UNIPD

Automated Essay Scoring: Methodologies, Limitations, and Knowledge Transfer

Lorenzo D’Antoni, Alessandro Canel, Davide Bassan, Hannaneh Kalantary and Hooman Sabzi

Automated essay scoring

What?

- A **computer-based** system tasked with **evaluating** student **essays**.
- nuanced and complex field

Why?

- Manual grading is **time-consuming**, **unreliable**, **expensive**, **not scalable**.
- Existing models fall short in terms of performance and generalizability.

Approach

- Test different approaches
- Assess if knowledge transfer is feasible

Related Work

The birth of AES

Project Essay Grader PEG (1966)

↓

Intelligent Essay Assessor IEA (1999)

↓

E-rater (2002)

Test Scoring System (2002)

Intellimetric (2006)

Deep Learning (2014)

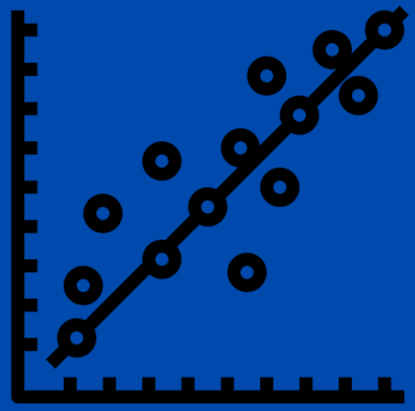
- both syntactic and semantic features
- performance improved

AES Technical Details

Grading Techniques

01

Regression

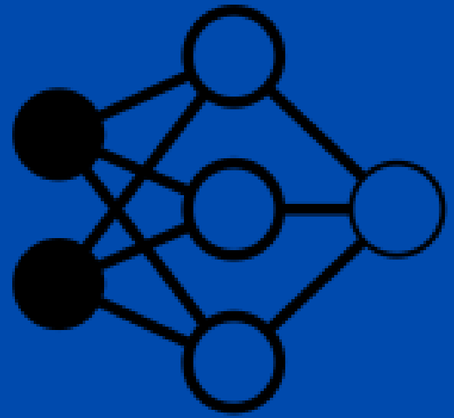


02

Classification

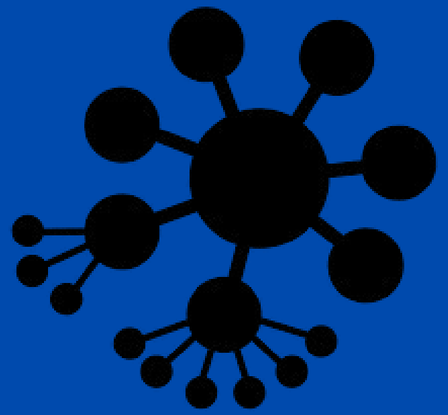
03

Neural networks

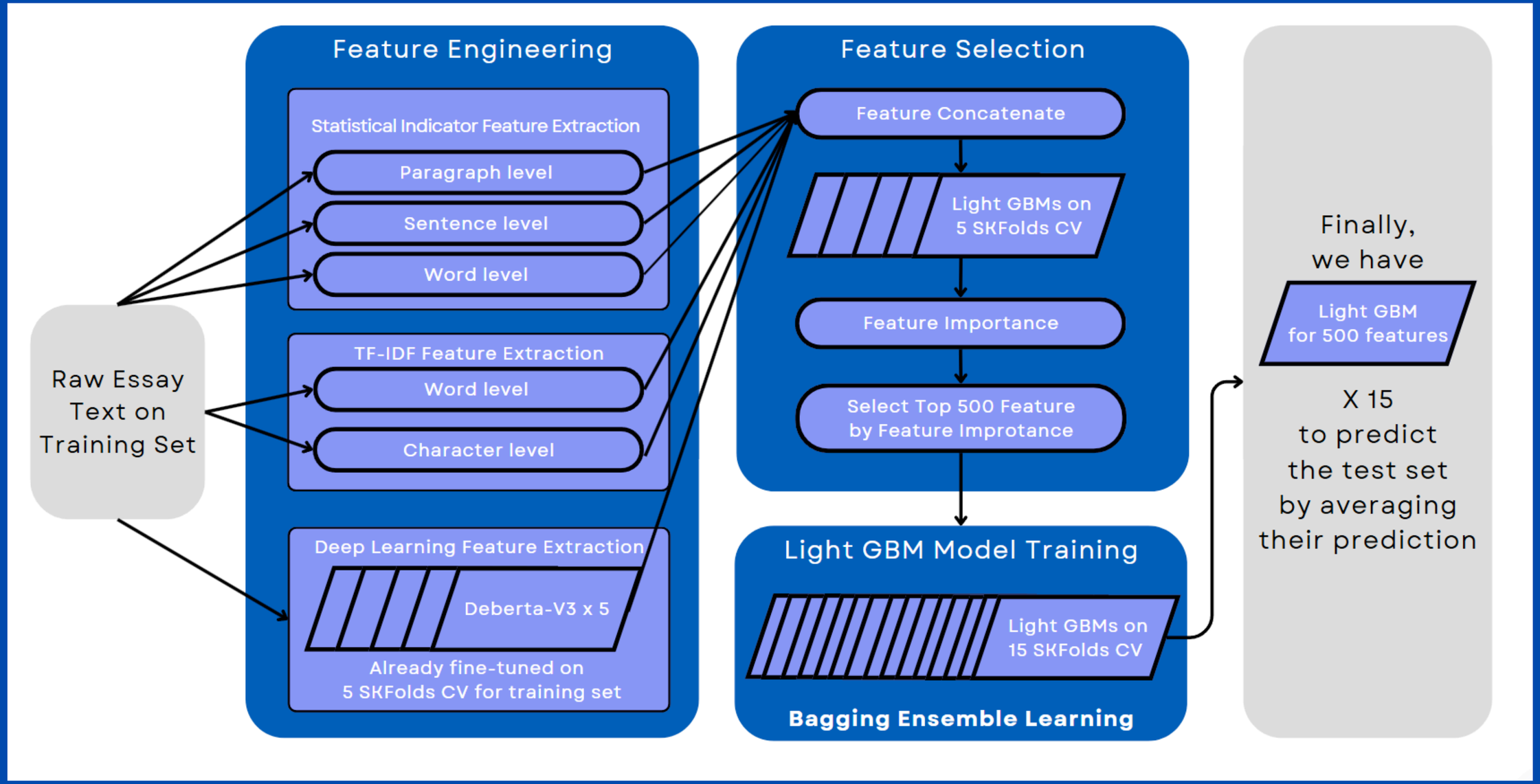


04

Ontology-based



Cutting-edge models



Challenges & limitations

Metrics

Quadratic Weighted Kappa does **not sufficiently** assess **feature extraction** or **detect irrelevant answers.**

Irrelevant content

Models often **fail to recognize** and appropriately score **irrelevant** or **adversarial content**.

