A Novel Algorithm for Automatic Essay Grading using Natural Language Processing Techniques

[1.] Vishwak Senan G, [2.] Pranav Kumar K, [3.] Sundaramoorthy T
[4.] Dr D Jayashree

[4.] Professor at Rajalakshmi Institute of Technology

[1.] vishwaksenang.2016.cse@ritchennai.edu.in,
[2.] pranavkumark.2016.cse@ritchennai.edu.in,
[3.] sundaramoorthyt.2016.cse@ritchennai.edu.in,
[4.] jayashree.d@ritchennai.edu.in,
Department of Computer Science and Engineering
Rajalakshmi Institute of Technology

Abstract - Evaluation of an English essay is one of the important and complex tasks which is done manually by skilled and efficient professors and faculties till date. The growth of science and technologies enables to automatic evaluation of an English essay using natural language processing (NLP) techniques. The intelligent system - built upon NLP multiple neural network model - gives out generic evaluation and the topic/question correlation for any given English essay.

Index Terms – Automatic evaluation, Natural Language Processing, generic evaluation, topic/question correlation.

I. Introduction

International examinations like GRE (Graduate Record Examination), IELTS (International English Language Testing Systems), etc., is gaining popularity day by day as this examination's results are considered as criteria for various universities and companies. Therefore, the number of international students across various counties are taking the exams, increasing the count day by day and there is a huge time buffer to evaluate their English essays and publish the results. To reduce the stress on the organization who are hosting these examinations and students to practice their writing skills, our project aims to evaluate the English essay so that the organization can focus their work in other aspects of examinations and students can practice at free will.

II. Literature Survey

There are various existing NLP techniques which are considered to be best for text classification which is essential for evaluation of the essay. The initial text classification techniques were to classify parts-of-speech of a sentence.

N-gram model

The n-gram model otherwise known as Shannon's Markov Chain [15], estimates the probability of the occurrence of the next word given the n-1 words. Therefore, this is a probabilistic model with a good accuracy if trained from a sophisticated dataset.

Perceptron Layer

This Model of finding the parts-of-speech tagger of a sentence uses 3-layer perceptron layer with n-inputs, n representing the total number of words in the dataset. While the computational cost of the training the network is drastically higher than the n-gram model but according the previously conducted experiment [14], the accuracy was 99.4% without over-fitting into the data. The accuracy was obtained by using the elastic hidden perceptron layer.

Long Short Term Memory (LSTM)

Long Short Term Memory is a type of Recurrent Neural Network (RNN) which is a very famous model to predict a series of a vector. This model has high computational requirement but yields a high accuracy for a large dataset. In a research, the low-cost hardware implementation of LSTM [16] made the neural network to be more optimized and efficient for any given low-cost machine. This invokes the possibility of training the neural network locally rather than renting GPUs online. The low-cost LSTM is achieved by applying stochastic function rather than hyperbolic activation function.

Bi-directional RNN

Bi-directional Recurrent Neural Network enables the RNN to work for dynamic dataflow and have multiple activation functions [17].

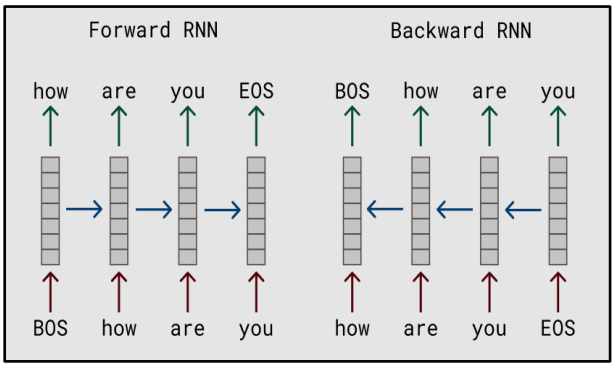

Figure: I Bidirectional RNN Flow

Clustering

Clustering algorithm is implemented to categorise the word component to particular to cluster, there exists two major types of cluster, one is soft cluster and other is hard cluster. Hard clusters are the concept of binary categorization, a component can exist in only one cluster. K-means is the algorithm utilised for clustering, where components are fed into nods as training dataset in an iterative methodology. K-means algorithm can be exhaustive and some components may not be segregated into one cluster and its left out. The process of k-means is to determine the distance between the centroid and the datapoints. The best centroid is where it has a mean of zero and standard deviation of one. For determining the centroid points, k-means follows Expectation minimization(EM) [21], this value assign a centroid point where the density of data points is high. In initial iterations, random set of data points are subjected to undergo Euclidean distance calculation [20] and the smallest distance obtained between any two datapoints, from them one of the datapoint is assigned as centroid and once the centroid is formed, in every iterations, its coordination is altered as to include the higher number of datapoints based on the RSS (residual sum of squares) [21] value.

