BOREALIS AI

Let's SOLVE it

U-Integrity

Tackling the issue of students abusing ChatGPT to conduct academic plagiarism

BOREALIS AI

Let's SOLVE it

Team Members

Yuxin (Katy) Chen

4th-year

UofT

Sharon E. Alex

4th-year

UofT

Maliha Lodi

4th-year

UofT

Sri Sai Pujitha Nallapati

4th-year

UofT

Omer Raza Khan

4th-year

UofT

Introduction

Meet Gary

- University student currently in the middle of final exam season
- A little bit stressed out
- Two take-home final essays due at midnight today
- Not enough time or energy to finish both of them

What should Gary do?

- Try and finish both essays as much as possible
- Use ChatGPT to write one for him, and write the other one himself

Introduction

What is ChatGPT?

- AI language model designed to understand and generate human-like responses to natural language inputs
- Recently caused an increase in academic integrity issues at educational institutions
- A Forbes conducted survey states:
  - **89%** of students used ChatGPT to help with homework assignments
  - **48%** of students used ChatGPT for take-home tests and quizzes

Source: https://www.forbes.com/sites/chriswestfall/2023/01/28/educators-battle-plagiarism-as-89-of-students-admit-to-using-open-ais-chatgpt-for-homework

Introduction

How advanced is ChatGPT?

- Can be difficult to differentiate between human-written and ChatGPT-written material
- Examples:
  - Peer reviewers approved an AI-generated research paper for publication, unaware it wasn't written by a person
  - Professors say it can be hard to catch students submitting ChatGPT-written assignments

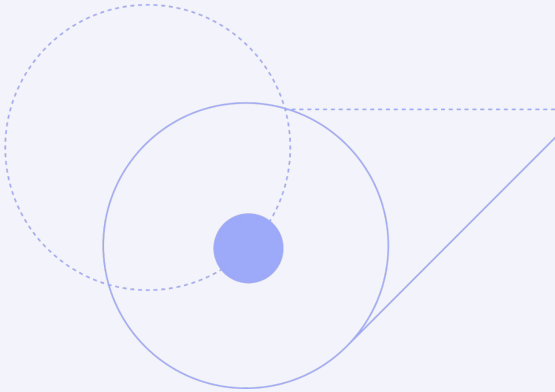
Sources:

- https://www.theguardian.com/technology/2023/mar/19/ai-makes-plagiarism-harder-to-detect-argue-academics-in-paper-written-by-chatbot
- https://www.forbes.com/sites/jasonwingard/2023/01/10/chatgpt-a-threat-to-higher-education

Motivation

How can ChatGPT be bad for student-use?

- Undermines the intended goal of education
  - Fail to benefit from learning opportunities and do not contribute their own work
  - Students rely on external tool instead of their own knowledge



How is this relevant for us?

- We are students
- UofT has changed its Academic Integrity policies to exclude the use of ChatGPT and other AI-tools
- Professors editing their course syllabus

Our Project

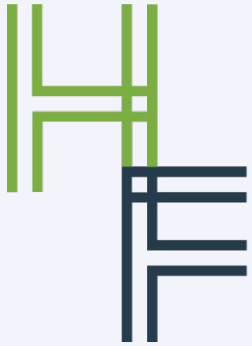
Problem Statement: We aim to compare the academic performance of students and ChatGPT by automatically grading student-written and ChatGPT-generated essays using a machine learning model

We want to address the following:

- Is it worthwhile for a student to use ChatGPT, knowing that they could potentially face academic penalties?

Kaggle Data

- Originally released as a part of a Kaggle competition sponsored by the Hewlett Foundation
  - **Automated Student Assessment Prize (ASAP)**
  - Develop an automated essay scoring algorithm
- Written by middle-school students
- Each consists of an essay prompt (question) along with rubric guidelines.
  - ~150-500 words in length
  - Dataset size: 14,000 essays



kaggle

Kaggle Data

Kaggle Essay Prompts Distribution

- Total of 8 essay prompts
- Excluding 4 prompts as they required background information to the texts we didn't have access to
  - 4 prompts remained: **prompts 1, 2, 7 and 8**
  - Total 6230 essays for these remaining prompts

Essay Prompt	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
No. Of Essays	1785	1800	1726	1772	1805	1800	1730	918

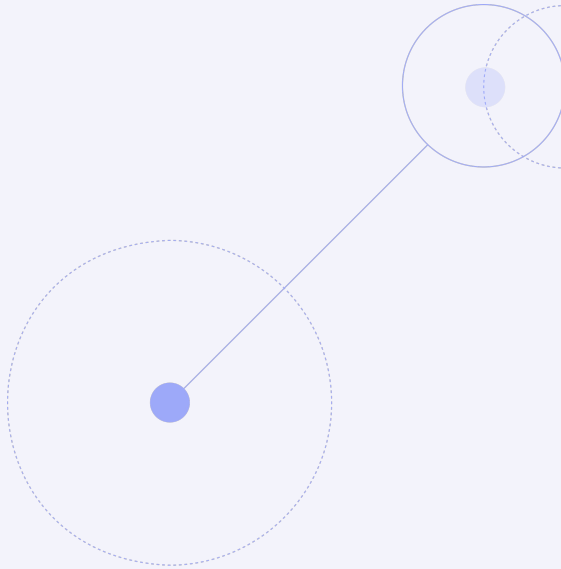
Kaggle Dataset

Column Name	Description
Essay_id	Essay id
Essay_set	Prompt number
Essay	Essay of that prompt
Rater1_domain1	Grader 1
Rater2_domain1	Grader 2
Rater3_domain1	Grader 3
Domain1_score	The sum of the scores given by individual graders for the domain 1 category
Rater1_domain2	Grader 1
Rater2_domain2	Grader 2
Domain2_score	The sum of the scores given by individual graders for the domain 2 category

ChatGPT Data

- **Dataset Generation:**
  - Used the OpenAI API to generate essays
  - Experimented with different OpenAI models available to the public
  - Final model chosen: "text-curie-001" as it's very capable, fast and low cost
  - Took ~24 mins to generate ~630 essays

