e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:04/April-2021 Impact Factor- 5.354 www.irjmets.com

SMART ESSAY GRADER SYSTEM

Aman Sharma*[1],Vaibhav Urkude*[2],Adhoksh Sonawane*[3], Prof. Nidhi Sharma*[4]

*[1,2,3]Student, Department Of Computer, Bharati Vidyapeeth College Of Engineering,
Navi Mumbai, Maharashtra, India.

*[4]Prof., Department Of Computer, Bharati Vidyapeeth College Of Engineering,
Navi Mumbai, Maharashtra, India.

ABSTRACT

In order to optimize Human-Machine agreement for automatic evaluation of textual summaries or essays, automated essay grading has been a research field. With a growing number of people taking multiple exams such as the GRE, TOEFL, and IELTS, grading each paper would become more challenging, not to mention the challenge for humans to maintain a consistent mindset. In this situation, it is extremely difficult to rate a large number of essays in a short amount of time. This project aims to address this issue by developing a stable interface that will aid humans in grading essays. This study served as a medium for us to extract features such as the Bag of Words, numerical features such as the count of sentences and words, as well as their average lengths, structure, and organization, in order to rate the essay with the highest level of precision. This algorithm was chosen because it works well for small datasets.

Keywords: NLP, Essay, Words, Sentences, NN.

I. INTRODUCTION

Purpose

Since the early 1960s, automated essay grading has been a research subject. It's a challenging job because we need to extract both quantifiable and nonquantifiable attributes, such as the writer's feelings, when writing on paper. It will appear that extracting is a simple procedure. The system's goal is to divide a large number of textual entities into a small number of distinct groups, each corresponding to a range of possible scores—for example, 1-100. The artificial environment we generate will recognize patterns and try to predict the next possible performance using a training dataset. This project investigates how text mining can aid in essay scoring. The method of grading student essays without human intervention is referred to as automated essay scoring. An AES system takes an essay written for a specific prompt as input and assigns a numeric score to the essay based on its material, grammar, and organization. Regression methods are normally applied to a collection of carefully constructed features in such AES systems. Humans find it difficult to understand all of the considerations that go into awarding a grade to an essay. Since the system is focused on recurrent neural networks, it can effectively encode the information needed for essay evaluation while also learning complex patterns in the data through non-linear neural layering. The findings show that the device outperforms a solid baseline in automated essay scoring and produces cutting-edge performance.

II. LITERATURE SURVEY

We have gone through a number of papers related to our project which had been pursued by number of researchers in the past. From these papers, we got to know the different technologies which were used by different researchers to implement this project. Here we have introduced two papers which we made base papers for our project. These base papers helped us to get a variety of past systems as well as the required information to support and flourish our research.

Table 1: Table of Base papers for our project

Sr No	Name of Paper	Author	Methodology	Limitations
1	Automated Essay Grading Using Feature Selection	Y.Harika, I.Srilatha, V.Lohith Sai, P.Sai Krishna , M.Suneetha	Existing Grading systems have been focusing on different NLP approaches and decided to improving score and performance.	System was quantifiable and accuracy was average.
2	A Neural Approach to Automated Essay Scoring	Kaveh Taghipour and Hwee Tou Ng	Worked on various systems like AES, EASE and compared the results for ideal solution.	Cons of the behavior of systems not able to fully adapt and extract in the model.

Base paper 1 : Automated Essay Grading Using Feature Selection

This project starts with the construction of a model using an input dataset containing essays. The extracted feature values are computed, and rubrics are generated. Stop word elimination, tokenizing, and stemming are all steps in the preprocessing of each essay. This involves 'case folding,' or converting to lower case, as in this example. Later, input test data is passed through, and it goes through the same steps as training data. Later computations are done for all features and score is generated as output within a range of 1 to 5. If necessary, these can be further discretized and graded.

Base paper 2 : A Neural Approach to Automated Essay Scoring

Researchers have devoted a substantial amount of effort to design effective features for automated essay scoring. Our system, on the other hand, takes an essay text as input and learns the features from the results. To do so, we devised a system based on recurrent neural networks for scoring essays from beginning to end. In this paper, we looked at a number of neural network models in order to find the best one. Our best model is a long short-term memory neural network (Hochreiter and Schmidhuber, 1997) and is trained as a regression method. Similar recurrent neural network approaches have recently been used successfully in a number of other NLP tasks.

III. TECHNOLOGY

NLP (Natural language processing) - NLP is a technique for deciphering the form and context of human speech. Here, it is used to study the performance of different existing models in grading the essay.

NN (Neural Networks) - NN is used to model complex relationships between inputs and outputs or to find patterns in data. Here it used as an alternative to NLP to actually mimic the working of human brain in grading the essay.

Django - It is a popular Python web framework. Here it is used to create the web app for our trained models for users to actually interact with our project.

Python - Python is an interpreted high-level general-purpose programming language. We have used Python because as it is having various powerful libraries which makes data visualization, training of models and creation of web app simpler.