t- Stochastic Neighbour Embedding

t-Distributed Stochastic neighbour embedding (t-SNE) is the algorithm implemented to determine the highest similarity between the subjected word component and the target component (subject component is obtained from variance of Gaussian [22]) by converting the high dimensional datapoints into low dimensional datapoints. Considering two parameters involving 2d graphical representation of multiple datapoints, these datapoints are the mapped and clustered with soft or hard clustering using the algorithms k-mean or gaussian mixture model, thus these clusters are needed to be evaluated for relevance or degree of similarity [22] by using

normal distribution to scale the datapoint's value against the distribution curve. The datapoints with highest distribution and with the lowest high dimensional Euclidean distance [22], gets assigned with vector value obtained from conditional probability distribution that represent similarities [22]. T-Sne algorithm forms a matrix based on the distance between dissimilar and similar datapoints representing the maximum similarity value is near the diagonal of the matrix, hence this matrix called diagonal matrix [22], since highest similarity is one, hence the similarity between a word to itself is neglected and similarity between word components are estimated n plotted in the matrix. Now, this algorithm has two parameters by which it can cluster the datapoints in low dimensional graph retaining the cluster similarity, word covariance and relevancy level. Similarity factor is implemented in low dimensional graph to cluster the datapoints at a cost of multiple iterations. Simply the algorithm gets one matrix in each iteration, thus the iteration is kept on progressed until it reaches the matrix value formed with high dimensional graph. In low dimensional graph, datapoints are mapped against t-distribution curve graph to obtain similarity between the datapoints.

Existing Works

There are currently many researches in progress to automate the evaluation of English essay. Each research accounts different factors to evaluate an essay.

The rubric based evaluation [18] is a hard-coded rule-based evaluation. The mentioned research explicitly aims to find and describe the narrative genre of the essay. The main limitation is that it is very topic focused and hard-coded.

Another research [19] aims to prepare answer documents for history subject. The goal is achieved by using an information retrieval technique, document pre-processing techniques, yielding good results related to history subject. The text correlation is well implemented, yet it particularly focuses on single subject which is the limitation of this research.

III. System Architecture

The main goal of the project is achieving the five crucial factors for evaluating an English essay. The factors are

1. Grammar and spell check.
2. Sentence complexity.
3. Style continuity.
4. Usage of advanced lexical resources.
5. Coherence and cohesion.

The above-mentioned factors are checked by developing a dedicated model for each factor having the input of the whole essay. The primary services of the project (evaluation engine which evaluates the essay) is hosted in any public clouds so that organizations can utilize our services too. Since there are 5 generic factors involved in evaluation of an essay, we will develop 5 neural network model which evaluates on its each of the factor.

Model 1 – Grammar and spell check.

Model 2 – Sentence complexity.

Model 3 – Style continuity.

Model 4 – Lexical Resources.

Model 5 – Coherence and cohesion.

Full architecture is implemented using python as the programming language. The architecture of the project is as given in *Figure 1*.

