# Automated Essay Grading System Using A Hybrid Model Of Natural Language Processing And Neural Network Approach

Avinash Pandey,Anshika Pandey,Sakshi Singh,Shiwani Gupta
*Thakur College of Engineering and Technology, Mumbai, India*
**pandeyavinash714@gmail.com,anshikapandey603@gmail.com,singhsakshi.0130@gmail.com,shiwanigupta3005@gmail.com**

*Abstract*-- **Essay Grading is a tedious process when done manually. As a result, over the years many models of Automated Essay Grading (AEG) systems have been proposed and implemented. These models extract various features using Natural Language Processing (NLP) from the essay corpus and grade the essays based on these features using Machine Learning algorithms. Recently, Neural Network is being used for prediction and have shown promising results. Hence, in this paper a hybrid model of NLP and Neural Network is proposed to get an effective result.**

***Keywords— Automated Essay Grading, Natural Language Processing, Neural Network***

## I. INTRODUCTION

Automated Essay Grading (AEG) or Scoring (AES) Systems are systems that make use of Natural Language Processing along with Machine Learning to grade essays based on various parameters. These parameters are termed as the features of the essay. There can many features used for the assessor but they are grouped mainly into 3 types viz. syntactic, semantic and sentiment. There are many AES systems that have been developed over the years. These models range from assessors that only use syntactic features to the ones that use a combination of all three features to the ones that use all of them. But these graders are either not available for use by common people or expensive. As a result, in this paper a model is proposed that makes use of NLP to extract the features and Neural Network to predict the score of the essays similar to the human grader. This essay grader takes into account a variety of linguistic features and uses them to judge the quality of the input essay. Here the features extracted are syntactic and sentiment parts of the essay. The model is then evaluated on the basis of Quadratic Weighted Kappa Score to determine its performance. The aim is to develop a model that performs either equally to the existing systems or better than them

## II. BACKGROUND

In many fields essays have been a part of curriculum as well as a metric to judge students since a long time. Every year, many students submit essays either as a part of their academics or as a part of their selection process for higher institutes or organizations. Grading these essays becomes a tedious and time consuming task especially when it is done manually let alone digitally on a computer. It can cause severe mental and physical strain on the grader. To avoid this, it is essential that essays should be graded in such a way that it requires minimum human efforts or inputs. Such grading of Essays is termed as Automated Essay Grading (AEG) or Automated Essay Scoring (AES). Automated essay grading (AEG) is a tool of educational assessment that makes use of computers, laptops or mobile phones for grading the written essay. Several such systems already exist named Project Essay Grader, Intelligent Essay Assessor, C-rater and many more. Project Essay Grader which was probably the first automatic assessment tool was developed in the 1960's and only took the style of writing in consideration. It was followed by Intelligent Essay Assessor and Educational Testing Service which used only the content of the essay. These were followed E-RATER, C-RATER, BETSY and many other such assessors which took both the style and content of the essay into consideration. The earlier systems made use of only syntactic features for the prediction purposes and as a result were less efficient. Due to research and development in these systems various other linguistic features are also used in recent times.

## III. LITERATURE REVIEW

Automated Essay Grading System has been under studies for a long time. There have been many different models proposed and developed over the years. Automated Essay Grading System has been under studies for a long time. There have been many different models proposed and developed over the years. Some either use only Natural Language Processing while some use Neural Network along with it. The models have their own pros and cons due to the type of machine learning model used or due to the features that are extracted. Nonetheless, AEGS has evolved over the years with different changes to it to enhance it and make better score prediction. There are numerous studies published over the years on AEGS some of which are listed below:

[1] A. Rokade, B. Patil, S. Rajani, S. Revandkar and R. Shedge proposed a model that grades hypothesis-based subjects using Natural Language Processing. They used Machine Learning to perform semantic analysis. They further made use of Ontology i.e. they extracted keywords and their synonyms related to the domain to match, along with the above-mentioned techniques on a data of technical answers. The scores were calculated based on the content of the answers of the students and based on the similarity of this answer with the correct answer that was provided.

[2] K. Srivastava, N. Dhanda, A. Shrivastava performed a comparative study on various existing Essay Grading Systems since their inception and how they have improved over the years. They compared the system based on various parameters viz. Technique, Performance, Application and Attributes. The observed how all these systems only focused on the style and content of the essays and not the meaning of the essay. As a result, they failed in achieving the accuracy of a human grader. Hence, they concluded this study by highlighting the lack of semantic attributes and its need in the future to enhance the systems and to detect correctness of the essays.