Jupyter Notebook - It is a document format based on JSON which is used to run Python code. Here it used to run different parts of project in different cells in order to make code readability simpler.

HTML, CSS, JavaScript - They are used to define the structure of a web page and make it interactive. Here they are used to make interactive webpages in Django framework.

IV. METHODOLOGY

4.1 Approaches Used

By NLP (Natural Language Processing): -

We started the process by taking a dummy dataset of essay answers from Kaggle which were divided and scored by different raters. The data cleaning process then began. The coding language used throughout the project was Python. The cleansing of data included the removal of unwanted rows and columns and unrelated bulk which included the Stop Words and punctuations using NLP. Now after cleaning we went to processing the data. The sentences were further divided to words for tokenization process. This process included the sentence, word, noun, adjective and other grammar items for the scoring of essay. The spell-check was included which negates the marking. Only those features that trigger changes in the score should be considered when processing data sets. Now the further process included the use of modules and Algorithms. Accuracy is observed by seeing the mean squared value of different NLP models. But in our project the resultant values were very low and thus the accuracy of the grade would be miles off the expected result. Thus, we shifted to NN (Neural Networks).

By NN (Neural Networks): -

Here we used the same but processed and cleaned data-set used in NLP processing. In the place of tokenization, the data here was pre-processed by vectorization so that it can be fed into the RNN. We also used the NLTK Toolkit which is used to process human language data and we used the 2-Layer LSTM model, a form of recurrent neural network approach. Users explore a broad dataset in an unstructured manner to discover initial trends, features, and points of interest in data discovery, which is the first step in data analysis. Word2Vec is a learning algorithm for word embedding that can be used on massive datasets. The conversion of text to vector format is required in order for the NLP module to read and process data. Feature selection is a process in which we try to fit a particular machine learning algorithm into a given dataset. It uses a greedy search method, evaluating all possible feature combinations against the evaluation criteria. We used Keras which is a Python Deep learning Application Programming Interface. We used the Hyper-effective 10-Fold-cross Validation technique to train the model for Good Kappa Score. Thus, we succeeded in bringing the highest Kappa score till date of 0.9691. The Correlation Score is calculated by taking into account all of the characteristics of an essay. Basically, the substance of the essay or the data that the user would enter into the user interface. After saving and sending the essay, it will be saved in text format in the database and sent to a text analyzer for normalization, function abstraction, and data optimization. Finally, the resultant score will be displayed after grading with maximum accuracy.

4.2 Data Visualization: -

Before designing any algorithm, we need to see what our data looks like, in what format we need to convert it so it is ideal for an algorithm to run on it. Figure 4.2.1 shows an example of the sample of part of an essay. We can see that essay has certain things like @Caps1 and @caps2. These represent external references and if look closely then they are not causing any problem and we can filter them out along with some different kinds of stop words that are not really important.

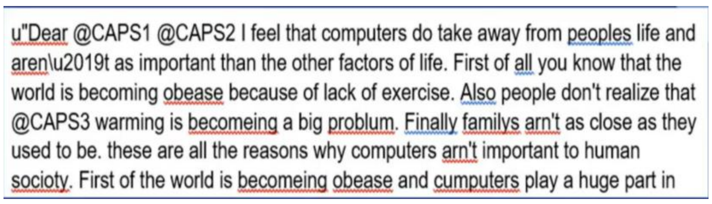
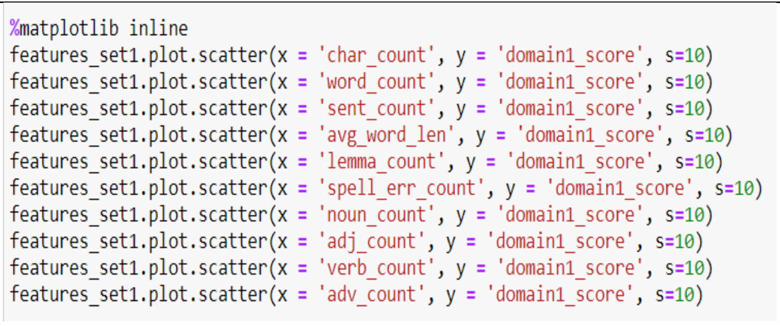

u"Dear @CAPS1 @CAPS2 I feel that computers do take away from peoples life and aren\u2019t as important than the other factors of life. First of all you know that the world is becoming obease because of lack of exercise. Also people don't realize that @CAPS3 warming is becomeing a big problum. Finally familys arn't as close as they used to be. these are all the reasons why computers arn't important to human socioty. First of the world is becomeing obease and cumputers play a huge part in

Figure 4.2.1 Sample part of Input Essay

So basically, we need to come up with parameters that would give us true insight into what should be considered to judge quality on an essay. And these features seemed right like length of essay, sentence length, average word length, sentiment of an essay, spelling and grammar. Now let us see how these affect the quality of essay.