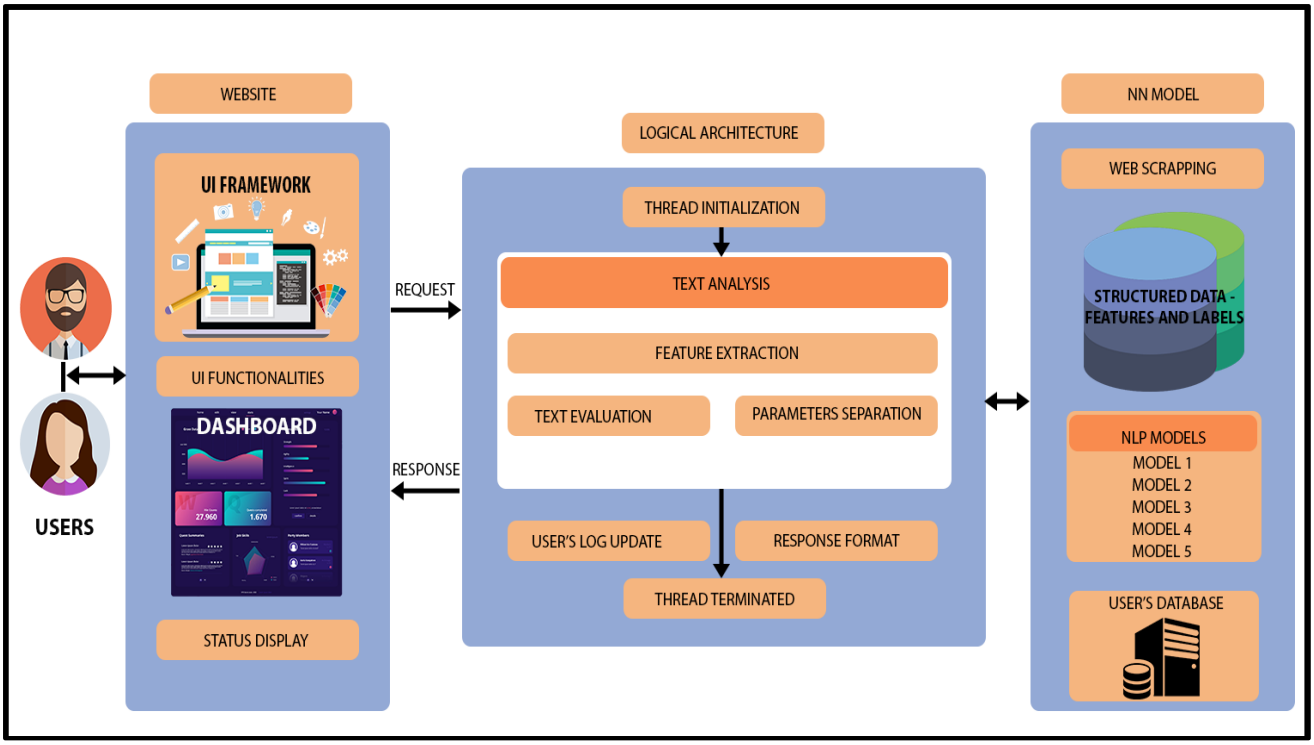

Figure I1: System Architecture

IV. Novel Algorithm

The Models which we are developing using different strategies to calculate the scores of each factor required by the project

Model 1 mainly classifies multiple sentences in the paragraph as correct or wrong. The classification mainly revolves around 2 factors. First factor is spelling mistake and second is grammatical errors in that sentence. Using rule based Gaussian Hidden Markov Model (HMM) [1] to find Parts-Of-Speech (POS) tags of a single English sentence from the essay. Using dictionary method and N-gram technique [2] to find the misspelled word and totally irrelated word which is not present in the English dictionary. When the sentence is classified as a perfect sentence or not, we calculate the percentage of correctness by

$$\frac{Number\ of\ perfect\ sentences}{Total\ number\ of\ sentences} * 10$$

Correctness score will range from [0-10].

Model 2 predicts how well the sentence is structured and expressed in higher standards. For example, take the phases "I know everything" and "No secret lies beyond my grasp" conveys the same meaning but the 2nd phase specifically conveys the message that he is a high level English user and thus 2nd phase should receive more score than the first phase. Data is scrapped from the internet and it segregated into 3 buckets namely bucket 1,2,3. The project utilizes Word level embedding, LSTM and Bidirectional Recurrent Neural Network (BRNN) [3]. Hence, we have 3 hidden layers each layer doing different mathematical operation to produce a necessary output. When the neural network is trained, a vector will be produced for the given input word. This vector

will be compared to the three bags and the nearest bag vector value is considered and the output is given. *Figure II* shows the consuming

of the neural network model and *Figure III* shows the training of the mode.