[3] H. K. Janda, A. Pawar, S. Du, V. Mago proposed a model that used rule-based grammar, surface level coherence and semantic similarity between the sentences to develop a graph-based relationship between the contents of the essay and calculated the polarity of the opinion expressions. This was then used to obtain features for the model. They used 23 features to judge the essay quality. They used a dataset of 13,000 essays. The model was evaluated on Quadratic Weighted Kappa and its value came out to be 0.793.

[4] A. H. Filho, F. Concatto, J. Nau, H. A. do Prado, D. Os. Imhof, E. Ferneda analyzed the performance of different types of statistical algorithms using machine learning and classification techniques, combined with different balancing method on essays of Brazil's National High School Examination (ENEM). They made use of 4 balancing (3 of oversampling and 1 of undersampling) technique viz. Synthetic Minority Over-sampling Technique (SMOTE), Adaptive Synthetic (ADASYN), Random Oversampling (ROS) and Random Undersampling (RUS). SMOTE and ADASYN resulted in less impact as they generated synthetic samples which indicates the presence of unusual pattern in the spatial distribution of feature vectors. The undersampling technique resulted in higher accuracy in minority classes but lower accuracy in majority classes. The machine learning algorithms used were, LASSO, Support Vector Regression, Support Vector Classification and Gradient Boosted Trees. Gradient Boosted Trees and Support Vector Classifiers yielded a correct prediction ratio of around 75% for all five classes. Future models can include neural networks and ensemble learning methods.

[5] S. M. Darwish and S. Kh. Mohamed proposed a model that uses Latent Semantic analysis and Fuzzy Ontology. Latent Semantic Analysis was used to check the semantics and Fuzzy Ontology was used to check the consistency and coherence of the essay apart from syntax and vocabulary to overcome the vagueness of language. The model provided score as well as feedback to the student. The specificity came out to be 100% while the use of the fuzzy ontology led to a sensitivity of 75.2%.

## IV. DATASET DESCRIPTION

The Automated Student Assessment Prize's (ASAP) Dataset from Kaggle.com, sponsored by Hewitt-Packard comprised of 13,000 essays, 8 different datasets of different genre. Each dataset was a collection of responses to its own prompt.

TABLE I

ESSAY SET DESCRIPTION

| Essay Set | Grade | No. of Essays | Genre | Avg. Length | Score |
|---|---|---|---|---|---|
| 1 | 8 | 1785 | Persuasive/ Narrative/ Expository | 350 | 2-12 |
| 2 | 10 | 1800 | Persuasive/ Narrative/ Expository | 350 | 1-6 |
| 3 | 10 | 1726 | Source-Dependent Responses | 150 | 0-3 |
| 4 | 10 | 1772 | Source-Dependent Responses | 150 | 0-3 |
| 5 | 8 | 1805 | Source-Dependent Responses | 150 | 0-4 |
| 6 | 10 | 1800 | Source-Dependent | 150 | 0-4 |
| 7 | 7 | 1730 | Persuasive/ Narrative/ Expository | 250 | 0-30 |
| 8 | 10 | 918 | Persuasive/ Narrative/ Expository | 650 | 0-60 |

The essays range from an average length of 150 to 550 words per response and were written by students of class 7 to 10. Each essay had one or more human scores and a final resolved score. Each essay set had a different technique to calculate the resolved score. Essays in each of the 8 sets have different attributes that was used for grading. Based on the type of essays, there was a different set of attributes for evaluation which are Ideas and Content, Organization, Word Choice, Voice and Language Conventions. This ensured that the automated grader was trained effectively across different types of essays.

There are three types of essays in the dataset:

1. Argumentative/Persuasive Essays: In these type of essays, the prompt is one in which the writer has to convince the reader about their stance for or against a topic. For example, free speech in public colleges.

2. Source-Dependent Responses: These essay response to a source text, where the writer responds to a question about the prompt. For instance, describing the writer's opinion about an incident that happened to him in the prompt.

3. Narrative/Descriptive Essays: These are essays where the prompt requires us to describe / narrate a story.

## V. FEATURE EXTRACTION

The features identified for the proposed model are mainly syntactic and sentiment.