Semantic features

Models do **not effectively evaluate content relevance** to the prompts and adequately address **cohesion and coherence**.

Understanding the Datasets

Analyzing and visualizing the datasets

Objectives

- Familiarize with the dataset
- Visualize dataset characteristics
- Gain insights for model development

Tasks

- Summarize dataset characteristics
- Analyze size, distribution and ...
- Create graphical representations

Chllenges:

- finding suitable datasets
- preparing datasets for training models

Datasets

We need a large and balanced dataset

Automated Essay Scoring Competition

The competition dataset comprises about 24000 student-written essays. Each essay was scored on a scale of 1 to 6. The goal is to predict the score of an essay.

IELTS Writing Scored Essays Dataset

A Comprehensive Dataset for IELTS Writing Tasks: Sample Essays and Scores

Automated Essay Scoring Competition

	essay_id	full_text	score
0	000d118	Many people have car where they live. The thin...	3
1	000fe60	I am a scientist at NASA that is discussing th...	3
2	001ab80	People always wish they had the same technolog...	4
3	001bdc0	We all heard about Venus, the planet without a...	4
4	002ba53	Dear, State Senator\n\nThis is a letter to arg...	3

Source: Kaggle

train, test and sample submission

train dataset:

- 17307 rows (Essays written by 2400 students)
- 3 columns
- score range (1 to 6)