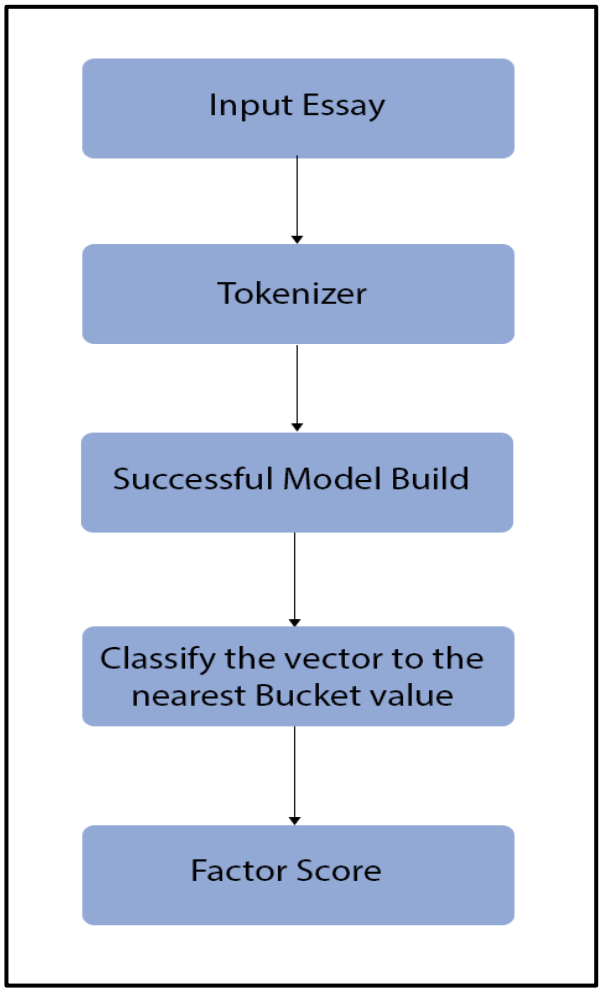

Figure IV: Consuming Model 2

Model 3, refers to as whether the user continuous the same English accent throughout the essay. A well-established English literature person will formulate sentences in particular accent so that it is uniform for the people reading the sentence can understand easily. Combining various English accents like American English, British English, Australian English in a single essay will create confusions in readers mind. Using I-vector approach and Gaussian Mixture Model (GMM) [4], we can get the consistency by dictionary vector [2] which will even work for a small dictionary dataset. The marks are calculated in such a way

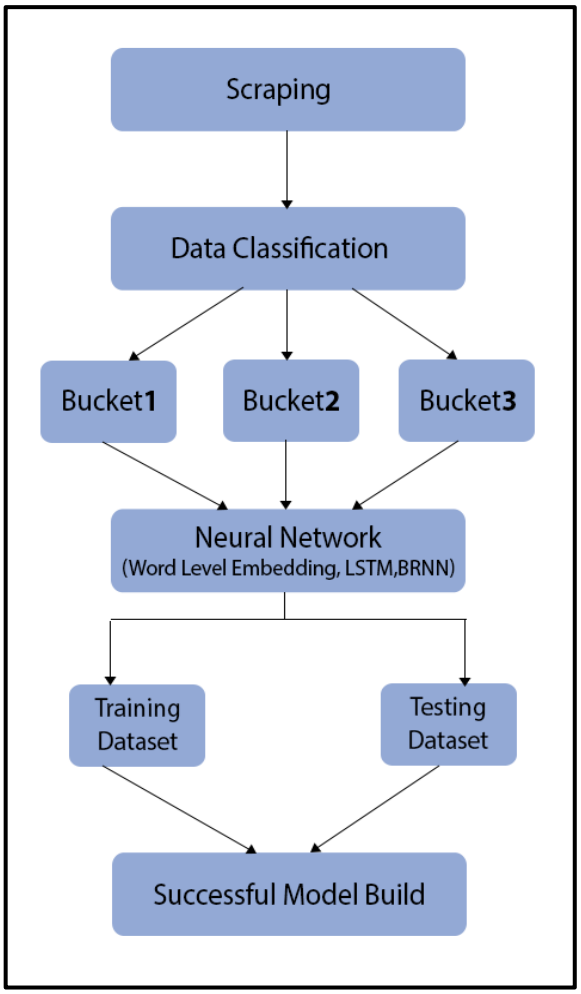

Figure III: Model 2 Training Procedure

that the same consistency is maintained throughout the essay. The score is calculated as

$$1 - \frac{Number\ of\ Accents\ Used}{Total\ available\ English\ Accents}$$

Model 4, categorizes the sophistication level of the user. Based on the input word provided, the word is mapped to its relevant word-set available in bucket and these word-set are word embedding obtained by Continuous bag-of-words (CBOW) [5]. Total of five buckets are utilised and each of them have a certain level of word standard. From bucket one to five, the word complexity increases and each bucket consists of fuzzy vector value based on the complexity of the word-set. Using word2vec [6,7] word embedding systems, cosine relevance between input corpus and bag-of-