Syntactical Features: Syntax refers to arrangement of content. Here majorly Natural Language Processing Toolkit (NLTK) is used, which is a python-based

platform to extract language-based data. It is helpful to extract syntactic features of a text such as tokenization, stemming, tagging, etc.
The model is using the following set of syntactic features to grade the essays:
A. Word, Character and Sentence Count: These are very basic features of any text and do influence the scoring of the essay as well. The variety of the length of sentences potentially reflect the complexity of syntactic.

B. Average Word Length: This feature helps to keep count of average words in an essay.

C. Unique Word Count: This feature indicates the number of unique words used in an essay.

D. Stop-word Count: This feature helps to get more meaningful results when one measures the lexical complexity of text as stop-words have little lexical richness and are only used to bind words in a sentence.

E. Lemma Count: This feature counts unique lemma forms using NLTK's part of speech tagger from the given essay and interface to the WordNet lemmatizer.

F. Part of Speech (POS) Tags: It is used to classify types of tokens according to their use in a sentence such as nouns, adjectives, verbs, adverbs, etc. It is crucial for evaluating the quality of content in the essay.

G. Spell Check: This step is aimed to check spelling mistakes. It is tested against the words in the installed corpus as students submit their answers. If the word is not found in the dictionary, it will be counted as a spelling mistake.

H. Punctuation Count: This feature is used to count the number of punctuations used in an essay.

I. Grammar Check: This feature checks the grammar mistakes in an essay and returns the total number of grammatical error in the essay.

J. Determiners: This feature calculates the number of determines in the essay.

K. Preposition: This feature calculates the number of prepositions in the essay.

L. Words ending with 'ing': This feature calculates the number of words in an essay that end with 'ing'.

M. Flesch Kincaid Grade Level: This feature calculates the minimum reading grade which is required to read the essay. Lower the grade, simpler the essay; higher the grade, complex the essay.
Sentiment Features: It highlights on opinions, attitudes and emotions of a writer as each writer has a unique thought towards the subject being written about.
In argumentative and persuasive essays, writer needs to defend and prove his/her point of view on the subject, their tone and the way of textual sentiment expression affect their writing. The sentiment analysis in the essays helps us know the polarity inclination they contribute to the text which is otherwise difficult to comprehend by the computer.
These features are calculated using open source tool VADER (Valence Aware Dictionary and sEntiment Reasoner) was used. It was preferred as it is fast and does not require any training dataset.
The sentiment-based set of features are:

A. Positive Polarity: This feature reflects the amount of positive polarity in an essay.

B. Negative Polarity: This feature reflects the amount of negative polarity in an essay.

C. Neutral Polarity: This feature reflects the amount of neutral polarity in an essay.

D. Compound Polarity: This feature reflects the overall polarity of the essay.

## VI. PERFORMANCE METRIC

The performance used here is Quadratic Weighted Kappa Score (QWK). It measures the agreement between two values. It is calculated between the predicted value and the actual value. It lies between 0 and 1. If the agreement is complete then the value is 1 and if there is no agreement then the value is 0. It is calculated as follow:

1. Create a multi class N-by-N confusion matrix 'O' between the actual and predicted value such that Oi,j corresponds to the number of actual values that have a value of i and received a predicted value j.
2. Construct an N-by-N weight matrix 'W' by subtracting the actual and predicted values.
3. Create two vectors one of actual values and one of predicted values consisting of number of values of each rating are present. The construct an N-by-N histogram matrix of expected ratings 'E' by taking the outer product between the actual values' histogram vector and the predicted values' histogram vector, normalized such that E and O have the same sum.
4. Weighted Kappa = 1 - sum($W_{i,j}$ $O_{i,j}$) / ($W_{i,j}$ $E_{i,j}$)

## VII. PROPOSED METHODOLOGY

The Automated Student Assessment Prize's (ASAP) dataset is being obtained from Kaggle.com. The following proposed model works in 2 phases. The first phase includes extracting the features and the second phase includes training the model with these features. In the feature extraction phase first the essay dataset is cleaned to remove stopwords, punctuations and other irrelevant data. Then this cleaned data is tokenized and is then passed to various functions to extract the features listed above using various libraries present in Python. The features added are common essay grading parameters like character count, word count, sentence count, lemma count, noun count, adjective count, verb count, adverb count, grammar and spelling mistakes etc. These features extracted now become the dataset which will be given as an input to the neural network model.
The neural network model used here is LSTM which stands for Long Short Term Memory as it has shown promising results in the past. The model is then trained with the dataset developed after the preprocessing and its efficiency is calculated. If the model performs poor then the hyperparameters are tuned and the model is

trained again. The final kappa score is calculated and this model is then used to predict the grade of new essays.