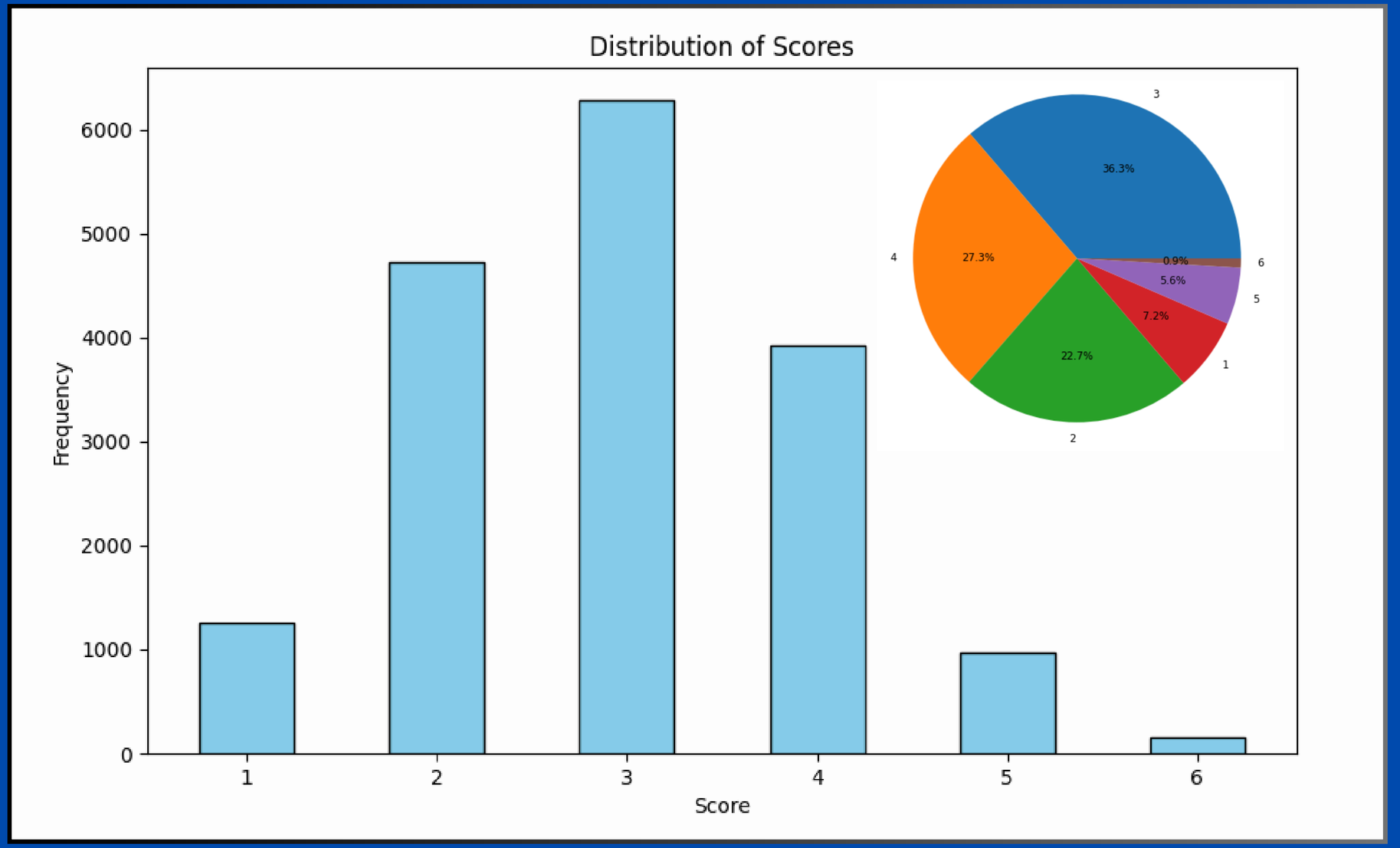
Rich but imbalance dataset

Score Distribution Plot

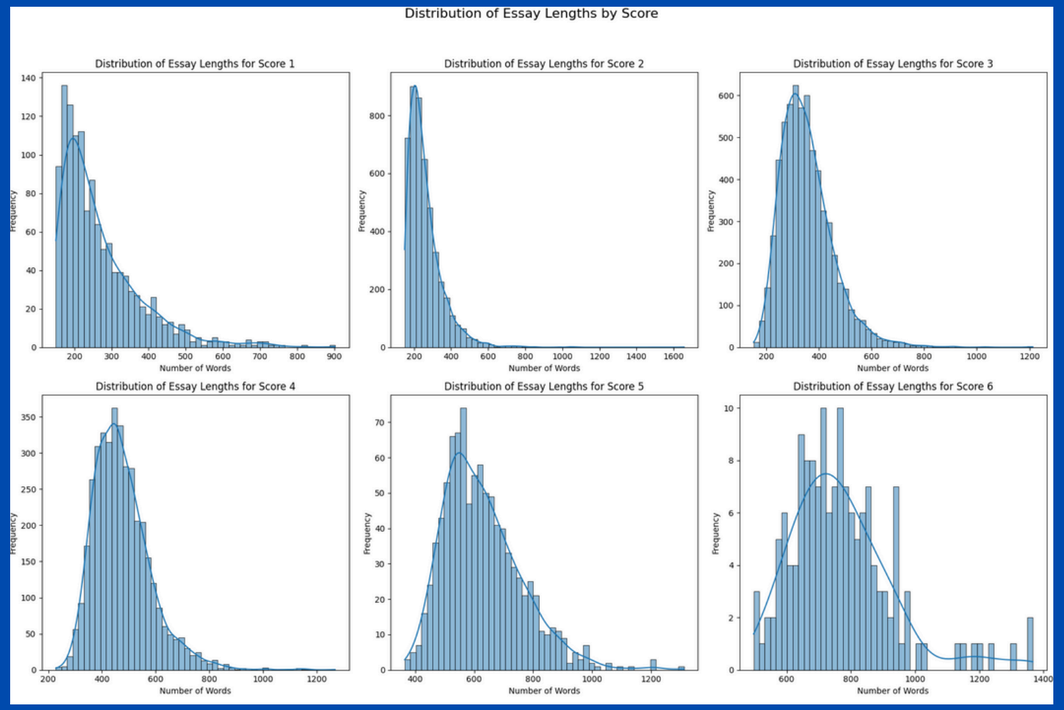
Stats:

- Mean Score: 2.95
- Median Score: 3.0
- Mode Score: 3
- Standard Deviation of Scores: 1.04

Imbalance between different classes

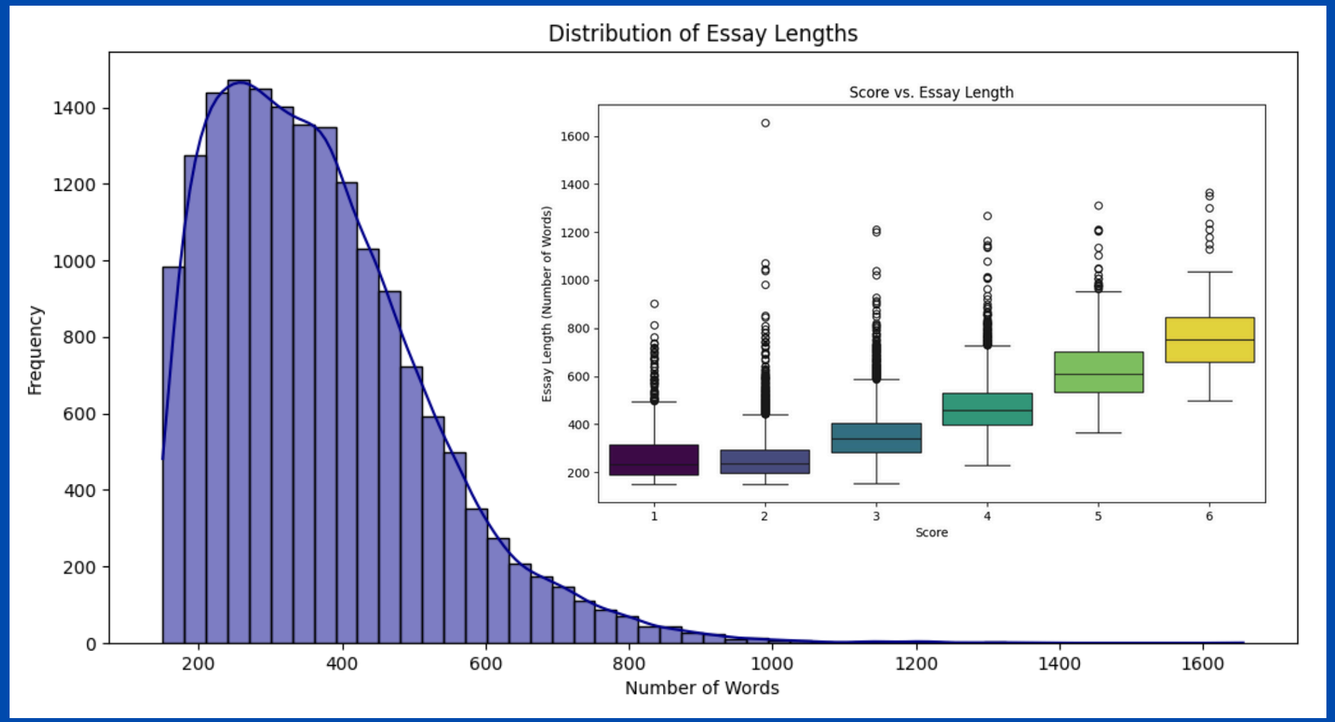


Distributions



- **Distributions of Essay Lengths by Score**

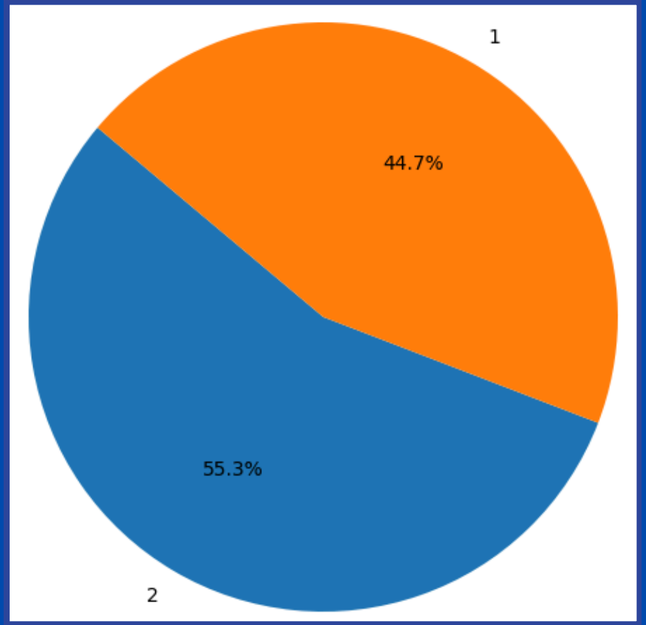
- **Distribution of Essay Lengths**
- **Score VS. Essay Lenths**