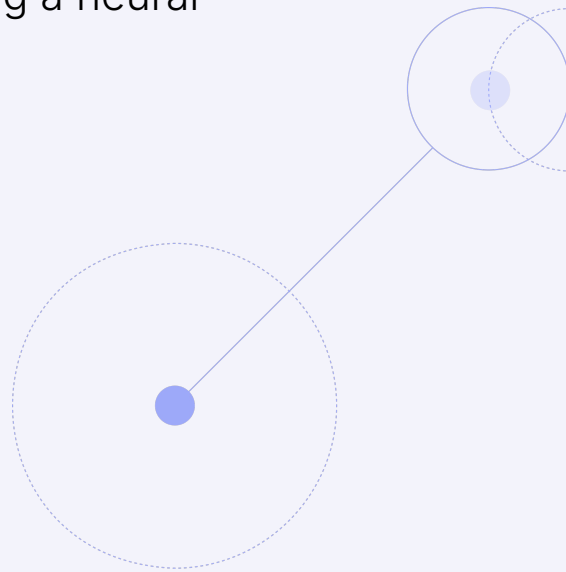
Essay Prompt	1	2	7	8
No. Of Essays	170	180	170	90



Data Pre-processing

Essays

- Pre-processed all essays before applying our model
  - Many ML models/algorithms require numerical input to perform calculations for their prediction
- We transformed all the textual data into word embeddings (i.e. numeric vectors)
  - **Word2Vec**: NLP algorithm for constructing vector representations of words using a neural network



Data Pre-processing

Score Normalization

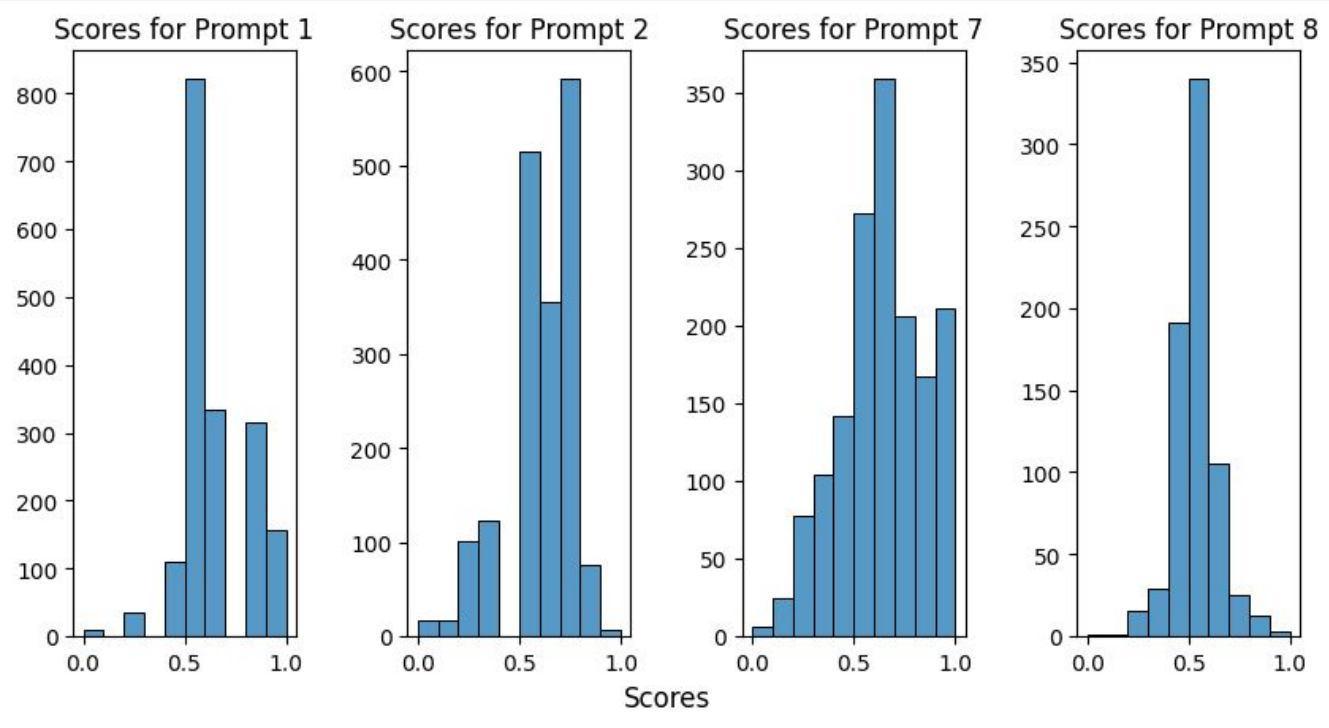
Dataset Scores

Prompt	Graders	Rubric Score	Final Score
1	2	1 - 6	***2 - 12***
2	2	1 - 4, 1 - 6	***1 - 10***
7	2	0 - 15	**0 - 30**
8	3	0 - 30	**0 - 60**