Code:

For plotting the graph between different features of the essay and the domain1 score, we used the python matplotlib library. Here domain1_score is the grade given by human scorer. We can see that in figure 4.2.2.



```
%matplotlib inline
features_set1.plot.scatter(x = 'char_count', y = 'domain1_score', s=10)
features_set1.plot.scatter(x = 'word_count', y = 'domain1_score', s=10)
features_set1.plot.scatter(x = 'sent_count', y = 'domain1_score', s=10)
features_set1.plot.scatter(x = 'avg_word_len', y = 'domain1_score', s=10)
features_set1.plot.scatter(x = 'lemma_count', y = 'domain1_score', s=10)
features_set1.plot.scatter(x = 'spell_err_count', y = 'domain1_score', s=10)
features_set1.plot.scatter(x = 'noun_count', y = 'domain1_score', s=10)
features_set1.plot.scatter(x = 'adj_count', y = 'domain1_score', s=10)
features_set1.plot.scatter(x = 'verb_count', y = 'domain1_score', s=10)
features_set1.plot.scatter(x = 'adv_count', y = 'domain1_score', s=10)
```

Figure 4.2.2 Code used for Graph Plotting

Graph:

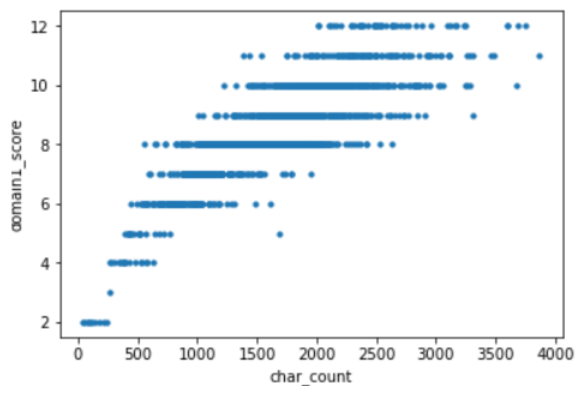


Figure 4.2.3 Char_count

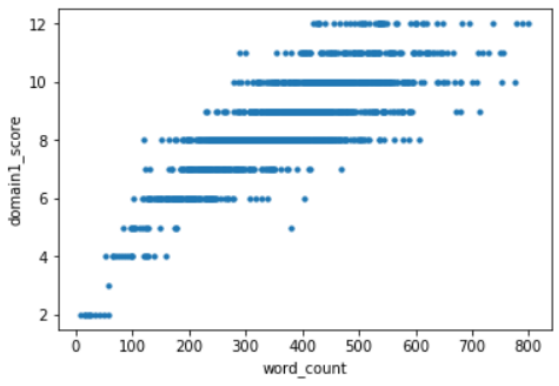


Figure 4.2.4 Word_count

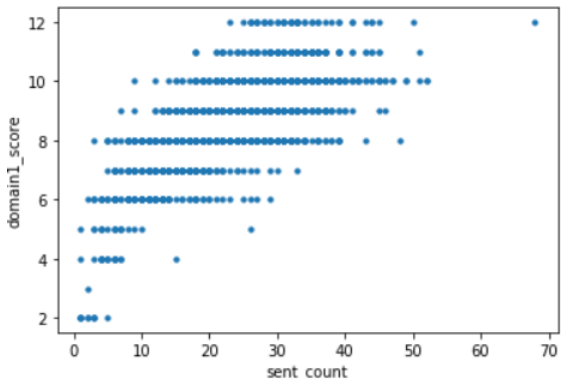


Figure 4.2.5 Sentence_count

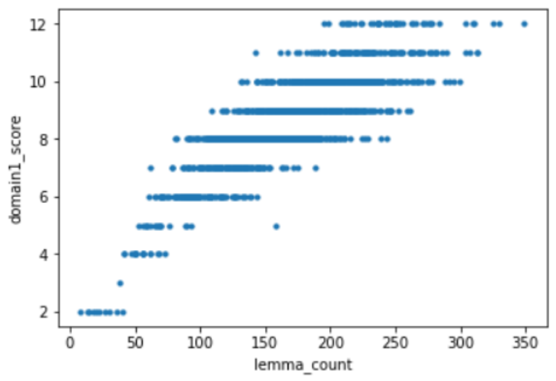


Figure 4.2.6 Lemma_count

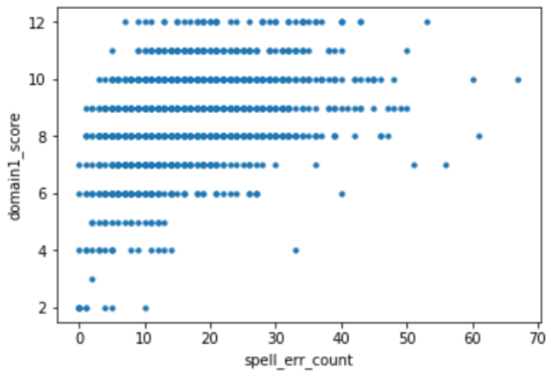


Figure 4.2.7 Spell_err_Count

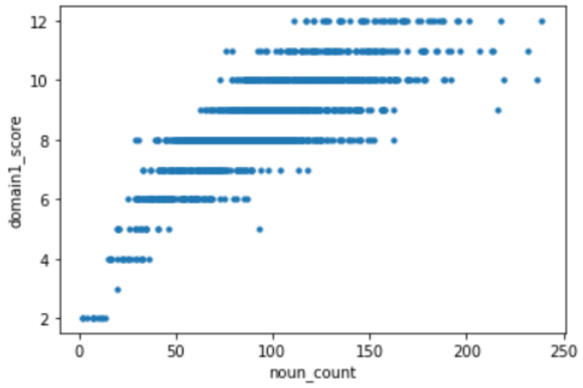


Figure 4.2.8 Noun_count

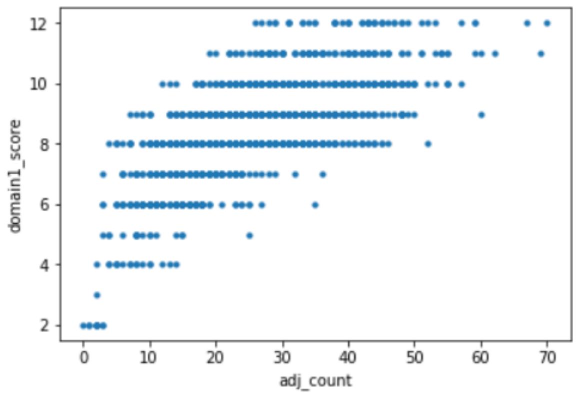


Figure 4.2.9 Adjective_count

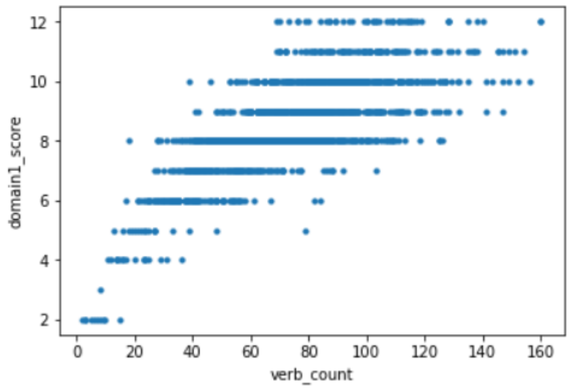


Figure 4.2.10 Verb_count

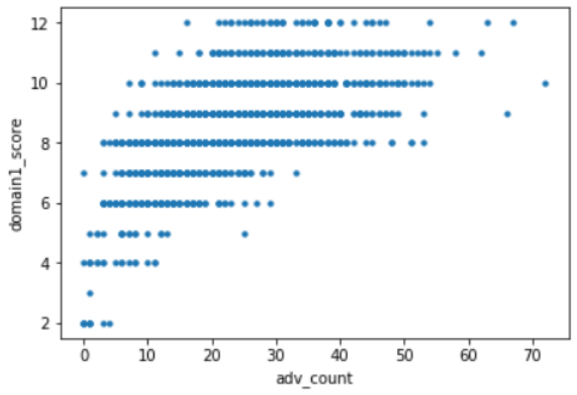


Figure 4.2.11 Adverb_count

