Automated Essay Grading

Deepak H R
170010026

Anudeep Tubati
170010039

IIT Dharwad

ABSTRACT

Automated Essay Grading is one of the lesser-known fields of NLP. This project aims to develop ML and DL models that can closely score essays to the scores given by human raters and also compare between the two types of models. It makes use of the latest NLP techniques to benchmark their performance against the standard methods. The project, in addition, investigates the working of the LSTM and BERT models by performing a series of experiments and suggests ways to improve them.

Keywords - NLP, Essay Grading, Kappa Score, Word Embeddings, LSTM, BERT

INTRODUCTION

Advancements in the NLP field in the past few years have made it possible for seamless models to *almost* perfectly understand human language. However, many systems still enforce rather conventional methods when it comes to text analysis. Essay grading is one of such areas. The estimate of enrolled students in Secondary education (2014) was *568 million*[1]. In CBSE alone, over *1.5 million*[2] students take the 10th class exam per year. Most of the Secondary education exams for language courses comprise essay writing. Due to a large number of students appearing for such exams, the written material generated from them is massive. Not to forget, a majority of these essays are graded by humans. Hence, they involve a huge amount of text corpus and human effort.

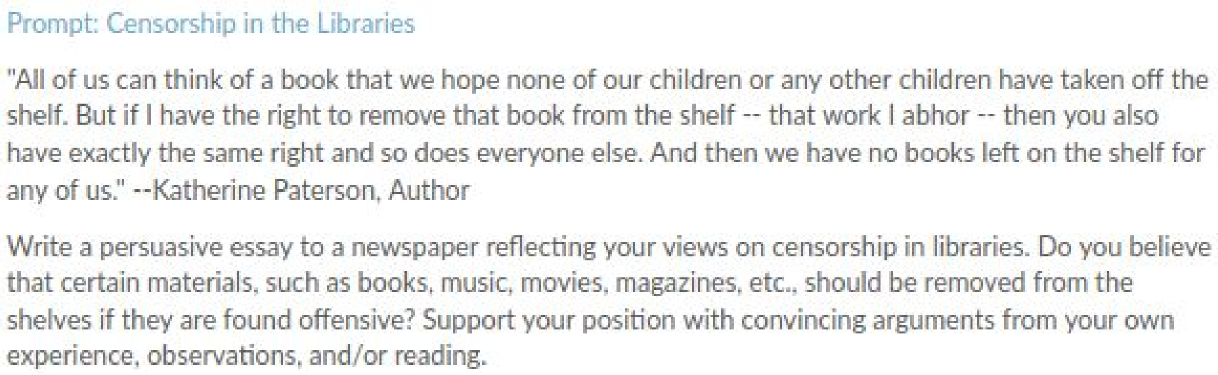
Though essays are scored based on defined rules, it requires some "practice" to analyse the parameters included in the rules. For example, a typical essay might be scored on the basis of creativity, vocabulary and fluency. Once the creativity, vocabulary and fluency are analysed, it is easy to combine them to get the final score. However, finding out these parameters takes "practice", as mentioned earlier.

This project aims to develop a model which can effectively score a given essay and for the inferred score to closely correspond with the one given by a human rater. Also, it shows a comparison of standard ML techniques and the contemporary DL models in the same task.

DATASET

1. The dataset used for this project is the one provided by the Kaggle-hosted ASAP AES competition[3] (hosted in 2012).
2. The data consists of 8 sets of essays, with each set of essays guided with a unique prompt.
3. Sets #1, #2, #7 and #8 have a persuasive/narrative tone whereas the remaining ones elicit source dependent responses.
4. Each essay has been rated by at least 1 rater and at most 3 raters.
5. Essays belonging to the set #2 have a Domain 2 score too, as opposed to just Domain 1 score for all the other sets. We chose to ignore this "Domain 2 Score" in our project as there wasn't sufficient data.
6. ***We work on the first 6 sets only as essay sets #7 and #8 have too big a range as compared to the other sets.***
7. We use essays from #7 **only** for verifying the patterns captured by our model.

Some examples

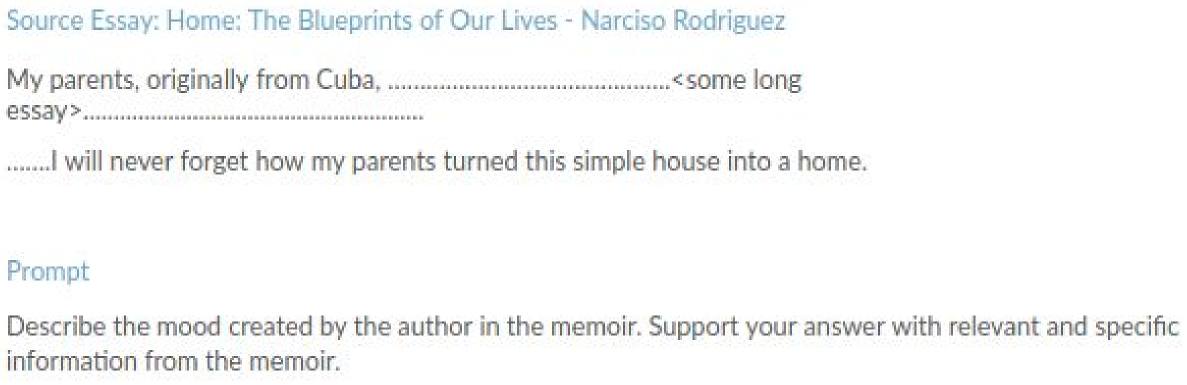


Prompt: Censorship in the Libraries

"All of us can think of a book that we hope none of our children or any other children have taken off the shelf. But if I have the right to remove that book from the shelf -- that work I abhor -- then you also have exactly the same right and so does everyone else. And then we have no books left on the shelf for any of us." --Katherine Paterson, Author

Write a persuasive essay to a newspaper reflecting your views on censorship in libraries. Do you believe that certain materials, such as books, music, movies, magazines, etc., should be removed from the shelves if they are found offensive? Support your position with convincing arguments from your own experience, observations, and/or reading.

Fig 1. An example of a Persuasive, Narrative or Expository essay prompt



Source Essay: Home: The Blueprints of Our Lives - Narciso Rodriguez

My parents, originally from Cuba, ...<some long essay>..

.......I will never forget how my parents turned this simple house into a home.

Prompt

Describe the mood created by the author in the memoir. Support your answer with relevant and specific information from the memoir.