Min-Max Scaling

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

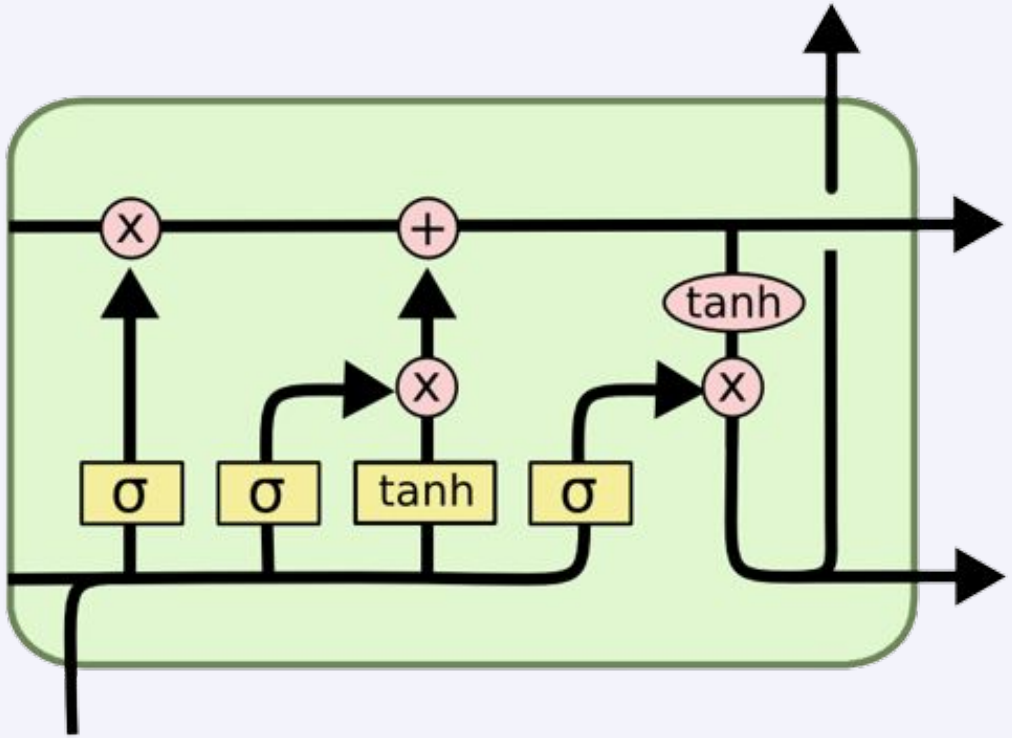
Final Distribution of Scores



Model Design

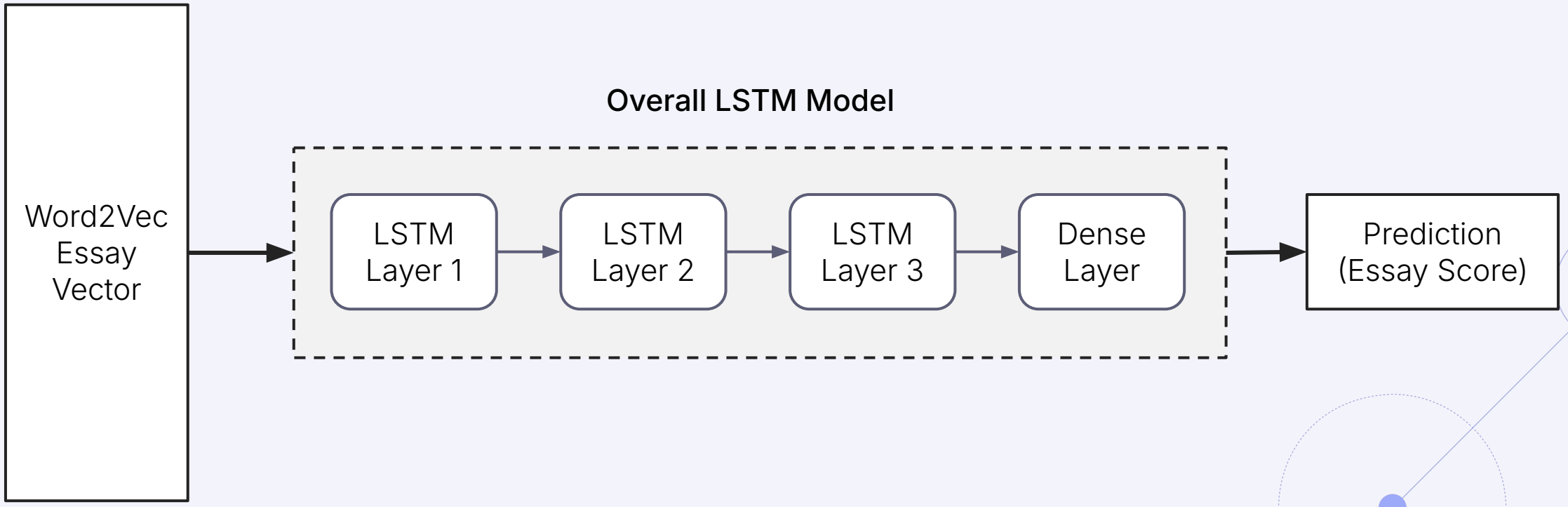
LSTM Network

- Type of deep learning algorithm
- Processes sequential data
- Able to learn long-term dependencies
- Specifically chosen as it performs well with natural language; will be able to understand the essays fairly easily