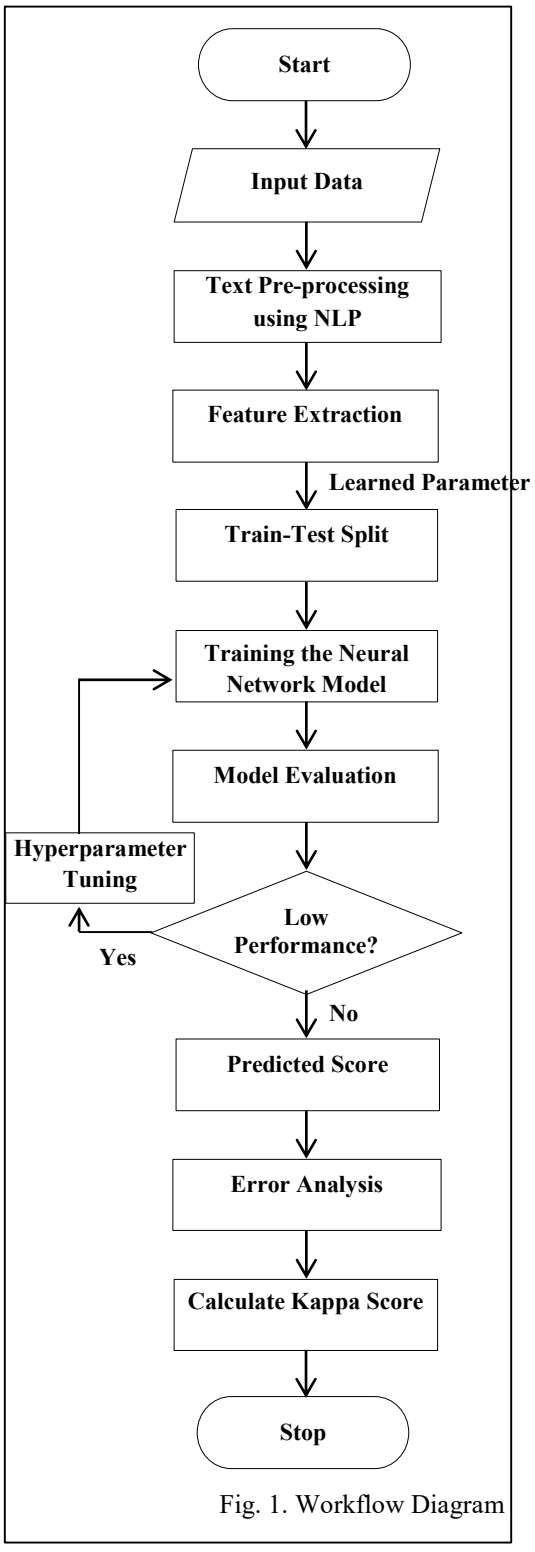


Fig. 1. Workflow Diagram

VIII. RESULT AND DISCUSSION

After the data cleaning and feature extraction processes the following dataset was obtained which will be used to train the model.

| char_count_before | punctuation_count | stopwords_count | char_count | word |
|---|---|---|---|---|
| 1538 | 55 | 157 | 1360 | |
| 1870 | 46 | 175 | 1819 | |
| 1263 | 34 | 129 | 1196 | |
| 2642 | 91 | 207 | 2382 | |
| 2105 | 55 | 211 | 1958 | |

| sent_count | average_word_length | spell_count | lemma_count | noun_count | adj |
|---|---|---|---|---|---|
| 16 | 4.237142857 | 41 | 162 | 83 | |
| 20 | 4.312056738 | 40 | 185 | 107 | |
| 14 | 4.342756184 | 27 | 145 | 82 | |
| 27 | 4.813207547 | 76 | 236 | 178 | |
| 30 | 4.334038055 | 48 | 190 | 114 | |

| verb_count | adv_count | ratio | grammar_error | flesch_kincaid_ |
|---|---|---|---|---|
| 74 | 24 | 0.88427 | 11 | |
| 85 | 19 | 0.97273 | 19 | |
| 52 | 16 | 0.94695 | 9 | |
| 97 | 29 | 0.90159 | 35 | |
| 90 | 36 | 0.93017 | 17 | |

