International Journal of Future Generation Communication and Networking
Vol. 13, No. 3s, (2020), pp. 173–179

Automated Essay Grading

Vinay Sanga, Shreyansh Patil, Abhishek Jagtap, Prashant Raut, Geeta S Navale

Department of Computer Engineering, Sinhgad Institute of Technology & Science, Savitribai Phule Pune University, Pune

vinayklind@gmail.com
shreyanshpatil98@gmail.com
abhishekjagtap88@gmail.com
rautprashant902@gmail.com
gsnavale_sits@sinhgad.edu

Abstract

In the task of essay grading a grade is assigned to an essay in the most human-like manner possible. Authors tackle this problem by building statistical models which try and predict how a human would evaluate an essay. In this paper, we propose to use Long Short Term Memory (LSTM) neural networks to create an essay grader to undertake this daunting task. The grader will give an essay a grade which would be expected from a human grader. The user will be able to see his score instantaneously which is an improvement as compared to human graders. There are some very challenging tasks such as recognizing the context of the essay. Also, the vocabulary and various versions of English will have to be considered while evaluating the essay. The grader will be able to grade an essay without any bias and high accuracy.

Keywords— Essay Grading, Deep Learning, LSTM, Computerised Testing, Text Grading.

1. INTRODUCTION

Essays are a tool for testing the students' fluency, vocabulary and grammatical correctness in a language. They are also useful to test one's creativity, originality and articulateness. The highly subjective and diverse nature of an individual in writing an essay makes it difficult to grade the essay uniformly across many human graders. In addition to this there are various other biasing factors in grading an essay. There has been research on automatic essay grading since the 1960s. The first systems based their grading on the surface information from essays. These systems were successful though they failed to capture aspects like grammatical correctness and language fluency. Much research has been conducted in the field most notably by Educational Testing Service (ETS). Clubbing this together with the resurgence in new technologies such as neural networks, deep neural networks, there is a whole new world of possibilities due to their capacity of modelling complex patterns in data. These methods do not depend on feature engineering so they are really useful for solving problems in an end-to-end fashion. With this intuition this project aims to the relevant knowledge in the field of education and try to create an essay grader which can make quality education more accessible. The work also explores methods of improving the quality and usability of the system.

I. MOTIVATION

The human graders unknowingly tend to grade an essay biasing towards the individual subject matter presented. Another major drawback is the time required to grade essays can be significantly high. The present technologies present an excellent opportunity to automate tedious tasks such as essay grading. Availability of powerful Deep Learning libraries is a major push towards the reliance and devising the system. This project will help in maintaining an unbiased and fair approach towards evaluating the written essay for competitive exams, tests, etc. which is in turn beneficial in multiple ways.

ISSN: 2233-7857 IJFGCN

II. RELATED WORK

Alex Adamson et. al. [1] have implemented an essay grader on the Hewlett Foundation dataset using different ML techniques such as SVM, Latent Semantic Analysis. Each essay was graded by at least two humans. It uses the Quadratic Weighted Kappa as closeness measure. The general pipeline involves extracting features from the raw essays, and iteratively training and using k- folds cross-validation on the model on selected essay sets in order to optimize hyperparameters. It is shown that for essays of intermediate writing level and given enough human graded training examples for a writing prompt, they can automate the grading process for that prompt with fairly good accuracy. However, the classifier was built for this task with limited success.

Derrick Higgins et. al. elaborated Support vector machine (SVM) for classification and regression [2]. SVM technique relies on kernel functions. There were multiple goals in this work. The authors wanted to introduce a concept of essay coherence comprising multiple aspects, and investigate what linguistic features drive each aspect in student essay writing. They have worked with writing experts to develop a comprehensive protocol that details how coherence in writing can be evaluated, either manually or automatically. Using this protocol, human annotators labeled a corpus of student essays, using the coherence dimensions. However, this approach is relatively new and though better than previous ones is valid only theoretically.

The LSA approach to detect major research topics and themes of a multidisciplinary field was applied by Gang Kou et. al. [3]. The LSA analysis can be summarized in three main steps. The first step is to set up a term-document matrix in which each row stands for a key word or term and each column stands for a document or context in which the key word appears. An entry in the matrix is the frequency of a key word in the corresponding document. The second step is to transform the term frequencies in a term-document matrix using various weighting schemes. The third step is to perform SVD on the matrix to reduce the dimensionality, which is the key feature of the LSA method.The limitations are - a. It is not possible to use LSA on an unestablished field of research; b. Only English journals are considered in this approach.