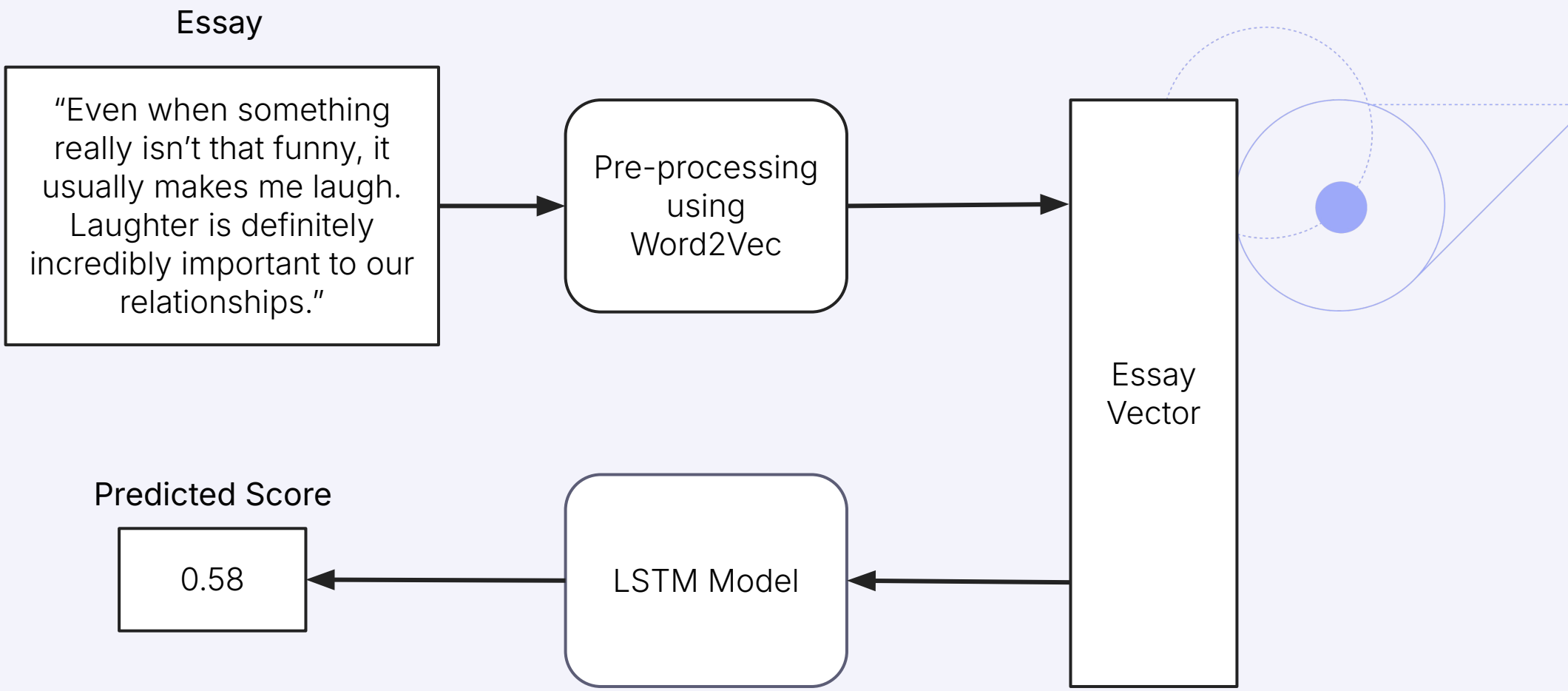


General LSTM Architecture

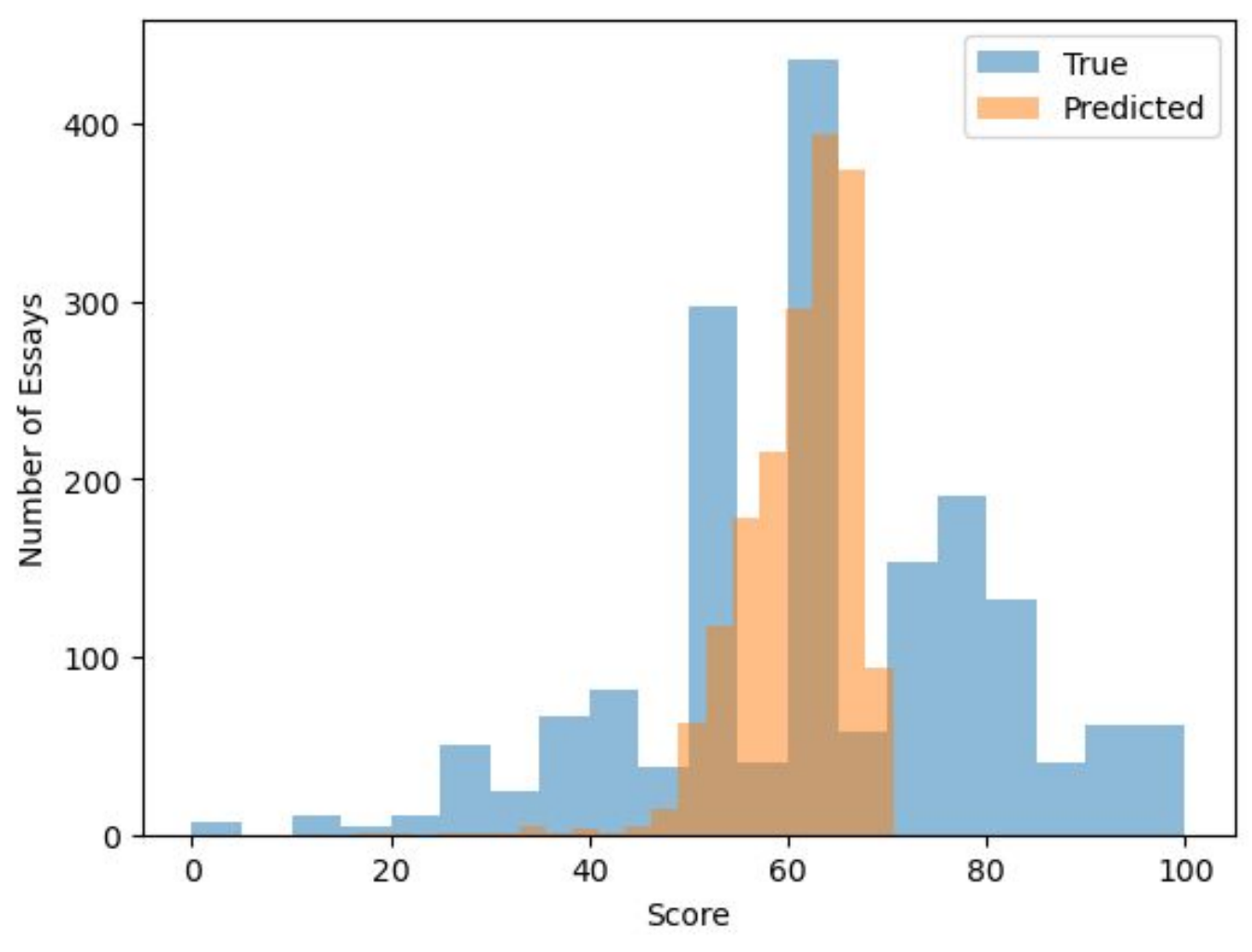
Final Model Architecture



Prediction Example



Outcomes - Model Validation

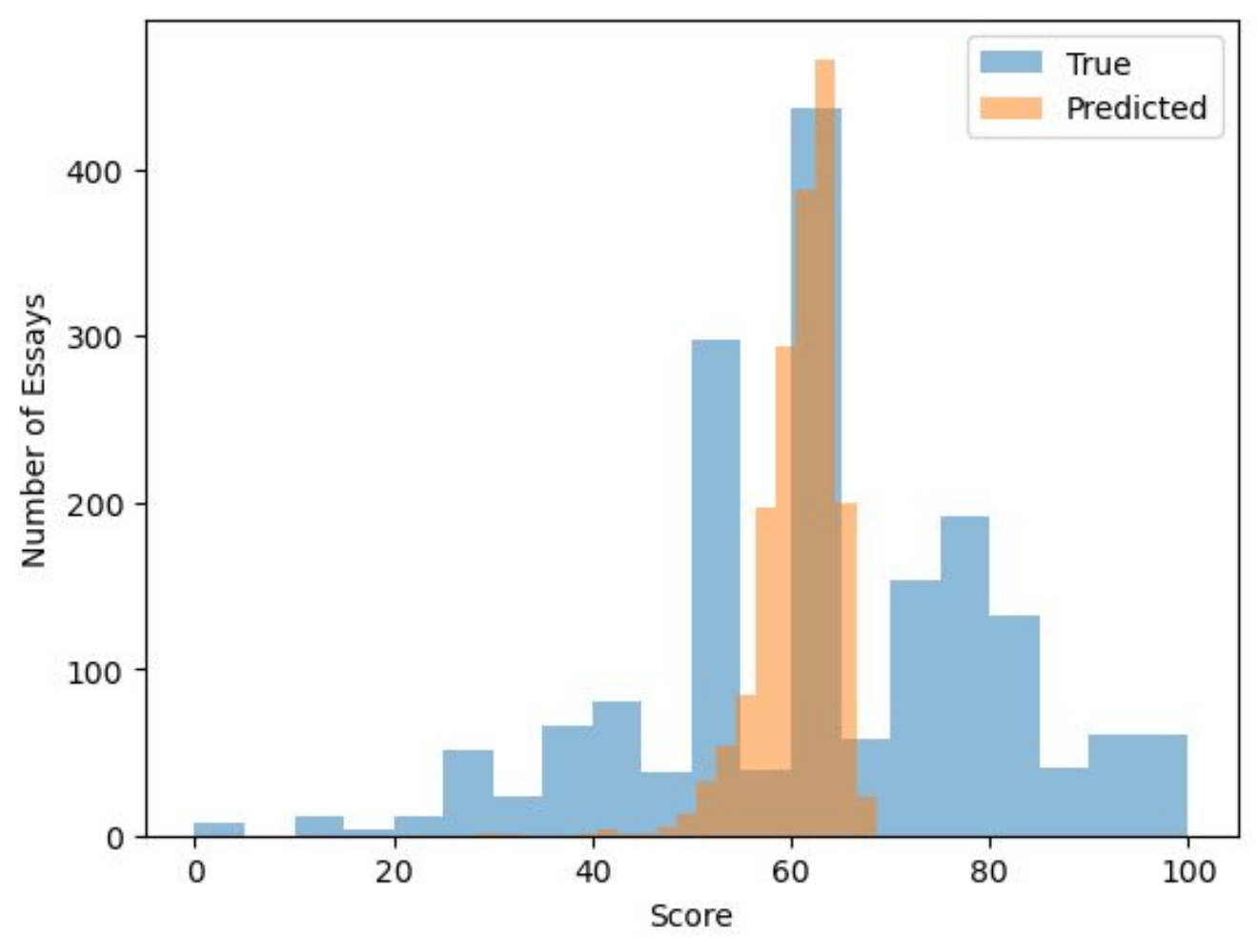


Mean Squared Error (MSE) Loss

- Loss: 0.0261
- MAE: 0.127