Fig 2. An example of a Source Dependent Response essay prompt

Overall Description

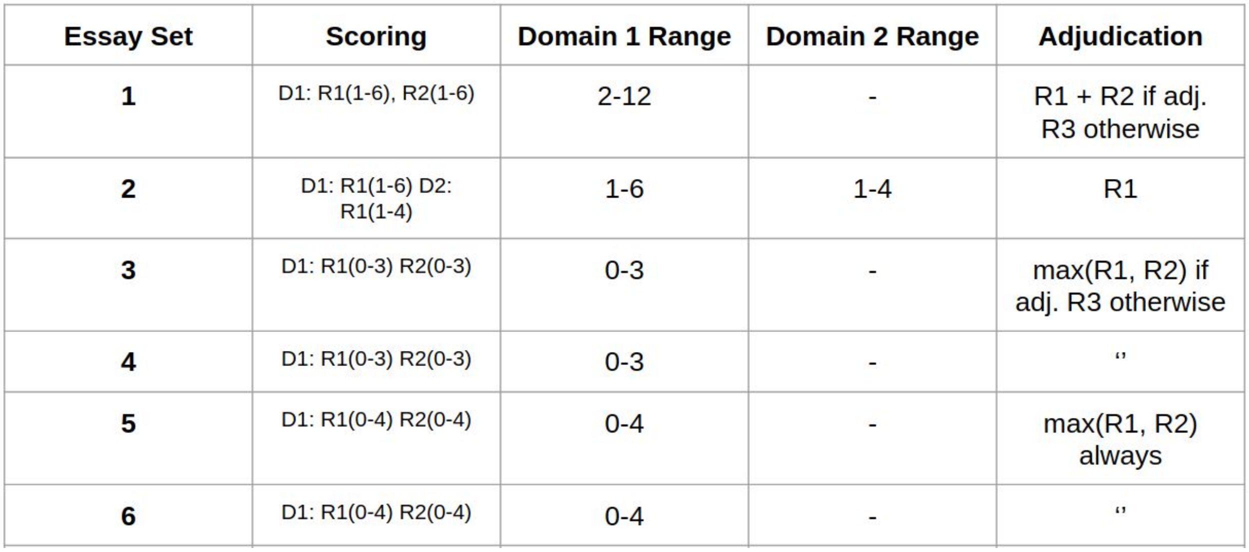
Table 1. General statistics of dataset

Set #	**Grade**	**Avg len**	**Train #**	**Validation #**	**Test #**
1	8th	350	1248	535	589
2	10th	350	1260	540	600
3	10th	150	1208	518	568
4	10th	150	1240	532	586
5	8th	150	1263	541	601
6	10th	150	1260	540	600
Total			**7480**	**3206**	**4218**

1. The training and validation set have been extracted from a 70-30 split from the file `training_set_rel3.xls.`
2. The test set has been extracted from `valid_set.xls`. The labels are included in `valid_sample_submission_1_column.csv` , which are only meant for illustrating format for submission at the kaggle competition.
3. All models will be trained and tested on the same split **(7479, 3206)**.

Detailed Description

Table 2. Details of each essay set



Essay Set	Scoring	Domain 1 Range	Domain 2 Range	Adjudication
1	D1: R1(1-6), R2(1-6)	2-12	-	R1 + R2 if adj. R3 otherwise
2	D1: R1(1-6) D2: R1(1-4)	1-6	1-4	R1
3	D1: R1(0-3) R2(0-3)	0-3	-	max(R1, R2) if adj. R3 otherwise
4	D1: R1(0-3) R2(0-3)	0-3	-	‘’
5	D1: R1(0-4) R2(0-4)	0-4	-	max(R1, R2) always
6	D1: R1(0-4) R2(0-4)	0-4	-	‘’

Evaluation Metric - Quadratic Weighted Kappa

Quadratic Weighted Kappa (QWK) is used as a metric for this project. This metric measures the agreeability of 2 given raters on the ratings given by them; one rating being the actual score of the essay and the other one, the score given by the model.

$$-1 \leq QWK \leq 1$$

A QWK of (-1) implies total disagreement of the raters, 0 implies a by-chance agreement and 1 implies perfect agreement.

To calculate QWK[4], consider 2 arrays of scores **u** and **v** and the possible scores being (1, 2, ….. , **K**). There are 3 components of the formula. **E**, an expectation matrix; **W**, a weighted matrix and **X**, the observed matrix.

Let **uf** and **vf** be **K**-length vectors such that

$$uf_i = number\ of\ items\ valued\ i\ in\ u$$

$$vf_i = number\ of\ items\ valued\ i\ in\ v$$

For calculating QWK,

$$w_{ij} = (i - j)^2\ \forall\ i, j\ \in \{1,\ 2,\ ....,\ K\}$$

$$E = uf * vf^T$$

$$x_{ij} = number\ of\ tuples\ (i,\ j,\ k)\ such\ that\ u_k = i\ and\ v_k = j\ \ \forall\ k \in \{1,\ 2,\ ....,\ K\}$$

$$QWK = 1 - \frac{1 - \sum_{i=1}^{K} \sum_{j=1}^{K} w_{ij} \cdot x_{ij}}{1 - \sum_{i=1}^{K} \sum_{j=1}^{K} w_{ij} \cdot e_{ij}}$$

A source of a visual example to calculate QWK is mentioned in the appendix[a].

Method of Comparing QWK

To compare our models performance with the human performance, we calculate QWK for the 2 raters (say H1/H2). Against this, we pit H1/(our models QWK). These scores signify the agreeability of a trained human H1 and a trained H2 as compared to that of a trained human H1 and our model.

QWK(H1, H2) = 0.905
(only sets 1-6)

OUR MODELS

Machine Learning Models

Preprocessing

- Standard stop words (from nltk library[5]) are removed from the essay.
- Vocabulary defined as a set of unique tokens that occur in all the essays in the dataset.
- One of two vectorizers is chosen to convert an essay to a feature vector.

➔ Count Vectorizer

We count the frequency of words in a given essay. The counts are made into an array such that nth word in the array is the frequency of the nth word in the vocabulary.

➔ TF-IDF Vectorizer

This was originally a term weighting scheme developed for information retrieval (as a ranking function for search engines results) that has also found good use in document classification and clustering.

```
TF-IDF(w, essay) = frequency(w, essay) * log((n_essays + 1)/(1+ document_freq(w))
```

For each word in the vocabulary, we use the [TF-IDF(word, essay) for every word in vocabulary] as the vector.

Models

The models are used as a baseline to compare with the neural networks discussed in the later sections.

Table 3. Results of ML Models

Model + Vectorizer	Train kappa	Train Accuracy	Test Kappa	Test Accuracy
GaussianNB + TFIDF	0.87	0.81	0.23	0.21
SVC + Count	0.92	0.91	0.54	0.48
MultinomialNB + Count	0.63	0.67	0.33	0.33
SVC + TF-IDF	0.78	0.73	0.39	0.36

Observations

- Although the training metrics are really good, the models fail to provide decent predictions on the test data.
- The actual score represents the fluency, transitional language and organization of words in the essay, whereas the predicted score is based on the frequency/ rarity of words in the dataset.
- These properties are a direct result of the order of words that appear in the essay.
- This information is lost in the step where the essays are vectorized using the bag of words approach.

Long Short Term Memory Networks

Since ML models failed to achieve good performance, DL models were next in line. RNNs are known to capture the sequential patterns in data very well. However, they have a problem with vanishing/exploding gradients. To mitigate that, LSTMs were used.

We used a single layer of 64 LSTMs, along with pre-trained GloVe word embeddings (300-dimension). Also, we used the default Keras Tokeniser on the essays. As expected, this model gave a performance jump compared to the ML models.

Time taken to train 50 epochs = 2 hrs (Google Colab GPU)

Table 4. Results for LSTM classifier with GloVe embeddings (50 epochs of training)

	Kappa			**Accuracy**		
Training	**Random**	**Before**	**After**	**Random**	**Before**	**After**
Train	0.0	0.070	0.851	0.14	0.04	0.63
Val	0.0	0.076	0.848	0.18	0.04	0.60

Neural network based on attention

The following lines are from the paper *Attention is all you need*[6].
"Recurrent models typically factor computation along the symbol positions of the input and output sequences. Aligning the positions to steps in computation time, they generate a sequence of hidden states h_t *, as a function of the previous hidden state* h_{t-1} *and the input for position* t*. This inherently sequential nature precludes parallelization within training examples, which becomes critical at longer sequence lengths, as memory constraints limit batching across examples."*

There is no backpropagation through time in the attention layer. The output at each timestep is computed parallely thus avoiding excessive application of the chain rule hence evading vanishing/exploding gradients associated with long sequences. This will be seen in the following section where the attention layer is explained in detail. *The ability to process data parallely more than compensates for the increased number of parameters we end up while dealing with training time.*

The transformer models implemented from the above paper showed very convincing results on machine translation. The paper goes on to explain how transformers which are deep models with attention are in fact better than LSTMs in modelling very long sequences.