IELTS Writing Scored Essays Dataset

	Task_Type	Question	Essay	Overall
0	1	The bar chart below describes some changes abo...	Between 1995 and 2010, a study was conducted r...	5.5
1	2	Rich countries often give money to poorer coun...	Poverty represents a worldwide crisis. It is t...	6.5
2	1	The bar chart below describes some changes abo...	The left chart shows the population change hap...	5.0
3	2	Rich countries often give money to poorer coun...	Human beings are facing many challenges nowada...	5.5
4	1	The graph below shows the number of overseas v...	Information about the thousands of visits from...	7.0

Proportion of each task type



Source: Kaggle

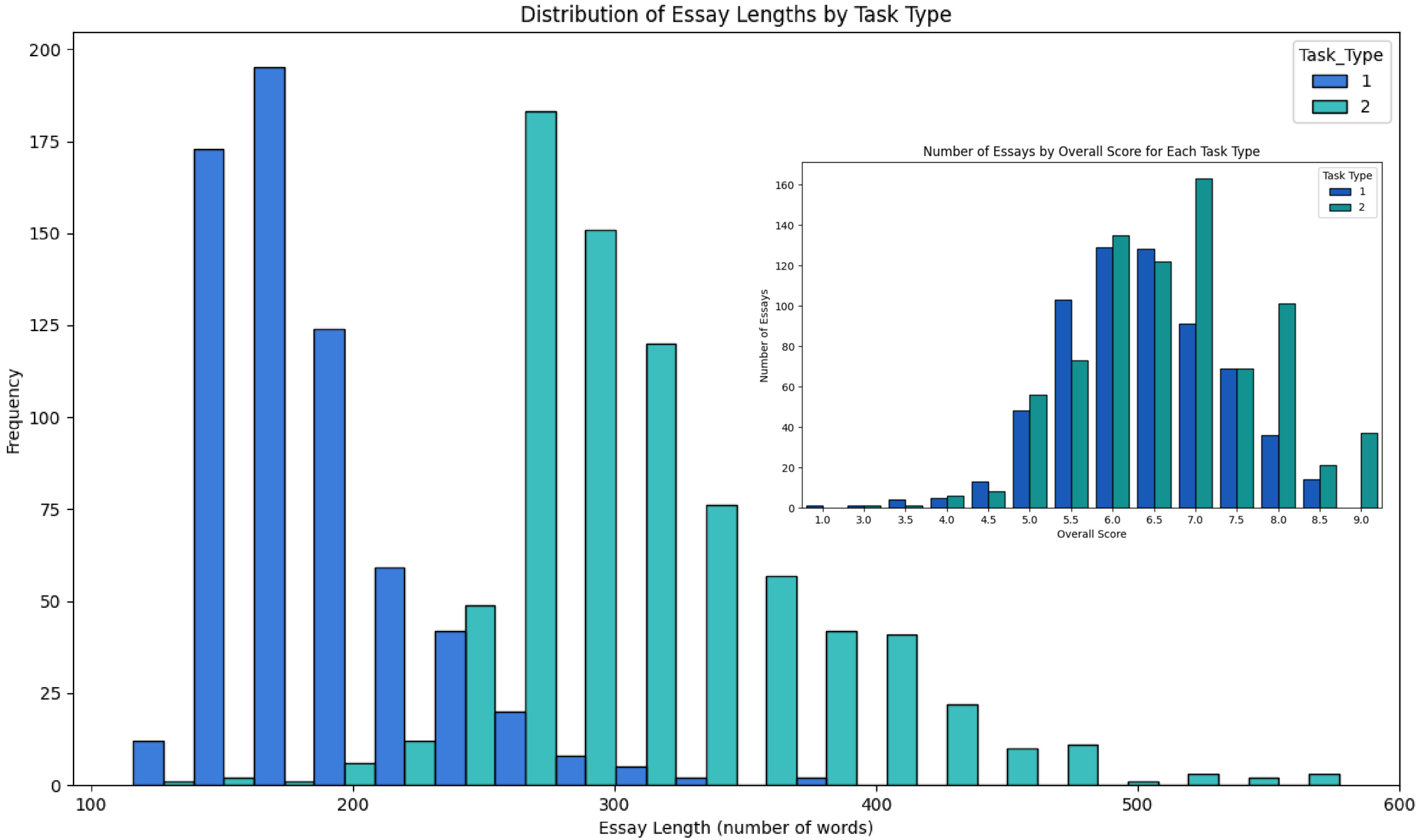
Size:

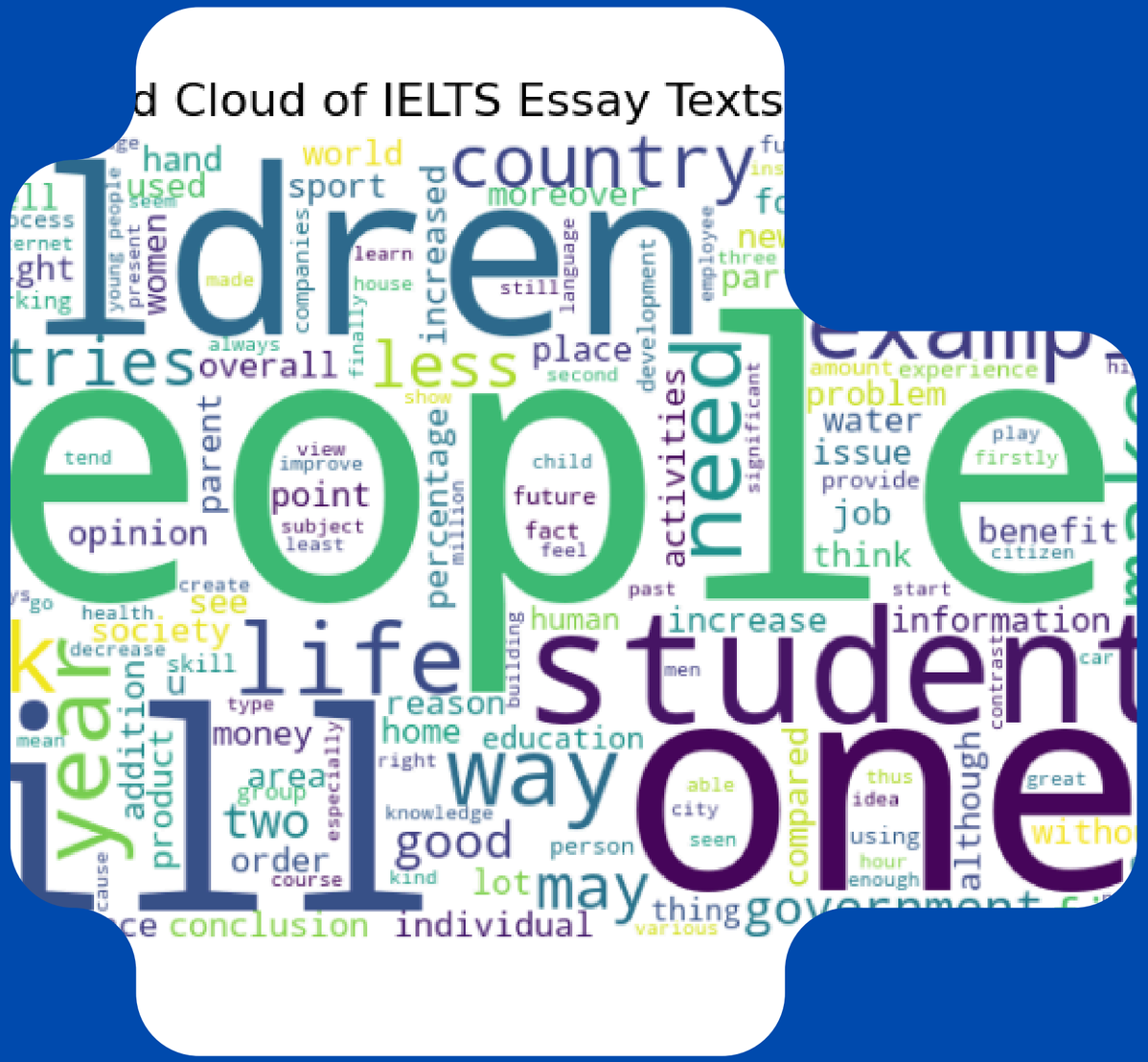
- **1435 rows**
- 4 useful columns
- The other 5 columns are NaN

Stats:

- Mean Score: 6.5
- Median Score: 3.0
- Mode Score: 6.0
- Standard Deviation of Scores: 1.06
- Minimum Score: 1.0
- Maximum Score: 9.0

Distributions





Preprocessing the Data

Ensured minimum of 3 essays per score category

Removed some scores and essays

Created a workable dataset for model development

Dataset imbalance posed a significant challenge

Our efforts in data selection and preprocessing enabled us to move forward with model training

The insights gained from visualizing and analyzing the dataset characteristics were crucial in understanding the data and refining our approach

Experiments

Baseline

1

k-NN (text+score)

Val. accuracy: 0.35

2

k-NN

(with more parameters)

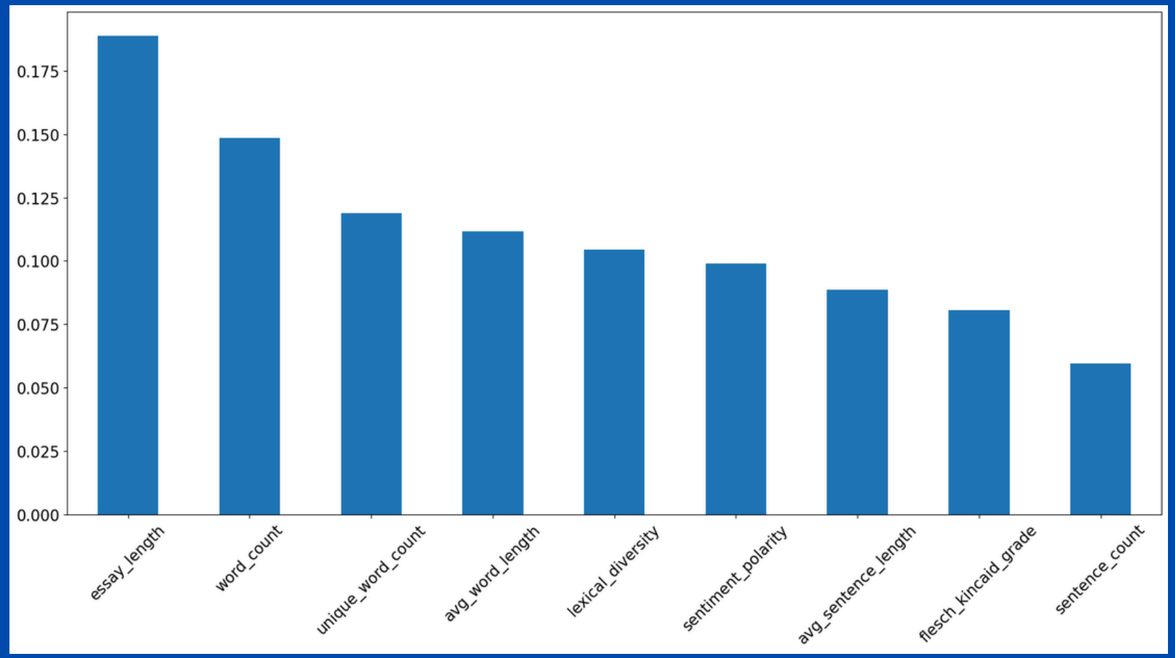
Val. accuracy: 0.48

Bias problem

3

optimize number of neighbors k

Val. accuracy: 0.55



5

ensemble (3 k-NNs)