This low value suggests that the model is performing well in terms of predicting the target variable.

Outcomes - Model Validation

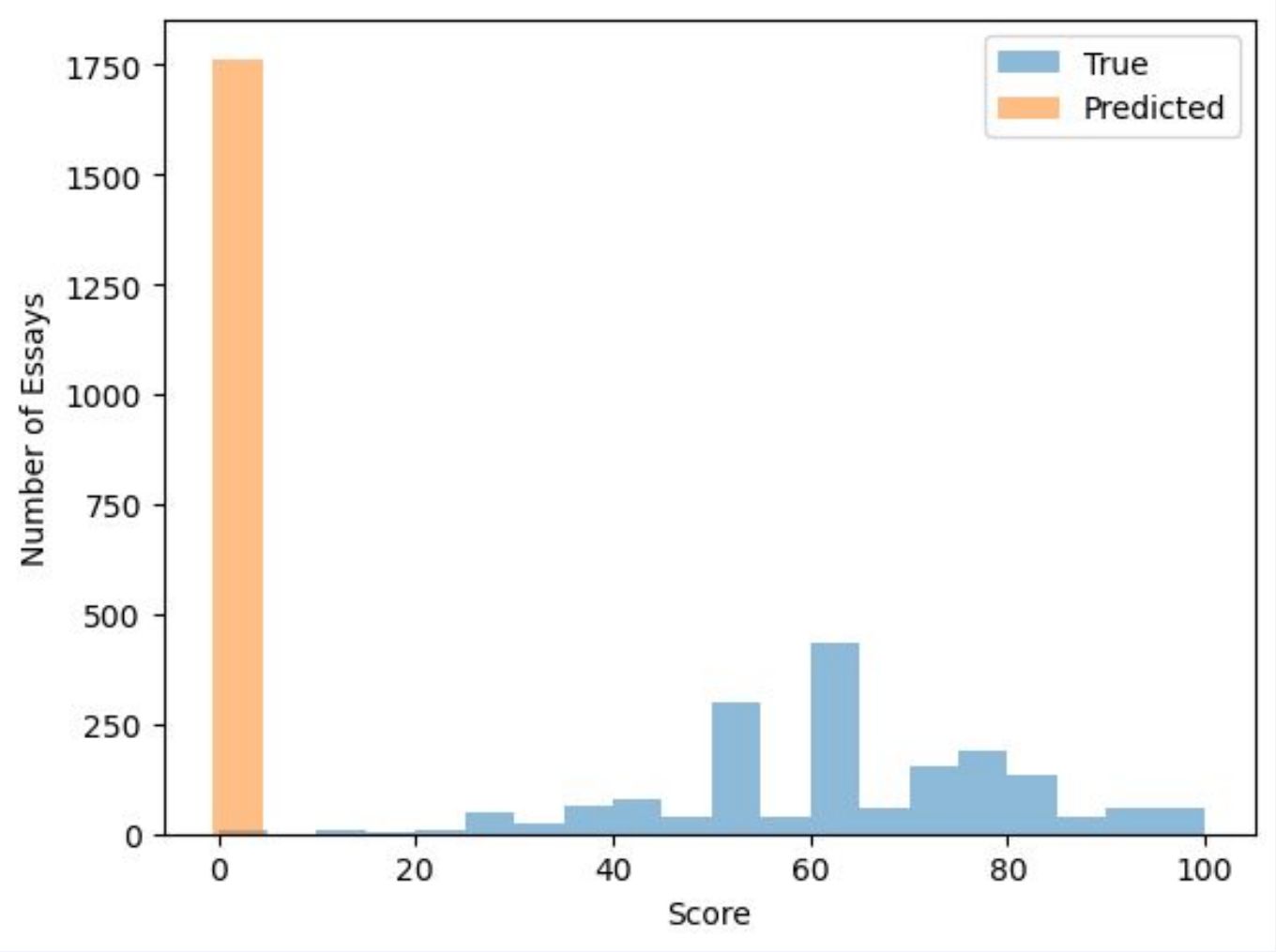


Mean Absolute Error (MAE) Loss

- Loss: 0.128

It is slightly higher than the MSE, it still suggests that the model is performing reasonably well.