The task of machine translation involves understanding the core meaning of the input sentence. This is akin to taking into account the transitional flow and use of language in the different essays. A model that is able to translate well should have the information regarding the core meaning of its input, enough to distinguish between essays of different scores.

Since this project has been a journey of learning new techniques and concepts in the field of NLP, we proceed to learn about the state of the art sequence processing models, the attention model.

Attention in Detail

Query, Key and Value vectors

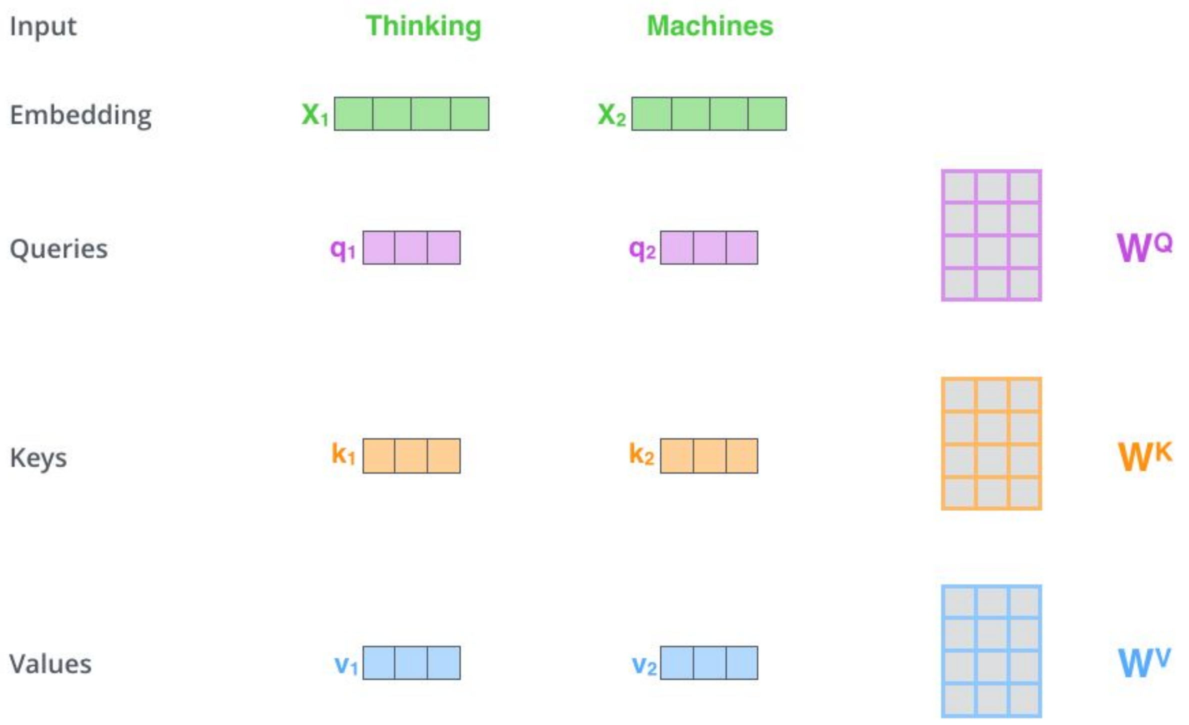


Fig 3. Query, Key and Value vectors

Attention allows the model to focus on the relevant parts of the input sequence as needed.

Multiplying x1 by the W^Q weight matrix produces q1, the "query" vector associated with that word. We end up creating a "query", a "key", and a "value" projection of each word in the input sentence.

Calculating attention values

Note that d_k is set as 64, the dimension used in the attention paper.

The attention values calculated here are 0.88 and 0.21 for the first timesteps. This means while we look at the word ‘Thinking’ we need to give 21% attention to the word ‘machines’.

The output vector z_1 will look 88% like v_1 and 12% like v_2.

The attention values are a result of the weights (query, key and value matrices W^Q, W^K and W^V) and are adjusted using the gradient back-propagated from the error in predictions made using the matrix $[z_i]$ by the subsequent layers.

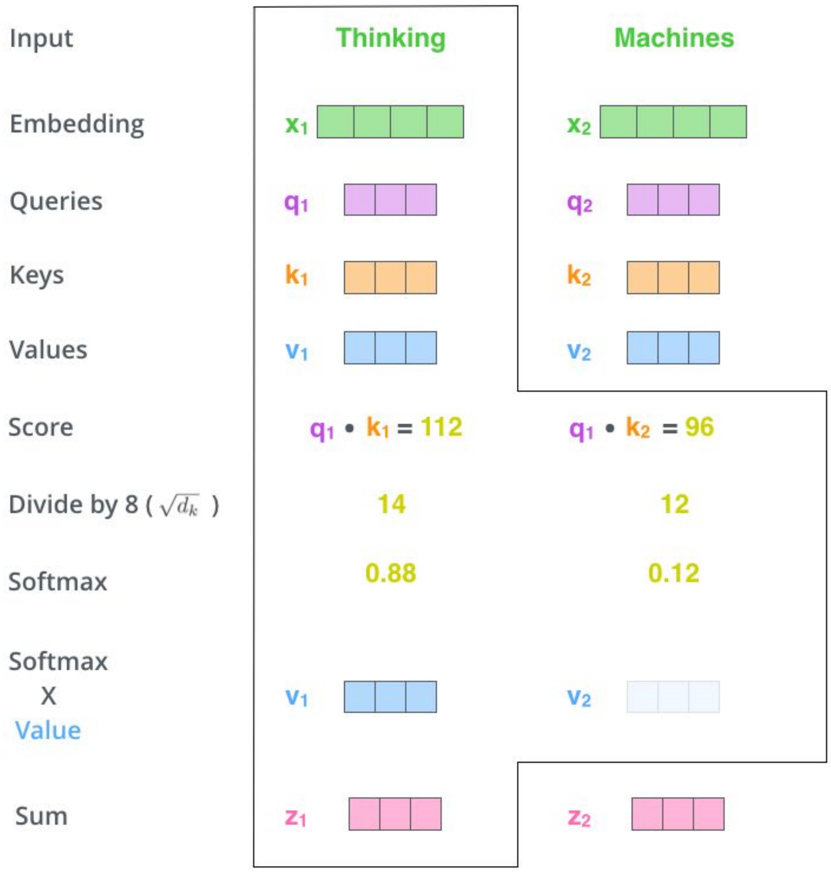


Fig 4. Calculation of Attention values

Visualizing attention values

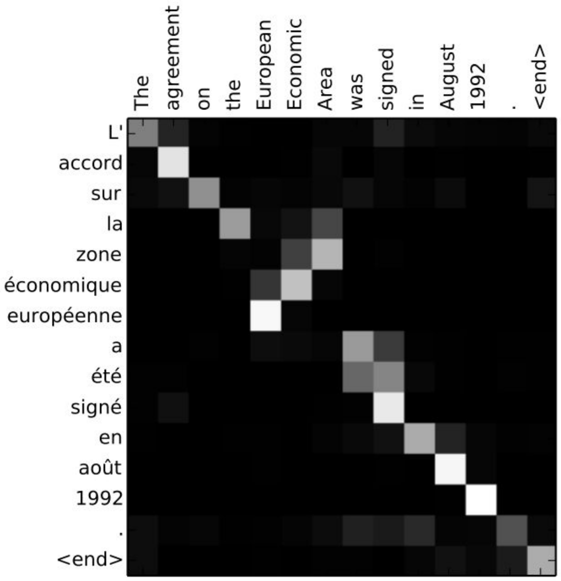


Fig 5. Attention matrix visualized for English to French translation for a sample sentence

Predictions at every timestep (corresponding to row labels in french) are made by giving attention ($q_i * k_j$ for every word pair i, j in the sentence).