Kaveh et. al discusses about various kinds of approaches available for constructing a Machine Learning model. They are also using the dataset mentioned in the first paper. They have selected LSTM model [4] as it works the best among all the available ML models. They have discussed their approach as well as the work done by other prominent researchers. Their model does not require feature engineering as compared to some other models. The model performs 5.6% better than the baseline performance. There is still quite a scope in improving the model as per their views.

In their paper, Hongbo Chen et. Al [5] argue that the current AES systems can be further improved by taking into accountthe agreement between human and machine raters. They have proposed a listwise learning to rank approach to automated essay scoring (AES) by directly incorporating the human-machine agreement into the loss function.

Furthermore, evidences supporting the use of Deep Learning are shown by Ronan Collbert et. Al[6]. They have shown how both multitask learning and semi-supervised learning improve the generalization of the shared tasks, resulting in state of the-art performance. They have used Part-Of-Speech Tagging (POS), Chunking, Named Entity Recognition (NER), Semantic Role Labeling (SRL), Language Models and Semantically Related Words (“Synonyms”) similar to previously seen approaches.

In an another approach[7], Peter Phandi et. Al proposed a novel domain adaptation technique based on Bayesian linear ridge regression and domain adaptation. Domain adaptation is the task of adapting knowledge learned in a source domain to a target domain.

TABLE I
LITERATURE SURVEY

ISSN: 2233-7857 IJFGCN

Ref No.	Highlights	Observations
1.	• Essays can be graded without the help of the teacher. • Different models are compared and best amongst them is used. • The grades are equivalent to human graders.	• The size of essay which can be graded is limited. • Good quality training examples are required as presently the accuracy is limited to the present set.
2.	• Hybrid approach is used to classify different coherence dimensions with a high- or low-quality rank. • In each of the experiments with less fold-cross validations results are reported for large set of essays.	• Choosing appropriate kernel function is difficult and complex task. • In case of using high dimension kernel it generates too many support vectors which reduces training speed.
3.	• Doesn't require feature engineering. • One of the best implementation of the grading system.	• Training time is more as it is a Deep Learning model.
4.	• LSA is capable of assuring decent results. It works well on dataset with diverse topics. • LSA can handle Synonymy problems to some extent. • It is faster as compared to other dimensionality reduction models.	• LSA depends on identifying frequent word usage patterns from a collection of text, it is difficult to capture a research area if it is not well established. • The research abstracts collected in this analysis include only English language journals.
5.	• A deep neural network model that is capable of representing local conceptual and usage information by using LSTM.	• ATS(automated text scoring) takes more time to generate text score using deep neural network.
6.	• A model for scoring essay dimension of argument strength which is important aspect of argument essays this model gives argument strength. • This model gives argument strength of essays which convince most reader.	• This model is only detecting the scores of argument essays automatically but not gives scores of all types of essays.
7.	• A general deep NN architecture for NLP. • This architecture is extremely fast enabling to take advantage of huge databases.	• When training the SRL task jointly with their language, model achieved state of-the-art performance in SRL without any explicit syntactic features.
8.	• Domain adaptation can achieve better results compared to using just the small number of target domain data or just using a large amount of data from a different domain. • This research will help reduce the amount of annotation work needed to be done by human graders to introduce a new prompt.	• The effectiveness of using domain adaptation when only have a small number of target domain essays is discussed.

III. PROPOSED SYSTEM

ISSN: 2233-7857 IJFGCN

The grading system is implemented as a Client – Server Architecture. The essay writers are the clients who can access the app through their web browsers. The model and other implementations as usual reside at the server end. The Essay will be passed to the Server for grading and then will be submitted for grading. After grading, the grade will be shown to the user on the screen. The representation of the system is shown in Figure no. 1

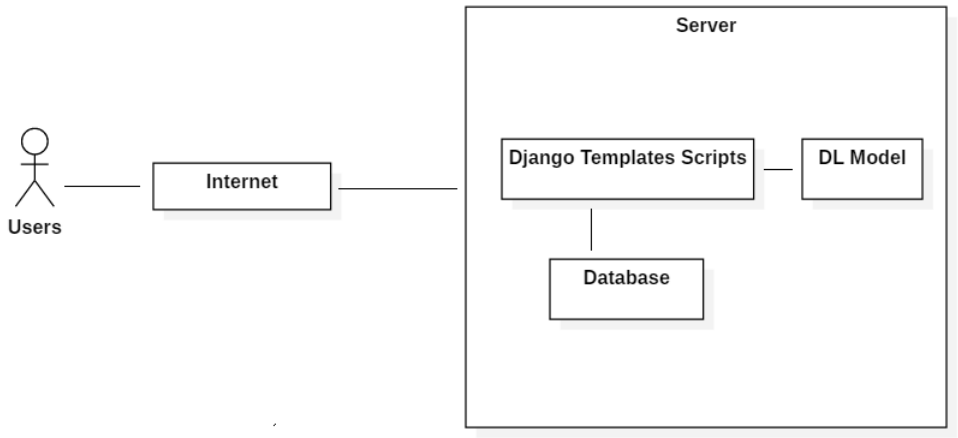

Figure no. 1: System Architecture

The communication between the LSTM predictor and the Backend script is the most important part. If the libraries are not compatible, it can cause issues due to the threading involved in the scripts.