Val. accuracy: 0.57

4

additional features

Val. accuracy: 0.51

Hugging Face's Transformers + **TensorFlow**

Experiments

Fixed errors in Github repo

missing test set

missing padding

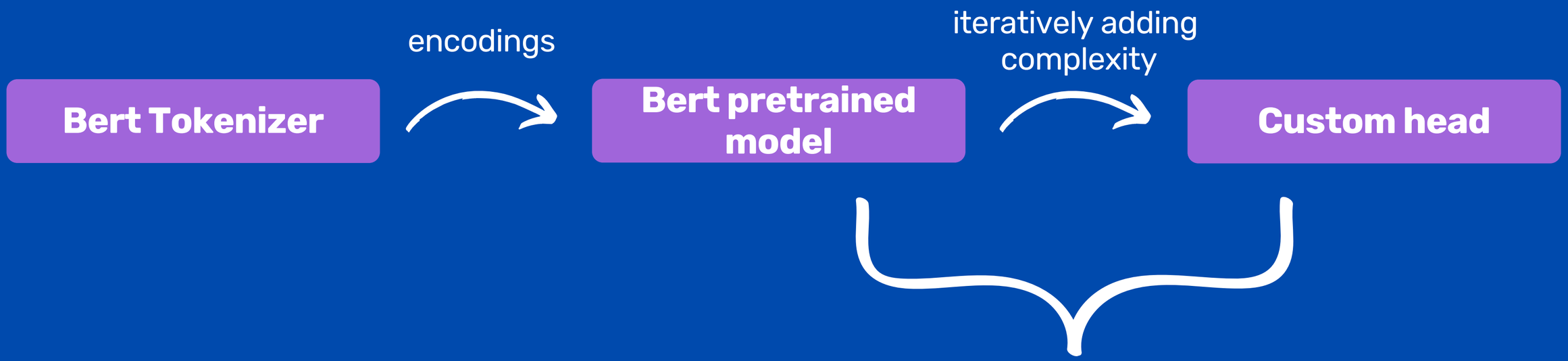
Max length problem

512 as max length for tokenized sequences

information loss vs. poor performance

From DeBERTa to BERT

Deep Learning Approach



Transfer learning

Gradually unfreeze a few **layers** at a time.
From the top (near the output) moving towards the input.

Classification approach

Dense(128, relu)
Dropout(0.3)
Dense(64, relu)
Dropout(0.3)
Dense(6, softmax)

Model Performance

Highest accuracy on the test set was **0.57**.
When classifying IELTS essays the accuracy dropped to **0.49**

With **ordinal data** a traditional classification approach that treats each misclassification with equal weight might not be appropriate, **misclassifying** a class by **one level** should be **penalized less** than **misclassifying** it by **several levels**

Regression approach

Ordinal regression

- Adopted KerasNLP
- From scores to ordinal matrix
- BERT variant
- QWK: 0.73, Accuracy: 0.58
- Problems with fine-tuning the IELTS dataset

3 -> [1, 1, 1, 0, 0, 0]

Linear approach

- linear activation function
- no need to scale scores
- allows for truly evaluating knowledge transfer
- training problems (see next slide)

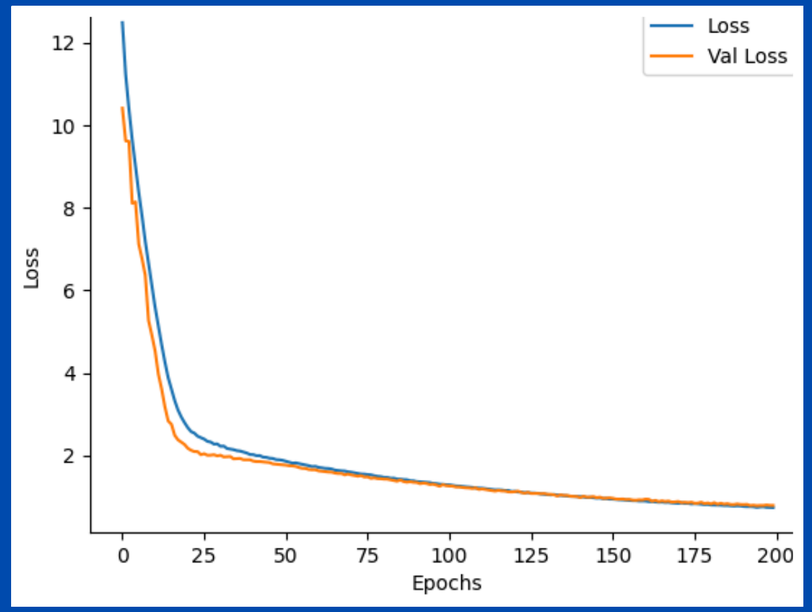
Tanh approach

- from [1, 6] to the [-1, 1]
- MSE on scaled data: 0.1
- MSE on original data 1
- high MSE on original data
- Problems with fine-tuning the IELTS dataset

Linear approach

it does **not leverage domain knowledge** of the scores, and **failing** to **normalize the scores** introduces potential issues with **exploding gradients** and **unstable learning**.

Dense(512, relu)
Dropout(0.3)
Dense(256, relu)
BatchNormalization()
Dropout(0.3)
Dense(256, relu)
BatchNormalization()
Dropout(0.3)
BatchNormalization()
Dense(128, relu)
BatchNormalization()
Dense(1, linear)



Test evaluation

Mean Squared Error	Root Mean Squared Error	Mean Absolute Error	R2
0.62	0.51	0.66	0.57