When the neural network predicted the word ‘la’ at time step 4, it paid attention to the input word ‘the’ at timestep 4 and also to the input word ‘Area’ at timestep 7, the next and only relevant word in the sentence.

We get different attention profiles from the same model for different input sequences.

Using BERT Encodings for essays

(Raffel *et al.*, 2019)[7] tells us that pretraining large transformers on data-rich tasks and fine-tuning them on a downstream task has been a powerful technique in NLP. We explored BERT encodings for essays, as BERT proved to be a revolution in the NLP field and it has many novel procedures to learn better than standard LSTMs. We benchmarked the LSTM + GloVe and BERT classifier models on 2 standard datasets. The datasets are Stanford Sentiment Treebank[8] and Disaster Tweets[9]. We explored different datasets to analyse and compare the performance of the two models (LSTM+GloVe and BERT + 2 layered ANN). The analysis of both models is elaborated next.

Stanford Sentiment Treebank (SST-2)

The task involved with the dataset is to associate a given sentence with a positive or a negative sentiment. This is essentially a binary sentence classification task.

Table 5. Description of SST-2 dataset

	Positive Examples	**Negative Examples**
Train	37569	29780
Test	444	428

Disaster Tweets (Kaggle competition)

The dataset contains tweets from various people and some of them are related to natural disasters. We need to understand the core meaning of an input tweet and predict whether it was associated with the occurrence of a natural disaster. This is yet another binary classification task that requires us to understand the core meaning of the text.

However, this is an ongoing competition. It has the test data whose labels are unknown.

Table 6. Description of Disaster Tweets dataset

	Positive Examples	Negative Examples
Train	3691	651
Validation	2780	491

Results

Table 7. Benchmarking various models on the datasets

	SST-2		Disaster Tweets	
	Train	**Test**	**Train**	**Test***
GloVe + LSTM	0.9942	0.8452	0.8524	0.79038
BERT + ANN classifier	0.9704	0.9248	0.9855	0.83435
SOTA	-	0.974**	-	1.0

* Disaster Tweets is a Kaggle competition. Performance is reported according to our submission on the competition page.

** According to paperswithcode[10], a transformer model that was first pre-trained on data-rich tasks and later fine-tuned on a downstreamed task.

Observations

1. The BERT+ANN classifier outperforms the GloVe+LSTM network on both the datasets.
2. The GloVe + LSTM model overfits on the SST2 dataset whereas under-fits on the Disaster-Tweets dataset.
3. The BERT+ANN classifier has a just ‘right fit’ on SST2. It overfits on the Disaster-tweets dataset, yet performs better than the LSTM model.

The benchmarking further showed that the BERT classifier can outperform its counterpart by a large margin. Therefore, we adopted a similar model comprising BERT + ANN classifier.

ANN Classifier with BERT Essay Encodings

We tokenized our essays (with BERT tokeniser) and then executed a forward pass with BERT to obtain a 768-length encoding for each essay. Those encodings were then made as training samples for a fully connected ANN. The ANN comprised 2 layers. First layer with 384 units and the next one (output) with 7 units. This gave a significant improvement over the previous model.

Time taken to train 200 epochs = 5 mins (Kaggle Kernel GPU)

Table 8. Results for ANN Classifier with BERT Encodings (200 epochs)

	Kappa		**Accuracy**	
Training	**Before**	**After**	**Before**	**After**
Train	0.086	0.844	0.24	0.63
Val	0.115	0.838	0.24	0.60

The results clearly indicate that the BERT model didn't really outperform the LSTM on this dataset. To investigate the reason behind this, we graphed the embeddings of each essay just before the classification layer for both BERT+ANN and GloVe+LSTM. They are graphed below (with the help of t-SNE[11]).