Outcomes - Model Validation



Mean Absolute Percentage Error (MAPE) Loss

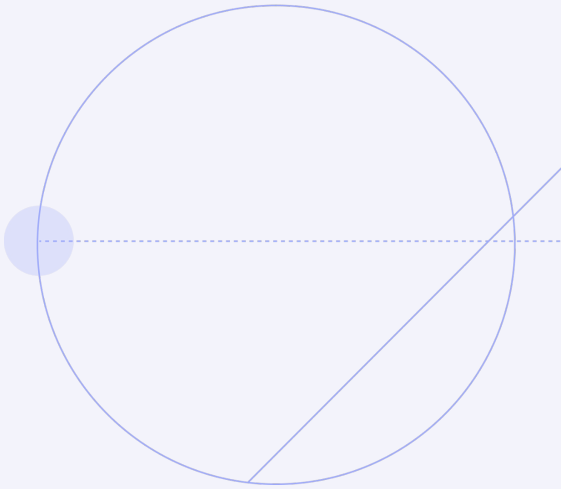
- Loss: 99.4894
- MAE: 0.6178

The model's predictions have relatively large percentage error

Outcomes - Metrics

MAE was used as metrics to compare the performance of model with different loss function.

Loss Function	MAE
MSE	0.127
MAE	0.128
MAPE	0.6178



- Small MAE indicates the model is making relatively smaller errors in its predictions
- Model using MSE loss is the best model among the others

Outcomes - Final Result

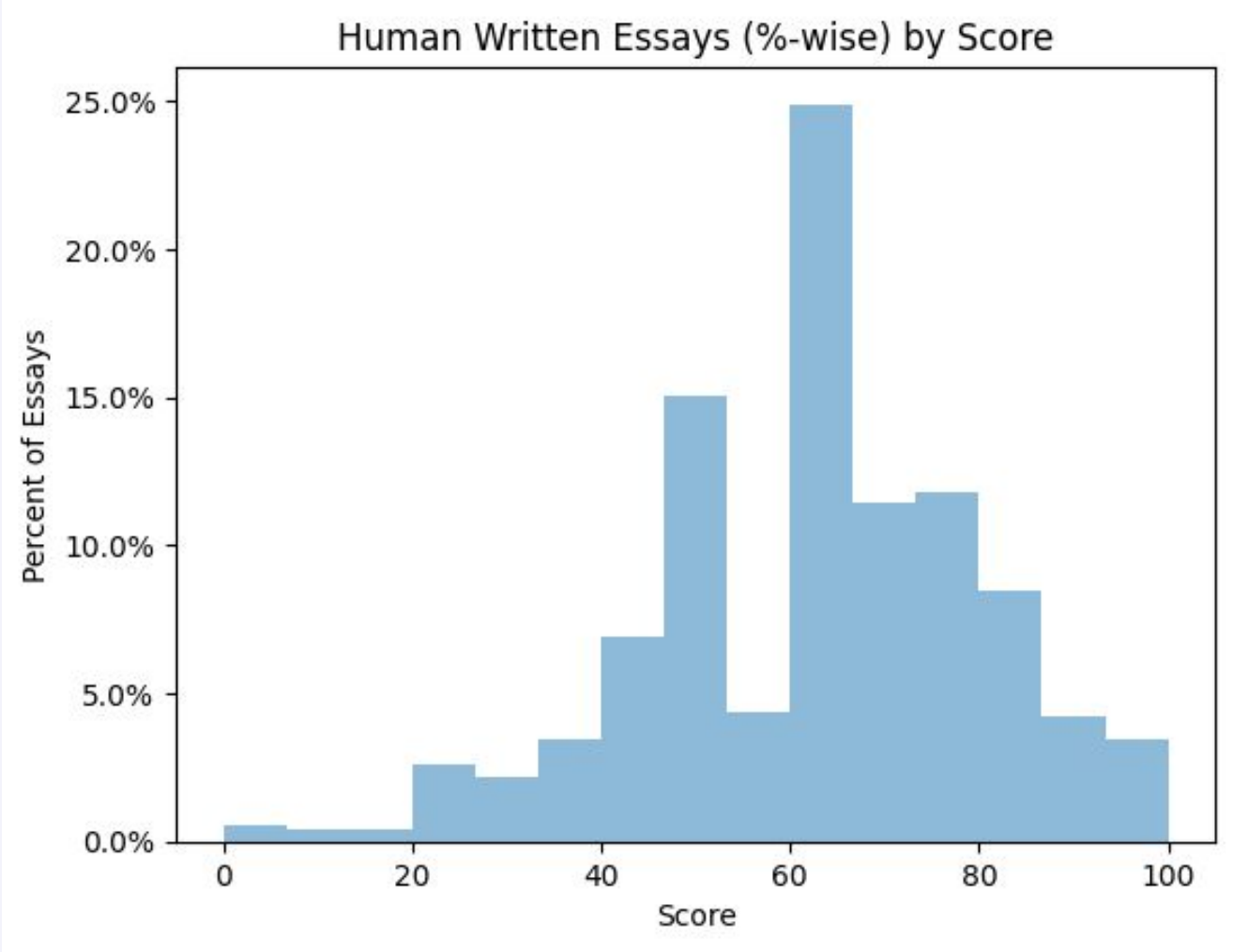
	Human Written Essay Scores	ChatGPT Essay Scores
count	5875	609
mean	62	65
std	18	8
min	0	41
25%	50	61
50%	60	63
75%	75	66
max	100	99