IV. RESULTS

The essay grader achieved an accuracy of 96.2% using a 3-folds cross validation. The model was trained and tested on both Google Colab as well as Local PC. The accuracy fluctuated between 96.17% -98.7% based on the random vector encodings but mostly it lingered around 96.2%. The obtained accuracy can be seen from Figure no. 2

ISSN: 2233-7857 IJFGCN

```
163/163 [==============================] - 3s 18ms/step - loss: 7.6775 - mae: 1.5711
Epoch 37/50
163/163 [==============================] - 3s 19ms/step - loss: 7.8863 - mae: 1.5725
Epoch 38/50
163/163 [==============================] - 3s 19ms/step - loss: 8.0074 - mae: 1.5718
Epoch 39/50
163/163 [==============================] - 3s 19ms/step - loss: 7.8068 - mae: 1.5676
Epoch 40/50
163/163 [==============================] - 3s 19ms/step - loss: 7.4385 - mae: 1.5597
Epoch 41/50
163/163 [==============================] - 3s 19ms/step - loss: 7.5482 - mae: 1.5364
Epoch 42/50
163/163 [==============================] - 3s 19ms/step - loss: 7.6092 - mae: 1.5582
Epoch 43/50
163/163 [==============================] - 3s 19ms/step - loss: 7.4036 - mae: 1.5433
Epoch 44/50
163/163 [==============================] - 3s 19ms/step - loss: 7.0982 - mae: 1.5263
Epoch 45/50
163/163 [==============================] - 3s 19ms/step - loss: 7.2797 - mae: 1.5259
Epoch 46/50
163/163 [==============================] - 3s 19ms/step - loss: 7.6733 - mae: 1.5462
Epoch 47/50
163/163 [==============================] - 3s 19ms/step - loss: 7.6982 - mae: 1.5389
Epoch 48/50
163/163 [==============================] - 3s 19ms/step - loss: 7.5287 - mae: 1.5333
Epoch 49/50
163/163 [==============================] - 3s 19ms/step - loss: 7.4795 - mae: 1.5172
Epoch 50/50
163/163 [==============================] - 3s 19ms/step - loss: 7.0886 - mae: 1.5117
Kappa Score: 0.9617094280868951
```

Figure no. 2: Accuracy measure in Kappa Score

The essay topic selection window and the grader score is displayed in figures 3, 4 & 5.