So, looking into this we can say that as the length increases there is certain trend that is moving towards the right as the grade is increasing. So, essay with less than 200 length really do not do that well very often and essay which were above 800 length really did well on an average. So let us look at some other parameters like average word length that do not seem to have very large impact on it. Grammar and spelling errors are very distributed and they don't show any trend. This indicated the fact that having the idea is very much important and primary as compared to just spelling errors. So, quality of an essay is basically determined by the content that user had put out rather than judging them on spelling errors. These insights made us realise what features are necessary to decide the grade of the essay the user will get.

V. PROPOSED ALGORITHMS

5.1 Natural Language Processing –

Logistic Regression:

The supervised learning classification algorithm logistic regression is used to estimate the likelihood of a variable to be measured. Since the existence of the target or dependent variable is dichotomous, there are only two classes. In simple terms, the dependent variable is binary in nature, with data coded as 1 (representing success/yes) or 0 (representing failure/no). A logistic regression model predicts P(Y=1) as a function of X mathematically. It's one of the most basic machine learning algorithms, and it can be used to solve a variety of classification problems including spam detection, diabetes prediction, cancer detection, and so on. For two-class problems, logistic regression is the go-to linear classification algorithm. The logistic function, which is at the heart of the system, is called logistic regression. Logistic regression, like linear regression, is represented by an equation. To predict an output value, input values (X) are linearly combined using weights or coefficient values (y). The output value being modelled is a binary value (0 or 1) rather than a numeric value, which is a crucial distinction from linear regression.