| determiner_count | preposition_count | words_ending_with_ing | domain1 |
|---|---|---|---|
| 20 | 54 | 15 | |
| 35 | 60 | 17 | |
| 27 | 32 | 9 | |
| 43 | 64 | 7 | |
| 54 | 44 | 7 | |

| positive | negative | neutral | compound | unique_word_c |
|---|---|---|---|---|
| 0.162 | 0 | 0.838 | 0.9951 | |
| 0.203 | 0.019 | 0.778 | 0.9981 | |
| 0.19 | 0.043 | 0.767 | 0.9947 | |
| 0.143 | 0.012 | 0.844 | 0.9977 | |
| 0.09 | 0.025 | 0.885 | 0.9735 | |

Fig. 2. Preprocessed Dataset

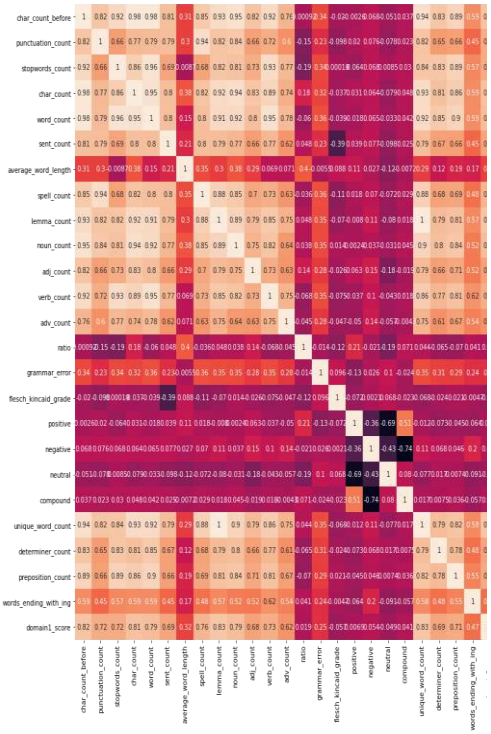

Fig. 3. Heatmap

A heatmap of the extracted features was generated and is shown above. This heatmap helped in establishing the correlation of the features with each other and their correlation with the grade feature that is to be predicted.

## CONCLUSION

Thus in this paper the cause of development of Automated Essay Grading System was identified and its applications were highlighted. Further the gaps in the existing system were identified and a methodology was proposed which will try to fill these gaps and provide a feasible system for essay evaluation. Additionally, various features that will be used by this system were listed and the output of the feature extraction along with the correlation of each was shown. The metric used for this system was also highlighted.

## FUTURE SCOPE

The proposed system can be used by testing agencies for test of English writing to take the burden off the human graders as around 1.8 million students participate every year in the competitive exams such as IELTS,TOEFL. It could be helpful to teachers in schools and institutions to grade the descriptive answers. The concept can be extended to other languages too if the dataset is available. It can be further expanded by developing a general purpose essay grader that considers many languages into account instead of just one.

**References**

[1] Rokade, B. Patil, S. Rajani, S. Revandkar and R. Shedge, "Automated Grading System Using Natural Language Processing", Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, 2018, pp. 1123- 1127.

[2] K. Srivastava, N. Dhanda , A. Shrivastava, "An Analysis of Automated Essay Grading Systems", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-6, March 2020.

[3] H. K. Janda, A. Pawar, S. Du and V. Mago, "Syntactic, Semantic and Sentiment Analysis: The Joint Effect on Automated Essay Evaluation" in IEEE Access, Vol. 7, 2019, pp. 108486-108503.

[4] A. H. Filhoa, F. Concattoa, J. Naua, H. A. do Pradob, D. O. Imhofa, E. Fernedab, "Imbalanced Learning Techniques for Improving the Performance of Statistical Models in Automated Essay Scoring", 23rd International Conference on Knowledge Based and Intelligent Information and Engineering Systems, Procedia Computer Science, Volume 159, 2019, Pages 764-773.

[5] S. M. Darwish, S. Kh. Mohamed, "Automated Essay Evaluation Based on Fusion of Fuzzy Ontology and Latent Semantic Analysis", The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2019), Vol. 921,2020,pp.566–57