- Overall the ChatGPT Essay is slightly better than human written one

Outcomes - Top 5 Most Common Scores

Score	Percentage
60.00	14.43%
50.00	13.24%
75.00	10.08%
70.00	6.21%
62.50	6.04%

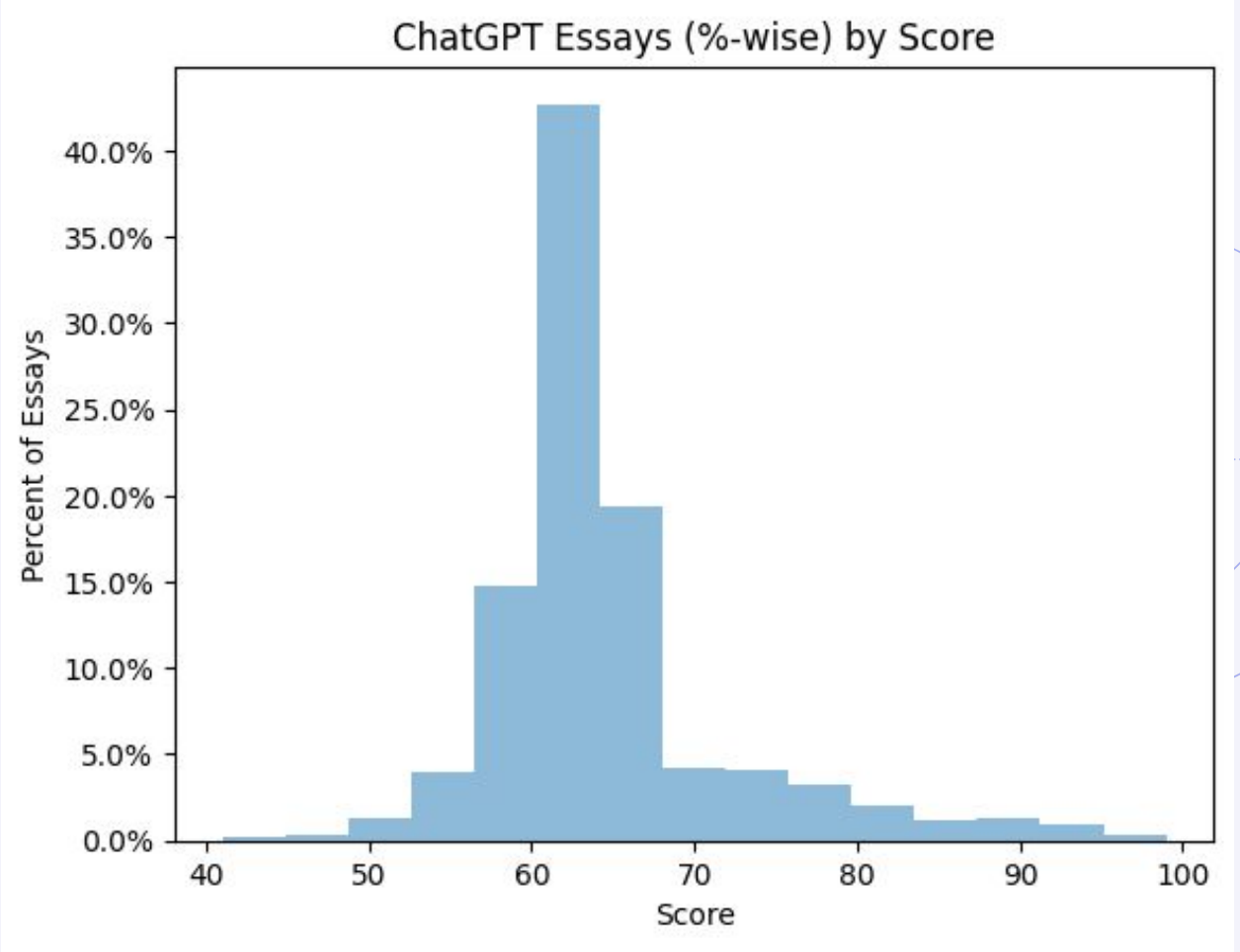
- Left skewed
- Two peaks at 60 and 50
- Highly variable