ARCHITECTURE

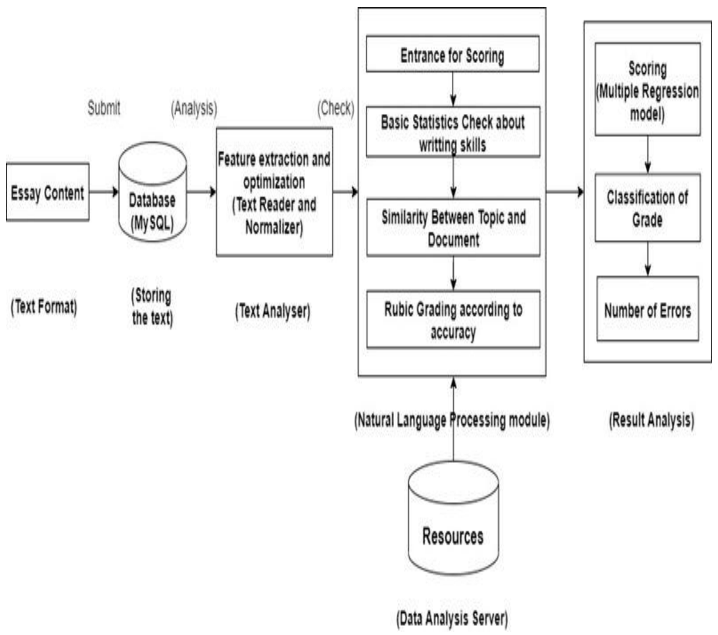

Submit
(Analysis)
(Check)
Essay Content
Database (MySQL)
Feature extraction and optimization (Text Reader and Normalizer)
Entrance for Scoring
Basic Statistics Check about writting skills
Similarity Between Topic and Document
Rubic Grading according to accuracy
Scoring (Multiple Regression model)
Classification of Grade
Number of Errors
(Text Format)
(Storing the text)
(Text Analyser)
(Natural Language Processing module)
(Result Analysis)
Resources
(Data Analysis Server)

Fig 5.1.1 Architecture of NLP

Here, we got a Mean squared error of 1.38194513377526200. Since Logistic Regression gave really bad results, we moved on to the models which do not map the features linearly like Support Vector Regression and Random Forest.

Support Vector Regression:

In classification problems, Support Vector Machines (SVMs) are well-known. However, the use of SVMs in regression is less well known. Support Vector Regression models are the name for these types of models (SVR). SVR allows us to determine how much error is reasonable in our model and will fit the data with an appropriate line (or hyper plane in higher dimensions). Unlike OLS, the aim of SVR is to minimize the coefficients, specifically the l2-norm of the coefficient vector, rather than the squared error. Instead, we treat the error term in the constraints, where we set the absolute error to be less than or equal to a given margin, called the maximum error (epsilon). We may adjust epsilon to achieve the model's desired accuracy.

Implementing Support Vector Regression (SVR) in Python

1. Step 1: Importing the libraries. import NumPy as np.
2. Step 2: Reading the dataset. dataset = pd.
3. Step 3: Feature Scaling. A real-world dataset contains features that vary in magnitudes, units, and range.
4. Step 4: Fitting SVR to the dataset.
5. Predicting a different outcome.

Minimize:

$$\text{MIN } ½ \, ||w||^2$$

Constraints:

$$|Yi - WiXi| <= \varepsilon$$

Here, we got a Mean squared error of 1.81 which is also not likely to be implemented.

Random Forest:

Random forest is a supervised learning algorithm that can be used to classify and predict data. However, it is mostly used to solve classification issues. The random forest algorithm constructs decision trees from data samples, extracts predictions from each, and then votes on the best solution. Random forests, also known as random decision forests, are an ensemble learning method for classification, regression, and other tasks that works by training a large number of decision trees and then getting output of the class that is the mode of the classes (classification) or the mean/average prediction (regression) of the individual trees. Random decision forests address the problem of decision trees overfitting their training collection. Random forests outperform decision trees in most cases, but they are less accurate than gradient boosted trees. Random Forest is a bagging extension that restricts the features that can be used to build trees, requiring trees to be special, in addition to creating trees based on multiple samples of your training data. As a consequence, better outcomes can be obtained. Limiting the number of features (rows) that the greedy algorithm can evaluate at each split point when building the tree can result in different decision trees. This approach is known as the Random Forest algorithm. Here too we got a Mean squared error of 1.81 which is not acceptable and thus from observing mean square error's result of all three models, we shifted to Neural Networks.

Algorithms Used:

- The algorithm starts with the null set and works its way forward, adding one function after another. Any time a new function is introduced, the output is calculated and compared to the previous value.
- The better the outcome, the more valuable the newly added function is regarded as, and it is kept in the package. Otherwise, it is disabled, and the algorithm moves on to the next function, repeating the process.