IELTS fine-tuning

Linear approach

Allows for **evaluating knowledge transfer**

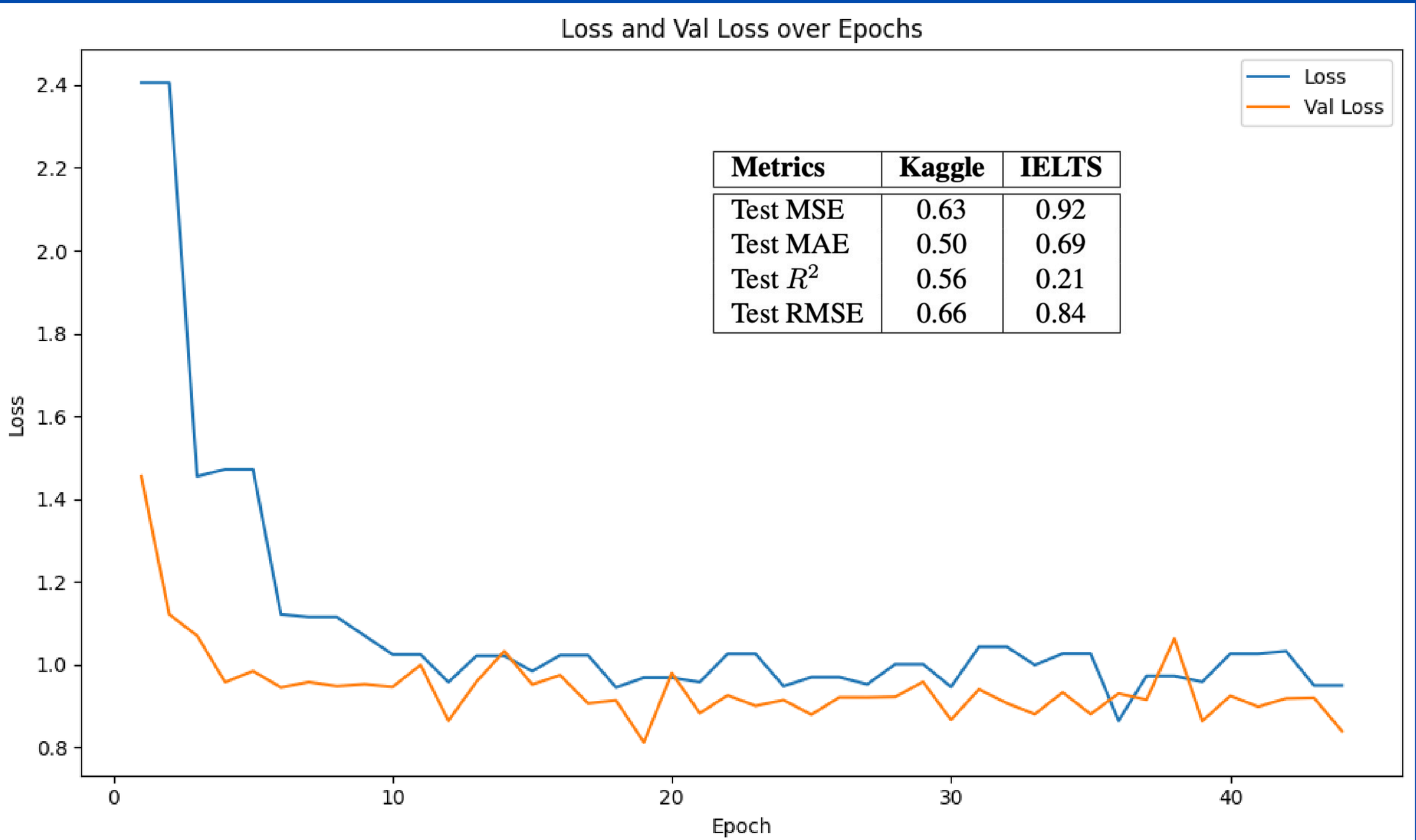
But there is **no noticeable transfer** of **knowledge**

Freezing & unfreezing weights

Tried various configurations but couldn’t improve performance

Learning rate scheduler (ReduceLROnPlateau)

Tested many hyperparameter values but did **not achieve significant improvement**



Metrics	Kaggle	IELTS
Test MSE	0.63	0.92
Test MAE	0.50	0.69
Test R^2	0.56	0.21
Test RMSE	0.66	0.84

Final Comparison

Baseline k-NN	Accuracy: 0.57
Classification approach	Accuracy: 0.57
Ordinal regression	QWK: 0.73, Accuracy: 0.58
Regression (tanh)	MSE: 1
Regression (linear)	MSE: 0.63

Performance is very low but in line with the state-of-the-art models

This suggests that deep learning models are not providing the big boost that they have in other areas of AI.

We could not get better results than the baseline.

Bad performance. Why?

Unbalanced & small datasets

Resource constraints

BERT input length

Semantic vs Syntactic

Conclusion

01

Key Findings

- Transfer learning between different datasets is highly constrained.
- Effective AES: syntactic & statistical features

02

Limitations

- Models fail to evaluate essay's relevance to prompts.
- Lack of coherence and cohesion analysis.

03

Future Research

- Prioritize content relevance, cohesion, and coherence.
- Provide constructive feedback to students.

Lorenzo D’Antoni
2073767

Davide Bassan
2076779

Thank You!

Canel Alessandro
2097570

Hannaneh Kalantary
2106028

Hooman Sabzi
2119061

Canel Alessandro
2097570

Lorenzo D’Antoni
2073767

Hannaneh Kalantary
2106028

Hooman Sabzi
2119061

Davide Bassan
2076779

Addressing Imbalancement with SMOTE

Challenges with Data Imbalancement

Solution: SMOTE (Synthetic Minority Over-sampling Technique)

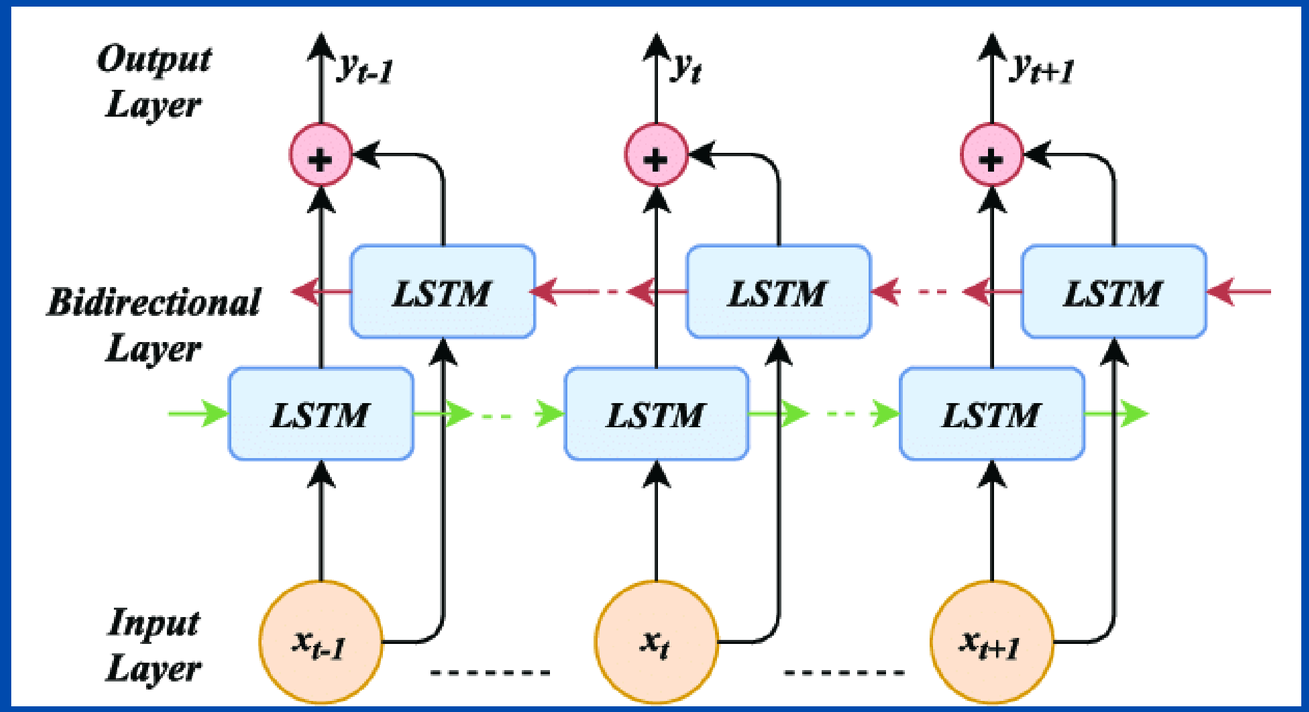
Generate samples for underrepresented classes

Outcome:

Failed to produce coherent, structured essays

SMOTE is not a solution for our problem; we need to search for better datasets.