words is determined and on the basis of IBW (Intimacy Between Words) [8], the highest intimacy fraction [8] obtained from input word and word-set determines to the bucket input word belongs, and mapped word is assigned with the vector value using term frequency(TF). When the frequency vector of the word is determined, then t-distributed Stochastic Neighbouring [11], the vector distribution gives optimizes the vector value for each back while we train our model. When the model is trained, the sentence is tokenized then the noun POS (parts of speech) tagged word by perceptron-based POS tagging [12] due to its high efficiency compared to another POS tagging algorithm. Then the model classifies the word to its nearest neighbour bucket and the evaluation is assigned for the sentence. Likewise, multiple sentences have their corresponding score and the scores are averaged out to get the final lexical resource credits.

In **model 5**, sentences are extracted as the candidate submits the essay. Then sentences are tokenized by letter string recognition and coding algorithm. The coded words will be given as input to the neural network for recognising the words. After passing as inputs, the coded words are transferred to the sentence syntax analysis module where it indexes the words before they are being processed. A 3-layer hamming neural network is being used to recognize the meaning of the sentence [9]. The model is trained in such a way that it can recognize the most related subject word for a sentence. Therefore, we can get the actual subject/topic of each sentence. Using Task based knowledge and collaborative filtering [13] techniques used to find relevance between the subjects of the consecutive sentence. In this way we can identify the cohesion and cohesion of the entire essay.

IV. Models Evaluation

There were variety of POS tagger produced various results, yet perceptron POS tagger yields out higher accuracy (*Figure IV*).

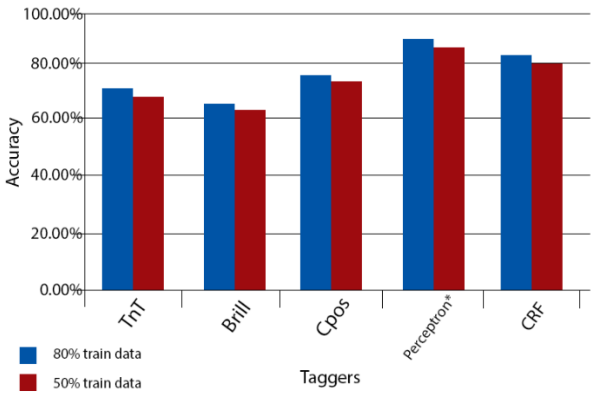

Figure IV: POS Tagger Comparison

Multinomial HMM (Hidden Markov Model) is a probability-based classification Model which yields out grammatical mistakes in the sentence. Dictionary Based n-gram model accuracy depends solely on the dataset we used. The dataset to find the correct spelling is scrapped from the WordNet Database.

Guassian Mixture Model is a probabilistic clustering algorithm which takes random centroid and clusters the datapoint with the probability regarding the centroid. This classifies the level the lexicon used in the sentence.

Since Bidirectional Neural Network is a generative neural network, it works for dynamic datapoints. The need for the dynamic datapoints is due to the complexity of the sentence varies from style to style.

V. Conclusion

Evaluation of the given English essay is up to date within the IELTS (International English Language Testing Systems) and GRE (Graduate Record Examination) standards and we aim to include various other models which

can evaluate other factors of the English essay. The current NLP models used are

1. Multinomial Hidden Markov Model.

2. Bidirectional Neural Network.

3. Long Short Term Memory.

4. Guassian Mixture Model.

5. Continous Bag of Words Method.

6. t-Distributed Stochastic Neighbour.

7. Hamming Neural Network.

8. Task Based Knowledge Filtering.

9. Collaborative Filtering.

IV. References

[1] F. Mandita, H. M. Abdullah, T. Anwar and P. Assawinjaiptech, "A novel algorithm for a grammar model checking using statistical Markov model," *2018 Seventh ICT International Student Project Conference (ICT-ISPC)*, Nakhonpathom, 2018, pp. 1-6.

[2] S. P. Singh, A. Kumar, L. Singh, M. Bhargava, K. Goyal and B. Sharma, "Frequency based spell checking and rule based grammar checking," *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, Chennai, 2016, pp. 4435-4439.