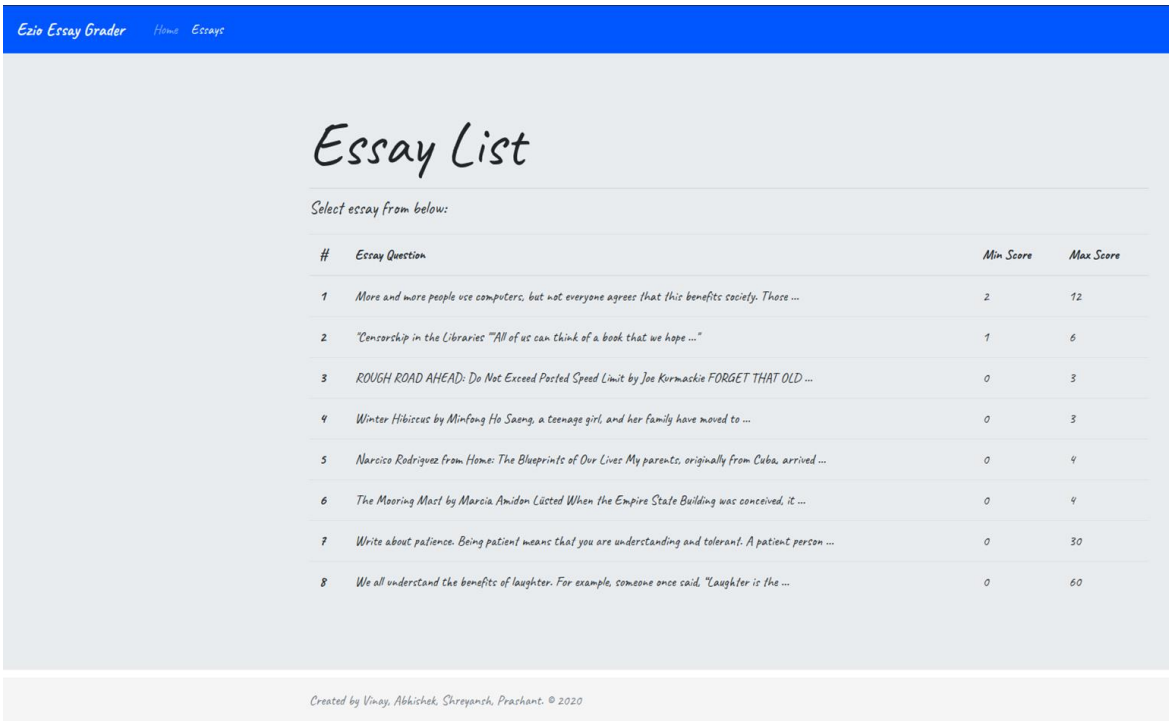

Figure no. 3: List of Essay Topics displayed

ISSN: 2233-7857 IJFGCN

Figure no. 4: The text box for writing the essay

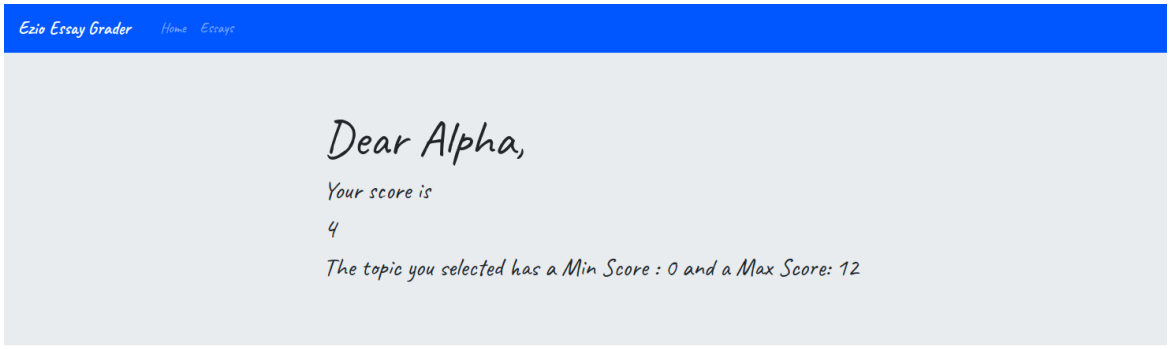

Figure no. 5: Score given by the grader

V. CONCLUSIONS

The essay grading task is a laborious task which can be automated with the help of Deep Learning. The grader which is proposed in the paper has an accuracy of 96.17% which is the highest that has been observed until now. The grader is very good in differentiating between ambiguous sentence formations and the new vector representation is very good in implementing the NLP translations.

VI. FUTURE SCOPE

The essay grader can be used by teachers for grading student essays. It can also be used by testing agencies for test of English writing to take the burden off the human graders. The concept can be extended to other languages too if the dataset is available. Lastly, the accuracy of the model can still be increased if more data is feeded to the LSTM network while training. More sophicated models can be developed using transfer learning at the cost of using more resources required to train the model.

ISSN: 2233-7857 IJFGCN

REFERENCES

1. Adamson Alex, Andrew Lamb, and Ralph Ma, Automated Essay Grading. 2014
2. Higgins Derrick, Jill Burstein, Daniel Marcu, and Claudia Gentile, Evaluating Multiple Aspects of Coherence in Student Essays. In HLT-NAACL, pp. 185-192. 2004.
3. Gang Kou, Yi Peng, An Application of Latent Semantic Analysis for Text Categorization In International Journal of Computers, Communications &
   Control (IJCCC) 10(3):357 April 2015
4. Kaveh Taghipour, Hwee Tou Ng, A Neural Approach to Automated Essay Scoring In Proceedings of the 2016 Conference on Empirical Methods in
   Natural Language Processing, November 2016
5. Isaac Persing and Vincent Ng , Modeling Argument Strength in Student Essays , 2015.
6. Dimitrios Alikaniotis , Helen Yannakoudakis and Marek Rei , Automatic Text Scoring Using Neural Networks , 16 Jun 2016.
7. Ronan Collobert and Jason Weston NJ 08540, A Unified Architecture for Natural Language Processing,2011.
8. Peter Phandi1 , Kian Ming A. Chai2 and Hwee Tou Ng1 , Flexible Domain Adaptation for Automated Essay Scoring Using Correlated Linear Regression , 17-21 September 2015.

ISSN: 2233-7857 IJFGCN