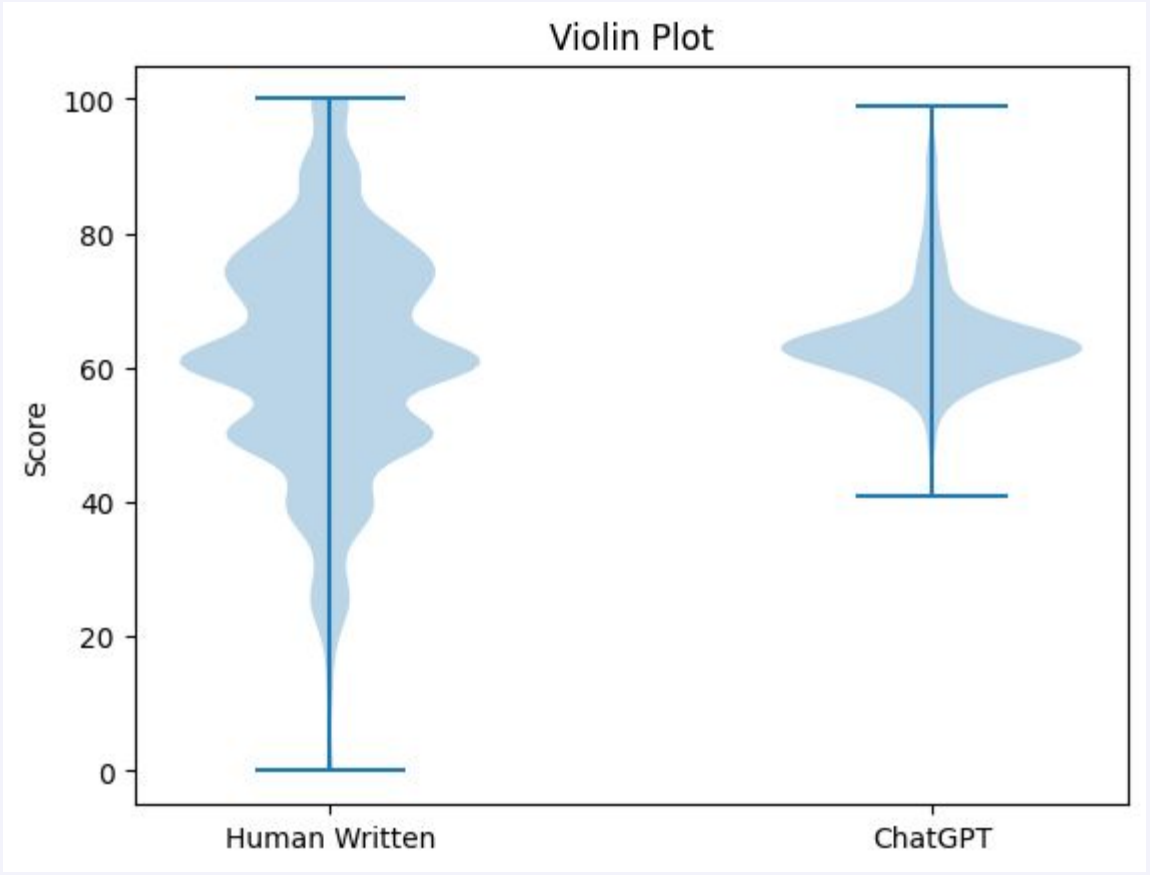
Outcomes - Top 5 Most Common Scores

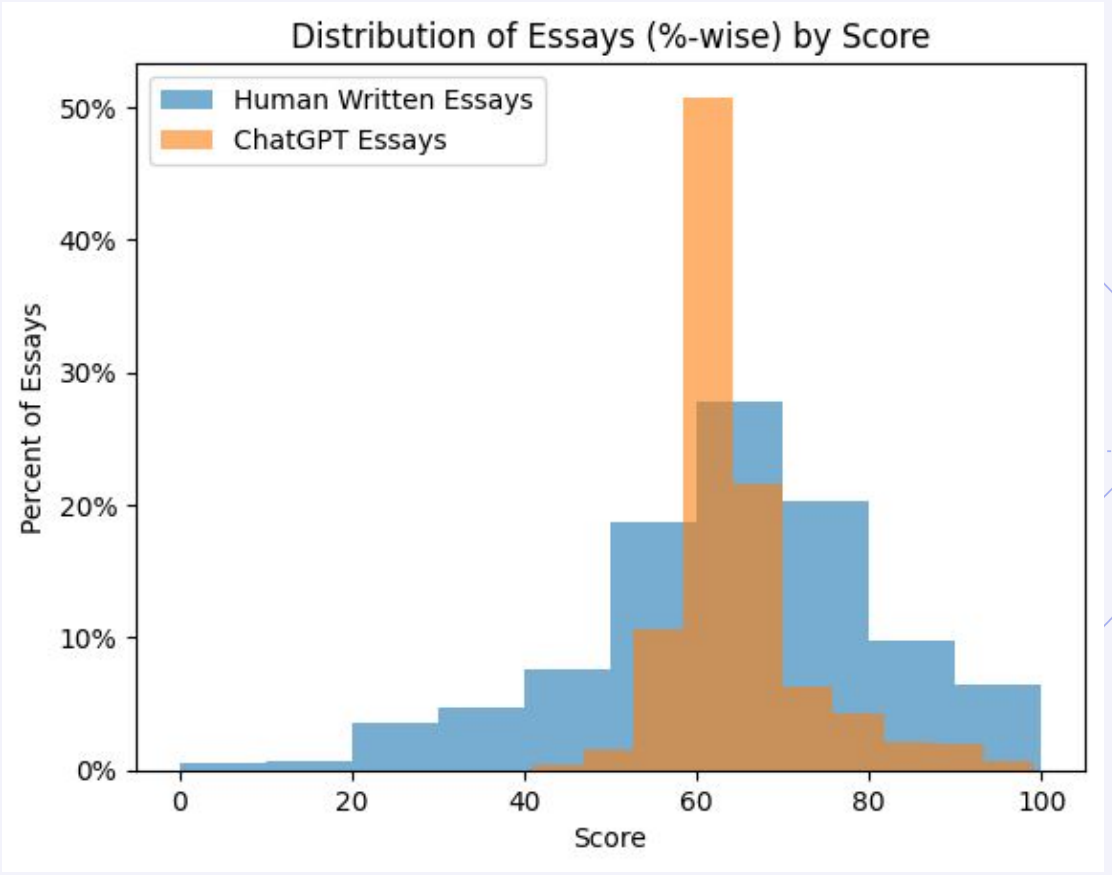
Score	Percentage
64.00	11.49%
63.00	11.17%
62.00	10.08%
61.00	9.69%
65.00	7.23%

- Right skewed
- Peaks at 60
- Less variable



Outcomes - Repercussions





- May be some underlying issues with the data
- ChatGPT is more capable of generating average essay

Outcomes

Overall Result: ChatGPT-generated essays scored **higher** than human-written essays on average

- Average ChatGPT-generated essay score: **65%**
- Average Human-written essay score: **62%**

What does this mean?

- Students may be willing to take the risk to use ChatGPT for their assignments
- Educational institutions must revise their Academic Integrity policies to include ChatGPT/other AI-tools
- ChatGPT's feedback loop allows it to continuously learn from its usage, meaning ChatGPT-generated essays could score higher in the future

Discussion

Challenges Faced

- Variation of scores lead to normalization research
- ChatGPT rate limits and using the OpenAI API
- Large loss: Mean Absolute Percentage Error (incompatible)

Achievements

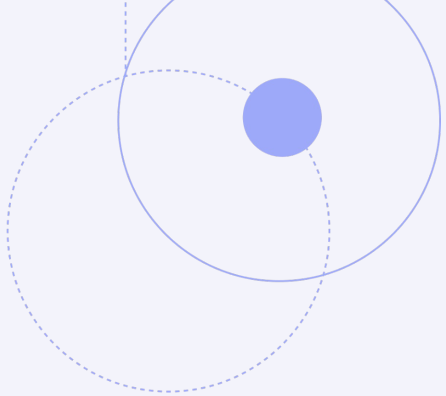
- Members with no prior knowledge: learned about machine learning, different models and the overall ML process
- Members with prior knowledge: gained experience working with textual data (i.e, how to preprocess it correctly) and neural networks
- Worked with messy, real-world data, on a ML project with Agile methodology

Conclusion

Future Directions

- Increase size of ChatGPT-generated dataset
  - currently 600
- Validate results across essays from different areas of study and levels of education
- Use GPT-4, potentially generating advanced essays

BOREALIS AI



Let's SOLVE it

Thank you! 🙂