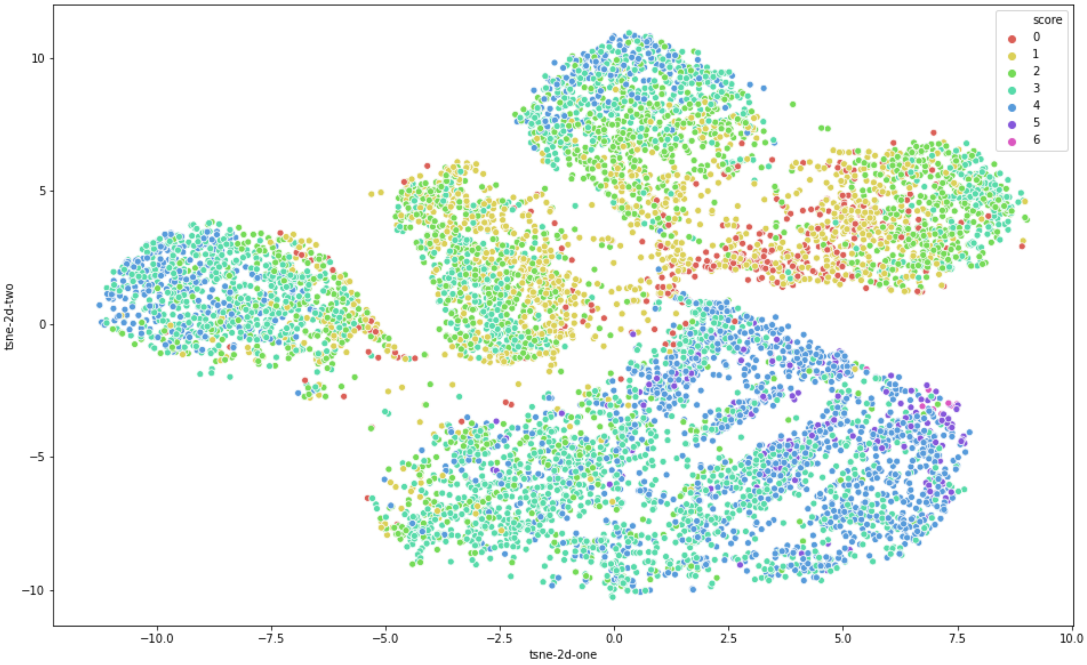


Fig 6. t-SNE Visualization for BERT+ANN essay encodings

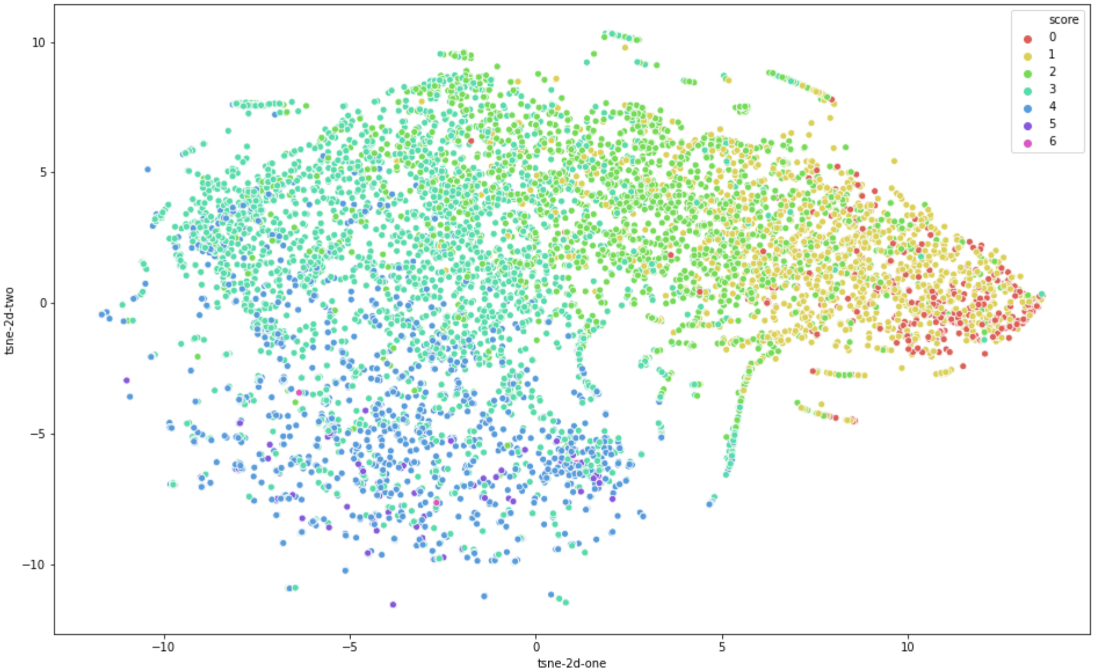


Fig 7. t-SNE Visualization for LSTM+GloVe essay encodings

The LSTM embeddings were much better distributed and separable, which explained the reason behind its performance. We anticipate the spelling mistakes and out-of-vocabulary words in the dataset to be the reason behind BERT showing a non-optimal performance. We also did not fine-tune BERT on this dataset.

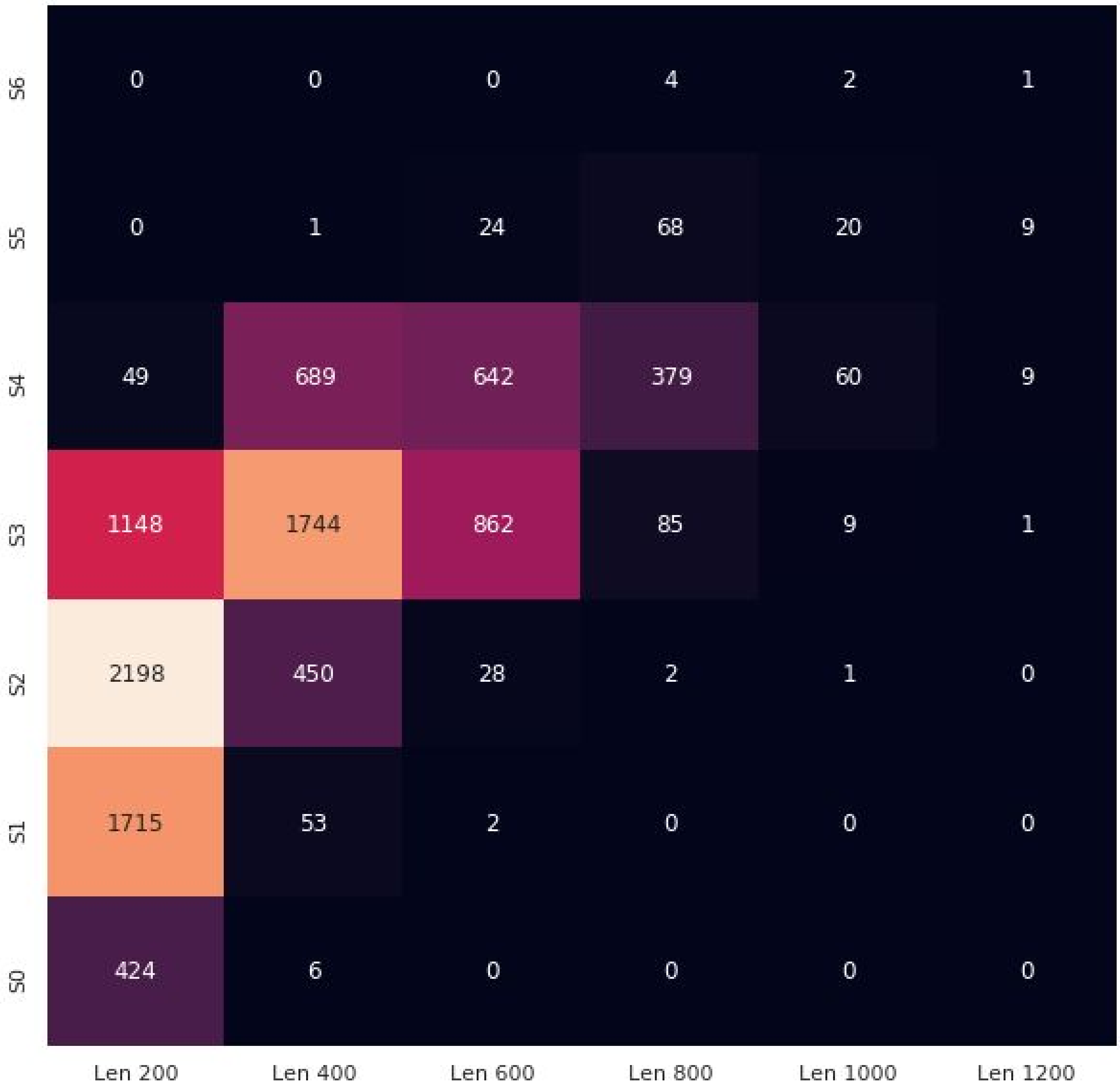
Apart from that, we found out that the BERT + ANN model had two major problems

i. Before returning the predictions, we upscale them from (0-5) to (2-12). That is, if x was a prediction for an essay belonging to set #1, we make it $(x * 2) + 2$. This upscaling cannot conceive odd-valued predictions. Hence, it produces a sort of inherent disadvantage.

ii. The tokens are capped at 512 as BERT accepts only those many tokens. This causes some information loss.

Joint Model comprising 2 ANN Regressors (one for short essays and one for long essays) with BERT encodings

To solve the inherent problem of odd-valued prediction omission, we chose a Regressor in place of the classifier. This allowed us to predict continuous values which could then become odd-valued predictions after upscaling.



	Len 200	Len 400	Len 600	Len 800	Len 1000	Len 1200
S6	0	0	0	4	2	1
S5	0	1	24	68	20	9
S4	49	689	642	379	60	9
S3	1148	1744	862	85	9	1
S2	2198	450	28	2	1	0
S1	1715	53	2	0	0	0
S0	424	6	0	0	0	0