The algorithm seems to be as follows:

```
1. Start with the empty set Y0={∅}
2. Assign a value of 'x' to the following best attribute.
   x= max_accuracy[Y,Y + x]
3. Update Y(k+1)=Y(k)+x;
   k=k+1
4. Go to Step 2.
```

5.2 Neural Networks –

The output of an AES system can be compared to the ratings assigned by human annotators using various measures of correlation or agreement (Yannakoudakis and Cummins, 2015). These measures include Pearson's correlation, Spearman's correlation, Kendall's Tau, and quadratic weighted Kappa (QWK). The ASAP competition adopted QWK as the official evaluation metric. Since we use the ASAP data set for evaluation in this paper, we also use QWK as the evaluation metric in our experiments. Quadratic weighted Kappa (Wij) is calculated as follows:

First, a weight matrix W is constructed according to Equation 1:

$$W_{i,j} = \left[(i-j)^2 / (N-1)^2 \right] \quad \ldots\text{Eqn (1)}$$

where i and j are the reference rating (assigned by a human annotator) and the hypothesis rating (assigned by an AES system), respectively, N refers to the total number of potential scores. A matrix O is calculated such that Oi, j denotes the number of essays that receive a rating i by the human annotator and a rating j by the AES system. An expected count matrix E is calculated as the outer product of histogram vectors of the two (reference and hypothesis) ratings. The matrix E is then normalized such that the sum of elements in E and the sum of elements in O are the same. Finally, given the matrices O and E, the QWK score is calculated according to Equation 2:

$$k = 1 - \left[\sum_{i,j} W_{i,j} O_{i,j} / \sum_{i,j} W_{i,j} E_{i,j} \right] \quad \ldots\text{Eqn (2)}$$

Smart Essay Grader (SEG) model implementation:

Pre-processing steps for neural networks are different from pre-processing steps for machine learning algorithms. Our training data is fed into the Word2Vec Embedding Layer. Word2Vec is a two-layer shallow neural network that has been trained to recreate linguistic contexts for terms. It takes a large corpus of words as input and outputs a vector space with hundreds of dimensions, with each specific word in the corpus assigned to a corresponding vector in the space. Word vectors are arranged in the vector space in such a way that terms with similar contexts in the corpus are close together in the space. For learning word embeddings from raw text, Word2Vec is a computationally efficient predictive model. Word2Vec features are fed into LSTM. We used a basic LSTM with a variable length and a scoring layer at the top. The LSTM takes a series of word vectors that correspond to the essay's terms and produces a vector that encapsulates the details in the essay. This vector is converted into a score in the appropriate range by the scoring layer at the end. The feed forward networks mentioned above use a bag of words model, which has limited capacity. As a result, the introduction of LSTM-based architecture was justified in order to collect sequence information in the dataset. LSTM will figure out which data in a series should be kept and which should be discarded. This greatly aids in the calculation of essay scores. Two Long Short-Term Memory (LSTM) layers and a Dense output layer make up the model architecture. The Relu activation feature is used in the final layer. The QWK is determined by using 10-fold cross validation to train a model on the dataset and taking the average of all ten folds.

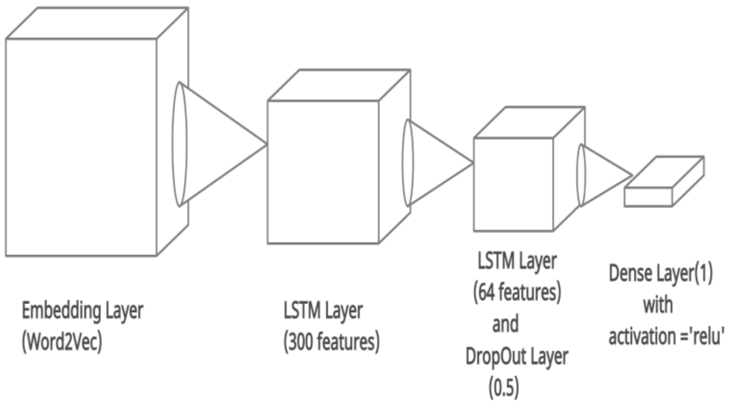

Embedding Layer (Word2Vec)
LSTM Layer (300 features)
LSTM Layer (64 features) and DropOut Layer (0.5)
Dense Layer(1) with activation ='relu'