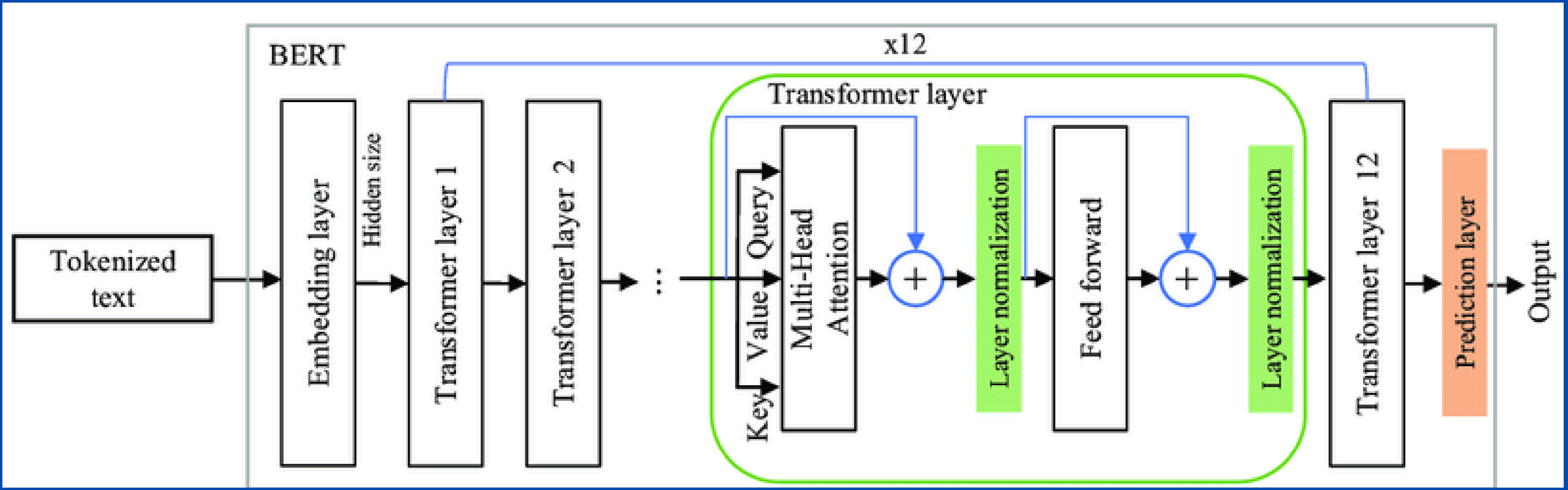
'has someone dont s families own home see " buy said think makes since what need thats about help much by looks taking always seen however technologies well certain even - system making expensive pay uses its support recognize job). based candidate directly going text trouble learn re such told away used 7 cameras 4 constantly skills feeling software facial expressions expression mona lisa experience normal created start lastly communicate shows understanding predicts cannot read valuable outside boredom flaws personal order easy issue doesnt handle happiness sadness imagery tracks movements situation afford complex privacy recognition watch effectively hack communication recorded watched ´ typically determine plenty stuent decode pc costly understood algorithms videos threats exemplifies woukld correlate ,"" many people have where they . the thing ' t know is that when you use a alot of can like get in or bad to on if but , because do not and sold move there are some only at it one for with this an example states from up good idea better all our been will was very why so how doing your new other should face be information tell just off could as now which possible any also would does make next whole them look something human ? who say says technology these may right being totally lot work want their take already problem more playing while able problems knowing possibly add feel less than time without almost give ( ) high ." actual conditions put things reasons likely sense another becoming computer children ; class student confused bored students then learning grades classroom computers understand child doesn wouldn program games safe mean cause d might having behind detect moment ," using due public company everyday times too majority looking techonlogy emotions dr huang emotion boring done lessons especially general probably looked past 6 alto smile companies emotional modify lesson effective online school struggling teachers society placed schools reading owning classrooms someones mistakes web ad appears screen follow instructor educational mostly happened rest fine interested caught alike ."(teacher parents tecnology implemented becasue outweigh positives google useful thier towards nature promote companys network purposes ads frown diffrent extra flaw brilliant intrests articles outcomes data becase anytime understandable computes generation childs reconize page logic childrens detecting profit diffent likey remove addressed simmilar honesty worksheet collect likelyhood wifi random technogly immune collected clarify users lovely slides diffrences computure targeted surronding lag gardians catering instructing entierly contrversy clicked tit poutcry'



BERT (Bidirectional Encoder Representations from Transformers)

- Developed by Google AI in 2018
- Purpose: Pre-training language representations for downstream tasks such as question answering and language inference.
- Key Innovation:
  - Bidirectional Training: Reads text in both directions (left-to-right and right-to-left) to capture context more effectively.
- Impact: State-of-the-art performance on various NLP tasks.

BERT Architecture



Layers:

- Transformer Layers: 12 (base) or 24 (large)
- Attention Heads: 12 (base) or 16 (large)
- Hidden Units: 768 (base) or 1024 (large)

Components:

- Input Embeddings: Token, Segment, and Position embeddings.
- Encoder: Multiple transformer encoders stacking self-attention and feed-forward layers.

Training Objectives:

- Masked Language Modeling (MLM): Randomly masks 15% of the input tokens and predicts them.
- Next Sentence Prediction (NSP): Predicts if two sentences appear sequentially in the text.

BERT's Functionality

Step-by-Step Workflow:

1. Tokenization
2. Embedding Creation
3. Passing Through Encoders
4. Output Generation

Use Cases:

- Sentiment Analysis
- Question Answering
- Named Entity Recognition