Fig 8. Score distribution of All Essays vs. Long Essays

Average value of scores of all essays = 2.493

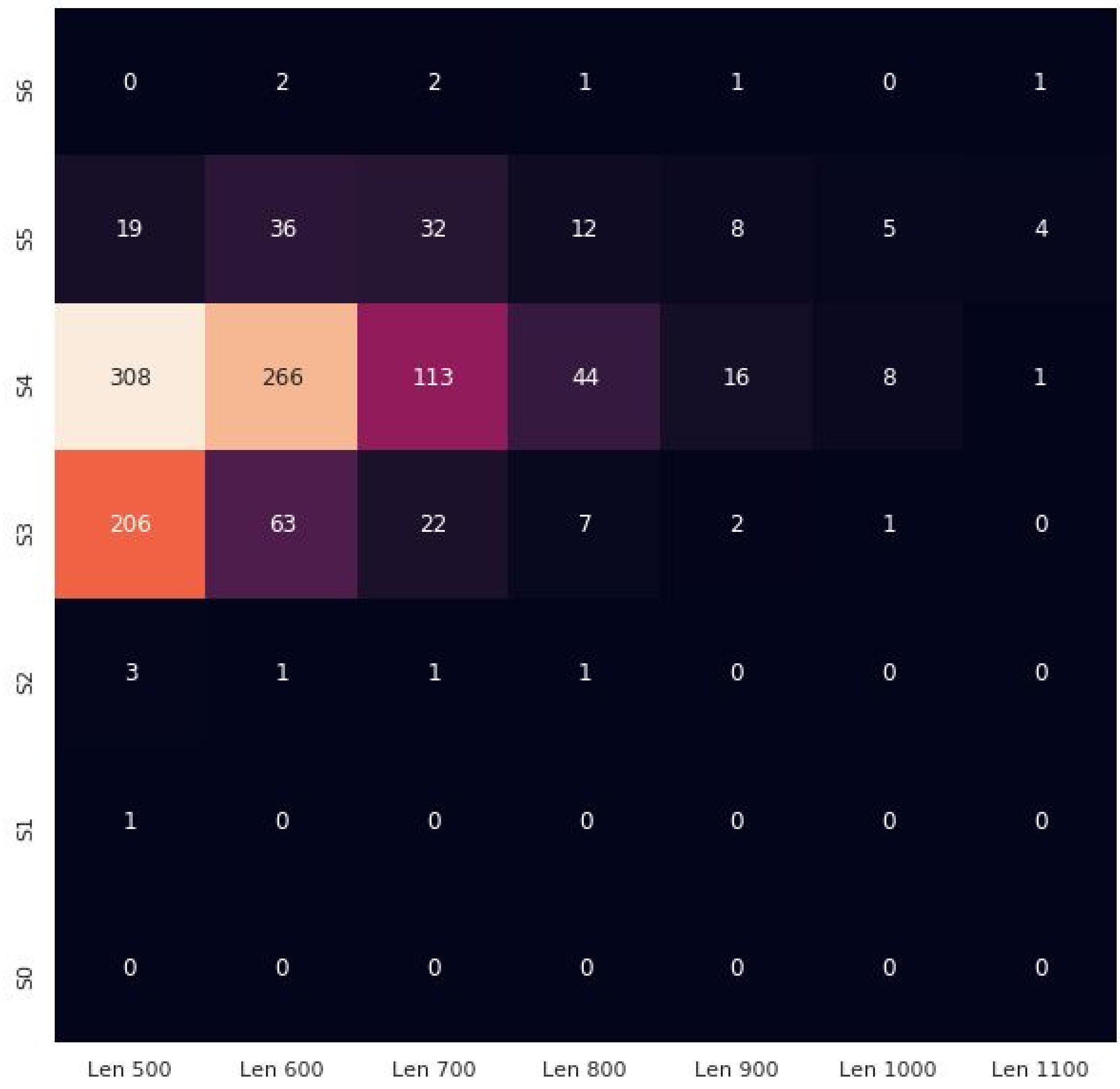


Fig 9. Score distribution of All Essays vs. Long Essays

Average value of scores of long essays = 3.843

From the score distribution of all the essays vs that of only long essays and the starkly different average value of their scores, it is evident that they do not share an identical distribution. This compelled us to use 2 ANN regressors, one for each type of essays. It also happened to solve the 512-tokens cap problem with BERT.

Note - Though we tried a similar approach (short and long essays) for the GloVe + LSTM network, it did not give any significant performance boost.

Time taken to train 200 epochs = 5 mins (Kaggle Kernel GPU)

Table 9. Results for Joint BERT Model comprising 2 ANN Regressors (200 epochs)

	Kappa	**Accuracy**

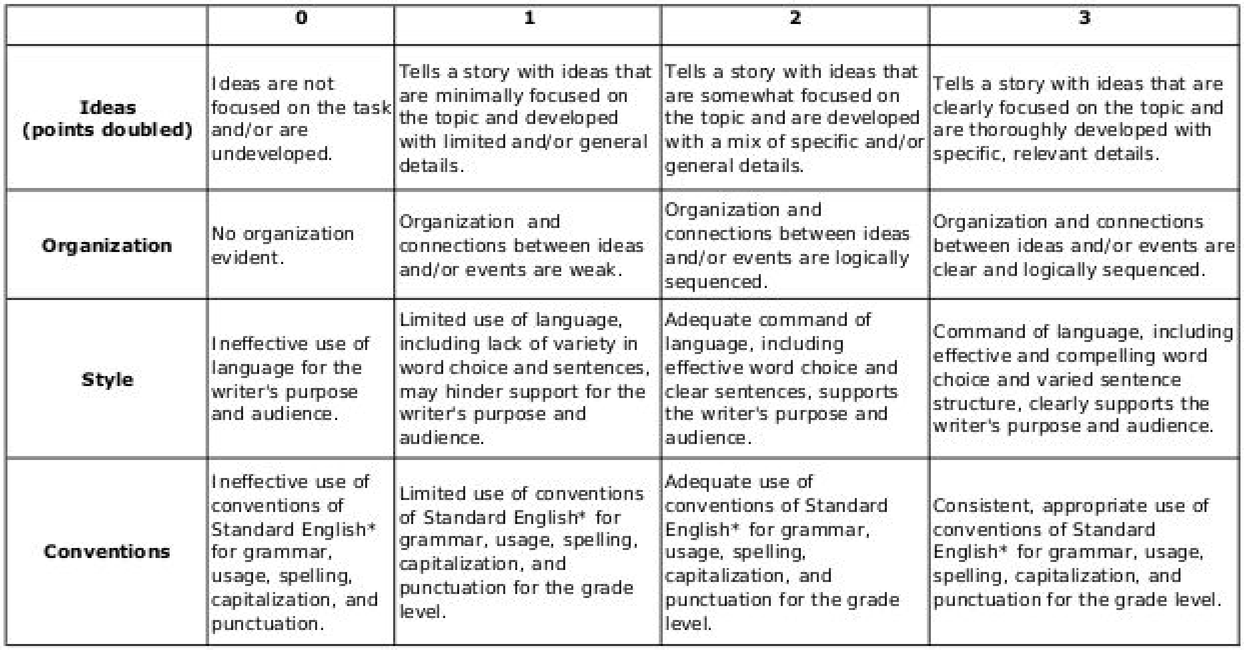
Training	Before	After	Before	After
Train	0.045	0.867	0.09	0.63
Val	0.053	0.864	0.09	0.63

PATTERNS CAPTURED BY OUR LSTM MODEL

Generating Embeddings for Unseen Essays

As mentioned in the earlier sections, humans essentially score essays on the basis of attributes like creativity, vocabulary, fluency etc. These features are “extracted” by the humans and then combined to get a final score. Upon experimenting with our LSTM model, we found that it tried to grade essays in a similar fashion.

We use essays from set #7 for this experiment as they have been scored on individual attributes explicitly. That is, the scoring guide for humans instructed them to score each essay on its **Ideas**, **Organisation**, **Style** and **Conventions**. These scores were then combined for a final score, which we won’t be requiring for this experiment.