Figure 5.2.1 4-Layer Neural network

VI. SEG SYSTEM DIAGRAM

Following depicts complete explanation of SEG model

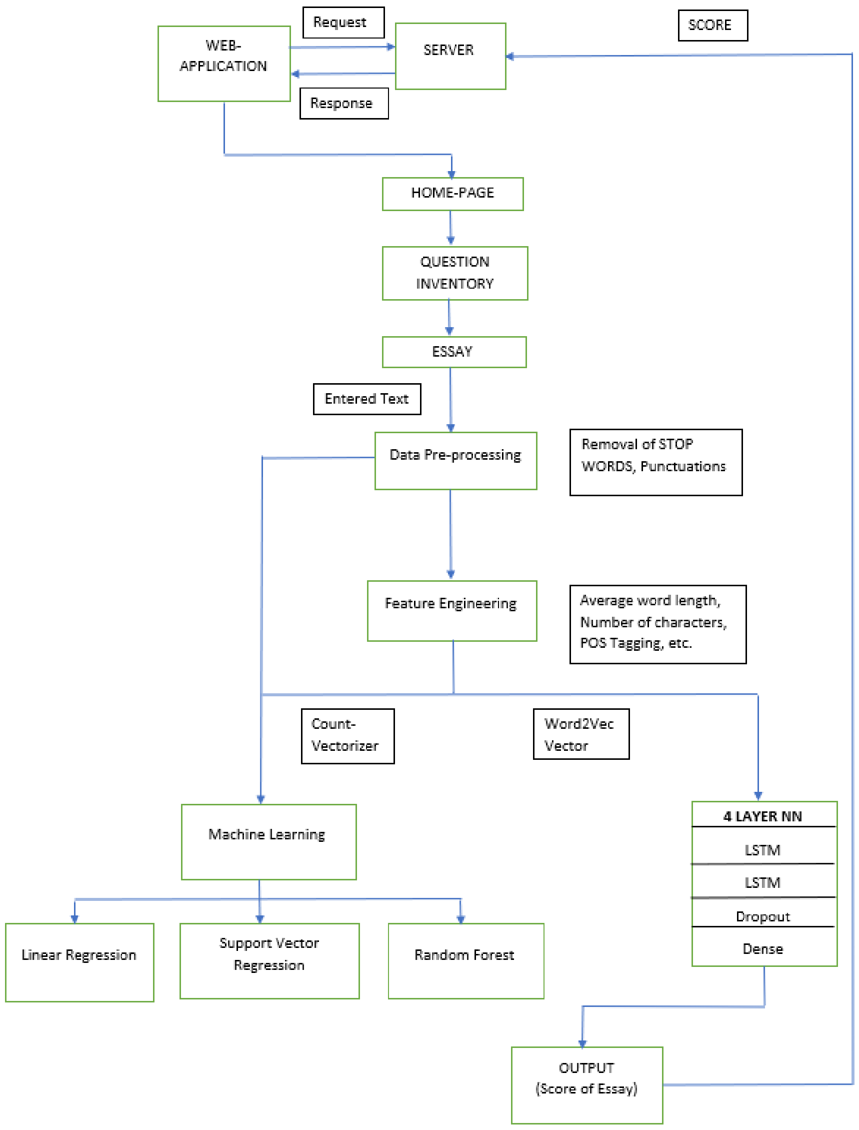
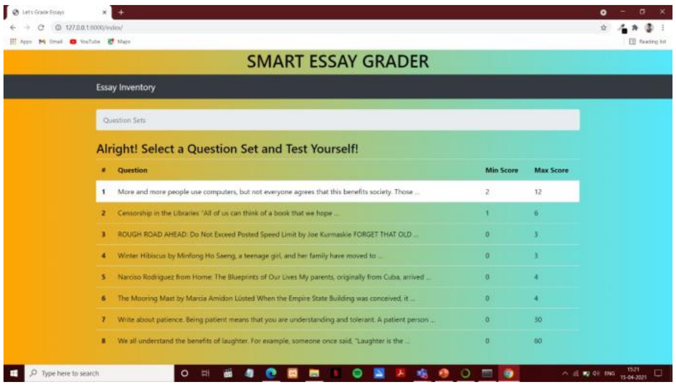


Figure 6.1 SEG System diagram

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:04/April-2021 Impact Factor- 5.354 www.irjmets.com

This is a step-by-step depiction of how SEG system works.

VII. SEG GUI AND OUTPUT


SMART ESSAY GRADER

Essay Inventory

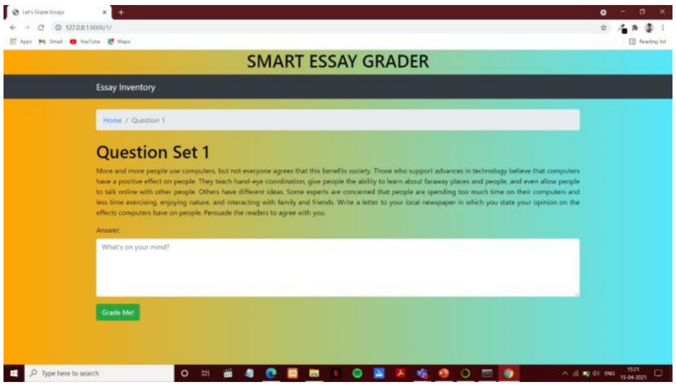
Question Sets

Alright! Select a Question Set and Test Yourself!

#	Question	Min Score	Max Score
1	More and more people use computers, but not everyone agrees that this benefits society. Those ...	2	12
2	Censorship in the Libraries "All of us can think of a book that we hope ...	1	6
3	ROUGH ROAD AHEAD: Do Not Exceed Posted Speed Limit by Joe Kurmaskie FORGET THAT OLD ...	0	3
4	Winter Hibiscus by Minfong Ho Saeng, a teenage girl, and her family have moved to ...	0	3
5	Narciso Rodriguez from Home: The Blueprints of Our Lives My parents, originally from Cuba, arrived ...	0	4
6	The Mooring Mast by Marcia Amidon Lüsted When the Empire State Building was conceived, it ...	0	4
7	Write about patience. Being patient means that you are understanding and tolerant. A patient person ...	0	30
8	We all understand the benefits of laughter. For example, someone once said, "Laughter is the ...	0	60

Figure 7.1: Essay questions with minimum and maximum score.

Figure 7.1 shows the inventory of questions with both minimum score and maximum score of the Essay.