[3] A. Hassan and A. Mahmood, "Efficient Deep Learning Model for Text Classification Based on Recurrent and Convolutional Layers," *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Cancun, 2017, pp. 1108-1113.

[4] O. P. Singh and R. Sinha, "Sparse representation classification over discriminatively learned dictionary for language recognition," *TENCON 2017 - 2017 IEEE Region 10 Conference*, Penang, 2017, pp. 2632-2636.

[5] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems. 26:3111–3119. 2013.

[6] Efficient estimation of word representations in vector space, 2013.

[7] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality[J]. In Advances in neural information processing systems. 26:3111–3119. 2013.

[8] R. Wang, W. Pan, J. Ma and H. Wang, "A Synonym Extraction Method Based on Intimacy*," *2019 IEEE International Conference on Agents (ICA)*, Jinan, China, 2019, pp. 57-60.

[9] Computational Intelligence: ImitatingLife, J.M. Zurada, R.J. Marks, C.J. Robinson, IEEE Press, New York, 1994.

[10] Sentence recognition using artificial neural networks, Maciej Majewski,Jacek M. Zurada,Elsevier B.V. 2008.

[11] N. Rogovschi, J. Kitazono, N. Grozavu, T. Omori and S. Ozawa, "t-Distributed stochastic neighbor embedding spectral clustering," *2017 International Joint Conference on Neural Networks (IJCNN)*, Anchorage, AK, 2017, pp. 1628-1632.

[12] R. Banga and P. Mehndiratta, "Tagging Efficiency Analysis on Part of Speech Taggers," *2017 International Conference on*

Information Technology (ICIT), Bhubaneswar, 2017, pp. 264-267.

[13] A. Enaanai, A. S. Doukkali, I. Saif, H. Moutachaouik and M. Hain, "The collaborative relevance in the distributed information retrieval," *2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA)*, Agadir, 2016, pp. 1-6.

[14] Q. Ma, K. Uchimoto, M. Murata and H. Isahara, "Elastic neural networks for part of speech tagging," *IJCNN'99. International Joint Conference on Neural Networks. Proceedings (Cat. No.99CH36339)*, Washington, DC, USA, 1999, pp. 2991-2996 vol.5.

[15] J. A. O'Sullivan, K. Mark and M. I. Miller, "Markov random fields on graphs for natural languages," *Proceedings of 1994 Workshop on Information Theory and Statistics*, Alexandria, VA, USA, 1994, pp. 47-.

[16] S. Li, Q. Wang, X. Liu and J. Chen, "Low Cost LSTM Implementation based on Stochastic Computing for Channel State Information Prediction," *2018 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)*, Chengdu, 2018, pp. 231-234.

[17] B. Roy and H. Cheung, "A Deep Learning Approach for Intrusion Detection in Internet of Things using Bi-Directional Long Short-Term Memory Recurrent Neural Network," *2018 28th International Telecommunication Networks and Applications Conference (ITNAC)*, Sydney, NSW, 2018, pp. 1-6.

[18] H. W. Lam, T. Dillon and E. Chang, "Determining Writing Genre: Towards a Rubric-based Approach to Automated Essay Grading," *2011 IEEE International Conference on Advanced Information Networking and Applications*, Singapore, 2011, pp. 270-274

[19] S. M. F. D. Mustapha, N. Idris and R. Abdullah, "Embedding Information Retrieval and Nearest-Neighbour Algorithm into Automated Essay Grading System," *Third International Conference on Information Technology and Applications (ICITA'05)*, Sydney, NSW, 2005, pp. 169-172.

[20] V. K. Singh, N. Tiwari and S. Garg, "Document Clustering Using K-Means, Heuristic K-Means and Fuzzy C-Means," 2011 International Conference on Computational Intelligence and Communication Networks, Gwalior, 2011, pp. 297-301.

[21] H. Tao, J. Li, T. Luo and C. Wang, "Research on topics trends based on weighted K-means," 2017 7th IEEE International Conference on Electronics Information and Emergency Communication (ICEIEC), Macau, 2017, pp. 457-460.

[22] N. Rogovschi, J. Kitazono, N. Grozavu, T. Omori and S. Ozawa, "t-Distributed stochastic neighbor embedding spectral clustering," 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, 2017, pp. 1628-1632.