	0	1	2	3
Ideas (points doubled)	Ideas are not focused on the task and/or are undeveloped.	Tells a story with ideas that are minimally focused on the topic and developed with limited and/or general details.	Tells a story with ideas that are somewhat focused on the topic and are developed with a mix of specific and/or general details.	Tells a story with ideas that are clearly focused on the topic and are thoroughly developed with specific, relevant details.
Organization	No organization evident.	Organization and connections between ideas and/or events are weak.	Organization and connections between ideas and/or events are logically sequenced.	Organization and connections between ideas and/or events are clear and logically sequenced.
Style	Ineffective use of language for the writer's purpose and audience.	Limited use of language, including lack of variety in word choice and sentences, may hinder support for the writer's purpose and audience.	Adequate command of language, including effective word choice and clear sentences, supports the writer's purpose and audience.	Command of language, including effective and compelling word choice and varied sentence structure, clearly supports the writer's purpose and audience.
Conventions	Ineffective use of conventions of Standard English* for grammar, usage, spelling, capitalization, and punctuation.	Limited use of conventions of Standard English* for grammar, usage, spelling, capitalization, and punctuation for the grade level.	Adequate use of conventions of Standard English* for grammar, usage, spelling, capitalization, and punctuation for the grade level.	Consistent, appropriate use of conventions of Standard English* for grammar, usage, spelling, capitalization, and punctuation for the grade level.

Fig 10. Scoring guide for humans (essays set #7)

We devised a simple network to test if our model really captures these defined attributes. We took only the trained input embedding and LSTM layers of our LSTM network. Since they

would finally produce 64-dimensional embeddings for each essay, we made a simple 2-layered (7 units each layer) ANN classifier to take the embeddings and output a grade.
We chose this approach of verification because it will show the quality of LSTM embeddings. If our model overfit on the essay sets 1-6, it would produce poorly distributed embeddings. This would in turn result in the ANN classifier not being able to detect the aforementioned attributes and hence failing to grade the essay based on the attributes.

The experiment was run 4 times, as there were 4 attributes to be graded. Each time, the model was made to predict the score for a particular attribute and compared against the score given by Human Rater #1. The results are tabulated below

Table 10. Accuracy of Classification for each attribute (200 epochs for each)

	Accuracy		
Attributes	**Random**	**Train**	**Test**
Ideas	0.25	0.59	0.52
Organisation	0.24	0.53	0.51
Style	0.24	0.65	0.63
Conventions	0.27	0.71	0.51

It is evident from the table that our model does actually capture the attributes to some extent. Though it doesn’t do an excellent job, it achieves much better accuracy than a random model. It is worth noting that the model performs better in Style and Conventions than the other 2. This is due to the fact that Style and Conventions majorly consist of grammar, spellings, punctuation and sentence structure, which are easier to comprehend than Idea and Organisation. The latter two require a higher understanding of the essays, even for a human. The model’s performance in scoring vocabulary, grammar and spellings is further demonstrated in the next experiment.

Tweaking a Sample Essay

We consider a sample essay from essay set #5 with a score of 4 (out of 4). We obtain the scores for the essay and its sabotaged derivatives, i.e., modified versions of the same essay with grammatical errors, poorer vocabulary and spelling errors. Also, we illustrate a particular aspect of our model with the help of a graph.

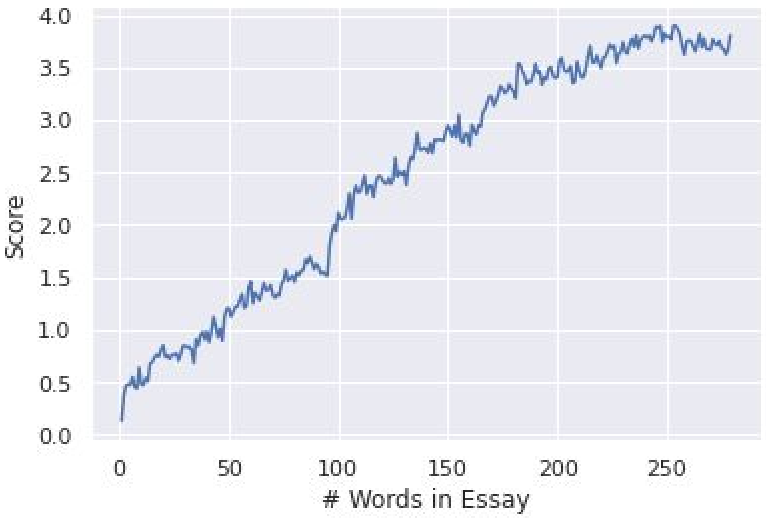


Fig 11. Score vs Length of Essay for an essay from set #5

The model has learnt to score an essay higher for more words on an average. This reflects the same trend in figures 8 and 9.

Table 11. Essay, its erroneous derivatives and their scores

Error Type	Essay	Score
No Error	The mood created by the author in this memoir is both happy and uplifting but also sad. The examples of how it is happy are the way the author describes in paragraph @NUM1 how his parents were able to afford to move into a better and larger house, "twenty-one-year-old @PERSON2 and twenty-seven-year-old Narciso Rodriguez, Sr., could afford to move into a modest, @NUM2-room apartment." @CAPS1 example of the happiness is when he is describing the time when his parents were decorating the appartment to look like a traditional-cuban home, "Within it's walls, my young parents created our traditional cuban home." @CAPS1 example of happiness is when he describes his community of united and hard-working immigrants, "In our neighborhood all of those cultures came together in great solidarity and friendship." And the final example is when he mentions that lot's of his family and friends grace their kitchen table more times than not. The mood changes about halfway through the memoir in paragraph @NUM3 when it says, "Even though it meant leaving behind their families, friends and carrers in the country they loved." This is when the parents are leaving Cuba and their families and friends to take on a life in @LOCATION1. @CAPS1 example of sadness is when it says "The barriers to work were strong and high, and my parents both had to accept that they might not be able to find the kind of jobs they deserved." This means that the parents will have a much harder time finding jobs in @LOCATION1, and they might not get the pay they deserve. Those of the various moods set by the author in this memoir.	3.83
Grammar	The mood created by the author ~~in~~ this memoir is both happy and uplifting ~~but~~ also sad. The examples of ~~how it is~~ happy are ~~the~~ way ~~the~~ author describes in paragraph @NUM1 how his parents were able to afford to move into a better and larger house, "twenty-one-year-old @PERSON2 and twenty-seven-year-old	3.75

