (QWK)

- Measures agreement between predicted and actual scores
- Weighted: Penalizes larger disagreements

Range: -1 to 1

1: Perfect agreement
0: Random agreement
< 0: disagreement

Interpretation:
0.8–1.0: Almost perfect agreement
0.6–0.8: Substantial agreement
0.4–0.6: Moderate agreement
< 0.4: Poor agreement