SMART ESSAY GRADER

Essay Inventory

Home / Question 1

Question Set 1

More and more people use computers, but not everyone agrees that this benefits society. Those who support advances in technology believe that computers have a positive effect on people. They teach hand-eye coordination, give people the ability to learn about faraway places and people, and even allow people to talk online with other people. Others have different ideas. Some experts are concerned that people are spending too much time on their computers and less time exercising, enjoying nature, and interacting with family and friends. Write a letter to your local newspaper in which you state your opinion on the effects computers have on people. Persuade the readers to agree with you.

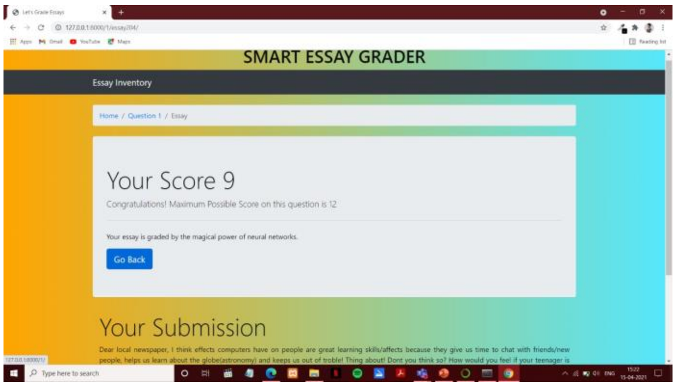
Answer:

What's on your mind?

Grade Me!

Figure 7.2: The Essay question format

Figure 7.2 shows the complete Question the user has selected, the input space answer for the user to write and grade button.


SMART ESSAY GRADER

Essay Inventory

Home / Question 1 / Essay

Your Score 9

Congratulations! Maximum Possible Score on this question is 12

Your essay is graded by the magical power of neural networks.

Go Back

Your Submission

Dear local newspaper, I think effects computers have on people are great learning skills/affects because they give us time to chat with friends/new people, helps us learn about the globe(astronomy) and keeps us out of troble! Thing about! Dont you think so? How would you feel if your teenager is

Figure 7.3: The Final result

Finally, in Figure 7.3 we get the output score out of the maximum score possible for the respective essay.

VIII. CONCLUSION

The use of various methods to study automated essay grading has been done a variety of times. This approach aims to model the language with the most useful features. The outcomes that can be achieved will be both inspiring and valid. We should be able to obtain an average absolute error that is slightly lower than the human standard deviation. As a result, we decided to express through our project SEG that it is possible to have an essay graded automatically, reducing the burden on the individual. All essays will be graded using the same criterion and given an accurate score. We'll do whatever we can to keep the best one up to date.

FUTURE SCOPE

- There is definitely room for improvement, particularly if we can identify the right features.
- Image processing can also be used to grade handwritten essays that have been graded offline.
- Also, we can further improve the user-friendly grading by specifying what went wrong in the essay and suggestions for a good grade.

IX. REFERENCES

[1] A Neural Approach to Automated Essay Scoring Kaveh Taghipour and Hwee Tou Ng Department of Computer Science National University of Singapore

[2] Automated Essay Grading Using Features Selection Y.Harika, Sri Latha, V.Lohith Sai, P.Sai Krishna , M.Suneetha (International Research Journal of Engineering and Technology (IRJET))

[3] PDF, Automated Essay Grading using Machine Learning Algorithm

[4] Valenti, S., Neri, F and Cucchiarelli, A. 2003 An Overview of Current Research on Automated Essay Grading

[5] Attali Y and Burstein, J 2006 Automated essay scoring with e-rater

[6] Automated essay evaluation with semantic analysis Kaza zupac, Zoran Bosnic.

[7] "Project Essay Grade: PEG", p. 43. In Shermis, Mark D., and Jill Burstein, eds., Automated Essay Scoring: A Cross-Disciplinary Perspective. Lawrence Erlbaum Associates, Mahwah, New Jersey,

[8] Larkey, Leah S., and W. Bruce Croft (2003). "A Text Categorization Approach to Automated Essay Grading", p. 55. In Shermis, Mark D., and Jill Burstein, eds. Automated Essay Scoring: A Cross-Disciplinary Perspective. Lawrence Erlbaum.

[9] "Neural Net or Neural Network - Gartner IT Glossary". www.gartner.com

[10] Russell, Ingrid. "Neural Networks Module". Archived from the original on 29 May 2014. Retrieved 2012.

[11] Graves, A.; Liwicki, M.; Fernandez, S.; Bertolami, R.; Bunke, H.; Schmidhuber, J. (2009). "A Novel Connectionist System for Improved Unconstrained Handwriting Recognition". IEEE Transactions on Pattern Analysis and Machine Intelligence. 31 (5): 855–868. CiteSeerX 10.1.1.139.4502

[12] Li, Xiangang; Wu, Xihong (2014-10-15). "Constructing Long Short-Term Memory based Deep Recurrent Neural Networks for Large Vocabulary Speech Recognition". arXiv:1410.4281.