	Narciso Rodriguez, Sr., could afford to move into ~~a~~ modest, @NUM2-room apartment." @CAPS1 example of the happiness is when he is describing the time when his parents were decorating the appartment ~~to~~ look like a traditional-cuban home, "Within it's walls, my young parents created our traditional cuban home." @CAPS1 example of happiness is when he describes his community ~~of~~ united and hard-working immigrants, "In our neighborhood all of those cultures came together ~~in~~ great solidarity and friendship." And the final example is when he mentions that lot's of his family and friends grace their kitchen table more times than not. The mood changes about halfway through the memoir in paragraph @NUM3 when it says, "Even though it meant leaving behind their families, friends and carrers in the country they loved." This ~~is~~ when the parents are leaving Cuba and their families and friends to take ~~on a~~ life in @LOCATION1. @CAPS1 example of sadness is when it says "The barriers to work were strong and high, and my parents both had to accept that they might not be able ~~to~~ find the kind of jobs they deserved." This means that the parents will have a much harder time finding jobs in @LOCATION1, and they might not get the pay they deserve. Those ~~of~~ the various moods set by the author in this memoir.	
Vocabulary	The mood created by the author in this memoir is ~~both~~ happy ~~and uplifting~~ but also sad. The examples of how it is happy are the way the author ~~describes~~ writes in paragraph @NUM1 how his parents were able to afford to move into a better and larger house, "twenty-one-year-old @PERSON2 and twenty-seven-year-old Narciso Rodriguez, Sr., could afford to move into ~~a modest~~ an okay, @NUM2-room apartment." @CAPS1 example of the happiness is when he is describing the time when his parents were decorating the appartment to look like a traditional-cuban home, "Within it's walls, my young parents created our traditional cuban home." @CAPS1 example of happiness is when he describes his ~~community~~ group of united and hard-working immigrants, "In our neighborhood all of those cultures came together in great ~~solidarity~~ togetherness and friendship." And the final example is when he mentions that lot's of his family and friends grace their kitchen table more times than not. The mood changes about halfway through the memoir in paragraph @NUM3 when it says, "Even though it meant leaving behind their families, friends and carrers in the country they loved." This is when the parents are leaving Cuba and their families and friends to take on a life in @LOCATION1. @CAPS1 example of sadness is when it says "The ~~barriers~~ difficulties to work were ~~strong and high~~ big, and my parents both had to accept that they might not be able to find the kind of jobs they deserved." This means that the parents will have a much harder time finding jobs in @LOCATION1, and they might not get the pay they deserve. Those of the various moods set by the author in this memoir.	3.65
Spellings	The mood created by the author in this memoir is both happy ~~and~~ nad uplifting but ~~also~~ aslo sad. The examples of how it is happy are the way the author describes in paragraph @NUM1 how his parents were ~~able~~ abel to afford to move into a better and ~~larger~~ largur house, "twenty-one-year-old @PERSON2 and twenty-seven-year-old Narciso Rodriguez, Sr., could afford to move into a modest, @NUM2-room ~~apartment~~ apatment." @CAPS1 example of the ~~happiness~~ happyness is when he is ~~describing~~ discribing the time when his parents were decorating the appartment to look like a ~~traditional-cuban~~ tradisional-cuban home, "Within it's walls, my young parents created our traditional cuban home." @CAPS1 example of happiness is when he describes his community of united and hard-working immigrants, "In our neighborhood all of those cultures came together in great solidarity and friendship." And the final example is when he ~~mentions~~ mensions that lot's of his family and friends grace their kitchen table more times than not. The mood changes about halfway through the memoir in paragraph @NUM3 when it says, "Even though it meant leaving ~~behind~~ bihind their ~~families~~ familys , friends and carrers in the country they loved." This is when the parents are ~~leaving~~ leeving Cuba and their families	3.7

	and friends to take on a life in @LOCATION1. @CAPS1 example of sadness is when it says "The barriers to work were strong and high, and my parents both had to accept that they might not be able to find the kind of jobs they deserved." This means that the parents will have a much harder time finding jobs in @LOCATION1, and they might not get the pay they deserve. Those of the various moods set by the author in this memoir.	
All combined	All combined	3.37

As seen in the table, our model was clearly able to detect the errors and appropriately reduce the score of the tweaked essays. Hence, this experiment concluded that the model can analyse the attributes like grammar, vocabulary and spelling in an essay. In addition, higher length essays are scored higher, on an average.

CONCLUSION

Table 12. Final Results of our project

Metrics □	**Kappa**		**Accuracy**	
Models ▯	**Train**	**Test**	**Train**	**Test**
GaussianNB + TFIDF	0.87	0.23	0.81	0.21
SVC + Count	0.92	0.54	0.91	0.48
MultinomialNB + Count	0.63	0.33	0.67	0.33
SVC + TF-IDF	0.78	0.39	0.73	0.36
GloVe + LSTM Classifier	0.851	0.848	0.63	0.60
ANN Classifier + BERT encodings	0.844	0.838	0.63	0.60
2 ANN Regressors (short and long) + BERT encodings*	**0.867**	**0.864**	**0.63**	**0.63**
GloVe + LSTM Regressor§	**0.877**	**0.873**	**0.64**	**0.63**

§ Best Kappa Model
*** Best Accuracy Model**

The best model achieves a kappa of 0.873. Though BERT hasn't been fine-tuned on this dataset, it achieves almost the performance as the LSTM network. Our models do not beat the human kappa (0.905), but achieve close results. The working of our LSTM model can be clearly explained by the 2 experiments we performed. Hence, we conclude that our model encodes even unseen essays from which the traits to score an essay can be extracted, to an extent. It also pays attention to details like grammar, vocabulary and spellings.

In the comparison of ML and DL models, DL ones largely outperform the former models. This clearly proves that the DL models provide immense leverage over the previously used ML models, in Essay Grading at least, and continue to expand.

What we have learnt

Deepak H R

This was my first project involving neural networks. On one end, I learnt different methods of preprocessing data such as tokenization, removing spelling mistakes, removing unlabelled data, splitting train-test sets. Although these are quite simple, they enable us to feed fairly clean data into the neural network, and how good data preprocessing could go a long way in attaining consistent results.

One the other end, I learnt to use TensorFlow and PyTorch libraries to implement an LSTM neural network and infer from sequential data. The project was an opening for me to explore my interests in neural networks and what lies beyond my project. This motivated me to read a few research papers along the way, and implement models extending the BERT model as a part of our project and also other datasets too. I also learnt about the attention layer and its types and what role it plays in the BERT model.

I learned a wide variety of visualization techniques like PCA and t-SNE and was able to see and understand the difference in the models. The productivity of the project majorly from the fact that it was pursued by a team and a mentor. I learned to work as a team and set my goals and expectations straight with the help of our mentor.

Anudeep Tubati

I learnt a number of things from this project, both related to the field we explored and about the general sense of formulating a research project. Firstly, I read research papers relating to our field which gave me a good understanding of how various models work and an intuition of what might be improved. I learnt about innovative models like Word2Vec, Attention, BERT and how they achieved good performance in their times. In addition to the theory part, I experimented a lot with the API of Keras and TensorFlow (not to forget Numpy and pandas) to best apply all the theory in code. As we progressed through the project, I learnt general skills like having better communication with the team/mentor, clearly explaining the motivation for new ideas and elaborating a project with well-defined inputs/outputs and visualisations to make things easier for a reader. As a whole, this project proved to grant me a holistic and scientific experience.

Future Work

We list some ideas that may improve the performance of our models. These ideas are just a part of our intuitions based on the experiments we've performed and give no guarantees to better our model.

- Encode Essay Prompts (or questions) along with the essays to supply more context to the model
- Use a better procedure to "join" the 2 models (long and short), for BERT
- Fine-tune BERT on this dataset
- Hyperparameter tuning (quite obviously)

APPENDIX

a. Quadratic Weighted Kappa calculation, https://docs.google.com/presentation/d/1A4-Um5dcuqe93karH2e9_L8msZ5SY9v4sI_g7NmZEXk/edit?usp=sharing

REFERENCES

1. UNESCO Institute for Statistics Data, http://data.uis.unesco.org. Accessed [22-05-2020]

2. CBSE NIC Press Note, http://cbse.nic.in/newsite/attach/PressNote_10_2019_X.pdf, Accessed [22-05-2020]

3. ASAP AES, The Hewlett Foundation, Kaggle Competitions, https://www.kaggle.com/c/asap-aes

4. Cohen's Kappa Wikipedia, https://en.wikipedia.org/wiki/Cohen%27s_kappa#Weighted_kappa. Accessed [24-05-2020]

5. NLTK, https://www.nltk.org/

6. Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017

7. Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *arXiv preprint arXiv:1910.10683* (2019)

8. Stanford Sentiment Treebank, Stanford NLP, https://nlp.stanford.edu/sentiment/index.html

9. NLP with Disaster Tweets, Kaggle Competitions, https://www.kaggle.com/c/nlp-getting-started

10. SOTA for SST-2, Paperswithcode, https://paperswithcode.com/sota/sentiment-analysis-on-sst-2-binary. Accessed [03-06-2020]

11. Maaten, Laurens van der, and Geoffrey Hinton. "Visualizing data using t-SNE." *Journal of machine learning research 9*. Nov (2008): 2579-2605.